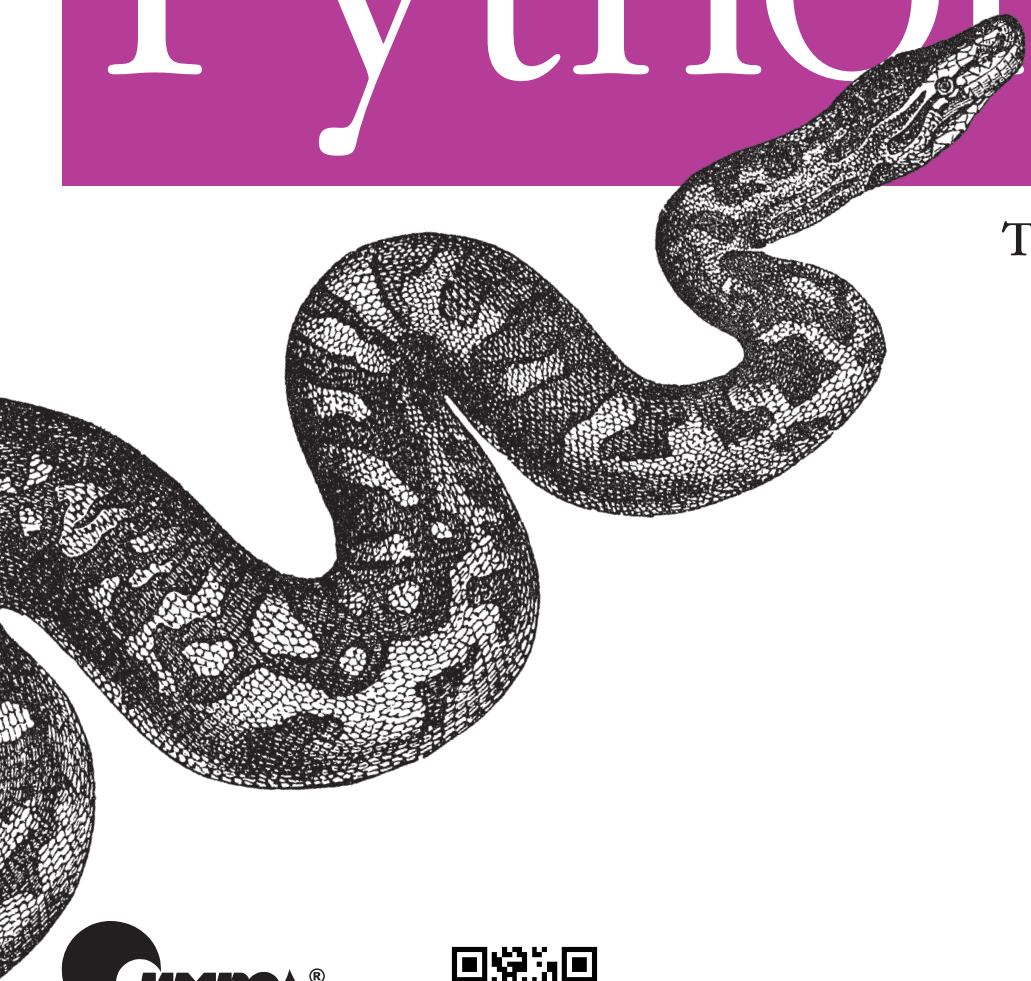
Эффективное объектно-ориентированное программирование

4-е издание
охватывает Python 3.x

Программирование на Python

том II



СИМВОЛ®
O'REILLY®

Марк Лутц

По договору между издательством «Символ-Плюс» и Интернет-магазином «Books.Ru – Книги России» единственный легальный способ получения данного файла с книгой ISBN 978-5-93286-211-7, название «Программирование на Python, 4-е издание, том 2» – покупка в Интернет-магазине «Books.Ru – Книги России». Если Вы получили данный файл каким-либо другим образом, Вы нарушили международное законодательство и законодательство Российской Федерации об охране авторского права. Вам необходимо удалить данный файл, а также сообщить издательству «Символ-Плюс» (piracy@symbol.ru), где именно Вы получили данный файл.

Programming Python

Fourth Edition

Mark Lutz

O'REILLY®

Программирование на Python

том II

Четвертое издание

Марк Лутц

Символ®

Санкт-Петербург — Москва
2011

Марк Лутц

Программирование на Python, том II, 4-е издание

Перевод А. Киселева

Главный редактор	*А. Галунов*
Зав. редакцией	*Н. Макарова*
Редактор	*Ю. Бочина*
Корректор	*С. Минин*
Верстка	*Д. Орлова*

Лутц М.

Программирование на Python, том II, 4-е издание. – Пер. с англ. – СПб.: Символ-Плюс, 2011. – 992 с., ил.

ISBN 978-5-93286-211-7

Монументальный труд Марка Лутца представляет собой учебник по применению языка Python в системном администрировании, для создания графических интерфейсов и веб-приложений. Исследуются приемы работы с базами данных, программирования сетевых взаимодействий, создания интерфейсов для сценариев, обработки текста и многие другие. Несмотря на то, что на протяжении всей книги используется язык Python, тем не менее основное внимание уделяется не основам языка, а приемам решения практических задач.

Второй том включает материалы по созданию сценариев для Интернета. Описывается порядок использования сетевых протоколов и инструментов электронной почты на стороне клиента, применение CGI-сценариев, рассматриваются приемы реализации веб-сайтов. Далее обсуждаются дополнительные темы, касающиеся разработки приложений на Python, а именно: технологии хранения информации между запусками программы – файлы DBM, сериализация объектов, хранилища объектов и интерфейсы Python к базам данных SQL; приемы реализации более сложных структур данных на Python – стеков, множеств, двоичных деревьев поиска, графов и др.; инструменты и приемы, используемые в языке Python для синтаксического анализа текстовой информации; приемы интеграции – расширение Python с помощью компилируемых библиотек и встраивание программного кода на Python в другие приложения.

ISBN 978-5-93286-211-7
ISBN 978-0-596-15810-1 (англ)

© Издательство Символ-Плюс, 2011
Authorized Russian translation of the English edition of Programming Python, Fourth Edition ISBN 9780596158101 © 2011 O'Reilly Media, Inc. This translation is published and sold by permission of O'Reilly Media, Inc., the owner of all rights to publish and sell the same.

Все права на данное издание защищены Законодательством РФ, включая право на полное или частичное воспроизведение в любой форме. Все товарные знаки или зарегистрированные товарные знаки, упоминаемые в настоящем издании, являются собственностью соответствующих фирм.

Издательство «Символ-Плюс». 199034, Санкт-Петербург, 16 линия, 7,
тел. (812) 380-5007, www.symbol.ru. Лицензия ЛП N 000054 от 25.12.98.
Подписано в печать 18.10.2011. Формат 70×100 1/16.
Печать офсетная. Объем 62 печ. л.

Оглавление

Часть IV. Создание сценариев для Интернета 15

Глава 12. Сетевые сценарии 17

«Подключись, зарегистрируйся и исчезни» 17

Темы, касающиеся разработки сценариев для Интернета. 19

Опробование примеров этой части книги 22

Другие возможности разработки сценариев для Интернета на языке Python 25

Трубопровод для Интернета 30

Слой сокетов 30

Слой протоколов 32

Библиотечные модули Python для Интернета 36

Программирование сокетов 38

Основы сокетов 40

Запуск программ, использующих сокеты, на локальном компьютере 47

Запуск программ, использующих сокеты, на удаленном компьютере. 48

Параллельный запуск нескольких клиентов 52

Подключение к зарезервированным портам 56

Обслуживание нескольких клиентов 58

Ветвление серверов. 58

Многопоточные серверы 73

Классы серверов в стандартной библиотеке 76

Мультиплексирование серверов с помощью select 79

Подводя итоги: выбор конструкции сервера 87

Придание сокетам внешнего вида файлов и потоков ввода-вывода 88

Вспомогательный модуль перенаправления потоков ввода-вывода. 89

Простой файловый сервер на Python 104

Запуск сервера файлов и клиентов 107

Добавляем графический интерфейс пользователя 108

Глава 13. Сценарии на стороне клиента 119
«Свяжись со мной!» 119
FTP: передача файлов по сети 120
Передача файлов с помощью ftplib 121
Использование пакета urllib для загрузки файлов 125
Утилиты FTP get и put 127
Добавляем пользовательский интерфейс 136
Передача каталогов с помощью ftplib 144
Загрузка каталогов сайта 145
Выгрузка каталогов сайтов 153
Реорганизация сценариев выгрузки и загрузки для многократного использования 158
Передача деревьев каталогов с помощью ftplib 168
Выгрузка локального дерева каталогов 168
Удаление деревьев каталогов на сервере 172
Загрузка деревьев каталогов с сервера 176
Обработка электронной почты 176
Поддержка Юникода в Python 3.X и инструменты электронной почты 177
POP: чтение электронной почты 179
Модуль настройки электронной почты 180
Сценарий чтения почты с сервера POP 183
Извлечение сообщений 186
Чтение почты из интерактивной оболочки 189
SMTP: отправка электронной почты 190
Сценарий отправки электронной почты по SMTP 192
Отправка сообщений 194
Отправка почты из интерактивной оболочки 202
Пакет email: анализ и составление электронных писем 203
Объекты Message 205
Базовые интерфейсы пакета email в действии 208
Юникод, интернационализация и пакет email в Python 3.1 211
Почтовый клиент командной строки 238
Работа с клиентом командной строки pymail 244
Вспомогательный пакет mailtools 249
Файл инициализации 250
Класс MailTool 252
Класс MailSender 252
Класс MailFetcher 262
Класс MailParser 274
Сценарий самотестирования 283
Обновление клиента командной строки pymail 286

NNTP: доступ к телеконференциям 293
HTTP: доступ к веб-сайтам. 296
Еще раз о пакете urllib 300
Другие интерфейсы urllib 302
Прочие возможности создания клиентских сценариев 306

Глава 14. Почтовый клиент PyMailGUI 309
«Пользуйся исходными текстами, Люк!» 309
Модули с исходными текстами и их объем 310
Зачем нужен PyMailGUI? 314
Запуск PyMailGUI 316
Стратегия представления 317
Основные изменения в PyMailGUI 318
Новое в версиях 2.1 и 2.0 (третье издание) 318
Новое в версии 3.0 (четвертое издание) 320
Демонстрация PyMailGUI. 330
Запуск 331
Загрузка почты 338
Многопоточная модель выполнения 339
Интерфейс загрузки с сервера 343
Обработка без подключения к Интернету, сохранение и открытие 344
Отправка почты и вложений 347
Просмотр электронных писем и вложений 351
Ответ на сообщения, пересылка и особенности адресации 359
Удаление сообщений 365
Номера POP-сообщений и синхронизация 367
Обработка содержимого электронной почты в формате HTML 369
Поддержка интернационализации содержимого 371
Альтернативные конфигурации и учетные записи 376
Многооконный интерфейс и сообщения о состоянии 377
Реализация PyMailGUI 380
PyMailGUI: главный модуль 382
SharedNames: глобальные переменные программы 385
ListWindows: окна со списками сообщений 387
ViewWindows: окна просмотра сообщений 409
messagecache: менеджер кэша сообщений 421
popuputil: диалоги общего назначения 425
wraplines: инструменты разбиения строк. 427
html2text: извлечение текста из разметки HTML (прототип, предварительное знакомство) 430
mailconfig: настройки пользователя 433

textConfig: настройка окон редактора PyEdit 440
PyMailGUIHelp: текст справки и ее отображение 440
altconfigs: настройка нескольких учетных записей 444
Идеи по усовершенствованию 447

Глава 15. Сценарии на стороне сервера 460
«До чего же запутанную паутину мы плетем...» 460
Что такое серверный CGI-сценарий? 462
Притаившийся сценарий 462
Создание CGI-сценариев на языке Python 464
Запуск примеров серверных сценариев 466
Выбор веб-сервера 467
Использование локального веб-сервера 467
Корневая страница с примерами на стороне сервера 470
Просмотр примеров серверных сценариев и их вывода 471
Вверх к познанию CGI 472
Первая веб-страница 472
Первый CGI-сценарий 479
Добавление картинок и создание таблиц 486
Добавление взаимодействия с пользователем 489
Табличная верстка форм 499
Добавление стандартных инструментов ввода 506
Изменение размещения элементов формы ввода 510
Передача параметров в жестко определенных адресах URL 513
Передача параметров в скрытых полях форм 516
Сохранение информации о состоянии в сценариях CGI 518
Параметры запроса в строке URL 520
Скрытые поля форм 521
HTTP «Cookies» 522
Базы данных на стороне сервера 527
Расширения модели CGI 528
Комбинирование приемов 529
Переключатель «Hello World» 530
Проверка отсутствующих или недопустимых данных 538
Рефакторинг программного кода с целью облегчения его сопровождения 540
Шаг 1: совместное использование объектов разными страницами – новая форма ввода 542
Шаг 2: многократно используемая утилита имитации формы 545
Шаг 3: объединим все вместе – новый сценарий ответа 549
Подробнее об экранировании HTML и URL 551

Соглашения по экранированию адресов URL 552
Инструменты Python для экранирования HTML и URL 553
Экранирование разметки HTML 554
Экранирование адресов URL 555
Экранирование адресов URL, встроенных в разметку HTML 556
Передача файлов между клиентами и серверами 561
Отображение произвольных файлов сервера на стороне клиента 563
Выгрузка файлов клиента на сервер 571
Как же все-таки протолкнуть биты через Сеть 583

Глава 16. Сервер PyMailCGI 585
«Список дел на поездку в Чикаго» 585
Веб-сайт PyMailCGI 586
Обзор реализации 586
Новое в версии для четвертого издания (версия 3.0) 590
Новое в версии для предыдущего издания (версия 2.0) 593
Обзорное представление программы 595
Опробование примеров из этой главы 595
Корневая страница 598
Настройка PyMailCGI 601
Отправка почты по SMTP 602
Страница составления сообщений 602
Сценарий отправки почты 603
Страницы с сообщениями об ошибках 607
Единство внешнего вида 608
Использование сценария отправки почты без броузера 609
Чтение электронной почты по протоколу POP 611
Страница ввода пароля POP 611
Страница выбора почты из списка 613
Передача информации о состоянии в параметрах URL-ссылки 617
Протоколы защиты данных 620
Страница просмотра сообщений 622
Передача информации о состоянии в скрытых полях форм HTML 626
Экранирование текста сообщения и паролей в HTML 628
Обработка загруженной почты 630
Ответ и пересылка 632
Удаление 633
Операция удаления и номера POP-сообщений 637
Вспомогательные модули 642

Внешние компоненты и настройки 643
Интерфейс к протоколу POP 644
Шифрование паролей 645
Общий вспомогательный модуль 655
Преимущества и недостатки сценариев CGI 661
PyMailGUI и PyMailCGI 662
Веб-приложения и настольные приложения 663
Другие подходы 667

Часть V. Инструменты и приемы 673

Глава 17. Базы данных и постоянное хранение 675
«Дайте мне приказ стоять до конца, но сохранить данные» 675
Возможности постоянного хранения данных в Python 676
Файлы DBM 677
Работа с файлами DBM 678
Особенности DBM: файлы, переносимость и необходимость закрытия 681
Сериализованные объекты 682
Применение сериализации объектов 683
Сериализация в действии 685
Особенности сериализации: протоколы, двоичные режимы и модуль _pickle 688
Файлы shelve 690
Использование хранилищ 691
Сохранение объектов встроенных типов в хранилищах 693
Сохранение экземпляров классов в хранилищах 694
Изменение классов хранимых объектов 696
Ограничения модуля shelve 698
Ограничения класса Pickler 700
Другие ограничения хранилищ модуля shelve 702
Объектно-ориентированная база данных ZODB 702
Сильно сокращенный учебник по ZODB 704
Интерфейсы баз данных SQL 707
Обзор интерфейса SQL 709
Учебник по API базы данных SQL на примере SQLite 712
Создание словарей записей 719
Объединяем все вместе 724
Загрузка таблиц базы данных из файлов 725
Вспомогательные сценарии SQL 729
Ресурсы SQL 737
ORM: механизмы объектно-реляционного отображения 738
PyForm: просмотр хранимых объектов (внешний пример) 740

Глава 18. Структуры данных . . . 743
«Розы – красные, фиалки – голубые; списки изменяемы, а также и класс Foo» . . . 743
Реализация стеков . . . 744
Встроенные возможности . . . 745
Модуль stack . . . 747
Класс Stack . . . 749
Индивидуальная настройка: мониторинг производительности . . . 752
Оптимизация: стеки в виде деревьев кортежей . . . 753
Оптимизация: непосредственная модификация списка в памяти . . . 755
Хронометраж усовершенствований . . . 757
Реализация множеств . . . 760
Встроенные возможности . . . 761
Функции множеств . . . 763
Классы множеств . . . 765
Оптимизация: перевод множеств на использование словарей . . . 766
Алгебра отношений для множеств (внешний пример) . . . 770
Создание подклассов встроенных типов . . . 771
Двоичные деревья поиска . . . 774
Встроенные возможности . . . 774
Реализация двоичных деревьев . . . 775
Деревья с ключами и значениями . . . 778
Поиск на графах . . . 779
Реализация поиска на графе . . . 780
Перевод графов на классы . . . 782
Перестановки последовательностей . . . 785
Обращение и сортировка последовательностей . . . 787
Реализация обращения . . . 788
Реализация сортировки . . . 789
Структуры данных в сравнении со встроенными типами: заключение . . . 791
PyTree: универсальное средство просмотра деревьев объектов . . . 793

Глава 19. Текст и язык . . . 796
«Пилите, Шура, пилите!» . . . 796
Стратегии обработки текста в Python . . . 797
Строковые методы . . . 798
Обработка шаблонов с помощью операций замены и форматирования . . . 800
Анализ текста с помощью методов split и join . . . 801

Суммирование по колонкам в файле 802
Синтаксический анализ строк правил и обратное преобразование 805
Поиск по шаблонам регулярных выражений 809
Модуль re 810
Первые примеры 810
Строковые операции и шаблоны 813
Использование модуля re 816
Дополнительные примеры шаблонов 822
Поиск совпадений с шаблонами в файлах заголовков C 824
Синтаксический анализ XML и HTML 826
Анализ XML 827
Анализ HTML 834
Дополнительные инструменты синтаксического анализа 837
Парсеры, написанные вручную 840
Грамматика выражений 841
Реализация парсера 842
Добавление интерпретатора дерева синтаксического анализа 850
Структура дерева синтаксического анализа 856
Исследование деревьев синтаксического анализа с помощью PyTree 858
Парсеры и возможности Python 859
PyCalc: программа/объект калькулятора 860
Графический интерфейс простого калькулятора 860
PyCalc – графический интерфейс «настоящего» калькулятора 866

Глава 20. Интеграция Python/C 889
«Я заблудился в C» 889
Расширение и встраивание 890
Расширения на C: обзор 893
Простой модуль расширения на C 894
Генератор интегрирующего программного кода SWIG 899
Простой пример SWIG 900
Создание оберток для функций окружения C 905
Добавление классов-оберток в простые библиотеки 909
Обертывание функций окружения C с помощью SWIG 910
Обертывание классов C++ с помощью SWIG 912
Простое расширение с классом C++ 913
Обертывание классов C++ с помощью SWIG 916
Использование класса C++ в Python 918
Другие инструменты создания расширений 923

Встраивание Python в C: обзор 928
Обзор API встраивания в C 928
Что представляет собой встроенный код? 930
Основные приемы встраивания 932
Выполнение простых строк программного кода 933
Выполнение строк программного кода
с использованием результатов и пространств имен 937
Вызов объектов Python 939
Выполнение строк в словарях 941
Предварительная компиляция строк в байт-код 943
Регистрация объектов для обработки обратных вызовов 945
Реализация регистрации 947
Использование классов Python в программах C 952
Другие темы интеграции 955

Часть VI. Финал 959

Глава 21. Заключение: Python и цикл разработки 961
«Книга заканчивается, пора уже и о смысле жизни» 962
«Как-то мы неправильно программируем компьютеры» 963
«Фактор Гиллигана» 963
Делать правильно 964
Цикл разработки для статических языков 965
Искусственные сложности 965
Одним языком не угодишь всем 965
И тут появляется Python 966
А как насчет того узкого места? 968
Python обеспечивает цикл разработки
без промежуточных стадий 968
Python является «выполняемым псевдокодом» 970
Python – это правильное ООП 970
Python способствует созданию гибридных приложений 971
По поводу потопления «Титаника» 973
Так что же такое Python: продолжение 975
Заключительный анализ... 976

Алфавитный указатель 978

IV

Создание сценариев для Интернета

В этой части книги рассматриваются роль Python как языка программирования приложений для Интернета и инструменты в его библиотеке для поддержки этой роли. Попутно привлекаются к использованию инструменты конструирования графических интерфейсов, представленные ранее в книге. Поскольку это популярная область применения Python, главы данной части охватывают все направления:

Глава 12

Здесь будут представлены основные понятия, связанные с Интернетом, низкоуровневые сетевые инструменты Python, такие как сокеты, а также основы программирования архитектуры клиент-сервер.

Глава 13

В этой главе показано, как сценарии могут использовать инструменты языка Python для доступа к стандартным сетевым протоколам клиента, таким как FTP, HTTP, протоколы электронной почты и другие.

Глава 14

Эта глава демонстрирует использование клиентских инструментов электронной почты, описанных в предыдущей главе, а также приемы конструирования графических интерфейсов из предыдущей части книги для реализации полнофункционального клиента электронной почты.

Глава 15

Эта глава освещает основы создания CGI-сценариев на языке Python, выполняемых на стороне сервера, – программ, используемых для реализации интерактивных веб-сайтов.

Глава 16

Эта глава демонстрирует приемы реализации веб-сайтов с помощью Python на примере реализации веб-интерфейса для доступа к электронной почте, отчасти в противовес и для сравнения с обычным решением, представленным в главе 14.

Хотя это и не имеет прямого отношения к данной книге, тем не менее в главе 12 также дается краткий обзор дополнительных инструментов Python для создания интернет-приложений, таких как Jython, Django, App Engine, Zope, PSP, pyjamas и HTMLgen, более полное описание которых вы найдете в соответствующих ресурсах. Здесь вы узнаете, что требуется знать, чтобы использовать такие инструменты, когда вы будете готовы перейти к ним.

Попутно мы также будем использовать общие концепции программирования, такие как объектно-ориентированное программирование (ООП), рефакторинг программного кода и повторное его использование. Как мы увидим далее, Python, графические интерфейсы и сетевые инструменты составляют мощную комбинацию.

12

Сетевые сценарии

«Подключись, зарегистрируйся и исчезни»

За последние 15 с лишним лет, прошедших с момента публикации первого издания этой книги, Интернет буквально вырвался на авансцену. Сеть быстро превратилась из простого средства обмена данными, используемого преимущественно учеными и исследователями, в средство массовой информации, ставшее почти таким же вездесущим, как телевидение и телефон. Социологи сравнивают Интернет с периодической печатью по культурному воздействию, а технические комментаторы считают, что все разработки программного обеспечения, заслуживающие внимания, связаны только с Интернетом. Конечно, только время окончательно рассудит, насколько справедливы такие заявления, но нет никаких сомнений, что Интернет является важным общественным фактором и одной из главных сфер приложения современных программных систем.

Интернет оказался также одной из основных областей применения языка программирования Python. За полтора десятилетия, прошедших с момента публикации первого издания этой книги, развитие Интернета неуклонно влияло на комплект инструментов языка Python и его роли. При наличии Python и компьютера, подключенного к Интернету через сокеты, можно написать сценарии Python для чтения и отправки электронной почты в любую точку земного шара, загрузки веб-страниц с удаленных сайтов, передачи файлов по FTP, программирования интерактивных сайтов, синтаксического анализа файлов HTML и XML, а также многого другого, просто используя модули поддержки Интернета, поставляемые в составе стандартной библиотеки Python.

В действительности многие компании со всего света так и действуют: Google, YouTube, Walt Disney, Hewlett-Packard, JPL и многие другие используют стандартные средства Python для обеспечения работоспособности своих сайтов. Например, поисковая система Google, известная своими усилиями по повышению удобства использования Интернета, широко использует Python в своей работе. Сайт видеороликов YouTube в значительной степени реализован на языке Python. Система BitTorrent обмена файлами, реализованная на языке Python и используемая десятками миллионов пользователей, эффективно использует сетевые инструменты языка Python для организации обмена файлами между клиентами и снятия нагрузки с серверов.

Многие также строят свои сайты и управляют ими с помощью крупных наборов инструментальных средств на основе Python. Например, одним из первых в этой области был сервер веб-приложений Zope, который сам написан на Python и может индивидуально настраиваться с его помощью. Другие создают сайты на основе системы управления содержимым Plone, которая построена на основе Zope, и делегируют управление содержимым сайта своим пользователям. Есть те, кто использует Python для управления веб-приложениями на языке Java посредством Jython (ранее был известен как JPython) – системы, которая компилирует программы Python в байт-код Java, экспортирует библиотеки Java для использования в сценариях на языке Python и позволяет программному коду на языке Python играть роль веб-апплетов, загружаемых и выполняемых в броузере.

В последние годы видное место в сфере разработки веб-приложений заняли новые приемы и системы. Например, интерфейсы XML-RPC и SOAP для Python обеспечивают возможность реализации веб-служб; появились мощные фреймворки для разработки веб-сайтов, такие как Google App Engine, Django и Turbo Gears; пакет XML в стандартной библиотеке Python, а также сторонние расширения предоставляют целый комплекс инструментов для работы с XML, а реализация IronPython обеспечивает тесную интеграцию Python с .NET/Mono подобно тому, как Jython обеспечивает интеграцию с библиотеками Java.

С ростом Интернета выросла и роль Python как инструмента Интернета. Язык Python хорошо подошел для создания сценариев, работающих с Интернетом, по тем же причинам, которые делают его идеальным в других областях. Его модульная архитектура и короткий цикл разработки хорошо соответствуют напряженным требованиям создания приложений для Интернета. В этой части книги мы увидим, что Python не просто поддерживает возможность создания сценариев для Интернета, но и благоприятствует высокой производительности труда разработчиков и легкости сопровождения, которые важны для интернет-проектов всех видов и масштабов.

Темы, касающиеся разработки сценариев для Интернета

Интернет-программирование охватывает много разных тем, поэтому, чтобы облегчить усвоение материала, этот предмет был разбит на пять глав. Ниже приводится краткий обзор содержимого глав этой части книги:

- Данная глава знакомит с основами Интернета и исследует *сокеты* – механизм взаимодействий, лежащий в основе Интернета. Мы уже кратко знакомились с сокетами как с инструментом IPC в главе 5 и второй раз встречались с ними в главе 10. Здесь мы рассмотрим их более детально и исследуем их роль в сетевых взаимодействиях.
- В главе 13 мы перейдем к обсуждению создания *клиентских сценариев* и протоколов Интернета. Мы исследуем поддержку протоколов FTP, электронной почты, HTTP, NNTP и других в стандартной библиотеке Python.
- В главе 14 будет представлен более крупный пример клиентского сценария: PyMailGUI – полнофункциональный клиент электронной почты.
- В главе 15 будут обсуждаться основы создания *серверных сценариев* и конструирования веб-сайтов. Здесь мы будем изучать приемы и понятия создания CGI-сценариев, которые лежат в основе большинства из того, что происходит во Всемирной паутине.
- В главе 16 будет представлен крупный пример серверного сценария: PyMailCGI – полнофункциональный клиент электронной почты с веб-интерфейсом.

Каждая глава предполагает знакомство с предыдущей, но вообще их можно читать в произвольном порядке, особенно при наличии некоторого опыта работы с Интернетом. Так как эти главы составляют значительную часть книги, в следующих разделах приводятся еще некоторые подробности о темах, которые нам предстоит изучать.

О чем будет рассказано

Концептуально Интернет можно представить себе состоящим из нескольких функциональных слоев:

Сетевые слои низкого уровня

Механизмы, такие как транспортный уровень TCP/IP, занимающиеся пересылкой байтов между устройствами, но не выполняющие их интерпретацию

Сокеты

Программный интерфейс доступа к сети, действующий поверх физических сетевых слоев типа TCP/IP и поддерживающий гибкие модели *клиент/сервер* для организации взаимодействий между процессами и обмена данными по сети.

Протоколы верхнего уровня

Структурированные схемы обмена информацией через Интернет, такие как FTP и электронная почта, действующие поверх сокетов и определяющие форматы сообщений и стандарты адресации

Серверные веб-сценарии (CGI)

Прикладные модели, такие как CGI, определяющие порядок взаимодействий между веб-броузерами и веб-серверами, также действующие поверх сокетов и поддерживающие понятие веб-программ

Фреймворки и инструменты высокого уровня

Системы сторонних разработчиков, такие как Django, App Engine, Jython и pyjamas, также использующие сокеты и протоколы обмена данными, но предназначенные для решения значительно более широкого круга задач

В данной книге рассматриваются *средние три слоя* из этого списка – сокеты, протоколы Интернета, основанные на них, и CGI-модель взаимодействий в Сети. Все, что мы будем изучать здесь, в равной степени относится и к более специализированным наборам инструментов, находящихся на последнем уровне, потому что все они, в конечном счете, опираются на те же самые основные принципы Интернета и Сети.

Основное внимание в этой и в следующей главе будет уделено программированию второго и третьего слоев: *сокетов* и *протоколов* высокого уровня. В этой главе мы начнем с самого низа и изучим модель программирования сетевых взаимодействий с применением сокетов. Программирование в Интернете может быть связано не только с сокетами, как мы видели в примерах организации взаимодействий между процессами в главе 5, но они представлены здесь так полно, потому что это их главная роль. Как будет показано, большая часть происходящего в Интернете осуществляется с помощью сокетов, даже когда это не бросается в глаза.

После знакомства с сокетами в следующих двух главах мы сделаем шаг вперед и познакомимся с интерфейсами Python на стороне клиента к протоколам более высокого уровня, таким как протоколы электронной почты и FTP, – действующим поверх сокетов. Оказывается, с помощью Python многое можно сделать на стороне клиента, и в главах 13 и 14 будут представлены образцы сценариев на языке Python, выполняемых на стороне клиента. В последних двух главах, завершающих эту часть, будут представлены сценарии, выполняемые на *стороне сервера*, – программы, выполняемые на компьютере-сервере и обычно запускаемые веб-броузерами.

О чем рассказываться не будет

Теперь, после рассказа о том, какие темы будут освещаться в книге, я должен ясно определить, о чем мы говорить не будем. Как и tkinter, Интернет – это обширная тема, и эта часть книги по большей части посвящена представлению базовых понятий и исследованию типичных

задач. Поскольку существует огромное количество модулей для Интернета, я не стремился превратить эту книгу в исчерпывающий справочник по этой области. Даже стандартная библиотека Python содержит слишком много модулей для Интернета, чтобы о каждом из них можно было рассказать в этой книге.

Кроме того, инструменты более высокого уровня, такие как Django, Jython и App Engine, являются самостоятельными большими системами, и для их изучения лучше обратиться к документации, в большей мере ориентированной именно на них. Поскольку по этим темам ныне существуют специальные книги, мы лишь слегка пройдемся по ним в кратком обзоре далее в этой главе. Кроме того, данная книга почти не затрагивает сетевые слои более низкого уровня, такие как TCP/IP. Если вам любопытно, что происходит в Интернете на уровне битов и проводов, обратитесь за подробностями к хорошему учебнику по сетям.

Иными словами, эта часть книги не является исчерпывающим справочником по разработке интернет- и веб-приложений на языке Python – сферы, которая бурно развивалась между предыдущими изданиями этой книги и несомненно продолжит свое развитие после выхода и этого издания. Вместо этого данная часть книги нацелена на то, чтобы служить вводным руководством в данную область программирования, позволяющим начать работать, а представленные здесь примеры помогут разобраться в документации к инструментам, которые вам может потребоваться исследовать после освоения основ.

Другие темы, рассматриваемые в этой части книги

Как и в других частях книги, в этой части также затрагиваются темы, отличные от основных. По ходу дела эта часть вовлекает в работу некоторые изученные ранее интерфейсы операционной системы и приемы создания графических интерфейсов (например, процессы, потоки выполнения, сигналы и tkinter). Мы также увидим применение языка Python при создании реалистичных программ и исследуем некоторые конструктивные идеи и серьезные задачи, порождаемые Интернетом.

В связи с последним заявлением следует сказать еще несколько слов. Разработка сценариев для Интернета, как и сценариев с графическим интерфейсом, – одна из наиболее заманчивых областей применения языка Python. Как и при работе с графическими интерфейсами, возникает неуловимое, но непосредственное чувство удовлетворения, когда видишь, как сетевая программа на языке Python распространяет информацию по всему свету. С другой стороны, сетевое программирование по самой своей природе влечет издержки, связанные со скоростью передачи, и ограничения в пользовательских интерфейсах. Некоторые приложения все же лучше не разворачивать в Сети, хотя такая позиция сегодня не в моде.

Традиционные «настольные» приложения с графическим интерфейсом, подобные тем, что были представлены в третьей части книги, способны совмещать в себе богатство и скорость отклика клиентских библиотек

с мощью сетевых взаимодействий. С другой стороны, веб-приложения предоставляют непревзойденную переносимость и простоту сопровождения. В этой части книги мы честно рассмотрим компромиссы, на которые приходится идти при работе в Сети, и исследуем примеры, иллюстрирующие преимущества обычных и веб-приложений. Фактически, крупные примеры приложений PyMailGUI и PyMailCGI, которые нам предстоит исследовать, отчасти служат именно этой цели.

Кроме того, многие считают Интернет чем-то вроде окончательной проверки идеи для инструментов с открытыми исходными текстами. Действительно, работа Сети в значительной мере основана на применении большого числа таких инструментов, как Python, Perl, веб-сервер Apache, программа sendmail, MySQL и Linux.[1] Более того, иногда кажется, что новые инструменты и технологии веб-программирования появляются быстрее, чем разработчики успевают их освоить.

Положительной стороной Python является нацеленность на интеграцию, делающая его самым подходящим инструментом в таком разнородном мире. В настоящее время программы на языке Python могут устанавливаться как инструменты на стороне клиента или сервера, использоваться в качестве апплетов и сервлетов в приложениях на языке Java, встраиваться в распределенные системы объектов, такие как CORBA, SOAP и XML-RPC, интегрироваться в приложения, использующие технологию AJAX, и так далее. Говоря более общим языком, основания для использования Python в Интернете точно такие же, как и в любых других областях, – акцент на качестве, производительности труда, переносимости и интеграции превращает язык Python в идеальное средство для написания программ для Интернета, которые общедоступны, просты в сопровождении и могут разрабатываться в очень сжатые сроки, характерные для этой области.

Опробование примеров этой части книги

Интернет-сценарии обычно предполагают контексты выполнения, которые не требовались для предыдущих примеров этой книги. Это означает, что опробование программ, взаимодействующих через сеть, часто оказывается несколько более сложным. Заранее приведем несколько практических замечаний относительно примеров из этой части книги:

- Для выполнения примеров из этой части книги не требуется загружать дополнительные пакеты. Все будущие примеры основываются

[1] Существует даже специальная аббревиатура LAMP для обозначения связки инструментов: операционная система Linux, веб-сервер Apache, система управления базами данных MySQL и языки сценариев Python, Perl и PHP. Не только возможно, но и часто встречается на практике, что веб-сервер уровня предприятия создается целиком на основе открытых инструментов. Пользователи Python могут также включить в этот список такие системы, как Zope, Django, Webware и CherryPy, при этом получившийся акроним может немного растянуться.

на стандартном наборе модулей поддержки Интернета, поставляемом вместе с Python, которые устанавливаются в каталог стандартной библиотеки Python.

- Для опробования большинства примеров из этой части книги не требуется сверхсовременного подключения к Сети или наличия учетной записи на веб-сервере. Хотя для демонстрации некоторые примеры использования сокетов будут запускаться удаленно, тем не менее, большинство из них могут запускаться на локальном компьютере. Примеры клиентских сценариев, демонстрирующих работу с такими протоколами, как FTP, требуют лишь простого доступа в Интернет, а примеры работы с электронной почты требуют лишь наличия POP и SMTP серверов.
- Вам не требуется иметь учетную запись на веб-сервере, чтобы запускать серверные сценарии, приведенные в последних главах, – они могут запускаться любым веб-броузером. Такая учетная запись может потребоваться, чтобы изменять эти сценарии, если вы решите хранить их на удаленном веб-сервере, но она не нужна, если вы просто используете локальный веб-сервер, как в примерах в этой книге.

В процессе обсуждения мы будем рассматривать детали настройки окружения, но вообще, когда сценарий Python открывает соединение с Интернетом (с помощью модуля `socket` или модулей поддержки протоколов Интернета), Python довольствуется любым соединением, которое существует на компьютере, будь то выделенная линия T1, линия DSL или простой модем. Например, открытие сокета на компьютере с Windows при необходимости автоматически инициирует соединение с поставщиком услуг Интернета.

Более того, если ваша платформа поддерживает сокеты, то, вероятно, она сможет выполнить многие из приведенных здесь примеров, даже если соединение с Интернетом вообще отсутствует. Как мы увидим, имя компьютера `localhost` или `""` (пустая строка) обычно означает сам локальный компьютер. Это позволяет тестировать как клиентские, так и серверные сценарии на одном и том же компьютере, не подключаясь к Сети. Например, на компьютере, работающем под управлением Windows, клиенты и серверы могут выполняться локально без выхода в Сеть. Иными словами, вы наверняка сможете опробовать программы, представленные здесь, независимо от наличия соединения с Интернетом.

В некоторых последующих примерах предполагается, что на компьютере сервера выполняется определенный тип сервера (например, FTP, POP, SMTP), но сценарии на стороне клиента работают на любом компьютере, подключенном к Интернету, с установленным на нем интерпретатором Python. Примеры серверных сценариев в главах 15 и 16 требуют большего: для разработки CGI-сценариев необходимо либо иметь учетную запись на веб-сервере, либо установить локальный веб-сервер (что на самом деле проще, чем вы думаете, – простой веб-сервер на языке Python будет представлен в главе 15). Дополнительные сторонние

системы, такие как Jython и Zope, естественно, придется загружать отдельно – некоторые из них мы кратко рассмотрим в этой главе, но оставим более подробное их описание за соответствующей документацией.

Вначале был Грааль

Помимо создания языка Python несколько лет тому назад Гвидо ван Россум написал на языке Python веб-броузер, названный (весьма уместно) Grail (Грааль). Отчасти Grail разрабатывался для демонстрации возможностей языка Python. Он дает пользователям возможность бродить по Сети, как с помощью Firefox или Internet Explorer, но может также расширяться апплетами Grail – программами Python/tkinter, загружаемыми с сервера, когда броузер клиента обращается к ним и запускает их. Апплеты Grail действуют во многом аналогично Java-апплетам в более популярных броузерах (подробнее об апплетах рассказывается в следующем разделе).

Хотя реализация этого броузера и была адаптирована для работы под управлением последних версий Python, тем не менее, Grail больше не развивается и используется сегодня в основном в исследовательских целях (в действительности, он является современником броузера Netscape). Но Python по-прежнему пожинает плоды проекта Grail в виде богатого набора инструментов для Интернета. Чтобы написать полноценный веб-броузер, необходимо обеспечить поддержку большого количества протоколов Интернета, и Гвидо оформил их поддержку в виде стандартных библиотечных модулей, поставляемых в настоящее время с языком Python.

Благодаря такому наследству в Python теперь имеется стандартная поддержка телеконференций Usenet (NNTP), обработки электронной почты (POP, SMTP, IMAP), пересылки файлов (FTP), веб-страниц и взаимодействий (HTTP, URL, HTML, CGI) и других часто используемых протоколов, таких как Telnet. Сценарии на языке Python могут соединяться со всеми этими компонентами Интернета, просто импортируя соответствующие библиотечные модули.

Уже после появления Grail в библиотеку Python были добавлены дополнительные средства синтаксического анализа файлов XML, защищенных сокетов OpenSSL и другие. Но в значительной мере поддержка Интернета в языке Python ведет свое происхождение от броузера Grail – еще один пример важности поддержки повторного использования программного кода в Python. Когда пишется эта книга, Grail все еще можно отыскать в Интернете, выполнив поиск по строке «Grail web browser».

Другие возможности разработки сценариев для Интернета на языке Python

Существует множество различных способов разработки веб-сценариев на языке Python, хотя описание многих из них выходит далеко за рамки этой книги. Как и в части книги, посвященной графическим интерфейсам, я хочу начать с краткого обзора наиболее популярных инструментов, используемых в этой области, прежде чем перейти к исследованию основ.

Сетевые инструменты

Как мы уже видели выше в этой главе, в состав Python входят инструменты поддержки простых сетевых взаимодействий, а также реализация некоторых видов сетевых серверов. В число этих инструментов входят *сокеты*, функция *select*, используемая для организации асинхронной работы серверов, а также готовые *классы серверов на основе сокетов*. Все эти инструменты сосредоточены в стандартных модулях `socket`, `select` и `socketserver`.

Инструменты поддержки протоколов на стороне клиента

Как мы увидим в следующей главе, арсенал инструментов для работы в Интернете языка Python также включает поддержку большинства стандартных протоколов Интернета на стороне клиента – сценарии легко могут использовать протоколы электронной почты, FTP, HTTP, Telnet и многие другие. В сочетании с настольными приложениями с графическим интерфейсом, подобными тем, с которыми мы встречались в предыдущей части книги, эти инструменты открывают путь к созданию полнофункциональных и отзывчивых приложений для работы с Сетью.

Серверные CGI-сценарии

CGI-сценарии, которые являются, пожалуй, самым простым способом реализации интерактивных веб-сайтов, представляют прикладную модель выполнения сценариев на стороне сервера для обработки данных форм ввода, выполнения операций на основе полученной информации и воспроизведения страниц с ответами. Мы будем использовать их далее в этой части книги. Данная модель имеет непосредственную поддержку в стандартной библиотеке Python, является основой большей части происходящего в Сети и вполне подходит для разработки простых сайтов. Сам механизм CGI не решает такие проблемы, как сохранение информации о состоянии между обращениями к страницам и параллельное обновление данных, но использование в CGI-сценариях таких инструментов, как cookies и базы данных, позволяет успешно решать эти проблемы.

Веб-фреймворки и «облака»

При создании более сложных веб-приложений фреймворки могут взять на себя значительную часть низкоуровневых операций и пре-

доставить более структурированные и мощные приемы реализации динамических сайтов. Кроме простых CGI-сценариев, мир Python полон сторонних веб-фреймворков, таких как *Django* – высокоуровневый фреймворк, поддерживающий быструю разработку, имеющий понятную и практичную архитектуру, обеспечивающий прикладной интерфейс динамического доступа к базе данных и свой собственный язык шаблонов для использования на стороне сервера; *Google App Engine* – фреймворк «облачных вычислений», предоставляющий инструменты уровня предприятия для использования в сценариях на языке Python и позволяющий сайтам использовать веб-инфраструктуру Google; и *Turbo Gears* – интегрированная коллекция инструментов, в число которых входят библиотека JavaScript, система шаблонов, инструмент веб-взаимодействий CherryPy и механизм SQLObject доступа к базам данных, использующий модель классов Python.

К категории фреймворков также относятся *Zope* – открытый сервер веб-приложений и набор инструментов, написанный на языке Python и расширяемый с его помощью, при использовании которого веб-сайты реализуются с применением базовой объектно-ориентированной модели; *Plone* – конструктор веб-сайтов на основе Zope, который предоставляет реализацию модели документооборота (называется системой управления содержимым), позволяющую авторам добавлять новое содержимое на сайт; и другие популярные системы конструирования веб-сайтов, включая pylons, web2py, CherryPy и Webware.

Многие из этих фреймворков основаны на архитектуре MVC (model-view-controller – модель-представление-контроллер), получившей широкое распространение, и большинство из них предоставляют решение проблемы сохранения информации о состоянии в базе данных. Некоторые из них используют модель ORM (*object relational mapping – объектно-реляционного отображения*), с которой мы познакомимся в следующей части книги. Эта модель обеспечивает отображение классов Python в таблицы реляционной базы данных, а фреймворк Zope хранит объекты сайта в объектно-ориентированной базе данных ZODB, которую мы также будем рассматривать в следующей части.

Полнофункциональные интернет-приложения (еще раз)

Обсуждавшиеся в начале главы 7 новейшие системы разработки «полнофункциональных интернет-приложений» (Rich Internet Application, RIA), такие как *Flex*, *Silverlight*, *JavaFX* и *pyjamas*, позволяют создавать пользовательские интерфейсы в веб-броузерах и обеспечивают более высокую динамичность и функциональность в сравнении с традиционными страницами HTML. Эти решения, применяемые на стороне клиента, основаны, как правило, на технологии AJAX и JavaScript и предоставляют наборы виджетов, которые способны конкурировать с традиционными «настольными» графическими интерфейсами и поддерживают механизмы асинхронных взаимо-

действий с веб-серверами. Согласно утверждениям некоторых наблюдателей такая интерактивность является важной составляющей модели «Web 2.0».

В конечном счете веб-броузер также является «настольным» приложением с графическим интерфейсом, но получившим весьма широкое распространение и позволяющим использовать его в качестве платформы для отображения других графических интерфейсов с помощью приемов полнофункциональных интернет-приложений и с использованием программных слоев, которые не опираются на использование какой-то определенной библиотеки создания графических интерфейсов. Благодаря этому технологии создания полнофункциональных интернет-приложений способны превратить веб-броузеры в расширяемые приложения с графическим интерфейсом.

По крайней мере, это основная их цель в настоящее время. По сравнению с традиционными графическими интерфейсами полнофункциональные интернет-приложения пользуются преимуществом переносимости и простоты развертывания в обмен на снижение производительности и увеличение сложности стека программного обеспечения. Кроме того, как и в мире традиционных графических интерфейсов, в сфере полнофункциональных интернет-приложений уже наблюдается конкуренция инструментов, которые могут добавлять дополнительные зависимости и оказывать влияние на переносимость. Пока не появится явный лидер, для использования полнофункционального интернет-приложения может требоваться выполнять этап установки, типичный для традиционных приложений.

Впрочем, продолжайте следить за событиями – история полнофункциональных интернет-приложений, как и Всемирной паутины в целом, продолжает развиваться. Появление стандарта HTML5, например, который, вероятнее всего, еще несколько лет не сможет занять господствующее положение, может, в конечном счете, ликвидировать необходимость в расширениях поддержки полнофункциональных интернет-приложений для броузеров.

Веб-службы: XML-RPC, SOAP

XML-RPC – это технология, обеспечивающая возможность вызова процедур удаленных компонентов через сеть. Она отправляет запросы по протоколу HTTP и передает данные туда и обратно в виде XML-документов. Для клиентов веб-серверы выглядят как простые функции – когда производится вызов функции, данные упаковываются в формат XML и передаются удаленному серверу с помощью протокола HTTP. Применение этой технологии упрощает реализацию взаимодействий клиентских программ с веб-серверами.

В более широком смысле технология XML-RPC привела к появлению понятия *веб-служб* – программных компонентов многократного пользования, которые выполняются в Веб. XML-RPC поддерживается модулем `xmlrpc.client` в стандартной библиотеке Python, реали-

зующим клиентскую часть этого протокола, и модулем `xmlrcp.server`, содержащим инструменты для реализации серверной части. SOAP – похожий, но в целом более тяжеловесный протокол веб-служб, поддержка которого в языке Python доступна в виде сторонних пакетов *SOAPy* и *ZSI*, среди прочих.

Брокеры объектных запросов CORBA

Более ранняя, но сопоставимая технология, CORBA – это архитектура программирования распределенных вычислений, в которой компоненты взаимодействуют через сеть, направляя вызовы через *брокер объектных запросов* (Object Request Broker, ORB). В языке Python поддержка CORBA доступна в виде стороннего пакета *OmniORB*, а также в виде (по-прежнему доступной, хотя давно уже не поддерживаемой) системы *ILU*.

Java и .NET: Jython и IronPython

Мы уже встречались с Jython и IronPython в начале главы 7, в контексте графических интерфейсов. За счет компиляции сценариев на языке Python в байт-код Java *Jython* позволяет использовать сценарии Python в любых контекстах, где могут использоваться программы Java, в том числе и в роли веб-сценариев, таких как апплеты, хранящихся на стороне сервера, но выполняющихся на стороне клиента при обращении к ним из веб-страниц. Система *IronPython*, также упоминавшаяся в главе 7, предлагает похожие возможности разработки веб-приложений, включая доступ к фреймворку Silverlight полнофункциональных интернет-приложений и его реализации Moonlight в системе Mono для Linux.

Инструменты синтаксического анализа XML и HTML

Хотя документы XML технически не привязаны к Интернету, тем не менее они часто используются на определенных этапах работы с ним. Следуя за многообразием применений XML, мы изучим базовую поддержку синтаксического анализа XML в Python, а также познакомимся со сторонними расширениями, выполняющими эту функцию, в следующей части книги, когда будем исследовать инструменты обработки текста, имеющиеся в Python. Как мы увидим, анализ XML обеспечивается пакетом `xml` из стандартной библиотеки Python, который предоставляет три вида парсеров – DOM, SAX и ElementTree, а среди открытого программного обеспечения можно, кроме того, найти расширения для поддержки XPath и многие другие инструменты. Библиотечный модуль `html.parser` также предоставляет инструмент синтаксического анализа разметки HTML, модель которого похожа на модель SAX для XML. Подобные инструменты используются для анализа и извлечения содержимого веб-страниц, загружаемых с помощью инструментов `urllib.request`.

COM и DCOM в Windows

Пакет *PyWin32* позволяет сценариям на языке Python использовать модель COM в Windows для выполнения таких задач, как редактирование документов Word и заполнение электронных таблиц Excel (имеются также дополнительные инструменты, поддерживающие возможность обработки документов Excel). Хотя это и не имеет прямого отношения к Интернету (и похоже, что в последние годы вытесняется .NET), тем не менее DCOM – распределенное расширение для COM – предлагает дополнительные возможности создания распределенных сетевых приложений.

Другие инструменты

Имеется множество других инструментов, которые играют более узкоспециализированные роли. Среди них: *mod_python* – система, оптимизирующая выполнение серверных сценариев на языке Python в окружении веб-сервера Apache; *Twisted* – асинхронный, управляемый событиями фреймворк для сетевых приложений, написанный на языке Python, который обеспечивает поддержку огромного количества сетевых протоколов и содержит готовые реализации типичных сетевых серверов; *HTMLgen* – легковесный инструмент, позволяющий генерировать разметку HTML из дерева объектов Python, описывающих структуру веб-страницы; и *Python Server Pages* (PSP) – механизм шаблонов, действующий на стороне сервера, позволяющий встраивать программный код Python в разметку HTML, выполнять его в ходе обработки запросов для отображения частей страниц и близко напоминающий PHP, ASP и JSP.

Как вы уже могли догадаться, учитывая особое положение Веб, для языка Python существует такое множество инструментов создания сценариев для Интернета, что обсудить их все просто невозможно в рамках этой книги. За дополнительной информацией по этой теме обращайтесь на веб-сайт PyPI, по адресу *http://python.org/*, или воспользуйтесь своей любимой поисковой системой (которая, возможно, тоже реализована с применением инструментов Python для Интернета).

Напомню, цель этой книги – достаточно подробно охватить основы, чтобы вы могли приступить к использованию инструментов, аналогичных тем, что перечислены выше, когда будете готовы перейти к применению более универсальных решений. Как вы увидите далее, базовая модель CGI-сценариев, с которой мы встретимся здесь, иллюстрирует механизмы, лежащие в основе любых веб-приложений, будь то простые сценарии или сложные фреймворки.

Однако прежде чем мы научимся бегать, нужно научиться ходить, поэтому давайте начнем с самого дна и посмотрим, чем в действительности является Интернет. Современный Интернет покоится на внушительном стеке программного обеспечения – инструменты позволяют

скрыть некоторые сложности, тем не менее, для грамотного программирования по-прежнему необходимо знать все его слои. Как мы увидим далее, развертывание Python в Сети, особенно средствами высокоуровневых веб-фреймворков, подобных тем, что были перечислены выше, возможно только потому, что мы действительно «путешествуем на плечах гигантов».

Трубопровод для Интернета

Если вы не провели последние десять-двадцать лет в пещере, то вы уже должны быть знакомы с Интернетом, по крайней мере, с точки зрения пользователя. Функционально мы используем его как коммуникационную и информационную среду, обмениваясь электронной почтой, просматривая веб-страницы, передавая файлы и так далее. Технически Интернет состоит из набора слоев, как абстрактных, так и функциональных – от реальных проводов для передачи битов по всему свету до веб-броузера, получающего эти биты и отображающего их на компьютере как текст, графику или звуки.

В данной книге нас в основном интересует интерфейс между программистом и Интернетом. Он тоже состоит из нескольких слоев: сокетов, являющихся программными интерфейсами к соединениям низкого уровня между компьютерами, и стандартных протоколов, которые структурируют обмен данными, производящийся через сокеты. Рассмотрим вначале вкратце каждый из этих слоев, а затем погрузимся в детали программирования.

Слой сокетов

Выражаясь простым языком, сокеты служат программным интерфейсом для организации соединений между программами, возможно выполняющимися на разных компьютерах в сети. Они также образуют основу и низкоуровневый «трубопровод» самого Интернета: все известные протоколы Сети верхнего уровня, такие как FTP, веб-страницы и электронная почта, в конечном итоге реализуются через сокеты. Сокеты иногда называют также конечными пунктами коммуникаций, так как они служат порталами, через которые программы посылают и принимают байты во время общения.

Несмотря на то, что сокеты часто используются для организации общения по сети, они точно так же могут использоваться, как механизм общения между программами, выполняющимися на одном и том же компьютере, принимая форму универсального механизма взаимодействий между процессами (Inter-Process Communication, IPC). Мы уже видели такой способ использования сокетов в главе 5. В отличие от некоторых других механизмов IPC, сокеты являются двунаправленными потоками данных: с их помощью программы могут и отправлять, и принимать данные.

Для программиста сокеты принимают форму группы вызовов, доступных через библиотеку. Эти вызовы сокетов умеют пересылать байты между компьютерами, используя низкоуровневые механизмы, такие как сетевой протокол TCP управления передачей данных. На своем уровне TCP умеет передавать байты, но его не заботит, что эти байты означают. Исходя из целей данной книги, мы опускаем вопрос о том, как осуществляется физическая передача байтов, посылаемых в сокеты. Однако для полного понимания сокетов нам потребуется представление о том, как компьютерам назначаются имена.

Идентификаторы компьютеров

Предположим на секунду, что вы хотите поговорить по телефону с кем-то, находящимся на другом конце света. В реальном мире вам, вероятно, потребуется номер телефона этого человека или справочник, в котором этот номер можно найти по его имени. То же справедливо для Интернета: прежде чем сценарий сможет общаться с каким-то другим компьютером в киберпространстве, ему нужно узнать номер или имя другого компьютера.

К счастью, в Интернете предусмотрены стандартные способы именования удаленных компьютеров и служб, предоставляемых этими компьютерами. Внутри сценария компьютерная программа, с которой нужно связаться через сокет, идентифицируется с помощью пары значений – имени компьютера и номера порта на этом компьютере:

Имена компьютеров

Имя компьютера может иметь вид строки из чисел, разделенных точками, называемой IP-адресом (например, `166.93.218.100`), или форму, известную как доменное имя, в более читаемом формате (например, *starship.python.net*). Доменные имена автоматически отображаются в соответствующие цифровые адреса с точками. Это отображение осуществляет сервер доменных имен (DNS-сервер) – программа в Сети, осуществляющая ту же функцию, что и ваша местная телефонная справочная служба. Особый случай представляет имя `localhost` и эквивалентный ему IP-адрес `127.0.0.1`, которые всегда соответствуют локальному компьютеру. Это позволяет обращаться к серверам, действующим на том же компьютере, где выполняется клиент.

Номера портов

Номер порта – это просто согласованный числовой идентификатор данной сетевой службы. Так как компьютеры в Сети могут предоставлять разнообразные услуги, для указания конкретной службы на данном компьютере используются номера портов. Чтобы два компьютера могли общаться через Сеть, при инициации сетевых соединений оба они должны связать сокеты с одним и тем же именем компьютера и номером порта. Как мы увидим далее, за протоколами Интернета, такими как электронная почта и HTTP, зарезервированы стандартные номера портов, благодаря чему клиенты могут за-

прашивать услугу независимо от того, какой компьютер предоставляет ее. Порт с номером 80, например, на любом компьютере веб-сервера обычно используется для обслуживания запросов на получение веб-страниц.

Комбинация из имени компьютера и номера порта однозначно идентифицирует каждую сетевую службу. Например, компьютер провайдера услуг Интернета может предоставлять клиентам различные услуги – веб-страницы, Telnet, передачу файлов по FTP, электронную почту и так далее. Каждой службе на компьютере присвоен уникальный номер порта, на который может посылаться запрос. Для получения веб-страниц с веб-сервера программы должны указывать IP-адрес или доменное имя веб-сервера и номер порта, на котором сервер ждет запросы веб-страниц.

Если все это далеко от вас, попробуйте представить себе ситуацию на бытовом языке. Например, чтобы поговорить по телефону с кем-то внутри компании, обычно требуется набрать номер телефона компании и дополнительный номер того сотрудника, который вам нужен. Если вы не знаете номера компании, то можете поискать его в телефонном справочнике по названию компании. В Сети все происходит почти так же – имена компьютеров идентифицируют наборы служб (как компанию), номера портов идентифицируют отдельные службы на конкретном компьютере (как добавочный номер), а доменные имена отображаются в IP-адреса серверами доменных имен (как телефонные книги).

Когда программы используют сокеты для организации взаимодействий с другим компьютером (или с другими процессами на том же компьютере) особыми способами, в них не должны использоваться порты с номерами, зарезервированными для стандартных протоколов, – числами в диапазоне от 0 до 1023, но чтобы понять, почему это так, нужно сначала разобраться с протоколами.

Слой протоколов

Сокеты образуют костяк Интернета, но значительная часть происходящего в Сети программируется с помощью протоколов[1], являющихся моделями сообщений более высокого уровня, действующими поверх сокетов. Если сказать коротко, то протоколы Интернета определяют структурированный способ общения через сокеты. Обычно они стандартизуют как форматы сообщений, так и номера портов:

- *Форматы сообщений* определяют структуру потока байтов, пересылаемых через сокеты во время обмена данными.

[1] В некоторых книгах термин *протокол* используется также для ссылки на транспортные схемы более низкого уровня, такие как TCP. В данной книге мы называем *протоколами* структуры более высокого уровня, создаваемые поверх сокетов; если вас интересует происходящее на более низких уровнях, обращайтесь к учебникам по сетям.

- *Номера портов* служат зарезервированными числовыми идентификаторами используемых сокетов, через которые происходит обмен сообщениями.

«Сырые» сокеты (raw sockets) все еще часто используются во многих системах, но чаще (и обычно проще) связь осуществляется с помощью одного из стандартных протоколов Интернета высокого уровня. Как мы увидим далее, в языке Python имеется поддержка стандартных протоколов, автоматизирующая большую часть операций по подготовке и отправке сообщений через сокеты.

Правила нумерации портов

Технически номер порта сокета может быть любым 16-битовым целым числом в диапазоне от 0 до 65535. Однако, чтобы облегчить программам поиск стандартных протоколов, порты с номерами 0–1023 зарезервированы и назначены стандартным протоколам высокого уровня. В табл. 12.1 перечислены номера портов, зарезервированные для многих стандартных протоколов; каждый из них получает один или более номеров из зарезервированного диапазона.

Таблица 12.1. Номера портов, зарезервированные для стандартных протоколов

Протокол	Стандартная функция	Номер порта	Модуль Python
HTTP	Веб-страницы	80	`http.client`, `http.server`
NNTP	Телеконференции Usenet	119	`nntplib`
FTP данные	Пересылка файлов	20	`ftplib`
FTP управление передачей	Пересылка файлов	21	`ftplib`
SMTP	Отправка электронной почты	25	`smtplib`
POP3	Получение электронной почты	110	`Poplib`
IMAP4	Получение электронной почты	143	`imaplib`
Finger	Информационный	79	Не поддерживается
SSH	Командная строка	22	Поддерживается сторонними расширениями
Telnet	Командная строка	23	`telnetlib`

Клиенты и серверы

Для программистов, использующих сокеты, наличие стандартных протоколов означает, что порты с номерами 0–1023 не должны использоваться в сценариях, если только не планируется действительное использование одного из этих протоколов верхнего уровня. Это соответствует стандартам и здравому смыслу. Например, программа Telnet может открыть диалог с любым компьютером, поддерживающим протокол Telnet, подключаясь к его порту 23; если бы не было предустановленных номеров портов, все серверы могли бы устанавливать службу Telnet на разные порты. Аналогично сайты стандартно ждут поступления запросов страниц от броузеров в порт с номером 80; если бы они этого не делали, то для посещения любого сайта в Сети требовалось бы знать и вводить номер порта HTTP.

В результате определения стандартных номеров портов для служб Сеть естественным образом приобретает архитектуру вида *клиент/сервер*. С одной стороны есть компьютеры, поддерживающие стандартные протоколы, на которых постоянно выполняется ряд программ, ожидающих запросов на соединение по зарезервированным портам. С другой стороны находятся компьютеры, которые связываются с этими программами, чтобы воспользоваться предоставляемыми ими услугами.

Программу, которая выполняется постоянно и ожидает запросы, обычно называют *сервером*, а соединяющуюся с ней программу – *клиентом*. В качестве примера возьмем знакомую модель обзора веб-страниц. Как показано в табл. 12.1, используемый в Сети протокол HTTP позволяет клиентам и серверам общаться через сокеты с номером порта 80:

Сервер

На компьютере, где хранятся сайты, обычно выполняется программа веб-сервера, постоянно ожидающая входящие запросы соединения на сокете, связанном с портом 80. Часто сам сервер не занимается ничем другим, кроме постоянного ожидания появления запросов к порту – обработка запросов передается порожденным процессам или потокам.

Клиенты

Программы, которым нужно поговорить с этим сервером, для инициации соединения указывают имя компьютера сервера и порт 80. Типичными клиентами веб-серверов являются веб-броузеры, такие как Firefox, Internet Explorer или Chrome, но открыть соединение со стороны клиента и получать веб-страницы с сервера может любой сценарий, указав номер порта 80. Именем компьютера сервера может быть также «localhost», если веб-сервер выполняется на том же компьютере, что и клиент.

В целом многие клиенты могут подключаться к серверу через сокеты независимо от того, реализован на нем стандартный протокол или нечто более специфическое для данного приложения. А в некоторых приложениях понятия клиента и сервера размыты – программы могут обмениваться между собой байтами скорее как равноправные участники, а не как главный и подчиненный. Например, агенты пиринговых сетей передачи файлов могут одновременно являться клиентами и серверами для разных участков передаваемых файлов.

Однако в данной книге программы, которые ждут появления запросов на сокетах, мы будем называть *серверами*, а программы, устанавливающие соединения, – *клиентами*. Иногда также мы будем называть *сервером* и *клиентом* компьютеры, на которых выполняются эти программы (например, компьютер, на котором выполняется программа веб-сервера, может быть назван *компьютером веб-сервера*), но это скорее относится к физической привязке, а не к функциональной природе.

Структуры протоколов

Функционально протоколы могут выполнять известную задачу, например чтение электронной почты или передачу сообщения в телеконференцию Usenet, но в конечном счете все сводится к пересылке байтов сообщений через сокеты. Структура этих сообщений зависит от протокола, сокрыта в библиотеке Python, и ее обсуждение по большей части находится за рамками данной книги, но несколько общих слов помогут развеять таинственность слоя протоколов.

Одни протоколы могут определять содержимое сообщений, пересылаемых через сокеты; другие могут задавать последовательность управляющих сообщений, которыми обмениваются стороны в процессе общения. Путем определения правильных схем связи протоколы делают ее более надежной. Они также могут препятствовать появлению блокировок, возникающих, когда компьютер ждет сообщения, которое никогда не поступит.

Например, протокол FTP предотвращает блокировку путем организации связи по двум сокетам: один служит только для обмена управляющими сообщениями, а другой – для передачи содержимого файла. Сервер FTP ждет управляющих сообщений (например, «передай мне файл») на одном порту, а содержимое файла передает по другому. Клиенты FTP открывают соединения с управляющим портом компьютера сервера, посылают запросы и передают или получают содержимое файлов через сокет, соединенный с портом данных на компьютере сервера. Протокол FTP определяет также стандартные структуры сообщений, передаваемые между клиентом и сервером. Например, управляющее сообщение запроса файла должно соответствовать стандартному формату.

Библиотечные модули Python для Интернета

Если все это показалось вам ужасно сложным, не унывайте: все детали обрабатываются стандартными модулями Python поддержки протоколов. Например, библиотечный модуль Python `ftplib` управляет установлением связи на уровне сокетов и сообщений, которое определено в протоколе FTP. Сценарии, импортирующие `ftplib`, получают доступ к интерфейсу пересылки файлов по FTP значительно более высокого уровня и могут в значительной мере оставаться в неведении относительно лежащего в основе протокола FTP и сокетов, на которых он выполняется.[1]

В действительности все поддерживаемые протоколы представлены в стандартной библиотеке Python либо пакетами модулей, имена которых соответствуют названию протокола, либо файлами модулей с именами в формате *xxxlib.py*, где *xxx* заменяется именем протокола. В последней колонке табл. 12.1 указано имя стандартного модуля поддержки протокола. Например, протокол FTP поддерживается файлом модуля *ftplib.py*, а протокол HTTP – пакетом `http.*`. Кроме того, в модулях протоколов имя объекта интерфейса верхнего уровня обычно совпадает с названием протокола. Так, например, чтобы начать сеанс FTP в сценарии Python, нужно выполнить инструкцию `import ftplib` и передать надлежащие параметры конструктору `ftplib.FTP`; для Telnet нужно создать экземпляр класса `telnetlib.Telnet`.

Помимо модулей реализации протоколов, указанных в табл. 12.1, в стандартной библиотеке Python есть модули для получения ответов веб-серверов (`urllib.request`), анализа и обработки данных, которые переданы через сокеты или протоколы (`html.parser`, пакеты `email.*` и `xml.*`), и многие другие. В табл. 12.2 перечислены наиболее часто используемые модули из этой категории.

Со многими из этих модулей мы встретимся в нескольких последующих главах, хотя и не со всеми. Кроме того, существует еще целый ряд модулей Python для работы с Интернетом, которые здесь не перечислены. Модули, демонстрируемые в этой книге, являются наиболее типичными, но, как обычно, за более полными и свежими данными обращайтесь к справочному руководству по стандартной библиотеке Python.

---

[1] Так как Python является системой, распространяемой с открытыми исходными текстами, можно прочесть исходный программный код модуля `ftplib`, если вас интересует, как действительно работает используемый протокол. Посмотрите файл *ftplib.py* в каталоге с исходными текстами стандартной библиотеки на своем компьютере. Его реализация сложна (он должен форматировать сообщения и управлять двумя сокетами), но, как и другие стандартные модули протоколов Интернета, он дает хороший пример низкоуровневого программирования сокетов.

Таблица 12.2. Стандартные модули, часто используемые для работы с Интернетом

Модули Python	Применение
socket, ssl	Поддержка сетевых взаимодействий (TCP/IP, UDP и другие) плюс безопасные сокеты SSL
cgi	Поддержка CGI-сценариев на стороне сервера: анализ входного потока, экранирование текста HTML и тому подобное.
urllib.request	Получение веб-страниц по их адресам (URL)
urllib.parse	Анализ и преобразование строк URL в компоненты, экранирование строк URL
http.client, ftplib, nntplib	Модули поддержки протоколов HTTP (Веб), FTP (пересылка файлов) и NNTP (телеконференции) на стороне клиента
http.cookies, http.cookiejar	Поддержка cookie протокола HTTP (блоки данных, сохраняемые на стороне клиента по запросу веб-сайта, поддержка на стороне клиента и сервера)
poplib, imaplib, smtplib	Модули поддержки протоколов POP, IMAP (получение почты) и SMTP (отправка почты)
telnetlib	Модуль поддержки протокола Telnet
html.parser, xml.*	Синтаксический анализ содержимого веб-страниц (документы HTML и XML)
xdrlib, socket	Переносимое кодирование передаваемых двоичных данных
struct, pickle	Кодирование объектов Python в пакеты двоичных данных или сериализованные строки байтов для передачи
email.*	Синтаксический анализ и составление сообщений электронной почты с заголовками, вложениями и кодировками
mailbox	Обработка почтовых ящиков на диске и сообщений в них
mimetypes	Определение типа содержимого файлов исходя из их имен и расширений
uu, binhex, base64, binascii, quopri, email.*	Кодирование и декодирование двоичных (или других) данных, передаваемых в виде текста (автоматически используется пакетом email)
socketserver	Фреймворк для создания стандартных серверов Сети
http.server	Базовая реализация сервера HTTP с обработчиками запросов для простых серверов и серверов с поддержкой CGI

О стандартах протоколов

Если необходимы полные сведения о протоколах и портах, то на момент написания этой книги полный список всех портов, зарезервированных для протоколов или зарегистрированных в качестве используемых различными стандартными системами, можно найти поиском по страницам веб, поддерживаемым организациями IETF (Internet Engineering Task Force – рабочей группой инженеров Интернета) и IANA (Internet Assigned Numbers Authority – полномочным комитетом по надзору за номерами, используемыми в Интернете). Организация IETF отвечает за сопровождение протоколов и стандартов Веб. Организация IANA является главным координатором назначения уникальных значений параметров для протоколов Интернета. Еще один орган стандартизации, консорциум W3C (от WWW), тоже сопровождает соответствующие документы. Подробности смотрите на следующих страницах:

http://www.ietf.org

http://www.iana.org/numbers.html

http://www.iana.org/assignments/port-numbers

http://www.w3.org

Вполне возможно, что за время жизни этой книги возникнут более свежие хранилища спецификаций стандартных протоколов, но в течение ближайшего времени главным авторитетом будет, по-видимому, служить веб-сайт IETF. Однако если вы соберетесь туда обратиться, то предупреждаем, что детали там, хм... слишком детализированы. Так как модули протоколов Python скрывают большую часть сложностей, связанных с сокетами и сообщениями и документированных в стандартах протоколов, обычно для работы в Сети с помощью Python не требуется держать в памяти эти документы.

Программирование сокетов

Теперь, когда мы знаем, какую роль играют сокеты в общей структуре Интернета, пойдем дальше и посмотрим, какие инструменты предоставляет Python для программирования сокетов в сценариях. В этом разделе будет показано, как использовать интерфейс Python к сокетам для организации низкоуровневых сетевых взаимодействий. В последующих главах мы будем использовать модули протоколов более высокого уровня, скрывающих операции с лежащими в их основе сокетами. При этом интерфейс Python к сокетам может использоваться непосред-

ственно для реализации собственных сетевых взаимодействий и организации доступа к стандартным протоколам вручную.

Как мы уже видели в главе 5, основным интерфейсом сокетов в Python является стандартный библиотечный модуль `socket`. Подобно модулю POSIX `os` модуль `socket` служит лишь тонкой оберткой (интерфейсным слоем) вокруг функций для работы с сокетами из библиотеки на языке C. Подобно файлам Python этот модуль основывается на объектах – методы объекта сокета, реализованные в этом модуле, после преобразования данных вызывают соответствующие операции библиотеки C. Например, функции `send` и `recv` в библиотеке C отображаются в методы объекта сокета в языке Python.

Модуль `socket` обеспечивает возможность выполнения операций с сокетами в любой системе, поддерживающей сокеты в стиле BSD – Windows, Mac OS, Linux, Unix и так далее, – и таким образом обеспечивает переносимый интерфейс сокетов. Кроме того, этот модуль поддерживает все стандартные типы сокетов – TCP/IP, UDP, дейтаграммы и доменные сокеты Unix – и может использоваться как прикладной интерфейс доступа к сети и как универсальный механизм взаимодействий между процессами, выполняющимися на одном и том же компьютере.

С функциональной точки зрения, сокеты являются программными инструментами передачи байтов между программами, которые могут выполняться на разных компьютерах. Хотя сокеты сами по себе способны передавать только строки байтов, тем не менее, вы можете также передавать через них объекты Python, используя модуль `pickle`. Этот модуль способен преобразовывать объекты Python, такие как списки, словари и экземпляры классов, в строки байтов и обратно и обеспечивает поддержку необходимого промежуточного этапа при передаче объектов высокого уровня через сокеты.

Для преобразования объектов Python в упакованные строки двоичных байтов, готовые к передаче, можно также использовать модуль `struct`, но в целом его возможности ограничиваются объектами, которые могут отображаться в типы данных языка программирования C. Модуль `pickle` поддерживает возможность передачи более крупных объектов, таких как словари и экземпляры классов. Для других задач, включая использование большинства стандартных протоколов Интернета, достаточно использовать простые строки байтов. Поближе с модулем `pickle` мы познакомимся в этой главе и далее в книге.

Помимо реализации простого обмена данными модуль `socket` также включает различные дополнительные инструменты. Например, в нем имеются функции для выполнения следующих и других задач:

- Переупорядочение байтов в стандартный сетевой порядок и обратно (`ntohl`, `htonl`).
- Определение имени компьютера и его сетевого адреса (`gethostname`, `gethostbyname`).

- Обертывание объектов-сокетов объектами файлов (sockobj.makefile).
- Перевод сокетов в неблокирующий режим (sockobj.setblocking).
- Установка предельного времени ожидания для сокета (sockobj.settimeout).

Если ваша версия Python была собрана с поддержкой протокола защищенных сокетов (Secure Sockets Layer, SSL), стандартный библиотечный модуль ssl обеспечит также возможность шифрования передаваемых данных с помощью функции ssl.wrap_socket. Эта функция обертывает объект сокета логикой SSL, которая в свою очередь будет использоваться другими стандартными библиотечными модулями для поддержки безопасного протокола HTTPS (http.client и urllib.request), безопасной передачи сообщений электронной почты (poplib и smtplib) и других видов взаимодействий. Мы встретимся с некоторыми из этих модулей далее в этой части книги, но не будем изучать все дополнительные особенности модуля socket – за подробностями, опущенными здесь, обращайтесь к руководству по стандартной библиотеке Python.

Основы сокетов

Мы не беремся за изучение дополнительных особенностей сокетов в этой главе, но простую передачу данных через сокеты удивительно легко реализовать на языке Python. Чтобы установить связь между компьютерами, программы на языке Python импортируют модуль socket, создают объект сокета и вызывают методы этого объекта для установления соединения, отправки и получения данных.

Сокеты по своей природе являются механизмами двунаправленной передачи данных, а методы объекта сокета прямо отображаются в вызовы функций сокетов библиотеки C. Так, сценарий в примере 12.1 реализует программу, которая просто ждет подключения к сокету и отправляет обратно через сокет все, что она через него получает, добавляя строку префикса Echo=>.

Пример 12.1. P4E\Internet\Sockets\echo-server.py

```
"""
На стороне сервера: открыть сокет TCP/IP с указанным номером порта,
ждать появления сообщения от клиента, отправить это же сообщение обратно;
это реализация простого одноступенчатого диалога вида запрос/ответ,
но сценарий выполняет бесконечный цикл и пока он действует, способен
обслужить множество клиентов; клиент может выполняться как на удаленном,
так и на локальном компьютере, если в качестве имени сервера
будет использовать 'localhost'
"""


from socket import *  # получить конструктор сокета и константы
myHost = ''           # '' = все доступные интерфейсы хоста
myPort = 50007        # использовать незарезервированный номер порта
```

```
sockobj = socket(AF_INET, SOCK_STREAM)    # создать объект сокета TCP
sockobj.bind((myHost, myPort))            # связать с номером порта сервера
sockobj.listen(5)                         # не более 5 ожидающих запросов

while True:                                  # пока процесс работает
    connection, address = sockobj.accept()   # ждать запроса
                                             # очередного клиента
    print('Server connected by', address)    # соединение является
                                             # новым сокетом
    while True:
        data = connection.recv(1024)      # читать следующую строку из сокета
        if not data: break                # отправить ответ клиенту
        connection.send(b'Echo=>' + data) # и так, пока от клиента поступают
    connection.close()                    # данные, после чего закрыть сокет
```

Как уже говорилось выше, обычно такие программы, которые ожидают входящих соединений, мы называем *серверами*, потому что они предоставляют сервис, доступный на данном компьютере и на данном порту через Интернет. Программы, подключающиеся к такому серверу для доступа к его услуге, обычно называются *клиентами*. В примере 12.2 приводится простой клиент, реализованный на языке Python.

Пример 12.2. PP4E\Internet\Sockets\echo-client.py

```
"""
На стороне клиента: использует сокеты для передачи данных серверу
и выводит ответ сервера на каждую строку сообщения; 'localhost' означает,
что сервер выполняется на одном компьютере с клиентом, что позволяет
тестировать клиента и сервер на одном компьютере; для тестирования
через Интернет запустите сервер на удаленном компьютере и установите
serverHost или argv[1] равными доменному имени компьютера или его IP-адресу;
сокеты Python являются переносимым интерфейсом к сокетам BSD,
с методами объектов для стандартных функций сокетов, доступных
в системной библиотеке C;
"""

import sys
from socket import *      # переносимый интерфейс сокетов плюс константы
serverHost = 'localhost'  # имя сервера, например: 'starship.python.net'
serverPort = 50007        # незарезервированный порт, используемый сервером

message = [b'Hello network world']   # текст, посылаемый серверу обязательно
                                     # типа bytes: b'' или str.encode()
if len(sys.argv) > 1:
    serverHost = sys.argv[1]      # сервер в аргументе 1 командной строки
    if len(sys.argv) > 2:         # текст в аргументах командной строки 2..n
        message = (x.encode() for x in sys.argv[2:])

sockobj = socket(AF_INET, SOCK_STREAM)      # создать объект сокета TCP/IP
sockobj.connect((serverHost, serverPort))   # соединение с сервером и портом
```

```
for line in message:
    sockobj.send(line)              # послать серверу строчку через сокет
    data = sockobj.recv(1024)       # получить строку от сервера: до 1k
    print('Client received:', data) # строка bytes выводится в кавычках,
                                    # было `x`, repr(x)
sockobj.close()             # закрыть сокет, чтобы послать серверу eof
```

Методы сокетов, используемые сервером

Прежде чем увидеть эти программы в действии, коротко рассмотрим, как здесь клиент и сервер выполняют свои функции. Оба они представляют достаточно простые примеры сценариев, использующих сокеты, но иллюстрируют общую схему последовательности вызовов методов, применяемую в большинстве программ, использующих сокеты. В действительности это достаточно стереотипный программный код: большинство программ сокетов обычно выполняет вызовы тех же методов сокетов, что и наши два сценария, поэтому разберем каждую строку в существенных местах этих сценариев.

Программы, подобные приведенной в примере 12.1, предоставляющие услуги другим программам с помощью сокетов, обычно начинают работу с такой последовательности вызовов:

`sockobj = socket(AF_INET, SOCK_STREAM)`

Здесь с помощью модуля `socket` создается объект сокета TCP. Имена `AF_INET` и `SOCK_STREAM` принадлежат предопределенным переменным, импортируемым из модуля `socket`; их совместное применение означает «создать сокет TCP/IP», стандартное средство связи для Интернета. Более точно, `AF_INET` означает протокол адресов IP, а `SOCK_STREAM` означает протокол передачи TCP. Комбинация `AF_INET/SOCK_STREAM` используется по умолчанию, потому что является наиболее типичной, но при этом обычно она указывается явно.

При использовании в этом вызове других имен можно создавать такие объекты, как сокеты UDP без логического соединения (второй параметр `SOCK_DGRAM`) и доменные сокеты Unix на локальном компьютере (первый параметр `AF_UNIX`), но в данной книге мы этого делать не будем. Смотрите подробности относительно этих и других параметров модуля `socket` в руководстве по библиотеке Python. Использование сокетов других типов предполагает использование иных стереотипных форм программного кода.

`sockobj.bind((myHost, myPort))`

Связывает объект сокета с адресом – для IP-адресов передается имя компьютера сервера и номер порта на этом компьютере. Здесь сервер идентифицирует компьютер и порт, связанные с сокетом. В серверных программах имя компьютера обычно задается пустой строкой (`''`), что означает компьютер, на котором выполняется сценарий (формально, все локальные и удаленные интерфейсы доступные на компьютере), а порт указывается как число за пределами диапазона 0–1023

(зарезервированного для стандартных протоколов, описывавшихся выше).

Обратите внимание, что у каждой поддерживаемой службы сокетов должен быть свой номер порта. Если попытаться открыть сокет на порту, который уже используется, Python возбудит исключение. Обратите также внимание на вложенные скобки в этом вызове – здесь для сокета с протоколом адресов `AF_INET` мы передаем методу `bind` адрес сокета хост/порт, как объект кортежа из двух элементов (для `AF_UNIX` передается строка). Технически метод `bind` принимает кортеж значений, соответствующий типу создаваемого сокета.

`sockobj.listen(5)`

Начинает ожидание входящих запросов на соединение от клиентов и позволяет помещать в очередь ожидания до пяти запросов. Передаваемое значение устанавливает количество входящих клиентских запросов, помещаемых операционной системой в очередь перед тем, как начать отклонять новые запросы (что происходит только тогда, когда сервер не успевает обрабатывать запросы и очередь переполняется). Для большинства программ, работающих с сокетами, обычно достаточно значения 5. Этому методу следует передавать число не менее 1.

После этого сервер готов к принятию запросов соединения от клиентских программ, выполняющихся на удаленных компьютерах (или том же самом компьютере), и входит в бесконечный цикл ожидания их поступления:

`connection, address = sockobj.accept()`

Ждет поступления от клиента нового запроса на соединение. Когда он поступит, метод `accept` вернет новый объект сокета, через который можно передавать данные соединившемуся клиенту и получать их от него. Соединение осуществляет объект `sockobj`, но связь с клиентом происходит через новый сокет, `connection`. Этот метод возвращает кортеж из двух элементов, где `address` является интернет-адресом соединившегося клиента. Метод `accept` может вызываться многократно, чтобы обслужить несколько клиентов. Поэтому каждый вызов возвращает новый сокет, через который происходит связь с конкретным клиентом.

Установив соединение с клиентом, мы попадаем в другой цикл, в котором получаем от клиента данные блоками по 1024 байта и отправляем каждый блок обратно клиенту:

`data = connection.recv(1024)`

Читает до 1024 байтов из очередного сообщения, посланного клиентом (то есть поступившего из сети или через соединение IPC), и возвращает их сценарию в виде строки. При завершении работы клиентом возвращается пустая строка байтов – когда клиент закрывает свой конец сокета, возвращается признак конца файла.

`connection.send('Echo=>' + data)`

Отправляет последний полученный блок данных обратно программе клиента, предварив его строкой `'Echo=>'`. Программа клиента получает эти отправленные ей данные с помощью метода `recv`. Технически, этот метод старается отправить максимально возможное количество данных и возвращает количество фактически отправленных байтов. Для обеспечения надежной передачи данных некоторые программы могут повторять передачу неотправленных фрагментов или использовать метод `connection.sendall` для принудительной передачи всех байтов.

`connection.close()`

Закрывает соединение с данным конкретным клиентом.

Передача строк байтов и объектов

Мы познакомились с методами, которые используются для передачи данных на стороне сервера, но что из себя представляют данные, передаваемые через сокет? Как мы узнали в главе 5, сокеты сами по себе всегда работают со *строками двоичных байтов*, а не с текстом. Для сценариев это означает, что они вынуждены передавать и принимать строки `bytes`, а не `str`, хотя вы можете предусмотреть кодирование текста и декодирование в текст с помощью методов `str.encode` и `bytes.decode`. Для удовлетворения требований сокетов мы будем использовать в наших сценариях литералы `bytes` вида `b'...'`. В других ситуациях можно использовать модули `struct` и `pickle`, автоматически возвращающие строки байтов, благодаря чему отпадает необходимость в выполнении дополнительных операций кодирования/декодирования.

Например, несмотря на то, что модель сокетов ограничивается передачей строк байтов, вы можете отправлять и принимать практически любые объекты Python с помощью стандартного модуля `pickle` сериализации объектов. Его функции `dumps` и `loads` преобразуют объекты Python в строки байтов и обратно, готовые к передаче через сокеты:

```
>>> import pickle
>>> x = pickle.dumps([99, 100])  # на передающей стороне...
                                 # преобразовать в строку байтов
>>> x                            # строка для отправки, возвращается
b'\x80\x03]q\x00(KcKde.'         # методом recv

>>> pickle.loads(x)              # на принимающей стороне...
[99, 100]                        # преобразовать обратно в объект
```

Для преобразования простых типов, соответствующих типам в языке C, можно также использовать модуль `struct`, который обеспечивает необходимую нам возможность преобразования в строку байтов:

```
>>> import struct
>>> x = struct.pack('>ii', 99, 100)   # преобразование данных простых типов
>>> x                                 # для передачи
```

```
b'\x00\x00\x00c\x00\x00\x00d'
>>> struct.unpack('>ii', x)
(99, 100)
```

Используя подобные преобразования, мы получаем возможность передавать объекты Python через сокеты. За дополнительной информацией о модуле struct обращайтесь к главе 4. Мы уже кратко рассматривали модуль pickle и сериализацию объектов в главе 1, но еще больше об этом модуле и о некоторых ограничениях сериализации объектов мы узнаем в главе 17, когда будем исследовать способы сохранения данных.

В действительности существуют различные способы расширения базовой модели передачи данных через сокеты. Например, подобно тому, как функция open способна обертывать дескрипторы файлов, полученные с помощью функции os.fdopen, о чем рассказывалось в главе 4, метод socket.makefile позволяет обертывать сокеты объектами файлов в текстовом режиме, которые выполняют кодирование и декодирование текста автоматически. Этот метод позволяет в Python 3.X указывать кодировки, отличные от кодировки по умолчанию, и определять параметры преобразования символов конца строки в виде дополнительных аргументов, как при использовании встроенной функции open. Поскольку результат вызова метода socket.makefile имитирует интерфейс файлов, его также можно использовать в вызовах функций модуля pickle, принимающих файлы, для неявной передачи объектов через сокеты. Дополнительные примеры обертывания сокетов объектами файлов будут представлены далее в этой главе.

В наших простых сценариях вся работа выполняется с помощью жестко определенных строк байтов и непосредственных вызовов методов сокетов. Завершив сеанс обмена с данным конкретным клиентом, сервер из примера 12.1 возвращается в свой бесконечный цикл и ждет следующего запроса на соединение от клиента. А теперь двинемся дальше и посмотрим, что происходит по другую сторону барьера.

Методы сокетов, используемые клиентом

Последовательность вызовов методов сокетов в клиентских программах вроде той, что показана в примере 12.2, имеет более простой вид, – фактически, половину сценария занимает логика подготовительных операций. Главное, о чем нужно помнить, – это то, что клиент и сервер при открытии своих сокетов должны указывать один и тот же номер порта, и дополнительно клиент должен идентифицировать компьютер, на котором выполняется сервер. В наших сценариях сервер и клиент договорились использовать для связи порт с номером 50007, вне диапазона стандартных протоколов. Ниже приводится последовательность вызова методов сокета на стороне клиента:

```
sockobj = socket(AF_INET, SOCK_STREAM)
```

Создает в клиентской программе объект сокета, так же как на сервере.

`sockobj.connect((serverHost, serverPort))`

Открывает соединение с компьютером и портом, где программа сервера ждет запросы на соединение от клиентов. В этом месте клиент указывает строку с именем службы, с которой ему необходимо связаться. Клиент может передать имя удаленного компьютера в виде доменного имени (например, *starship.python.net*) или числового IP-адреса. Можно также определить имя сервера как `localhost` (или использовать эквивалентный ему IP-адрес `127.0.0.1`), указав тем самым, что программа сервера выполняется на том же компьютере, что и клиент. Это удобно для отладки серверов без подключения к Сети. И снова номер порта клиента должен в точности соответствовать номеру порта сервера. Обратите еще раз внимание на вложенные скобки – так же, как в вызове метода `bind` на сервере, мы передаем методу `connect` адрес хоста/порта сервера в виде кортежа.

Установив соединение с сервером, клиент попадает в цикл, посылая сообщения построчно и выводя то, что возвращает сервер после передачи каждой строки:

`sockobj.send(line)`

Пересылает серверу очередную строку байтов сообщения через сокет. Обратите внимание, что список сообщений по умолчанию содержит строки байтов (`b'...'`). Так же, как и на стороне сервера, данные, передаваемые через сокет, должны иметь вид строк байтов, впрочем, при необходимости это может быть результат кодирования текста вручную вызовом метода `str.encode` или результат преобразования с помощью модуля `pickle` или `struct`. Когда строки для передачи передаются в аргументах командной строки, они должны быть преобразованы из типа `str` в тип `bytes` – эта операция выполняется клиентом с помощью выражения-генератора (тот же эффект можно было бы получить вызовом функции `map(str.encode, sys.argv[2:])`).

`data = sockobj.recv(1024)`

Читает очередную строку ответа, переданную программой-сервером. Технически этот вызов читает до 1024 байтов очередного ответа, которые возвращаются как строка байтов.

`sockobj.close()`

Закрывает соединение с сервером, посылая сигнал «конец файла».

Вот и все. Сервер обменивается одной или несколькими строками текста с каждым подключившимся клиентом. Операционная система обеспечивает поиск удаленных компьютеров, направляя пересылаемые между программами байты через Интернет и (с помощью TCP) обеспечивая доставку сообщений в целости. При этом может выполняться еще масса работы – по пути наши строки могут путешествовать по всему свету, переходя из телефонных линий в спутниковые каналы и так далее. Но при программировании на языке Python мы можем оставаться

в счастливом неведении относительно того, что происходит ниже уровня вызовов сокетов.

Запуск программ, использующих сокеты, на локальном компьютере

А теперь заставим эти сценарии поработать. Есть два способа запустить их – на одном и том же компьютере или на разных. Чтобы запустить клиент и сервер на одном компьютере, откройте две консоли командной строки, запустите в одной программу сервера, а в другой несколько раз запустите клиента. Сервер работает постоянно и отвечает на запросы, которые происходят при каждом запуске сценария клиента в другом окне.

Например, ниже приводится текст, который появляется в окне консоли MS-DOS, где я запустил сценарий сервера:

```
C:\...\PP4E\Internet\Sockets> python echo-server.py
Server connected by ('127.0.0.1', 57666)
Server connected by ('127.0.0.1', 57667)
Server connected by ('127.0.0.1', 57668)
```

В выводе указан адрес (IP-имя компьютера и номер порта) каждого соединившегося клиента. Как и большинство серверов, этот сервер выполняется вечно, ожидая запросов на соединение от клиентов. Здесь он получил три запроса, но чтобы понять их значение, нужно показать текст в окне клиента:

```
C:\...\PP4E\Internet\Sockets> python echo-client.py
Client received: b'Echo=>Hello network world'

C:\...\PP4E\Internet\Sockets> python echo-client.py localhost spam Spam SPAM
Client received: b'Echo=>spam'
Client received: b'Echo=>Spam'
Client received: b'Echo=>SPAM'

C:\...\PP4E\Internet\Sockets> python echo-client.py localhost Shrubbery
Client received: b'Echo=>Shrubbery'
```

Здесь сценарий клиента был запущен три раза, в то время как сценарий сервера постоянно выполнялся в другом окне. Каждый клиент, соединявшийся с сервером, посылал ему сообщение из одной или нескольких строк текста и читал ответ, возвращаемый сервером, – эхо каждой строки текста, отправленной клиентом. И при запуске каждого клиента в окне сервера появлялось новое сообщение о соединении (вот почему там их три). Поскольку сервер выполняет бесконечный цикл, в Windows вам, вероятно, придется завершить его с помощью Диспетчера задач (Task Manager) по окончании тестирования, потому что нажатие комбинации Ctrl-C в консоли, где был запущен сервер, будет проигнорировано – на других платформах ситуация выглядит несколько лучше.

Важно отметить, что клиенты и сервер выполняются здесь на одном и том же компьютере (в Windows). Сервер и клиент используют один и тот же номер порта, но разные имена компьютеров – `''` и `localhost`, соответственно, при ссылке на компьютер, на котором они выполняются. В действительности здесь нет никакого соединения через Интернет. Это всего лишь механизм IPC, вроде тех, с которыми мы встречались в главе 5: сокеты прекрасно справляются с ролью средства связи между программами, выполняющимися на одном компьютере.

Запуск программ, использующих сокеты, на удаленном компьютере

Чтобы заставить эти сценарии общаться через Интернет, а не в пределах одного компьютера, необходимо проделать некоторую дополнительную работу, чтобы запустить серверный сценарий на другом компьютере. Во-первых, нужно загрузить файл с исходным программным кодом сервера на удаленный компьютер, где у вас есть учетная запись и Python. Ниже показано, как я выгружаю этот сценарий через FTP на сайт, который располагается на компьютере с доменным именем *learning-python.com*, принадлежащем мне. Большая часть информационных строк в следующем примере сеанса была удалена, имя вашего сервера и детали интерфейса загрузки могут отличаться, а кроме того, есть другие способы копирования файлов на компьютер (например FTP-клиенты с графическим интерфейсом, электронная почта, формы передачи данных на веб-страницах и другие – смотрите врезку «Советы по использованию удаленных серверов» ниже, где приводятся некоторые подсказки по использованию удаленных серверов):

```
C:\...\PP4E\Internet\Sockets> ftp learning-python.com
Connected to learning-python.com.
User (learning-python.com:(none)): xxxxxxxx
Password: yyyyyyyy
ftp> mkdir scripts
ftp> cd scripts
ftp> put echo-server.py
ftp> quit
```

После пересылки программы сервера на другой компьютер нужно запустить ее там. Соединитесь с этим компьютером и запустите программу сервера. Обычно я подключаюсь к компьютеру своего сервера через Telnet или SSH и запускаю программу сервера из командной строки как постоянно выполняющийся процесс. Для запуска сценария сервера в фоновом режиме из командных оболочек Unix/Linux может использоваться синтаксис с `&`; можно также сделать сервер непосредственно исполняемым с помощью строки `#!` и команды `chmod` (подробности в главе 3).

Ниже приводится текст, который появляется в окне свободного клиента PuTTY на моем PC при открытии сеанса SSH на сервере Linux, где

у меня есть учетная запись (опять же опущены несколько информационных строк):

```
login as: xxxxxxxx
XXXXXXXX@learning-python.com's password: yyyyyyyy
Last login: Fri Apr 23 07:46:33 2010 from 72.236.109.185
[...]$ cd scripts
[...]$ python echo-server.py &
[1] 23016
```

Теперь, когда сервер ждет соединений через Сеть, снова несколько раз запустите клиент на своем локальном компьютере. На этот раз клиент выполняется на компьютере, отличном от сервера, поэтому передадим доменное имя или IP-адрес сервера, как аргумент командной строки клиента. Сервер по-прежнему использует имя компьютера `''`, так как он всегда должен прослушать сокет, на каком бы компьютере ни выполнялся. Ниже показано, что появляется в окне сеанса SSH с сервером *learning-python.com* на моем PC:

```
[...]$ Server connected by ('72.236.109.185', 57697)
Server connected by ('72.236.109.185', 57698)
Server connected by ('72.236.109.185', 57699)
Server connected by ('72.236.109.185', 57700)
```

А далее показано, что выводится в окно консоли MS-DOS, когда я запускаю клиента. Сообщение «connected by» появляется в окне сеанса SSH с сервером каждый раз, когда сценарий клиента запускается в окне клиента:

```
C:\...\PP4E\Internet\Sockets> python echo-client.py learning-python.com
Client received: b'Echo=>Hello network world'

C:\...\PP4E\Internet\Sockets> python echo-client.py learning-python.com
                              ni Ni NI
Client received: b'Echo=>ni'
Client received: b'Echo=>Ni'
Client received: b'Echo=>NI'

C:\...\PP4E\Internet\Sockets> python echo-client.py learning-python.com
                              Shrubbery
Client received: b'Echo=>Shrubbery'
```

Получить IP-адрес компьютера по доменному имени можно с помощью команды `ping`; для подключения клиент может использовать любую из этих форм имени компьютера:

```
C:\...\PP4E\Internet\Sockets> ping learning-python.com
Pinging learning-python.com [97.74.215.115] with 32 bytes of data:
Reply from 97.74.215.115: bytes=32 time=94ms TTL=47
Ctrl-C

C:\...\PP4E\Internet\Sockets> python echo-client.py 97.74.215.115 Brave
                              Sir Robin
```

```
Client received: b'Echo=>Brave'
Client received: b'Echo=>Sir'
Client received: b'Echo=>Robin'
```

Этот вывод, возможно, излишне сдержан – за кулисами много чего происходит. Клиент, выполняющийся в Windows на моем ноутбуке, соединяется с программой сервера, выполняемой на компьютере, работающем под управлением Linux, находящемся, возможно, на расстоянии тысяч миль, и обменивается с ней данными. Это происходит почти так же быстро, как если бы клиент и сервер выполнялись на одном ноутбуке, при этом используются одни и те же библиотечные вызовы. Изменяется только имя сервера, передаваемое клиентам.

Несмотря на свою простоту, этот пример иллюстрирует одно из основных преимуществ использования сокетов для организации взаимодействий между программами: они по своей природе поддерживают возможность общения программ, выполняющихся на разных компьютерах, причем для этого требуется внести в сценарии минимум изменений, а иногда можно обойтись вообще без изменений. При этом сокеты обеспечивают легкость разделения и распределения частей системы по сети, когда это необходимо.

Практическое использование сокетов

Прежде чем двигаться дальше, следует сказать о трех аспектах практического использования сокетов. Во-первых, такие клиент и сервер могут выполняться на любых двух подключенных к Интернету компьютерах, где установлен Python. Конечно, чтобы запускать клиенты и сервер на разных компьютерах, необходимы действующее соединение с Интернетом и доступ к тому компьютеру, на котором должен быть запущен сервер.

При этом не требуется дорогостоящего высокоскоростного соединения – при работе с сокетами Python довольствуется любым соединением, которое существует на компьютере, будь то выделенная линия T1, беспроводное подключение, кабельный модем или простое коммутируемое соединение. Кроме того, если у вас нет собственной учетной записи на собственном сервере, как у меня на сервере *learning-python.com*, запускайте примеры клиента и сервера на одном компьютере, `localhost`, как было показано выше, – для этого лишь требуется, чтобы компьютер разрешал использовать сокеты, что бывает почти всегда.

Во-вторых, модуль `socket` обычно возбуждает исключение при запросе чего-либо недопустимого. Например, неудачной будет попытка подключения к несуществующему серверу (или недоступному, если нет связи с Интернетом):

```
C:\...\PP4E\Internet\Sockets> python echo-client.py www.nonesuch.com hello
Traceback (most recent call last):
  File "echo-client.py", line 24, in <module>
    sockobj.connect((serverHost, serverPort))   # соединение с сервером и...
```

```
socket.error: [Errno 10060] A connection attempt failed because the
connected
party did not properly respond after a period of time, or established
connection failed because connected host has failed to respond
```
(socket.error: [Ошибка 10060] Попытка соединения потерпела неудачу, потому что противоположная сторона не ответила в течение заданного интервала времени, или установленное соединение было разорвано, потому что другая сторона не смогла ответить)

Наконец, следите за тем, чтобы остановить процесс сервера, прежде чем запустить его заново, потому что иначе порт окажется занятым и вы получите другое исключение, как на моем удаленном компьютере сервера:

```
[...]$ ps -x
  PID TTY      STAT   TIME COMMAND
 5378 pts/0    S      0:00 python echo-server.py
22017 pts/0    Ss     0:00 -bash
26805 pts/0    R+     0:00 ps -x

[...]$ python echo-server.py
Traceback (most recent call last):
  File "echo-server.py", line 14, in <module>
    sockobj.bind((myHost, myPort))       # связать с номером порта сервера
socket.error: [Errno 10048] Only one usage of each socket address (protocol/
network address/port) is normally permitted
```
(socket.error: [Ошибка 10048] Только один сокет может быть связан с каждым адресом (протокол/сетевой адрес/порт))

В Linux сервер можно остановить несколькими нажатиями комбинации клавиш Ctrl-C (если он был запущен в фоновом режиме с &, сначала нужно перевести его на передний план командой fg):

```
[...]$ fg
python echo-server.py
Traceback (most recent call last):
  File "echo-server.py", line 18, in <module>
    connection, address = sockobj.accept()   # ждать запроса
                                             # очередного клиента
KeyboardInterrupt
```

Как упоминалось выше, комбинация клавиш Ctrl-C не завершает сервер на моем компьютере с Windows 7. Чтобы завершить локальный постоянно выполняющийся процесс сервера в Windows, может потребоваться запустить Диспетчер задач (Task Manager) (то есть нажать комбинацию клавиш Ctrl-Alt-Delete), а затем завершить задачу Python, выбрав ее в появившемся списке процессов. Кроме того, в Windows можно просто закрыть окно консоли, где был запущен сервер, чтобы остановить его, но при этом будет потеряна история команд. В Linux можно также завершить работу сервера, запущенного в другом окне или в фоновом режиме, с помощью команды оболочки `kill -9 pid`, но использование комбинации Ctrl-C требует меньше нажатий на клавиши.

Советы по использованию удаленных серверов

В некоторых примерах этой главы предлагается запускать сервер на удаленном компьютере. Хотя вы можете запускать примеры локально, используя имя `localhost`, однако использование удаленного компьютера лучше отражает гибкость и мощь сокетов. Чтобы запустить сервер на удаленном компьютере, вам потребуется удаленный компьютер с доступом в Интернет и с установленным интерпретатором Python, куда вы смогли бы выгружать свои сценарии и запускать их. Вам также необходим будет доступ к удаленному серверу с вашего ПК. Чтобы помочь вам выполнить этот последний шаг, ниже приводятся несколько советов для тех из вас, кто впервые использует удаленные серверы.

Чтобы выгрузить свои сценарии на удаленный компьютер, можно воспользоваться стандартной командой *FTP*, имеющейся в Windows и в большинстве других операционных систем. В Windows просто введите ее в окне консоли, чтобы соединиться с сервером FTP, или запустите свою любимую программу клиента FTP с графическим интерфейсом. В Linux введите команду FTP в окне `xterm`. Для подключения к неанонимному FTP-сайту потребуется ввести имя учетной записи и пароль. Для анонимного FTP в качестве имени пользователя укажите «anonymous», а в качестве пароля – свой адрес электронной почты.

Чтобы запустить сценарии удаленно из командной строки, можно воспользоваться командой *Telnet*, которая является стандартной командой в Unix-подобных системах. В Windows можно найти клиента с графическим интерфейсом. Для подключения к некоторым серверам вместо Telnet может потребоваться использовать безопасную командную оболочку *SSH*, чтобы получить доступ к командной строке. В Интернете можно найти различные утилиты SSH, включая PuTTY, используемую в этой книге. В составе самого языка Python имеется модуль `telnetlib`, а поискав в Интернете, можно найти инструменты поддержки SSH для использования в сценариях на языке Python, включая `ssh.py`, *paramiko*, *Twisted*, *Pexpect* и даже `subprocess.Popen`.

Параллельный запуск нескольких клиентов

До настоящего момента мы запускали сервер локально и удаленно и выполняли сценарии клиентов вручную. Настоящие серверы обычно предусматривают возможность обслуживания множества клиентов, причем одновременно. Чтобы посмотреть, как наш сервер справляется с нагрузкой, запустим параллельно восемь экземпляров сценария клиента

с помощью сценария, представленного в примере 12.3. Описание реализации модуля launchmodes, используемого здесь для запуска клиентов, приведено в конце главы 5; там же вы найдете альтернативные приемы на основе модулей multiprocessing и subprocess.

Пример 12.3. PP4E\Internet\Sockets\testecho.py

```
import sys
from PP4E.launchmodes import QuietPortableLauncher

numclients = 8
def start(cmdline):
    QuietPortableLauncher(cmdline, cmdline)()

# start('echo-server.py')              # запустить сервер локально,
                                       # если еще не запущен

args = ' '.join(sys.argv[1:])          # передать имя сервера,
                                       # если он запущен удаленно
for i in range(numclients):
    start('echo-client.py %s' % args)  # запустить 8? клиентов
                                       # для тестирования сервера
```

Если запустить этот сценарий без аргументов, клиенты будут общаться с сервером, выполняющимся на локальном компьютере, по порту 50007. Если передать сценарию действительное имя компьютера, будет установлено соединение с удаленным сервером. В этом эксперименте участвуют три окна консоли – для клиентов, для локального сервера и для удаленного сервера. В Windows при запуске клиентов этим сценарием их вывод отбрасывается, но он был бы аналогичен тому, что мы уже видели выше. Ниже приводится диалог из окна клиента – 8 клиентов запускаются локально и взаимодействуют с локальным и удаленным серверами:

```
C:\...\PP4E\Internet\Sockets> set PYTHONPATH=C:\...\dev\Examples

C:\...\PP4E\Internet\Sockets> python testecho.py

C:\...\PP4E\Internet\Sockets> python testecho.py learning-python.com
```

Если запускаемые клиенты соединяются с сервером, выполняющимся локально (первая команда запуска клиентов), в окне локального сервера появляются сообщения о соединениях:

```
C:\...\PP4E\Internet\Sockets> python echo-server.py
Server connected by ('127.0.0.1', 57721)
Server connected by ('127.0.0.1', 57722)
Server connected by ('127.0.0.1', 57723)
Server connected by ('127.0.0.1', 57724)
Server connected by ('127.0.0.1', 57725)
Server connected by ('127.0.0.1', 57726)
Server connected by ('127.0.0.1', 57727)
Server connected by ('127.0.0.1', 57728)
```

Если сервер выполняется удаленно, сообщения о соединениях клиентов появляются в окне, которое отображает SSH (или другое) соединение с удаленным компьютером, в данном случае – с *learning-python.com*:

```
[...]$ python echo-server.py
Server connected by ('72.236.109.185', 57729)
Server connected by ('72.236.109.185', 57730)
Server connected by ('72.236.109.185', 57731)
Server connected by ('72.236.109.185', 57732)
Server connected by ('72.236.109.185', 57733)
Server connected by ('72.236.109.185', 57734)
Server connected by ('72.236.109.185', 57735)
Server connected by ('72.236.109.185', 57736)
```

Отклонение запросов клиентов на соединение

Таким образом, наш эхо-сервер, действующий локально или удаленно, способен общаться со множеством клиентов. Имейте, однако, в виду, что это верно только для наших простых сценариев, поскольку серверу не требуется много времени для ответа на каждый запрос клиента – он может вовремя вернуться в начало внешнего цикла `while`, чтобы обработать запрос следующего клиента. Если бы он не мог этого сделать, то, возможно, потребовалось бы изменить сервер так, чтобы все клиенты обрабатывались параллельно, иначе некоторым из них пришлось бы отказать в соединении.

Технически попытки подключения клиентов будут завершаться неудачей, когда уже есть пять клиентов, ожидающих, пока сервер обратит на них свое внимание, как определено в вызове метода `listen` в реализации сервера. Чтобы убедиться в правоте этих слов, добавьте где-нибудь внутри главного цикла сервера в примере 12.1, после операции принятия соединения, вызов `time.sleep`, имитирующий продолжительную операцию (если у вас появится желание поэкспериментировать с этим вариантом сервера, вы найдете его в файле *echo-serversleep.py*, в дереве примеров):

```
while True:                                  # пока процесс работает,
    connection, address = sockobj.accept()   # ждать запроса очередного
                                             # клиента
    while True:
        data = connection.recv(1024)   # читать следующую строку из сокета
        time.sleep(3)                  # время, необходимое
        ...                            # на обработку запроса
```

Если затем запустить этот сервер и сценарий `testecho`, запускающий клиентов, вы заметите, что не все 8 клиентов смогли подключиться к серверу, потому что сервер оказался слишком загруженным, чтобы вовремя освободить очередь ожидающих запросов. Когда я запустил этот сервер в Windows, он смог обслужить только 6 клиентов – запрос от одного из них был принят тотчас же, а 5 запросов были помещены в очередь ожидания. Запросы двух последних клиентов были отвергнуты.

В следующем листинге приводятся сообщения, воспроизводимые данным сервером и клиентами, включая сообщения об ошибках, которые были выведены двумя клиентами, получившими отказ. Чтобы в Windows увидеть сообщения, которые было выведены клиентами, можно изменить сценарий testecho так, чтобы он использовал инструмент запуска StartArgs с ключом /B в начале командной строки, для переадресации сообщений в окно консоли (смотрите файл *testecho-messages.py* в дереве примеров):

```
C:\...\PP4E\dev\Examples\PP4E\Internet\Sockets> echo-server-sleep.py
Server connected by ('127.0.0.1', 59625)
Server connected by ('127.0.0.1', 59626)
Server connected by ('127.0.0.1', 59627)
Server connected by ('127.0.0.1', 59628)
Server connected by ('127.0.0.1', 59629)
Server connected by ('127.0.0.1', 59630)

C:\...\PP4E\dev\Examples\PP4E\Internet\Sockets> testecho-messages.py
/B echo-client.py
/B echo-client.py
/B echo-client.py
/B echo-client.py
/B echo-client.py
/B echo-client.py
/B echo-client.py
/B echo-client.py
Client received: b'Echo=>Hello network world'

Traceback (most recent call last):
  File "C:\...\PP4E\Internet\Sockets\echo-client.py", line 24, in <module>
    sockobj.connect((serverHost, serverPort))   # соединение с сервером и...
socket.error: [Errno 10061] No connection could be made because the target
machine actively refused it
```

(socket.error: [Ошибка 10061] Невозможно установить соединение, потому что оно отвергнуто другой стороной)

```
Traceback (most recent call last):
  File "C:\...\PP4E\Internet\Sockets\echo-client.py", line 24, in <module>
    sockobj.connect((serverHost, serverPort))   # соединение с сервером и...
socket.error: [Errno 10061] No connection could be made because the target
machine actively refused it
```

(socket.error: [Ошибка 10061] Невозможно установить соединение, потому что оно отвергнуто другой стороной)

```
Client received: b'Echo=>Hello network world'
Client received: b'Echo=>Hello network world'
Client received: b'Echo=>Hello network world'
Client received: b'Echo=>Hello network world'
Client received: b'Echo=>Hello network world'
```

Как видите, в этом примере было запущено 8 клиентов, но только 6 из них смогли воспользоваться услугами такого неповоротливого сервера, а 2 потерпели неудачу и возбудили исключения. Если нельзя быть уверенным, что удовлетворение запросов клиентов требует очень немного внимания от сервера, то для обслуживания множества запросов, перекрывающихся во времени, нам необходимо предусмотреть некоторый механизм, который обеспечил бы параллельное их обслуживание. Чуть ниже мы увидим, каким образом серверы могут надежно обрабатывать несколько клиентов, однако для начала поэкспериментируем с некоторыми специальными портами.

Подключение к зарезервированным портам

Важно также знать, что эти клиент и сервер участвуют в диалоге частного свойства и потому используют порт с номером 50007 – вне диапазона, зарезервированного для стандартных протоколов (0–1023). Однако ничто не мешает клиенту открыть сокет на одном из этих выделенных портов. Например, следующий программный код соединяется с программами, ожидающими запросов на соединение на стандартных портах электронной почты, FTP и веб-сервера HTTP на трех разных компьютерах:

```
C:\...\PP4E\Internet\Sockets> python
>>> from socket import *
>>> sock = socket(AF_INET, SOCK_STREAM)
>>> sock.connect(('pop.secureserver.net', 110)) # подключиться к POP-серверу
>>> print(sock.recv(70))
b'+OK <14654.1272040794@p3pop01-09.prod.phx3.gdg>\r\n'
>>> sock.close()

>>> sock = socket(AF_INET, SOCK_STREAM)
>>> sock.connect(('learning-python.com', 21))  # подключиться к FTP-серверу
>>> print(sock.recv(70))
b'220---------- Welcome to Pure-FTPd [privsep] [TLS] ----------\r\n220-You'
>>> sock.close()

>>> sock = socket(AF_INET, SOCK_STREAM)
>>> sock.connect(('www.python.net', 80))       # подключиться к HTTP-серверу

>>> sock.send(b'GET /\r\n')                    # получить корневую страницу
7
>>> sock.recv(70)
b'<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"\r\n "http://'
>>> sock.recv(70)
b'www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">\r\n<html xmlns="http://www.'
```

Если уметь интерпретировать вывод, возвращаемый серверами этих портов, то можно непосредственно использовать такие сокеты для получения почты, передачи файлов, загрузки веб-страниц и запуска сценариев на сервере. К счастью, нет надобности беспокоиться о деталях про-

исходящего – модули Python poplib, ftplib, http.client и urllib.request предоставляют интерфейсы более высокого уровня для связи с серверами через эти порты. Существуют также другие модули протоколов Python, которые осуществляют то же самое для других стандартных портов (например, NNTP, Telnet и так далее). С некоторыми из этих клиентских модулей протоколов мы познакомимся в следующей главе.[1]

Привязка сокетов к зарезервированным портам

Если говорить о зарезервированных портах, то со стороны клиента нет ограничений на открытие соединения с такими портами, как это было продемонстрировано в предыдущем разделе, но для установки собственных серверных сценариев для этих портов необходимо иметь особые права доступа. На сервере, где находится мой сайт *learning-python.com*, например, порт 80 веб-сервера запрещен для использования простыми сценариями (если не входить в командную оболочку с использованием специальной учетной записи):

```
[...]$ python
>>> from socket import *
>>> sock = socket(AF_INET,SOCK_STREAM)    # попробовать привязать порт 80
>>> sock.bind(('', 80))            # на общем компьютере learning-python.com
Traceback (most recent call last):
  File "<stdin>", line 1, in

  File "<string>", line 1, in bind
socket.error: (13, 'Permission denied')
(socket.error: (13,'Недостаточно прав'))
```

Даже если у пользователя будут все необходимые права, при выполнении этих инструкций будет возбуждено исключение, если порт уже используется действующим веб-сервером. Компьютеры, используемые как общие серверы, действительно резервируют эти порты. Это одна из причин, по которым для тестирования мы будем запускать собственный веб-сервер локально, когда начнем писать серверные сценарии далее в этой книге – программный код, представленный выше, выполняется без ошибок на компьютере в Windows, что позволяет нам экспериментировать с локальными веб-сайтами на отдельном компьютере:

---

[1] Вам может быть интересно узнать, что последняя часть этого примера, обращающаяся к порту 80, представляет в точности то, что делает ваш веб-броузер при просмотре страниц веб-сайтов: переход по ссылкам заставляет его загружать веб-страницы через этот порт. На практике этот скромный порт составляет основу основ всего Веб. В главе 15 мы увидим целую среду приложений, основывающуюся на пересылке форматированных данных через порт 80, – серверные CGI-сценарии. Тем не менее, в самой глубине Веб – это всего только передача байтов через сокеты плюс пользовательский интерфейс. Фокус без покрова таинственности уже не производит такого сильного впечатления!

```
C:\...\PP4E\Internet\Sockets> python
>>> from socket import *
>>> sock = socket(AF_INET, SOCK_STREAM)  # в Windows можно привязать порт 80
>>> sock.bind(('', 80))                  # что позволяет запустить сервер
>>>                                      # на компьютере localhost
```

Подробнее установку веб-сервера мы будем рассматривать в главе 15. А в этой главе нам необходимо представить себе, как в реальности серверы сокетов обслуживают клиентов.

Обслуживание нескольких клиентов

Показанные выше программы клиента и сервера echo иллюстрируют основы использования сокетов. Но реализация сервера страдает довольно существенным недостатком. Как описывалось выше, если соединиться с сервером попытаются сразу несколько клиентов и обработка запроса каждого клиента занимает длительное время, то происходит отказ сервера. Более точно, если трудоемкость обработки данного запроса не позволит серверу вовремя вернуться в цикл, проверяющий наличие новых запросов от клиентов, сервер не сможет удовлетворить все запросы и некоторым клиентам будет отказано в соединении.

В реальных программах клиент/сервер сервер чаще реализуется так, чтобы избежать блокировки новых запросов во время обработки текущего запроса клиента. Вероятно, проще всего достичь этого путем параллельной обработки всех запросов клиентов – в новом процессе, новом потоке выполнения или путем переключения (мультиплексирования) между клиентами вручную в цикле событий. Эта проблема не связана с сокетами как таковыми, и мы уже научились запускать процессы и потоки в главе 5. Но так как эти схемы реализации типичны для серверов, работающих с сокетами, рассмотрим здесь все три способа параллельной обработки запросов клиентов.

Ветвление серверов

Сценарий, представленный в примере 12.4, действует подобно оригинальному серверу echo, но для обработки каждого нового соединения с клиентом ответвляет новый процесс. Так как функция handleClient выполняется в новом процессе, функция dispatcher может сразу продолжить выполнение своего главного цикла, чтобы обнаружить и обслужить новый поступивший запрос.

Пример 12.4. PP4E\Internet\Sockets\fork-server.py

```
"""
На стороне сервера: открывает сокет на указанном порту, ожидает
поступления сообщения от клиента и отправляет его обратно; порождает
дочерний процесс для обслуживания каждого соединения с клиентом;
```

```
дочерние процессы совместно используют дескрипторы родительских сокетов;
прием ветвления процессов менее переносим, чем прием на основе потоков
выполнения, – он не поддерживается в Windows, если не используется Cygwin
или подобная ей оболочка;
"""

import os, time, sys
from socket import *   # получить конструктор сокетов и константы
myHost = ''            # компьютер-сервер, '' означает локальный хост
myPort = 50007         # использовать незарезервированный номер порта

sockobj = socket(AF_INET, SOCK_STREAM)   # создать объект сокета TCP
sockobj.bind((myHost, myPort))           # связать с номером порта сервера
sockobj.listen(5)                        # не более 5 ожидающих запросов

def now():                               # текущее время на сервере
    return time.ctime(time.time())

activeChildren = []
def reapChildren():           # убрать завершившиеся дочерние процессы,
    while activeChildren:     # иначе может переполниться системная таблица
        pid, stat = os.waitpid(0, os.WNOHANG)  # не блокировать сервер, если
        if not pid: break                # дочерний процесс не завершился
        activeChildren.remove(pid)

def handleClient(connection):            # дочерний процесс: ответить, выйти
    time.sleep(5)                        # имитировать блокирующие действия
    while True:                          # чтение, запись в сокет клиента
        data = connection.recv(1024)     # до получения признака eof, когда
        if not data: break               # сокет будет закрыт клиентом
        reply = 'Echo=>%s at %s' % (data, now())
        connection.send(reply.encode())
    connection.close()
    os._exit(0)

def dispatcher():                      # пока процесс работает
    while True:                        # ждать запроса очередного клиента,
        connection, address = sockobj.accept()          # передать процессу
        print('Server connected by', address, end=' ')  # для обслуживания
        print('at', now())
        reapChildren()                 # теперь убрать завершившиеся потомки
        childPid = os.fork()           # копировать этот процесс
        if childPid == 0:              # в дочернем процессе: обслужить
            handleClient(connection)
        else:                          # иначе: ждать следующего запроса
            activeChildren.append(childPid)  # добавить в список
                                             # активных потомков
dispatcher()
```

Запуск ветвящегося сервера

Некоторые части этого сценария написаны довольно замысловато, и большинство библиотечных вызовов в нем работает только в Unix-подобных системах. Важно, что в Windows он может выполняться под управлением Python для Cygwin, но не под управлением стандартной версии Python для Windows. Однако, прежде чем подробно вникать во все детали ветвления, рассмотрим, как наш сервер обрабатывает несколько клиентских запросов.

Прежде всего, обратите внимание, что для имитации продолжительных операций (таких, как обновление базы данных или пересылки информации по сети) этот сервер добавляет пятисекундную задержку с помощью `time.sleep` внутри функции `handleClient` обработки запроса клиента. После задержки клиенту возвращается ответ, как и раньше. Это значит, что на этот раз клиенты будут получать ответ не ранее, чем через 5 секунд после отправки запроса серверу.

Чтобы помочь следить за запросами и ответами, сервер выводит свое системное время при каждом получении запроса от клиента и добавляет свое системное время к ответу. Клиенты выводят время ответа, полученное с сервера, а не свое собственное – часы на сервере и у клиента могут быть установлены совершенно по-разному, поэтому, чтобы складывать яблоки с яблоками, все действия отмечаются временем сервера. Из-за имитируемой задержки в Windows обычно приходится запускать каждый сценарий клиента в собственном окне консоли (на некоторых платформах клиенты остаются в заблокированном состоянии, пока не получат свой ответ).

Но самое важное здесь, что сценарий выполняет на компьютере сервера один главный родительский процесс, единственной функцией которого является ожидание запросов на соединение (в функции `dispatcher`), плюс один дочерний процесс на каждое активное соединение с клиентом, выполняемый параллельно с главным родительским процессом и другими клиентскими процессами (в функции `handleClient`). В принципе, сервер может обрабатывать запросы от любого количества клиентов без заминок.

Для проверки запустим сервер удаленно в окне SSH или Telnet и запустим три клиента локально в трех разных окнах консоли. Как мы увидим чуть ниже, этот сервер можно также запускать локально, в оболочке Cygwin, если у вас есть она, но нет учетной записи на удаленном сервере, таком как *learning-python.com*, используемый здесь:

[окно сервера (SSH или Telnet)]

```
[...]$ uname -p -o
i686 GNU/Linux
[...]$ python fork-server.py
Server connected by ('72.236.109.185', 58395) at Sat Apr 24 06:46:45 2010
Server connected by ('72.236.109.185', 58396) at Sat Apr 24 06:46:49 2010
Server connected by ('72.236.109.185', 58397) at Sat Apr 24 06:46:51 2010
```

[окно клиента 1]

```
C:\...\PP4E\Internet\Sockets> python echo-client.py learning-python.com
Client received: b"Echo=>b'Hello network world' at Sat Apr 24 06:46:50 2010"
```

[окно клиента 2]

```
C:\...\PP4E\Internet\Sockets> python echo-client.py learning-python.com
                              Bruce
Client received: b"Echo=>b'Bruce' at Sat Apr 24 06:46:54 2010"
```

[окно клиента 3]

```
C:\...\Sockets> python echo-client.py learning-python.com The Meaning
                of Life
Client received: b"Echo=>b'The' at Sat Apr 24 06:46:56 2010"
Client received: b"Echo=>b'Meaning' at Sat Apr 24 06:46:56 2010"
Client received: b"Echo=>b'of' at Sat Apr 24 06:46:56 2010"
Client received: b"Echo=>b'Life' at Sat Apr 24 06:46:57 2010"
```

И снова все значения времени соответствуют времени на сервере. Это может показаться немного странным, поскольку участвуют четыре окна. На обычном языке этот тест можно описать так:

1. На удаленном компьютере запускается сервер.
2. Запускаются все три клиента, которые соединяются с сервером примерно в одно и то же время.
3. На сервере три клиентских запроса запускают три дочерних процесса, которые сразу приостанавливаются на пять секунд (изображая занятость чем-то полезным).
4. Каждый клиент ждет ответа сервера, который генерируется через пять секунд после получения запроса.

Иными словами, клиенты обслуживаются дочерними процессами, запущенными в одно и то же время, при этом главный родительский процесс продолжает ждать новых клиентских запросов. Если бы клиенты не обслуживались параллельно, ни один из них не смог бы соединиться до истечения пятисекундной задержки, вызванной обработкой текущего клиента.

В действующем приложении такая задержка могла бы оказаться роковой, если бы к серверу попытались подключиться сразу нескольких клиентов – сервер застрял бы на операции, которую мы имитируем с помощью `time.sleep`, и не вернулся бы в главный цикл, чтобы принять новые запросы клиентов. При ветвлении, при котором на каждый запрос отводится по процессу, все клиенты могут обслуживаться параллельно.

Обратите внимание, что здесь используется прежний сценарий клиента (*echo-client.py* из примера 12.2), а сценарий сервера – другой. Клиенты просто посылают свои данные компьютеру сервера в указанный порт и получают их оттуда, не зная, каким образом обслуживаются их запросы на сервере. Отображаемый результат содержит строку байтов, вложенную в другую строку байтов. Это обусловлено тем, что клиент

отправляет серверу какую-то строку байтов, а сервер возвращает какую-то строку обратно – сервер использует операции форматирования строк и кодирования вместо конкатенации строк байтов, поэтому клиентское сообщение отображается здесь явно, как строка байтов.

Другие способы запуска: локальные серверы в Cygwin и удаленные клиенты

Обратите также внимание, что сервер удаленно выполняется на компьютере с операционной системой Linux. Как мы узнали в главе 5, на момент написания книги функция `fork` не поддерживается в Python для Windows. Однако сервер может выполняться под управлением Python для Cygwin, что позволяет запустить его локально на компьютере `localhost`, где запускаются клиенты:

```
[окно оболочки Cygwin]
[C:\...\PP4E\Internet\Socekts]$ python fork-server.py
Server connected by ('127.0.0.1', 58258) at Sat Apr 24 07:50:15 2010
Server connected by ('127.0.0.1', 58259) at Sat Apr 24 07:50:17 2010

[консоль Windows, тот же компьютер]
C:\...\PP4E\Internet\Sockets> python echo-client.py localhost bright side
                              of life
Client received: b"Echo=>b'bright' at Sat Apr 24 07:50:20 2010"
Client received: b"Echo=>b'side' at Sat Apr 24 07:50:20 2010"
Client received: b"Echo=>b'of' at Sat Apr 24 07:50:20 2010"
Client received: b"Echo=>b'life' at Sat Apr 24 07:50:20 2010"

[консоль Windows, тот же компьютер]
C:\...\PP4E\Internet\Sockets> python echo-client.py
Client received: b"Echo=>b'Hello network world' at Sat Apr 24 07:50:22 2010"
```

Можно запустить этот тест целиком на удаленном сервере Linux посредством двух окон SSH или Telnet. Он будет действовать примерно так же, как при запуске клиентов на локальном компьютере, в окне консоли DOS, но здесь «локальный» означает удаленный компьютер, с которым вы работаете локально. Забавы ради попробуем также соединиться с удаленным сервером из клиента, запущенного локально, чтобы показать, что сервер может быть доступным из Интернета в целом – когда серверы запрограммированы с сокетами и ветвятся подобным образом, клиенты могут подключаться к ним, находясь на любых компьютерах, и их запросы могут поступать одновременно:

```
[одно окно SSH (или Telnet)]
[...]$ python fork-server.py
Server connected by ('127.0.0.1', 55743) at Sat Apr 24 07:15:14 2010
Server connected by ('127.0.0.1', 55854) at Sat Apr 24 07:15:26 2010
Server connected by ('127.0.0.1', 55950) at Sat Apr 24 07:15:36 2010
Server connected by ('72.236.109.185', 58414) at Sat Apr 24 07:19:50 2010
```

[другое окно SSH, тот же компьютер]

```
[...]$ python echo-client.py
Client received: b"Echo=>b'Hello network world' at Sat Apr 24 07:15:19 2010"
[...]$ python echo-client.py localhost niNiNI!
Client received: b"Echo=>b'niNiNI!' at Sat Apr 24 07:15:31 2010"
[...]$ python echo-client.py localhost Say no more!
Client received: b"Echo=>b'Say' at Sat Apr 24 07:15:41 2010"
Client received: b"Echo=>b'no' at Sat Apr 24 07:15:41 2010"
Client received: b"Echo=>b'more!' at Sat Apr 24 07:15:41 2010"
```

[консоль Windows, локальный компьютер]

```
C:\...\Internet\Sockets> python echo-client.py learning-python.com Blue,
                         no yellow!
Client received: b"Echo=>b'Blue,' at Sat Apr 24 07:19:55 2010"
Client received: b"Echo=>b'no' at Sat Apr 24 07:19:55 2010"
Client received: b"Echo=>b'yellow!' at Sat Apr 24 07:19:55 2010"
```

Теперь, когда мы достигли понимания принципов работы основной модели, перейдем к рассмотрению некоторых хитростей. Реализация сценария сервера, организующего ветвление, достаточно проста, но следует сказать несколько слов об использовании некоторых библиотечных инструментов.

Ветвление процессов и сокеты

Мы уже познакомились с функцией os.fork в главе 5, тем не менее напомню, что ответвленные дочерние процессы в сущности являются копией породившего их процесса и наследуют от родительского процесса дескрипторы файлов и сокетов. Благодаря этому новый дочерний процесс, выполняющий функцию handleClient, имеет доступ к сокету соединения, созданному в родительском процессе. Именно поэтому оказывается возможной работа дочерних процессов – для общения с клиентом дочерний процесс использует тот же сокет, который был создан вызовом метода accept в родительском процессе. Программы узнают о том, что они выполняются в ответвленном дочернем процессе, если вызов fork возвращает 0 – в родительском процессе эта функция возвращает идентификатор нового дочернего процесса.

Завершение дочерних процессов

В предшествующих примерах ветвления дочерние процессы обычно вызывали одну из функций семейства exec для запуска новой программы в дочернем процессе. Здесь же дочерний процесс просто вызывает функцию в той же программе и завершается с помощью функции os._exit. Здесь необходимо вызывать os._exit – если этого не сделать, дочерний процесс продолжит существовать после возврата из handleClient и также примет участие в приеме новых запросов от клиентов.

На самом деле без вызова os._exit мы получили бы столько вечных процессов сервера, сколько было обслужено запросов – уберите вызов

os._exit, выполните команду оболочки ps после запуска нескольких клиентов, и вы поймете, что я имею в виду. При наличии вызова этой функции только родительский процесс будет ждать новые запросы. Функция os._exit похожа на sys.exit, но завершает вызвавший его процесс сразу, не выполняя заключительных операций. Обычно он используется только в дочерних процессах, а sys.exit используется во всех остальных случаях.

Удаление зомби: не бойтесь грязной работы

Заметьте, однако, что недостаточно просто убедиться в завершении дочернего процесса. В таких системах, как Linux, но не в Cygwin, родительский процесс должен также выполнить системный вызов wait, чтобы удалить записи, касающиеся завершившихся дочерних процессов, из системной таблицы процессов. Если этого не сделать, то дочерние процессы выполняться не будут, но будут занимать место в системной таблице процессов. Для серверов, выполняющихся длительное время, такие фальшивые записи могут вызвать неприятности.

Такие недействующие, но числящиеся в строю процессы обычно называют *зомби*: они продолжают использовать системные ресурсы даже после возврата в операционную систему. Для освобождения ресурсов, занимаемых завершившимися дочерними процессами, наш сервер ведет список activeChildren, содержащий идентификаторы всех порожденных им дочерних процессов. При получении нового запроса от клиента сервер вызывает функцию reapChildren, чтобы вызвать wait для всех завершившихся дочерних процессов путем вызова стандартной функции Python os.waitpid(0,os.WNOHANG).

Функция os.waitpid пытается дождаться завершения дочернего процесса и возвращает идентификатор этого процесса и код завершения. При передаче 0 в первом аргументе ожидается завершение любого дочернего процесса. При передаче значения WNOHANG во втором аргументе функция ничего не делает, если к этому моменту никакой дочерний процесс не завершился (то есть вызвавший процесс не блокируется и не приостанавливается). В итоге данный вызов просто запрашивает у операционной системы идентификатор любого завершившегося дочернего процесса. Если такой процесс есть, полученный идентификатор удаляется из системной таблицы процессов и из списка activeChildren этого сценария.

Чтобы понять, для чего нужны такие сложности, закомментируйте в этом сценарии вызов функции reapChildren, запустите его на сервере, где проявляются описанные выше проблемы, а затем запустите несколько клиентов. На моем сервере Linux команда ps -f, которая выводит полный список процессов, показывает, что все завершившиеся дочерние процессы сохраняются в системной таблице процессов (помечены как <defunct>):

```
[...]$ ps -f
UID        PID  PPID  C STIME TTY          TIME CMD
```

```
5693094   9990 30778  0 04:34 pts/0   00:00:00 python fork-server.py
5693094  10844  9990  0 04:35 pts/0   00:00:00 [python] <defunct>
5693094  10869  9990  0 04:35 pts/0   00:00:00 [python] <defunct>
5693094  11130  9990  0 04:36 pts/0   00:00:00 [python] <defunct>
5693094  11151  9990  0 04:36 pts/0   00:00:00 [python] <defunct>
5693094  11482 30778  0 04:36 pts/0   00:00:00 ps -f
5693094  30778 30772  0 04:23 pts/0   00:00:00 -bash
```

Если снова раскомментировать вызов функции `reapChildren`, записи о завершившихся дочерних зомби будут удаляться всякий раз, когда сервер будет получать от клиента новый запрос на соединение, путем вызова функции `os.waitpid`. Если сервер сильно загружен, может накопиться несколько зомби, но они сохранятся только до получения нового запроса на соединение от клиента:

```
[...]$ python fork-server.py &
[1] 20515
[...]$ ps -f
UID        PID  PPID  C STIME TTY         TIME CMD
5693094  20515 30778  0 04:43 pts/0   00:00:00 python fork-server.py
5693094  20777 30778  0 04:43 pts/0   00:00:00 ps -f
5693094  30778 30772  0 04:23 pts/0   00:00:00 -bash
[...]$
Server connected by ('72.236.109.185', 58672) at Sun Apr 25 04:43:51 2010
Server connected by ('72.236.109.185', 58673) at Sun Apr 25 04:43:54 2010
[...]$ ps -f
UID        PID  PPID  C STIME TTY         TIME CMD
5693094  20515 30778  0 04:43 pts/0   00:00:00 python fork-server.py
5693094  21339 20515  0 04:43 pts/0   00:00:00 [python] <defunct>
5693094  21398 20515  0 04:43 pts/0   00:00:00 [python] <defunct>
5693094  21573 30778  0 04:44 pts/0   00:00:00 ps -f
5693094  30778 30772  0 04:23 pts/0   00:00:00 -bash
[...]$
Server connected by ('72.236.109.185', 58674) at Sun Apr 25 04:44:07 2010
[...]$ ps -f
UID        PID  PPID  C STIME TTY         TIME CMD
5693094  20515 30778  0 04:43 pts/0   00:00:00 python fork-server.py
5693094  21646 20515  0 04:44 pts/0   00:00:00 [python] <defunct>
5693094  21813 30778  0 04:44 pts/0   00:00:00 ps -f
5693094  30778 30772  0 04:23 pts/0   00:00:00 -bash
```

Фактически, если вы печатаете очень быстро, можно успеть увидеть, как дочерний процесс превращается из выполняющейся программы в зомби. Здесь, например, дочерний процесс, порожденный для обработки нового запроса, при выходе превращается в `<defunct>`. Его подключение удаляет оставшиеся зомби, а его собственная запись о процессе будет полностью удалена при получении следующего запроса:

```
[...]$
Server connected by ('72.236.109.185', 58676) at Sun Apr 25 04:48:22 2010
[...] ps -f
UID        PID  PPID  C STIME TTY         TIME CMD
```

```
5693094  20515 30778  0 04:43 pts/0   00:00:00 python fork-server.py
5693094  27120 20515  0 04:48 pts/0   00:00:00 python fork-server.py
5693094  27174 30778  0 04:48 pts/0   00:00:00 ps -f
5693094  30778 30772  0 04:23 pts/0   00:00:00 -bash
[...]$ ps -f
UID        PID  PPID  C STIME TTY         TIME CMD
5693094  20515 30778  0 04:43 pts/0   00:00:00 python fork-server.py
5693094  27120 20515  0 04:48 pts/0   00:00:00 [python] <defunct>
5693094  27234 30778  0 04:48 pts/0   00:00:00 ps -f
5693094  30778 30772  0 04:23 pts/0   00:00:00 -bash
```

Предотвращение появления зомби с помощью обработчиков сигналов

В некоторых системах можно также удалять дочерние процессы-зомби путем переустановки обработчика сигнала SIGCHLD, отправляемого операционной системой родительскому процессу по завершении дочернего процесса. Если сценарий Python определит в качестве обработчика сигнала SIGCHLD действие SIG_IGN (игнорировать), зомби будут удаляться автоматически и немедленно по завершении дочерних процессов – родительскому процессу не придется выполнять вызовы wait, чтобы освободить ресурсы, занимаемые ими. Благодаря этому такая схема служит более простой альтернативой ручному удалению зомби на платформах, где она поддерживается.

Если вы прочли главу 5, то знаете, что обработчики сигналов, программно-генерируемых событий, можно устанавливать с помощью стандартного модуля Python signal. В качестве демонстрации ниже приводится небольшой пример, который показывает, как это можно использовать для удаления зомби. Сценарий в примере 12.5 устанавливает функцию обработчика сигналов, написанную на языке Python, реагирующую на номер сигнала, вводимый в командной строке.

Пример 12.5. PP4E\Internet\Sockets\signal-demo.py

```
……
Демонстрация модуля signal; номер сигнала передается в аргументе командной
строки, а отправить сигнал этому процессу можно с помощью команды оболочки
¨kill -N pid¨; на моем компьютере с Linux SIGUSR1=10, SIGUSR2=12, SIGCHLD=17
и обработчик SIGCHLD остается действующим, даже если не восстанавливается
в исходное состояние: все остальные обработчики сигналов переустанавливаются
интерпретатором Python после получения сигнала, но поведение сигнала SIGCHLD
не регламентируется и его реализация оставлена за платформой;
модуль signal можно также использовать в Windows, но в ней доступны
лишь несколько типов сигналов; в целом сигналы не очень хорошо переносимы;
……


import sys, signal, time

def now():
    return time.asctime()
```

```
def onSignal(signum, stackframe):             # обработчик сигнала на Python
    print('Got signal', signum, 'at', now())  # большинство обработчиков
    if signum == signal.SIGCHLD:            # не требуется переустанавливать,
        print('sigchld caught')             # кроме обработчика sigchld
        #signal.signal(signal.SIGCHLD, onSignal)

signum = int(sys.argv[1])
signal.signal(signum, onSignal)             # установить обработчик сигнала
while True: signal.pause()                  # ждать появления сигнала
```

Чтобы опробовать этот сценарий, просто запустите его в фоновом режиме и посылайте ему сигналы, вводя команду `kill -номер-сигнала id-процесса` в командной строке – это эквивалент функции `os.kill` в языке Python, доступной только в Unix-подобных системах. Идентификаторы процессов перечислены в колонке PID результатов выполнения команды `ps`. Ниже показано, как действует этот сценарий, перехватывая сигналы с номерами 10 (зарезервирован для общего использования) и 9 (безусловный сигнал завершения):

```
[...]$ python signal-demo.py 10 &
[1] 10141
[...]$ ps -f
UID        PID  PPID  C STIME TTY         TIME CMD
5693094  10141 30778  0 05:00 pts/0   00:00:00 python signal-demo.py 10
5693094  10228 30778  0 05:00 pts/0   00:00:00 ps -f
5693094  30778 30772  0 04:23 pts/0   00:00:00 -bash

[...]$ kill -10 10141
Got signal 10 at Sun Apr 25 05:00:31 2010

[...]$ kill -10 10141
Got signal 10 at Sun Apr 25 05:00:34 2010

[...]$ kill -9 10141
[1]+  Killed                  python signal-demo.py 10
```

А в следующем примере сценарий перехватывает сигнал с номером 17, который на моем сервере с Linux соответствует сигналу `SIGCHLD`. Номера сигналов зависят от используемой операционной системы, поэтому обычно следует пользоваться именами сигналов, а не номерами. Поведение сигнала `SIGCHLD` тоже может зависеть от платформы. У меня в установленной оболочке Cygwin, например, сигнал с номером 10 может иметь другое назначение, а сигнал `SIGCHLD` имеет номер 20. В Cygwin данный сценарий обрабатывает сигнал 10 так же, как в Linux, но при попытке установить обработчик сигнала 17 возбуждает исключение (впрочем, в Cygwin нет необходимости удалять зомби). Дополнительные подробности смотрите в руководстве по библиотеке, в разделе с описанием модуля `signal`:

```
[...]$ python signal-demo.py 17 &
[1] 11592
```

```
[...]$ ps -f
UID        PID  PPID  C STIME TTY         TIME CMD
5693094  11592 30778  0 05:00 pts/0   00:00:00 python signal-demo.py 17
5693094  11728 30778  0 05:01 pts/0   00:00:00 ps -f
5693094  30778 30772  0 04:23 pts/0   00:00:00 -bash

[...]$ kill -17 11592
Got signal 17 at Sun Apr 25 05:01:28 2010
sigchld caught

[...]$ kill -17 11592
Got signal 17 at Sun Apr 25 05:01:35 2010
sigchld caught

[...]$ kill -9 11592
[1]+  Killed                  python signal-demo.py 17
```

Теперь, чтобы применить все эти знания для удаления зомби, просто установим в качестве обработчика сигнала SIGCHLD действие SIG_IGN — в системах, где поддерживается такое назначение, дочерние процессы будут удаляться сразу же по их завершении. Вариант ветвящегося сервера, представленный в примере 12.6, использует этот прием для управления своими дочерними процессами.

Пример 12.6. PP4E\Internet\Sockets\fork-server-signal.py

```
"""
То же, что и fork-server.py, но использует модуль signal, чтобы обеспечить
автоматическое удаление дочерних процессов-зомби после их завершения вместо
явного удаления перед приемом каждого нового соединения; действие SIG_IGN
означает игнорирование и может действовать с сигналом SIGCHLD завершения
дочерних процессов не на всех платформах; смотрите документацию
к операционной системе Linux, где описывается возможность перезапуска
вызова socket.accept, прерванного сигналом;
"""

import os, time, sys, signal, signal
from socket import * # получить конструктор сокета и константы
myHost = ''          # компьютер сервера, '' означает локальный хост
myPort = 50007       # использовать незарезервированный номер порта

sockobj = socket(AF_INET, SOCK_STREAM)  # создать объект сокета TCP
sockobj.bind((myHost, myPort))          # связать с номером порта сервера
sockobj.listen(5)                       # не более 5 ожидающих запросов
signal.signal(signal.SIGCHLD, signal.SIG_IGN)  # автоматически удалять
                                               # дочерние процессы-зомби
def now():                                     # текущее время на сервере
    return time.ctime(time.time())

def handleClient(connection):         # дочерний процесс: ответить, выйти
    time.sleep(5)                     # имитировать блокирующие действия
    while True:                       # чтение, запись в сокет клиента
```

```
        data = connection.recv(1024)
        if not data: break
        reply = 'Echo=>%s at %s' % (data, now())
        connection.send(reply.encode())
    connection.close()
    os._exit(0)

def dispatcher():                       # пока процесс работает
    while True:                         # ждать запроса очередного клиента,
        connection, address = sockobj.accept()          # передать процессу
        print('Server connected by', address, end=' ')  # для обслуживания
        print('at', now())
        childPid = os.fork()            # копировать этот процесс
        if childPid == 0:               # в дочернем процессе: обслужить
            handleClient(connection)    # иначе: ждать следующего запроса

dispatcher()
```

Там, где возможно его применение, такой прием:

- Гораздо проще – не нужно следить за дочерними процессами и вручную убирать их.
- Более точный – нет зомби, временно присутствующих в промежутке между запросами клиентов.

На самом деле обработкой зомби здесь занимается всего одна строка программного кода: вызов функции `signal.signal` в начале сценария, устанавливающий обработчик. К сожалению, данная версия еще в меньшей степени переносима, чем первая с использованием `os.fork`, поскольку действие сигналов может несколько различаться в зависимости от платформы. Например, на некоторых платформах вообще не разрешается использовать `SIG_IGN` в качестве действия для `SIGCHLD`. Однако в системах Linux этот более простой сервер с ветвлением действует замечательно:

```
[...]$ python fork-server-signal.py &
[1] 3837
Server connected by ('72.236.109.185', 58817) at Sun Apr 25 08:11:12 2010

[...] ps -f
UID        PID  PPID  C STIME TTY         TIME CMD
5693094   3837 30778  0 08:10 pts/0   00:00:00 python fork-server-signal.py
5693094   4378  3837  0 08:11 pts/0   00:00:00 python fork-server-signal.py
5693094   4413 30778  0 08:11 pts/0   00:00:00 ps -f
5693094  30778 30772  0 04:23 pts/0   00:00:00 -bash

[...]$ ps -f
UID        PID  PPID  C STIME TTY         TIME CMD
5693094   3837 30778  0 08:10 pts/0   00:00:00 python fork-server-signal.py
5693094   4584 30778  0 08:11 pts/0   00:00:00 ps -f
5693094  30778 30772  0 04:23 pts/0   00:00:00 -bash
```

Обратите внимание, что в этой версии запись о дочернем процессе исчезает сразу, как только он завершается, даже раньше, чем будет получен новый клиентский запрос. Никаких зомби с пометкой «defunct» не возникает. Еще более знаменательно, что если теперь запустить наш более старый сценарий, порождающий восемь параллельных клиентов (*testecho.py*), соединяющихся с сервером, то все они появляются на сервере при выполнении и немедленно удаляются после завершения:

```
[окно клиента]
C:\...\PP4E\Internet\Sockets> testecho.py learning-python.com

[окно сервера]
[...]$
Server connected by ('72.236.109.185', 58829) at Sun Apr 25 08:16:34 2010
Server connected by ('72.236.109.185', 58830) at Sun Apr 25 08:16:34 2010
Server connected by ('72.236.109.185', 58831) at Sun Apr 25 08:16:34 2010
Server connected by ('72.236.109.185', 58832) at Sun Apr 25 08:16:34 2010
Server connected by ('72.236.109.185', 58833) at Sun Apr 25 08:16:34 2010
Server connected by ('72.236.109.185', 58834) at Sun Apr 25 08:16:34 2010
Server connected by ('72.236.109.185', 58835) at Sun Apr 25 08:16:34 2010
Server connected by ('72.236.109.185', 58836) at Sun Apr 25 08:16:34 2010

[...]$ ps -f
UID        PID  PPID  C STIME TTY         TIME CMD
5693094   3837 30778  0 08:10 pts/0   00:00:00 python fork-server-signal.py
5693094   9666  3837  0 08:16 pts/0   00:00:00 python fork-server-signal.py
5693094   9667  3837  0 08:16 pts/0   00:00:00 python fork-server-signal.py
5693094   9668  3837  0 08:16 pts/0   00:00:00 python fork-server-signal.py
5693094   9670  3837  0 08:16 pts/0   00:00:00 python fork-server-signal.py
5693094   9674  3837  0 08:16 pts/0   00:00:00 python fork-server-signal.py
5693094   9678  3837  0 08:16 pts/0   00:00:00 python fork-server-signal.py
5693094   9681  3837  0 08:16 pts/0   00:00:00 python fork-server-signal.py
5693094   9682  3837  0 08:16 pts/0   00:00:00 python fork-server-signal.py
5693094   9722 30778  0 08:16 pts/0   00:00:00 ps -f
5693094  30778 30772  0 04:23 pts/0   00:00:00 -bash

[...]$ ps -f
UID        PID  PPID  C STIME TTY         TIME CMD
5693094   3837 30778  0 08:10 pts/0   00:00:00 python fork-server-signal.py
5693094  10045 30778  0 08:16 pts/0   00:00:00 ps -f
5693094  30778 30772  0 04:23 pts/0   00:00:00 -bash
```

Теперь, когда я показал вам, как использовать обработчики сигналов для автоматического удаления записей о дочерних процессах в Linux, я должен подчеркнуть, что этот прием не является универсальным и поддерживается не всеми версиями Unix. Если переносимость имеет важное значение, предпочтительнее использовать прием удаления дочерних процессов вручную, использовавшийся в примере 12.4.

Почему модуль multiprocessing не обеспечивает переносимость серверов сокетов

В главе 5 мы познакомились с новым модулем multiprocessing. Как мы видели, он обеспечивает более переносимую возможность выполнения функций в новых процессах, чем функция os.fork, использованная в реализации этого сервера, и выполняет их не в потоках, а в отдельных процессах, чтобы обойти ограничения, накладываемые глобальной блокировкой GIL. В частности, модуль multiprocessing можно использовать также в стандартной версии Python для Windows, в отличие от функции os.fork.

Я решил поэкспериментировать с версией сервера, опирающегося на этот модуль, чтобы посмотреть, сможет ли он помочь повысить переносимость серверов сокетов. Полный программный код этого сервера можно найти в файле *multi-server.py* в дереве примеров, а ниже приводятся несколько наиболее важных отличительных фрагментов:

```
...остальной программный код не отличается от fork-server.py...
from multiprocessing import Process

def handleClient(connection):
    print('Child:', os.getpid())      # дочерний процесс: ответить, выйти
    time.sleep(5)                     # имитировать блокирующие действия
    while True:                       # чтение, запись в сокет клиента
        data = connection.recv(1024)  # продолжать, пока сокет
                                      # не будет закрыт
        ... остальной программный код не отличается...

def dispatcher():                     # пока процесс работает
    while True:                       # ждать запроса очередного клиента
        connection, address = sockobj.accept()   # передать процессу
        print('Server connected by', address, end=' ')  # для обслуживания
        print('at', now())
        Process(target=handleClient, args=(connection,)).start()

if __name__ == '__main__':
    print('Parent:', os.getpid())
    sockobj = socket(AF_INET, SOCK_STREAM)  # создать объект сокета TCP
    sockobj.bind((myHost, myPort))       # связать с номером порта сервера
    sockobj.listen(5)                    # не более 5 ожидающих запросов
    dispatcher()
```

Эта версия сервера заметно проще. Подобно ветвящемуся серверу, версией которого он является, данный сервер отлично работает на компьютере localhost под управлением Python для Cygwin в Windows. Вероятно, он также будет работать в других Unix-подобных системах, потому что в таких системах модуль multiprocessing использует прием ветвления процессов, при котором дескрипторы файлов и сокетов наследуются дочерними процессами как обычно. Следовательно, дочерний процесс будет использовать тот же самый подключенный сокет, что и роди-

тельский процесс. Ниже демонстрируется картина, наблюдаемая в окне оболочки Cygwin в Windows, где запущен сервер, и в двух окнах с клиентами:

```
[окно сервера]
[C:\...\PP4E\Internet\Sockets]$ python multi-server.py
Parent: 8388
Server connected by ('127.0.0.1', 58271) at Sat Apr 24 08:13:27 2010
Child: 8144
Server connected by ('127.0.0.1', 58272) at Sat Apr 24 08:13:29 2010
Child: 8036

[два окна клиентов]
C:\...\PP4E\Internet\Sockets> python echo-client.py
Client received: b"Echo=>b'Hello network world' at Sat Apr 24 08:13:33 2010"

C:\...\PP4E\Internet\Sockets> python echo-client.py localhost Brave
                              Sir Robin
Client received: b"Echo=>b'Brave' at Sat Apr 24 08:13:35 2010"
Client received: b"Echo=>b'Sir' at Sat Apr 24 08:13:35 2010"
Client received: b"Echo=>b'Robin' at Sat Apr 24 08:13:35 2010"
```

Однако этот сервер не работает под управлением стандартной версии Python для Windows – из-за попытки использовать модуль `multiprocessing` в этом контексте – потому что открытые сокеты некорректно сериализуются при передаче новому процессу в виде аргументов. Ниже показано, что происходит в окне сервера в Windows 7, где установлена версия Python 3.1:

```
C:\...\PP4E\Internet\Sockets> python multi-server.py
Parent: 9140
Server connected by ('127.0.0.1', 58276) at Sat Apr 24 08:17:41 2010
Child: 9628
Process Process-1:
Traceback (most recent call last):
  File "C:\Python31\lib\multiprocessing\process.py", line 233, in _bootstrap
    self.run()
  File "C:\Python31\lib\multiprocessing\process.py", line 88, in run
    self._target(*self._args, **self._kwargs)
  File "C:\...\PP4E\Internet\Sockets\multi-server.py", line 38, in
    handleClient data = connection.recv(1024)   # продолжать, пока сокет...
socket.error: [Errno 10038] An operation was attempted on something that is
not a socket
(socket.error: [Ошибка 10038] Попытка выполнить операцию с объектом,
не являющимся сокетом)
```

Как рассказывалось в главе 5, в Windows модуль `multiprocessing` передает контекст новому процессу интерпретатора Python, сериализуя его с помощью модуля `pickle`, поэтому аргументы конструктора `Process` при вызове в Windows должны поддерживать возможность сериализации. При попытке сериализовать сокеты в Python 3.1 ошибки не возникает,

благодаря тому, что они являются экземплярами классов, но сама сериализация выполняется некорректно:

```
>>> from pickle import *
>>> from socket import *
>>> s = socket()
>>> x = dumps(s)
>>> s
<socket.socket object, fd=180, family=2, type=1, proto=0>
>>> loads(x)
<socket.socket object, fd=-1, family=0, type=0, proto=0>
>>> x
b'\x80\x03csocket\nsocket\nq\x00)\x81q\x01N}q\x02(X\x08\x00\x00\x00_io_
refsq\x03K\x00X\x07\x00\x00\x00_closedq\x04\x89u\x86q\x05b.'
```

Как мы видели в главе 5, модуль `multiprocessing` имеет другие инструменты IPC, такие как его собственные каналы и очереди, которые могут использоваться вместо сокетов для решения этой проблемы. Но тогда их должны были бы использовать и клиенты – получившийся в результате сервер оказался бы не так широко доступен, как сервер на основе сокетов Интернета.

Но даже если бы модуль `multiprocessing` работал в Windows, необходимость запускать новый процесс интерпретатора Python сделала бы сервер более медленным, чем более традиционные приемы порождения дочерних потоков выполнения для общения с клиентами. Что, по случайному совпадению, является темой следующего раздела.

Многопоточные серверы

Только что описанная модель ветвления в целом хорошо работает на Unix-подобных платформах, но потенциально страдает существенными ограничениями:

Производительность

На некоторых компьютерах запуск нового процесса обходится довольно дорого в отношении ресурсов времени и памяти.

Переносимость

Ветвление процессов – это инструмент Unix. Как мы уже знаем, функция `os.fork` в настоящее время не работает на платформах, отличных от Unix, таких как Windows под управлением стандартной версии Python. Как мы также узнали ранее, функцию `fork` можно использовать в Windows, в версии Python для Cygwin, но она может быть недостаточно эффективной и не точно соответствовать версии `fork` в Unix. И, как мы только что обнаружили, модуль `multiprocessing` не способен решить проблему в Windows, потому что подключенные сокеты не могут передаваться в сериализованном виде через границы процессов.

Сложность

Если вам кажется, что организация серверов на основе ветвления процессов может оказаться сложной, то вы правы. Как мы только что видели, ветвление приводит ко всей этой мороке по управлению и удалению зомби, – к зачистке после дочерних процессов, завершающихся раньше, чем их родители.

После прочтения главы 5 вам должно быть известно, что обычным решением этих проблем является использование *потоков выполнения* вместо процессов. Потоки выполняются параллельно и совместно используют глобальную память (то есть память модуля и интерпретатора).

Поскольку все потоки выполняются в пределах одного процесса и в той же области памяти, они автоматически получают сокеты в общее пользование и могут передавать их друг другу, примерно так же, как дочерние процессы получают в наследство дескрипторы сокетов. Однако, в отличие от процессов, запуск потоков обычно требует меньших издержек, а работать в настоящее время они могут и в Unix-подобных системах, и в Windows, под управлением стандартных версий Python. Кроме того, многие (хотя и не все) считают потоки более простыми в программировании – дочерние потоки завершаются тихо, не оставляя за собой зомби, преследующих сервер.

В примере 12.7 представлена еще одна версия эхо-сервера, в которой параллельная обработка клиентских запросов выполняется в потоках, а не в процессах.

Пример 12.7. PP4E\Internet\Sockets\thread-server.py

```
"""
На стороне сервера: открывает сокет с указанным номером порта, ожидает
появления сообщения от клиента и отправляет это же сообщение обратно;
продолжает возвращать сообщения клиенту, пока не будет получен признак eof
при закрытии сокета на стороне клиента; для обслуживания клиентов порождает
дочерние потоки выполнения; потоки используют глобальную память совместно
с главным потоком; этот прием является более переносимым, чем ветвление:
потоки выполнения действуют в стандартных версиях Python для Windows,
тогда как прием ветвления - нет;
"""


import time, _thread as thread # или использовать threading.Thread().start()
from socket import *        # получить конструктор сокетов и константы
myHost = ''                 # компьютер-сервер, '' означает локальный хост
myPort = 50007              # использовать незарезервированный номер порта

sockobj = socket(AF_INET, SOCK_STREAM)  # создать объект сокета TCP
sockobj.bind((myHost, myPort))          # связать с номером порта сервера
sockobj.listen(5)                       # не более 5 ожидающих запросов

def now():
    return time.ctime(time.time())      # текущее время на сервере
```

```
def handleClient(connection):           # в дочернем потоке: ответить
    time.sleep(5)                       # имитировать блокирующие действия
    while True:                         # чтение, запись в сокет клиента
        data = connection.recv(1024)
        if not data: break
        reply = 'Echo=>%s at %s' % (data, now())
        connection.send(reply.encode())
    connection.close()

def dispatcher():                       # пока процесс работает,
    while True:                         # ждать запроса очередного клиента,
        connection, address = sockobj.accept()         # передать потоку
        print('Server connected by', address, end=' ') # для обслуживания
        print('at', now())
        thread.start_new_thread(handleClient, (connection,))

dispatcher()
```

Эта функция `dispatcher` передает каждый входящий клиентский запрос в новый порождаемый поток, выполняющий функцию `handleClient`. Благодаря этому данный сервер может одновременно обрабатывать несколько клиентов, а главный цикл диспетчера может быстро вернуться в начало и проверить поступление новых запросов. В результате новым клиентам не будет отказано в обслуживании из-за занятости сервера.

Функционально эта версия аналогична решению на основе функции `fork` (клиенты обрабатываются параллельно), но может работать в любой системе, поддерживающей потоки выполнения, в том числе в Windows и Linux. Проверим ее работу в обеих системах. Сначала запустим сервер в Linux, а клиентские сценарии – в Linux и в Windows:

[окно 1: серверный процесс, использующий потоки; сервер продолжает принимать запросы клиентов и при этом обслуживание предыдущих запросов производится в дочерних потоках]

```
[...]$ python thread-server.py
Server connected by ('127.0.0.1', 37335) at Sun Apr 25 08:59:05 2010
Server connected by ('72.236.109.185', 58866) at Sun Apr 25 08:59:54 2010
Server connected by ('72.236.109.185', 58867) at Sun Apr 25 08:59:56 2010
Server connected by ('72.236.109.185', 58868) at Sun Apr 25 08:59:58 2010
```

[окно 2: клиент, выполняющийся на компьютере сервера]

```
[...]$ python echo-client.py
Client received: b"Echo=>b'Hello network world' at Sun Apr 25 08:59:10 2010"
```

[окна 3–5: локальные клиенты, ПК]

```
C:\...\PP4E\Internet\Sockets> python echo-client.py learning-python.com
Client received: b"Echo=>b'Hello network world' at Sun Apr 25 08:59:59 2010"

C:\...\PP4E\Internet\Sockets> python echo-clie nt.py learning-python.com
                              Bruce
Client received: b"Echo=>b'Bruce' at Sun Apr 25 09:00:01 2010"
```

```
C:\...\Sockets> python echo-client.py learning-python.com The Meaning
                of life
Client received: b"Echo=>b'The' at Sun Apr 25 09:00:03 2010"
Client received: b"Echo=>b'Meaning' at Sun Apr 25 09:00:03 2010"
Client received: b"Echo=>b'of' at Sun Apr 25 09:00:03 2010"
Client received: b"Echo=>b'life' at Sun Apr 25 09:00:03 2010"
```

Поскольку вместо ветвящихся процессов этот сервер использует потоки выполнения, его можно запускать переносимым образом в Linux и в Windows. Снова запустим его, на этот раз на одном и том же локальном компьютере с Windows, вместе с клиентами. Главным, на что следует обратить внимание, здесь является то, что во время обслуживания предшествующих клиентов новые запросы могут приниматься и обслуживаться параллельно с другими клиентами и главным потоком (во время 5-секундной задержки):

[окно 1: сервер на локальном PC]

```
C:\...\PP4E\Internet\Sockets> python thread-server.py
Server connected by ('127.0.0.1', 58987) at Sun Apr 25 12:41:46 2010
Server connected by ('127.0.0.1', 58988) at Sun Apr 25 12:41:47 2010
Server connected by ('127.0.0.1', 58989) at Sun Apr 25 12:41:49 2010
```

[окна 2–4: клиенты на локальном PC]

```
C:\...\PP4E\Internet\Sockets> python echo-client.py
Client received: b"Echo=>b'Hello network world' at Sun Apr 25 12:41:51 2010"

C:\...\PP4E\Internet\Sockets> python echo-client.py localhost Brian
Client received: b"Echo=>b'Brian' at Sun Apr 25 12:41:52 2010"

C:\...\PP4E\Internet\Sockets> python echo-client.py localhost Bright side
                              of life
Client received: b"Echo=>b'Bright' at Sun Apr 25 12:41:54 2010"
Client received: b"Echo=>b'side' at Sun Apr 25 12:41:54 2010"
Client received: b"Echo=>b'of' at Sun Apr 25 12:41:54 2010"
Client received: b"Echo=>b'life' at Sun Apr 25 12:41:54 2010"
```

Напомню, что поток просто завершается при возврате из функции, которую он выполняет, – в отличие от версии с ветвлением процессов, в функции обслуживания клиента не вызывается ничего похожего на `os._exit` (этого и нельзя делать – можно завершить все потоки в процессе!). Благодаря этому версия с потоками не только более переносима, но и проще.

Классы серверов в стандартной библиотеке

Теперь, когда я показал, как писать серверы с применением ветвления или потоков для обслуживания клиентов без блокирования входящих запросов, следует сказать, что в библиотеке Python имеются стандартные инструменты, облегчающие этот процесс. В частности, в модуле `socketserver` определены классы, реализующие практически все виды

серверов, реализующих прием ветвления или использующих потоки выполнения, которые могут вас заинтересовать.

Подобно серверам, созданным вручную, которые мы только что рассмотрели, основные классы в этом модуле реализуют серверы, обеспечивающие одновременное (или асинхронное) обслуживание нескольких клиентов, ликвидируя угрозу отказа в обслуживании новых запросов при выполнении продолжительных операций с другими клиентами. Основное их назначение состоит в том, чтобы автоматизировать реализацию наиболее типичных разновидностей серверов. При использовании этого модуля достаточно просто создать объект сервера нужного импортируемого типа и передать ему объект обработчика с собственным методом обратного вызова, как показано в примере 12.8 реализации многопоточного сервера TCP.

Пример 12.8. PP4E\Internet\Sockets\class-server.py

```
"""
На стороне сервера: открывает сокет на указанном порту, ожидает поступления
сообщения от клиента и отправляет его обратно; эта версия использует
стандартный модуль socketserver; модуль socketserver предоставляет классы
TCPServer, ThreadingTCPServer, ForkingTCPServer, их варианты для протокола
UDP и многое другое, передает каждый запрос клиента на соединение методу
handle нового экземпляра указанного объекта обработчика; кроме того,
модуль socketserver поддерживает доменные сокеты Unix, но только
в Unix-подобных системах; смотрите руководство по стандартной
библиотеке Python.
"""

import socketserver, time  # получить серверы сокетов, объекты-обработчики
myHost = ''                # компьютер-сервер, '' означает локальный хост
myPort = 50007             # использовать незарезервированный номер порта
def now():
    return time.ctime(time.time())

class MyClientHandler(socketserver.BaseRequestHandler):
    def handle(self):                       # для каждого клиента
        print(self.client_address, now())   # показать адрес этого клиента
        time.sleep(5)                       # имитировать блокирующие действия
        while True:                         # self.request - сокет клиента
            data = self.request.recv(1024) # чтение, запись в сокет клиента
            if not data: break
            reply = 'Echo=>%s at %s' % (data, now())
            self.request.send(reply.encode())
        self.request.close()

# создать сервер с поддержкой многопоточной модели выполнения,
# слушать/обслуживать клиентов непрерывно
myaddr = (myHost, myPort)
server = socketserver.ThreadingTCPServer(myaddr, MyClientHandler)
server.serve_forever()
```

Этот сервер действует так же, как сервер с потоками выполнения, написанный нами вручную в предыдущем разделе, но здесь усилия сосредоточены на реализации услуги (индивидуальной реализации метода handle), а не на деталях поддержки многопоточной модели выполнения. И выполняется он точно так же – ниже приводится результат обработки трех клиентов, созданных вручную, и восьми, порожденных сценарием testecho из примера 12.3:

[окно 1: сервер, serverHost='localhost' в echo-client.py]

```
C:\...\PP4E\Internet\Sockets> python class-server.py
('127.0.0.1', 59036) Sun Apr 25 13:50:23 2010
('127.0.0.1', 59037) Sun Apr 25 13:50:25 2010
('127.0.0.1', 59038) Sun Apr 25 13:50:26 2010
('127.0.0.1', 59039) Sun Apr 25 13:51:05 2010
('127.0.0.1', 59040) Sun Apr 25 13:51:05 2010
('127.0.0.1', 59041) Sun Apr 25 13:51:06 2010
('127.0.0.1', 59042) Sun Apr 25 13:51:06 2010
('127.0.0.1', 59043) Sun Apr 25 13:51:06 2010
('127.0.0.1', 59044) Sun Apr 25 13:51:06 2010
('127.0.0.1', 59045) Sun Apr 25 13:51:06 2010
('127.0.0.1', 59046) Sun Apr 25 13:51:06 2010
```

[окна 2-4: клиент, тот же компьютер]

```
C:\...\PP4E\Internet\Sockets> python echo-client.py
Client received: b"Echo=>b'Hello network world' at Sun Apr 25 13:50:28 2010"

C:\...\PP4E\Internet\Sockets> python echo-client.py localhost Arthur
Client received: b"Echo=>b'Arthur' at Sun Apr 25 13:50:30 2010"

C:\...\PP4E\Internet\Sockets> python echo-client.py localhost Brave
                              Sir Robin
Client received: b"Echo=>b'Brave' at Sun Apr 25 13:50:31 2010"
Client received: b"Echo=>b'Sir' at Sun Apr 25 13:50:31 2010"
Client received: b"Echo=>b'Robin' at Sun Apr 25 13:50:31 2010"

C:\...\PP4E\Internet\Sockets> python testecho.py
```

Чтобы создать ветвящийся сервер, достаточно при создании объекта сервера просто использовать имя класса ForkingTCPServer. Модуль socketserver является более мощным, чем может показаться из данного примера: он поддерживает также непараллельные (последовательные или синхронные) серверы, UDP и доменные сокеты Unix и прерывание работы серверов комбинацией клавиш Ctrl-C в Windows. Подробности ищите в руководстве по библиотеке Python.

Для удовлетворения более сложных потребностей в составе стандартной библиотеки Python имеются также инструменты, которые используют представленные здесь серверы и позволяют в нескольких строках реализовать простой, но полнофункциональный сервер HTTP, который знает, как запускать серверные CGI-сценарии. Мы исследуем эти инструменты в главе 15.

Мультиплексирование серверов с помощью select

К настоящему времени мы узнали, как одновременно обслуживать несколько клиентов с помощью ветвления процессов и дочерних потоков, и рассмотрели библиотечный класс, инкапсулирующий обе эти схемы. В обоих подходах обработчики, обслуживающие клиентов, выполняются параллельно друг с другом и с главным циклом, продолжающим ожидать новые входящие запросы. Так как все эти задачи выполняются параллельно (то есть одновременно), сервер не блокируется при получении новых запросов или при выполнении продолжительных операций во время обслуживания клиентов.

Однако технически потоки и процессы на самом деле не выполняются одновременно, если только у вас на компьютере нет очень большого количества процессоров. На практике операционная система проделывает фокус – она делит вычислительную мощность процессора между всеми активными задачами, выполняя часть одной, затем часть другой и так далее. Кажется, что все задачи выполняются параллельно, но это происходит только потому, что операционная система переключается между выполнением разных задач так быстро, что обычно это незаметно. Такой процесс переключения между задачами, осуществляемый операционной системой, иногда называют *квантованием времени* (*time-slicing*). Более общим его названием является *мультиплексирование* (*multiplexing*).

При порождении потоков и процессов мы рассчитываем, что операционная система так будет жонглировать активными задачами, что ни одна из них не будет ущемляться в вычислительных ресурсах, особенно главный поток сервера. Однако нет никаких причин, по которым этим не мог бы заниматься также сценарий Python. Например, сценарий может разделять задачи на несколько этапов – выполнить этап одной задачи, затем другой и так далее, пока все они не будут завершены. Чтобы самостоятельно осуществлять мультиплексирование, сценарий должен лишь уметь распределять свое внимание среди нескольких активных задач.

Серверы могут с помощью этого приема осуществить еще один способ одновременной обработки клиентов, при котором не требуются ни потоки, ни ветвление. Мультиплексируя соединения клиентов и главного диспетчера с помощью системного вызова `select`, можно обслуживать клиентов и принимать новые соединения параллельно (или близко к тому, избегая задержек). Такие серверы иногда называются *асинхронными*, поскольку они обслуживают клиентов импульсно, по мере их готовности к общению. В асинхронных серверах один главный цикл, выполняемый в одном процессе и потоке, решает каждый раз, которому из клиентов должно быть уделено внимание. Запросы клиентов и главный диспетчер получают небольшой квант внимания сервера, если они готовы к общению.

Основное волшебство при такой организации сервера обеспечивается вызовом select операционной системы, доступным в Python на всех основных платформах через стандартный модуль select. Приблизительно действие select заключается в том, что его просят следить за списком источников входных данных, выходных данных и исключительных ситуаций, а он сообщает, какие из источников готовы к обработке. Можно заставить его просто опрашивать все источники, чтобы находить те, которые готовы; ждать готовности источников в течение некоторого предельного времени или ждать неограниченное время готовности к обработке одного или нескольких источников.

При любом режиме использования select позволяет направлять внимание на сокеты, которые готовы к обмену данными, чтобы избежать блокирования при обращении к тем, которые не готовы. Это означает, что, когда источники, передаваемые вызову select, являются сокетами, можно быть уверенными, что такие методы сокетов, как accept, recv и send, не заблокируют (не остановят) сервер при применении к объектам, возвращаемым вызовом select. Благодаря этому сервер с единственным циклом, но использующий select, не застрянет при обслуживании какого-то одного клиента или в ожидании новых, в то время как остальные клиенты будут обделены его вниманием.

Поскольку такая разновидность серверов не требует запускать потоки или процессы, она может оказаться более эффективной, когда обслуживание клиентов занимает относительно короткое время. Однако при этом также требуется, чтобы обмен данными с клиентами выполнялся очень быстро. В противном случае возникает риск застрять в ожидании окончания диалога с каким-то определенным клиентом, если не предусмотреть использование потоков или ветвления процессов для выполнения продолжительных операций.[1]

Эхо-сервер на базе select

Посмотрим, как все это можно воплотить в программный код. Сценарий в примере 12.9 реализует еще один эхо-сервер, который может обрабатывать несколько клиентов, не запуская новые процессы или потоки.

---

[1] Странно, но *асинхронными* часто называют серверы на базе вызова select – чтобы выделить применяемый в них прием мультиплексирования коротких операций. Однако в действительности классические серверы, основанные на ветвлении или многопоточной модели выполнения, представленные ранее, тоже являются асинхронными, так как они не останавливаются в ожидании, пока завершится обработка запроса того или иного клиента. Между последовательными и параллельными серверами имеется четкое отличие: первые обслуживают по одному клиенту за раз, а последние – нет. Термины «синхронный» и «асинхронный» по сути используются как синонимы для терминов «последовательный» и «параллельный». Согласно этому определению ветвление, многопоточная модель выполнения и циклы на основе вызова select являются тремя вариантами реализации параллельных, асинхронных серверов.

Пример 12.9. PP4E\Internet\Sockets\select-server.py

```
"""
Сервер: обслуживает параллельно несколько клиентов с помощью select.
Использует модуль select для мультиплексирования в группе сокетов:
главных сокетов, принимающих от клиентов новые запросы на соединение,
и входных сокетов, связанных с клиентами, запрос на соединение от которых
был удовлетворен; вызов select может принимать необязательный 4-й аргумент -
0 означает "опрашивать", число n.m означает "ждать n.m секунд", отсутствие
аргумента означает "ждать готовности к обработке любого сокета".
"""

import sys, time
from select import select
from socket import socket, AF_INET, SOCK_STREAM
def now(): return time.ctime(time.time())

myHost = ''             # компьютер-сервер, '' означает локальный хост
myPort = 50007          # использовать незарезервированный номер порта
if len(sys.argv) == 3:  # хост/порт можно указать в командной строке
    myHost, myPort = sys.argv[1:]
numPortSocks = 2        # количество портов для подключения клиентов

# создать главные сокеты для приема новых запросов на соединение от клиентов
mainsocks, readsocks, writesocks = [], [], []
for i in range(numPortSocks):
    portsock = socket(AF_INET, SOCK_STREAM) # создать объект сокета TCP
    portsock.bind((myHost, myPort))  # связать с номером порта сервера
    portsock.listen(5)               # не более 5 ожидающих запросов
    mainsocks.append(portsock)       # добавить в главный список
                                     # для идентификации
    readsocks.append(portsock)       # добавить в список источников select
    myPort += 1                      # привязка выполняется к смежным портам

# цикл событий: слушать и мультиплексировать, пока процесс не завершится
print('select-server loop starting')
while True:
    #print(readsocks)
    readables, writeables, exceptions = select(readsocks, writesocks, [])
    for sockobj in readables:
        if sockobj in mainsocks:           # для готовых входных сокетов
            # сокет порта: принять соединение от нового клиента
            newsock, address = sockobj.accept() # accept не должен
                                                # блокировать
            print('Connect:', address, id(newsock)) # newsock - новый сокет
            readsocks.append(newsock)      # добавить в список select, ждать
        else:
            # сокет клиента: читать следующую строку
            data = sockobj.recv(1024)        # recv не должен блокировать
            print('\tgot', data, 'on', id(sockobj))
            if not data:                     # если закрыто клиентом
                sockobj.close()              # закрыть и удалить из списка
```

```
        readsocks.remove(sockobj)    # иначе повторно будет
    else:                            # обслуживаться вызовом select
        # может блокировать: в действительности для операции записи
        # тоже следовало бы использовать вызов select
        reply = 'Echo=>%s at %s' % (data, now())
        sockobj.send(reply.encode())
```

Основу этого сценария составляет большой цикл событий `while`, в котором вызывается функция `select`, чтобы определить, какие сокеты готовы к обработке (в том числе главные сокеты, к которым могут подключаться клиенты, и открытые соединения с клиентами). Затем все готовые сокеты перебираются в цикле, при этом для главных сокетов выполняется прием соединений, а для клиентских сокетов, готовых к вводу, производится чтение или эхо-вывод. Методы `accept` и `recv`, используемые здесь, гарантированно не будут блокировать процесс сервера после возврата из `select`. Благодаря этому сервер может быстро вернуться в начало цикла и обработать вновь поступившие клиентские запросы на соединение и входные данные, отправленные уже подключенными клиентами. В итоге все новые запросы и клиенты обслуживаются псевдопараллельным образом.

Чтобы этот процесс мог действовать, сервер добавляет все сокеты, подключенные к клиентам, в список `readables`, передаваемый функции `select`, и просто ждет, когда какой-либо сокет появится в списке, который возвращается этой функцией. Для иллюстрации мы задали больше одного порта, на которых этот сервер слушает клиентов, – в наших примерах это порты 50007 и 50008. Так как главные сокеты портов также опрашиваются функцией `select`, запросы на соединение по любому порту могут быть приняты без блокирования уже подключившихся клиентов или новых запросов на соединение, появляющихся на другом порту. Вызов `select` возвращает сокеты из списка `readables`, которые готовы к обработке. Это могут быть и главные сокеты портов, и сокеты, соединенные с обслуживаемыми в данный момент клиентами.

Запуск сервера на базе select

Запустим этот сервер локально и посмотрим, как он работает (клиент и сервер могут запускаться на разных компьютерах, как в предыдущих примерах). Сначала предположим, что этот сервер уже был запущен на локальном компьютере в одном окне, и запустим несколько клиентов, которые попробуют пообщаться с ним. В следующем листинге приводится диалог в двух клиентских окнах консолей в Windows. В первом окне просто дважды был запущен сценарий `echo-client`, подключающийся к серверу, а во втором запускался сценарий `testecho`, запускающий восемь программ `echo-client`, выполняющихся параллельно.

Как и ранее, сервер просто возвращает любой текст, отправленный клиентом, однако здесь не выполняется приостановка с помощью функции `time.sleep`. Заметьте, что во втором окне клиента в действительности выполняется сценарий с именем `echo-client-50008`, который соединяется

с сокетом второго порта на сервере, – это тот же самый сценарий echo-client, но в нем жестко определен другой номер порта (к сожалению, первоначальный сценарий не позволяет передавать ему номер порта):

```
[окно клиента 1]
C:\...\PP4E\Internet\Sockets> python echo-client.py
Client received: b"Echo=>b'Hello network world' at Sun Apr 25 14:51:21 2010"

C:\...\PP4E\Internet\Sockets> python echo-client.py
Client received: b"Echo=>b'Hello network world' at Sun Apr 25 14:51:27 2010"

[окно клиента 2]
C:\...\PP4E\Internet\Sockets> python echo-client-5008.py localhost
                              Sir Galahad
Client received: b"Echo=>b'Sir' at Sun Apr 25 14:51:22 2010"
Client received: b"Echo=>b'Galahad' at Sun Apr 25 14:51:22 2010"

C:\...\PP4E\Internet\Sockets> python testecho.py
```

В следующем листинге приводится вывод в окне, где был запущен сервер. Первые три сообщения о соединении соответствуют запущенным клиентам echo-client; остальные – результат взаимодействия с восемью программами, порожденными сценарием testecho во втором клиентском окне. Этот сервер можно также запустить в Windows, потому что вызов select доступен на этой платформе. Сопоставьте эти результаты с программным кодом в сценарии сервера, чтобы понять, как он действует.

Обратите внимание, что для сценария testecho подключение новых клиентов и ввод данных мультиплексируются вместе. Если внимательно изучить вывод, можно заметить, что эти операции перекрываются во времени, потому что все они управляются единственным циклом событий на сервере. На практике вывод сервера наверняка каждый раз будет выглядеть по-разному. Сообщения о подключении и обслуживании клиентов будут перемешиваться почти случайным образом из-за различных временных задержек на разных компьютерах. Тот же эффект можно наблюдать в серверах, поддерживающих ветвление и многопоточную модель выполнения, но в них переключение между циклом диспетчера и функциями обслуживания клиентов выполняется автоматически самой операционной системой.

Заметьте также, что когда клиент закрывает сокет, сервер получает пустую строку. Мы следим за тем, чтобы сразу же закрывать и удалять такие сокеты, иначе они будут без нужды снова и снова попадать в список, просматриваемый вызовом select, в каждой итерации главного цикла:

```
[окно сервера]
C:\...\PP4E\Internet\Sockets> python select-server.py
select-server loop starting
Connect: ('127.0.0.1', 59080) 21339352
```

```
         got b'Hello network world' on 21339352
         got b'' on 21339352
Connect: ('127.0.0.1', 59081) 21338128
         got b'Sir' on 21338128
         got b'Galahad' on 21338128
         got b'' on 21338128
Connect: ('127.0.0.1', 59082) 21339352
         got b'Hello network world' on 21339352
         got b'' on 21339352
```

[результаты testecho]

```
Connect: ('127.0.0.1', 59083) 21338128
         got b'Hello network world' on 21338128
         got b'' on 21338128
Connect: ('127.0.0.1', 59084) 21339352
         got b'Hello network world' on 21339352
         got b'' on 21339352
Connect: ('127.0.0.1', 59085) 21338128
         got b'Hello network world' on 21338128
         got b'' on 21338128
Connect: ('127.0.0.1', 59086) 21339352
         got b'Hello network world' on 21339352
         got b'' on 21339352
Connect: ('127.0.0.1', 59087) 21338128
         got b'Hello network world' on 21338128
         got b'' on 21338128
Connect: ('127.0.0.1', 59088) 21339352
Connect: ('127.0.0.1', 59089) 21338128
         got b'Hello network world' on 21339352
         got b'Hello network world' on 21338128
Connect: ('127.0.0.1', 59090) 21338056
         got b'' on 21339352
         got b'' on 21338128
         got b'Hello network world' on 21338056
         got b'' on 21338056
```

Помимо большей подробности этого вывода есть еще одно тонкое, но важное отличие, на которое следует обратить внимание: в данной реализации было бы неразумным вызывать функцию `time.sleep` для имитации выполнения продолжительной операции – так как все клиенты обрабатываются в одном и том же цикле, задержка остановит их *все*, и расстроит весь смысл мультиплексирующего сервера. Напомню, что серверы, выполняющие мультиплексирование вручную, как в данном примере, лучше всего работают, когда обслуживание клиентов занимает минимальный промежуток времени, в противном случае необходимо предусматривать специальные способы обслуживания.

Прежде чем двинуться дальше, необходимо сделать еще несколько замечаний:

Особенности вызова `select`

Формально функции `select` передается три списка выбираемых объектов (входные источники, выходные источники и источники исключительных ситуаций), а также необязательное предельное время ожидания. Значением аргумента времени ожидания может быть действительное значение времени ожидания в секундах (чтобы определить доли секунды, используются числа с плавающей точкой); нулевое значение указывает, что должен выполняться простой опрос с немедленным возвратом; а отсутствие этого аргумента определяет необходимость ожидания готовности, по крайней мере, одного объекта (как сделано выше в нашем сценарии). Функция возвращает тройку готовых объектов – подмножеств первых трех аргументов, причем все или некоторые из них могут быть пустыми, если предельное время ожидания было превышено раньше, чем источники оказались готовы.

Переносимость `select`

Подобно многопоточным серверам и в отличие ветвящихся серверов, серверы на основе вызова `select` способны также выполняться в Windows. Технически функция `select` в Windows может работать только с сокетами, но в Unix и Macintosh она может обслуживать также такие объекты, как файлы и каналы. Конечно, для серверов, работающих в Интернете, основным интересующим нас инструментом являются сокеты.

Неблокирующие сокеты

Функция `select` гарантирует, что вызовы методов сокетов, таких как `accept` и `recv`, не будут блокировать (приостанавливать) вызывающую программу, но в языке Python есть также возможность сделать сокеты неблокирующими в целом. С помощью метода `setblocking` объектов сокетов они устанавливаются в блокирующий или неблокирующий режим. Например, после вызова `sock.setblocking(flag)` сокет `sock` устанавливается в неблокирующий режим, если флаг `flag` равен нулю, и в блокирующий режим – в противном случае. Все сокеты изначально открываются в блокирующем режиме, поэтому вызовы методов сокетов всегда могут привести к приостановке вызывающей программы.

Но при работе в неблокирующем режиме, когда метод `recv` не находит данных или метод `send` не может немедленно передать данные, возбуждается исключение `socket.error`. Сценарий может перехватить это исключение, чтобы определить, готов ли сокет к обработке. В блокирующем режиме эти методы всегда блокируют вызывающую программу, пока не смогут продолжить работу. Конечно, обработка запроса клиента может не ограничиваться пересылкой данных (обработка запроса может потребовать длительных расчетов), поэтому неблокирующие сокеты не гарантируют отсутствие задержки на

сервере в целом. Они лишь предоставляют еще один способ реализации серверов с мультиплексированием. Подобно select они лучше подходят для случаев, когда запросы клиентов могут быть обслужены быстро.

Инструменты модуля asyncore

Если вас заинтересовало использование функции select, то, вероятно, вам будет интересно обратить внимание на модуль asyncore.py из стандартной библиотеки Python. Он реализует модель обратного вызова, основанную на классах, в которой обратные вызовы для ввода и вывода переадресуются методам класса, уже реализованным циклом событий select. Таким образом, он позволяет строить серверы без потоков выполнения и ветвлений и является альтернативой на основе вызова select рассматривавшемуся в предыдущих разделах модулю socketserver, использующему потоки выполнения и ветвление. Как и для любых других серверов этого типа, модуль asyncore лучше всего использовать в ситуациях, когда обслуживание клиента занимает короткий промежуток времени, то есть когда основная работа связана с выполнением операций ввода-вывода, а не с вычислениями, так как в последнем случае необходимо использовать потоки или ветвление. Подробное описание и примеры использования этого модуля вы найдете в руководстве по стандартной библиотеке Python.

Twisted

Еще один способ реализации серверов предоставляет открытая система Twisted (*http://twistedmatrix.com*). Twisted – это асинхронный сетевой фреймворк, написанный на языке Python и поддерживающий протоколы TCP, UDP, SSL/TLS, IP Multicast, взаимодействие через последовательный порт и многое другое. Он может использоваться для создания клиентов и серверов и включает реализацию множества наиболее типичных сетевых служб, таких как веб-сервер, чат-сервер IRC, почтовый сервер, интерфейс к реляционной базе данных и брокер объектов.

Фреймворк Twisted поддерживает не только возможность запуска процессов и потоков для выполнения продолжительных операций, но и позволяет использовать асинхронную модель обслуживания клиентов, управляемую событиями, напоминающую цикл событий в библиотеках графического интерфейса, подобных библиотеке tkinter. Он реализует цикл событий, выполняющий переключения среди множества сокетов открытых соединений, автоматизирует множество операций, обычно выполняемых асинхронным сервером, и служит управляемой событиями основой сценариев, предназначенных для решения разнообразных прикладных задач. Внутренний механизм событий фреймворка Twisted своим принципом действия напоминает наш сервер на основе вызова select и модуль asyncore, но по общему признанию считается более совершенным. Twisted – это сис-

тема, реализованная сторонними разработчиками, она не является инструментом стандартной библиотеки. За дополнительными подробностями обращайтесь к соответствующей документации и на веб-сайт проекта.

Подводя итоги: выбор конструкции сервера

Так в каких же случаях для создания сервера следует использовать вызов select вместо потоков выполнения или ветвления? Конечно, в каждом приложении могут быть свои потребности, но обычно считается, что серверы, основанные на вызове select, очень хорошо работают, когда обслуживание клиента занимает относительно короткие интервалы времени. Если обслуживание может занимать продолжительное время, потоки или ветвление могут оказаться более удачным способом распределения обработки нескольких клиентов. Потоки и ветвление особенно полезно применять, если помимо передачи данных клиентам требуется длительная их обработка. Однако возможны также комбинации этих приемов – ничто не мешает запускать потоки выполнения из цикла опроса на основе вызова select.

Важно помнить, что схемы, основанные на select (и неблокирующих сокетах), не вполне защищены от блокирования. В примере 12.9, представленном выше, вызов метода send, который возвращает текст клиенту, тоже может оказаться блокирующим и задержать работу всего сервера. Можно было бы преодолеть эту опасность блокирования, применяя select для проверки готовности к операции вывода перед попыткой выполнить ее (например, использовать список writesocks и добавить еще один цикл для отправки ответов готовым выходным сокетам), но это существенно уменьшило бы ясность программы.

Однако в целом, когда нельзя разделить обработку клиентского запроса на этапы, чтобы ее можно было мультиплексировать с другими запросами, и не заблокировать основной цикл сервера, select может оказаться не лучшим способом построения сервера. Эти ограничения учитываются далеко не всеми существующими сетевыми серверами.

Кроме того, реализация на основе вызова select оказывается также более сложной, чем реализация на основе ветвления процессов или потоков выполнения, поскольку требует вручную передавать управление всем участвующим задачам (сравните, например, версии этого сервера с потоками и с select, даже без применения select для выполнения операций записи). Как всегда, степень этой сложности зависит от конкретного приложения. Модуль asyncore из стандартной библиотеки, упоминавшийся выше, может упростить реализацию цикла событий сервера на основе вызова select, а фреймворк Twisted предлагает дополнительные, гибридные решения.

Придание сокетам внешнего вида файлов и потоков ввода-вывода

До сих пор в этой главе мы рассматривали сокеты с точки зрения классической модели сетевых взаимодействий клиент-сервер. Это одно из основных их предназначений, но они могут также использоваться и в других распространенных ситуациях.

В главе 5, например, мы рассматривали сокеты как один из основных механизмов взаимодействий между процессами и потоками, выполняющимися на одном компьютере. А в главе 10 при исследовании возможностей организации взаимодействий между сценариями командной строки и графическими интерфейсами мы написали вспомогательный модуль (пример 10.23), подключающий стандартный поток вывода вызывающей программы к сокету, который мог бы использоваться графическим интерфейсом для приема и отображения вывода. Там я обещал, что мы исследуем подробности, касающиеся дополнительных способов передачи данных, как только достаточно близко познакомимся с сокетами. Теперь, после такого знакомства, в этом разделе мы ненадолго покинем мир сетевых серверов и познакомимся с остальной частью этой истории.

Некоторые программы можно писать или переписывать, явно реализуя в них возможность обмена данными через сокеты, но такое решение подходит не для всех случаев. Для изменения существующих сценариев может потребоваться приложить значительные усилия, а иногда такой подход может воспрепятствовать использованию нужных режимов, не связанных с сокетами. В некоторых случаях лучше было бы позволить сценарию использовать стандартные инструменты работы с потоками ввода-вывода, такие как встроенные функции `print` и `input` или методы файлов из модуля `sys` (например, `sys.stdout.write`), и подключать их к сокетам только при необходимости.

Поскольку такие инструменты потоков ввода-вывода предназначены для работы с файлами в текстовом режиме, самая большая сложность здесь состоит в том, чтобы хитростью заставить эти инструменты действовать в двоичном режиме, свойственном сокетам, и использовать совершенно иной интерфейс их методов. К счастью, сокеты обладают методом, позволяющим добиться такой подмены.

Метод `makefile` объекта сокета приходит на помощь всякий раз, когда возникает необходимость работать с сокетами с помощью обычных методов файлов или когда необходимо передать сокет существующему интерфейсу или программе, ожидающей получить файл. Объект-обертка, возвращаемый сокетом, позволяет сценариям передавать данные через сокет с помощью привычных методов `read` и `write` вместо `recv` и `send`. Поскольку встроенные функции `input` и `print` используют первую пару методов, они с таким же успехом могут взаимодействовать с сокетами, обернутыми вызовом этого метода.

Метод makefile позволяет также интерпретировать двоичные данные сокета как текст, а не как строки байтов, и может принимать дополнительные аргументы, такие как encoding, что позволяет определять кодировки при передаче текста, отличные от кодировки по умолчанию, как это делают встроенные функции open и os.fdopen, с которыми мы встречались в главе 4, при работе с файловыми дескрипторами. Конечно, при работе с двоичными сокетами текст всегда можно кодировать и декодировать вручную, однако метод makefile снимает это бремя с вашего программного кода и переносит его на обертывающий объект файла.

Такое сходство с файлами может пригодиться, когда необходимо использовать программное обеспечение, поддерживающее интерфейс файлов. Например, методы load и dump из модуля pickle ожидают получить объект с интерфейсом файла (например, с методами read и write), но не требуют, чтобы этот объект соответствовал физическому файлу. Передача сокета TCP/IP, обернутого вызовом метода makefile, методам из модуля pickle позволяет передавать сериализованные объекты Python через Интернет без необходимости выполнять преобразование в строки байтов и вызывать методы сокетов вручную. Этот прием обеспечивает альтернативу использованию строковых методов модуля pickle (dumps, loads) вместе с методами сокетов send и recv и может предложить большую гибкость для программного обеспечения, поддерживающего различные механизмы обмена данными. Дополнительные подробности о сериализации объектов вы найдете в главе 17.

В общем случае любой компонент, ожидающий поддержку протокола методов файлов, благополучно сможет работать с сокетом, обернутым вызовом метода makefile. Такие интерфейсы смогут также принимать строки, обернутые встроенным классом io.StringIO, и любые другие объекты, поддерживающие те же самые методы, что и встроенные объекты файлов. Как это принято в языке Python, мы реализуем протоколы – интерфейсы объектов, – а не определенные типы данных.

Вспомогательный модуль перенаправления потоков ввода-вывода

Для иллюстрации действия метода makefile в примере 12.10 приводится реализация различных схем перенаправления потоков ввода-вывода вызывающей программы в сокеты, которые могут использоваться для взаимодействий с другим процессом. Первая из этих функций перенаправляет стандартный поток вывода, и именно ее мы использовали в главе 10; другие функции перенаправляют стандартный поток ввода и оба потока, ввода и вывода, тремя различными способами.

Объект обертки, возвращаемый методом socket.makefile, по своей природе допускает возможность прямого использования файловых методов read и write, и независимо от стандартных потоков ввода-вывода. Представленный пример также использует эти методы, хотя и неявно,

через встроенные функции print и input доступа к потокам ввода вывода, и отражает типичные способы использования данного инструмента.

Пример 12.10. PP4E\Internet\Sockets\socket_stream_redirect.py

```
"""
###############################################################################
Инструменты для подключения стандартных потоков ввода-вывода программ без ГИ
к сокетам, которые программы с графическими интерфейсами (и другие)
могут использовать для взаимодействий с программами без ГИ.
###############################################################################
"""

import sys
from socket import *
port = 50008         # передать другой порт, если этот
                     # уже занят другой службой
host = 'localhost'   # передать другое имя хоста для подключения
                     # к удаленным слушателям

def initListenerSocket(port=port):
    """
    инициализирует подключенный сокет для вызывающих сценариев,
    которые играют роль сервера
    """
    sock = socket(AF_INET, SOCK_STREAM)
    sock.bind(('', port))       # слушать порт с этим номером
    sock.listen(5)              # длина очереди ожидающих запросов
    conn, addr = sock.accept()  # ждать подключения клиента
    return conn                 # вернуть подключенный сокет

def redirectOut(port=port, host=host):
    """
    подключает стандартный поток вывода вызывающей программы к сокету
    для графического интерфейса, уже ожидающего запуск вызывающей программы,
    иначе попытка соединения потерпит неудачу перед вызовом метода accept
    """
    sock = socket(AF_INET, SOCK_STREAM)
    sock.connect((host, port))  # вызывающий сценарий действует как клиент
    file = sock.makefile('w')   # интерфейс файла: текстовый режим, буфериз.
    sys.stdout = file           # обеспечить вызов sock.send при выводе
    return sock                 # на случай, если вызывающему сценарию
                                # потребуется работать с сокетом напрямую

def redirectIn(port=port, host=host):
    """
    подключает стандартный поток ввода вызывающей программы к сокету
    для получения данных из графического интерфейса
    """
    sock = socket(AF_INET, SOCK_STREAM)
    sock.connect((host, port))
    file = sock.makefile('r')   # обертка с интерфейсом файла
```

```
    sys.stdin = file            # обеспечить вызов sock.recv при вводе
    return sock                 # возвращаемое значение можно игнорировать

def redirectBothAsClient(port=port, host=host):
    """
    подключает стандартные потоки ввода и вывода вызывающей
    программы к одному и тому же сокету;
    в этом режиме вызывающая программа играет роль клиента:
    отправляет сообщение и получает ответ
    """
    sock = socket(AF_INET, SOCK_STREAM)
    sock.connect((host, port))  # открыть в режиме 'rw'
    ofile = sock.makefile('w')  # интерфейс файла: текстовый режим, буфериз.
    ifile = sock.makefile('r')  # два объекта файла, обертывающих один сокет
    sys.stdout = ofile          # обеспечить вызов sock.send при выводе
    sys.stdin  = ifile          # обеспечить вызов sock.recv при вводе
    return sock

def redirectBothAsServer(port=port, host=host):
    """
    подключает стандартные потоки ввода и вывода вызывающей
    программы к одному и тому же сокету;
    в этом режиме вызывающая программа играет роль сервера:
    получает сообщение и отправляет ответ
    """
    sock = socket(AF_INET, SOCK_STREAM)
    sock.bind((host, port))     # вызывающий сценарий - сервер
    sock.listen(5)
    conn, addr = sock.accept()
    ofile = conn.makefile('w')  # обертка с интерфейсом файла
    ifile = conn.makefile('r')  # два объекта файла, обертывающих один сокет
    sys.stdout = ofile          # обеспечить вызов sock.send при выводе
    sys.stdin  = ifile          # обеспечить вызов sock.recv при вводе
    return conn
```

Чтобы протестировать этот сценарий, в примере 12.11 определяется пять групп функций, реализующих клиентов и серверы. Функции-клиенты выполняются в этом же процессе, а функции-серверы запускаются в отдельных процессах переносимым способом с помощью модуля `multiprocessing`, с которым мы познакомились в главе 5. Таким образом, функции клиентов и серверов выполняются в различных процессах, но общаются между собой посредством сокета, подключенного к стандартным потокам ввода-вывода внутри процесса тестового сценария.

Пример 12.11. PP4E\Internet\Sockets\test-socket_stream_redirect.py

```
"""
##############################################################################
тестирование режимов socket_stream_redirection.py
##############################################################################
"""
```

```
import sys, os, multiprocessing
from socket_stream_redirect import *

##############################################################################
# перенаправление вывода в клиенте
##############################################################################

def server1():
    mypid = os.getpid()
    conn = initListenerSocket()         # блокируется до подключения клиента
    file = conn.makefile('r')
    for i in range(3):                  # читать вывод клиента
        data = file.readline().rstrip() # блокируется до поступления данных
        print('server %s got [%s]' % (mypid, data)) # вывод в окно терминала

def client1():
    mypid = os.getpid()
    redirectOut()
    for i in range(3):
        print('client %s: %s' % (mypid, i))  # вывод в сокет
        sys.stdout.flush()                   # иначе останется в буфере
                                             # до завершения!

##############################################################################
# перенаправление ввода в клиенте
##############################################################################

def server2():
    mypid = os.getpid()          # простой сокет без буферизации
    conn = initListenerSocket()  # отправляет в поток ввода клиента
    for i in range(3):
        conn.send(('server %s: %s\n' % (mypid, i)).encode())

def client2():
    mypid = os.getpid()
    redirectIn()
    for i in range(3):
        data = input()                               # ввод из сокета
        print('client %s got [%s]' % (mypid, data)) # вывод в окно терминала

##############################################################################
# перенаправление ввода и вывода в клиенте, клиент является
# клиентом для сокета
##############################################################################

def server3():
    mypid = os.getpid()
    conn = initListenerSocket()  # ждать подключения клиента
    file = conn.makefile('r')    # принимает от print(), передает в input()
    for i in range(3):           # readline блокируется до появления данных
```

```
        data = file.readline().rstrip()
        conn.send(('server %s got [%s]\n' % (mypid, data)).encode())

def client3():
    mypid = os.getpid()
    redirectBothAsClient()
    for i in range(3):
        print('client %s: %s' % (mypid, i)) # вывод в сокет
        data = input()                      # ввод из сокета: выталкивает!
        sys.stderr.write('client %s got [%s]\n' % (mypid, data)) # не был
                                                                  # перенаправлен

##############################################################################
# перенаправление ввода и вывода в клиенте, клиент является
# сервером для сокета
##############################################################################

def server4():
    mypid = os.getpid()
    sock = socket(AF_INET, SOCK_STREAM)
    sock.connect((host, port))
    file = sock.makefile('r')
    for i in range(3):
        sock.send(('server %s: %s\n'%(mypid,i)).encode())  # передать
                                                           # в input()
        data = file.readline().rstrip()               # принять от print()
        print('server %s got [%s]' % (mypid, data))   # результат в терминал

def client4():
    mypid = os.getpid()
    redirectBothAsServer()   # играет роль сервера в этом режиме
    for i in range(3):
        data = input()       # ввод из сокета: выталкивает выходной буфер!
        print('client %s got [%s]' % (mypid, data))  # вывод в сокет
        sys.stdout.flush()   # иначе последняя порция данных останется
                             # в буфере до завершения!

##############################################################################
# перенаправление ввода и вывода в клиенте, клиент является клиентом
# для сокета, сервер первым инициирует обмен
##############################################################################

def server5():
    mypid = os.getpid()          # тест № 4, но соединение принимает сервер
    conn = initListenerSocket()  # ждать подключения клиента
    file = conn.makefile('r')    # принимает от print(), передает в input()
    for i in range(3):
        conn.send(('server %s: %s\n' % (mypid, i)).encode())
        data = file.readline().rstrip()
        print('server %s got [%s]' % (mypid, data))
```

```
def client5():
    mypid = os.getpid()
    s = redirectBothAsClient() # играет роль клиента в этом режиме
    for i in range(3):
        data = input()         # ввод из сокета: выталкивает выходной буфер!
        print('client %s got [%s]' % (mypid, data))  # вывод в сокет
        sys.stdout.flush()     # иначе последняя порция данных останется
                               # в буфере до завершения!

##############################################################################
# номер выполняемого теста определяется аргументом командной строки
##############################################################################

if __name__ == '__main__':
    server = eval('server' + sys.argv[1])
    client = eval('client' + sys.argv[1])  # клиент - в этом процессе
    multiprocessing.Process(target=server).start()  # сервер -
                                                    # в новом процессе
    client()                                # переустановить потоки в клиенте
    #import time; time.sleep(5)             # проверка эффекта выталкивания
                                            # буферов при выходе
```

Запустим тестовый сценарий, указав номер клиента и сервера в командной строке, чтобы проверить работу инструментов модуля. В сообщениях отображаются числовые идентификаторы процессов, а в квадратных скобках отображаются сообщения, переданные (дважды, если вложенные) через потоки ввода-вывода, подключенные к сокетам:

```
C:\...\PP4E\Internet\Sockets> test-socket_stream_redirect.py 1
server 3844 got [client 1112: 0]
server 3844 got [client 1112: 1]
server 3844 got [client 1112: 2]

C:\...\PP4E\Internet\Sockets> test-socket_stream_redirect.py 2
client 5188 got [server 2020: 0]
client 5188 got [server 2020: 1]
client 5188 got [server 2020: 2]

C:\...\PP4E\Internet\Sockets> test-socket_stream_redirect.py 3
client 7796 got [server 2780 got [client 7796: 0]]
client 7796 got [server 2780 got [client 7796: 1]]
client 7796 got [server 2780 got [client 7796: 2]]

C:\...\PP4E\Internet\Sockets> test-socket_stream_redirect.py 4
server 4288 got [client 3852 got [server 4288: 0]]
server 4288 got [client 3852 got [server 4288: 1]]
server 4288 got [client 3852 got [server 4288: 2]]

C:\...\PP4E\Internet\Sockets> test-socket_stream_redirect.py 5
server 6040 got [client 7728 got [server 6040: 0]]
server 6040 got [client 7728 got [server 6040: 1]]
server 6040 got [client 7728 got [server 6040: 2]]
```

Если сопоставить вывод сценария с программным кодом, чтобы понять, как передаются сообщения между клиентом и сервером, можно будет увидеть, что функции `print` и `input` в функциях клиентов в конечном итоге обращаются к сокетам в другом процессе. В функциях клиентов связь с сокетами остается практически незаметной.

Текстовые файлы и буферизация потоков вывода

Прежде чем двинуться дальше, необходимо осветить два тонких аспекта, касающихся реализации примера:

Преобразование двоичных данных в текст

Простые сокеты передают данные в виде строк двоичных байтов, но благодаря тому, что файлы-обертки открываются в текстовом режиме, при выполнении операций ввода-вывода их содержимое автоматически преобразуется в текст. Файлы-обертки должны открываться в текстовом режиме, когда доступ к ним осуществляется с применением инструментов для работы со стандартными потоками ввода-вывода, таких как встроенная функция `print`, которая выводит текстовые строки (как мы уже знаем, файлы, открытые в двоичном режиме, напротив, работают со строками байтов). Однако при непосредственном использовании сокетов, текст по-прежнему требуется кодировать в строки байтов вручную, как это делается в большинстве тестов в примере 12.11.

Буферизация потоков ввода-вывода, вывод программы и взаимоблокировки

Как мы узнали в главах 5 и 10, стандартные потоки вывода обычно буферизуются и при выводе текста его может потребоваться выталкивать их выходного буфера, чтобы он появился в сокете, подключенном к стандартному потоку вывода процесса. Действительно, в некоторых тестах в примере 12.11 необходимо явно или неявно выталкивать буферы, чтобы они могли работать. В противном случае вывод может оказаться неполным или вообще оставаться в выходном буфере до завершения программы. В самых тяжелых случаях это может привести к взаимоблокировке, когда один процесс ждет вывода другого процесса, который никогда не появится. В других ситуациях может также возникнуть ошибка чтения из сокета, если пишущий процесс завершится слишком рано, особенно при двустороннем диалоге.

Например, если `client1` и `client4` не будут периодически выталкивать выходной буфер, как они это делают, то единственной гарантией их работы стало бы автоматическое выталкивание потоков вывода при завершении их процессов. Без выталкивания буфера вручную передача данных в `client1` не производилась бы до момента завершения процесса (в этот момент все выходные данные были бы отправлены в виде единого сообщения), а данные, генерируемые функцией `client4`, были бы отправлены не полностью, до момента завершения

процесса (последнее выводимое сообщение задержалось бы в выходном буфере).

Еще более тонкий аспект: обе функции, `client3` и `client4`, полагаются на то, что встроенная функция `input` сначала автоматически выталкивает `sys.stdout`, чтобы гарантировать вывод строки приглашения к вводу, и тем самым обеспечивает отправку данных, записанных в выходной буфер предыдущими вызовами функции `print`. Без такого неявного выталкивания буферов (или без дополнительных операций выталкивания, выполняемых вручную) функция `client3` была бы немедленно *заблокирована*. Как и `client4`, если из нее удалить операцию выталкивания буфера вручную (даже с учетом автоматического выталкивания буфера функцией `input` удаление операции выталкивания буфера из функции `client4` приведет к тому, что заключительное сообщение, выводимое функцией `print`, не будет отправлено, пока процесс не завершится). Функция `client5` проявляет те же черты поведения, что и `client4`, потому что она просто меняет местами процессы – ожидающий и устанавливающий соединение.

В общем случае, если необходимо обеспечить отправку данных по мере их вывода, а не в момент завершения программы или при переполнении буфера вывода, необходимо либо периодически вызывать `sys.stdout.flush`, либо отключать буферизацию потоков вывода с помощью ключа `-u` командной строки, если это возможно, как описано в главе 5.

Мы, конечно, можем открывать файлы-обертки сокетов в *небуферизованном* режиме, передав методу `makefile` нулевое значение во втором аргументе (как обычной функции `open`), однако это не позволит обертке работать в текстовом режиме, необходимом для функции `print` и желательном для функции `input`. Фактически, такая попытка открыть файл-обертку для сокета в текстовом режиме с отключенной буферизацией приведет к исключению, потому что Python 3.X больше не поддерживает небуферизованный режим для текстовых файлов (в настоящее время он допускается только для двоичных файлов). Иными словами, из-за того, что функция `print` требует использование текстового режима, режим буферизации для файлов потоков вывода предполагается по умолчанию. Кроме того, возможность открытия файлов-оберток для сокетов в режиме *построчной буферизации*, похоже, больше не поддерживается в Python 3.X (подробнее об этом рассказывается ниже).

Пока режим буферизации может зависеть от особенностей платформы или от реализации библиотеки, иногда может оказаться необходимым выталкивать выходные буферы вручную или напрямую обращаться к сокетам. Обратите внимание, что с помощью вызова метода `setblocking(0)` сокеты могут переводиться в неблокирующий режим, но это позволит всего лишь избежать приостановки при отправке данных и никак не решает проблему передачи буферизованного вывода.

Требования потоков ввода-вывода

Чтобы конкретизировать вышеизложенное, в примере 12.12 показано, как некоторые из описанных сложностей влияют на перенаправление стандартных потоков ввода-вывода. В нем выполняется попытка связать потоки ввода-вывода с файлами в текстовом и двоичном режимах, создаваемых функцией open, и обратиться к ним с помощью встроенных функций print и input, как это обычно делается в сценариях.

Пример 12.12. PP4E\Internet\Sockets\test-stream-modes.py

```
"""
проверка эффекта связывания стандартных потоков ввода-вывода с файлами,
открытыми в текстовом и двоичном режимах; то же справедливо и для socket.
makefile: функция print требует текстовый режим, а текстовый режим,
в свою очередь, препятствует отключению буферизации –
используйте ключ -u или вызывайте метод sys.stdout.flush()
"""

import sys

def reader(F):
    tmp, sys.stdin = sys.stdin, F
    line = input()
    print(line)
    sys.stdin = tmp

reader(open('test-stream-modes.py'))        # работает: input() возвращает
                                            # текст
reader(open('test-stream-modes.py', 'rb'))  # работает: но input()
                                            # возвращает байты

def writer(F):
    tmp, sys.stdout = sys.stdout, F
    print(99, 'spam')
    sys.stdout = tmp

writer( open('temp', 'w') )     # работает: print() передает .write()
                                # текст str
print( open('temp').read() )

writer( open('temp', 'wb') )    # ОШИБКА в print: двоичный режим
                                # требует байты
writer( open('temp', 'w', 0) )  # ОШИБКА в open: буферизация в текстовом
                                # режиме обязательна
```

Если запустить этот сценарий, последние две инструкции в нем потерпят неудачу – вторая с конца терпит неудачу потому, что print пытается вывести текстовую строку в двоичный файл (что вообще недопустимо для файлов). А последняя инструкция терпит неудачу потому, что в Python 3.X не допускается открывать текстовые файлы в небуферизован-

ном режиме (текстовый режим предполагает кодирование символов Юникода). Ниже приводятся сообщения об ошибках, которые выводятся при попытке запустить этот сценарий: первое сообщение выводится, если запустить сценарий в приведенном виде, а второе появляется, если закомментировать вторую с конца инструкцию (я немного отредактировал текст исключения для большей наглядности):

```
C:\...\PP4E\Internet\Sockets> test-stream-modes.py
"""
b'"""\r'
99 spam

Traceback (most recent call last):
  File "C:\...\PP4E\Internet\Sockets\test-stream-modes.py", line 26,
       in <module>
    writer( open('temp', 'wb') )    # ОШИБКА в print: двоичный режим...
  File "C:\...\PP4E\Internet\Sockets\test-stream-modes.py", line 20,
       in writer
    print(99, 'spam')
TypeError: must be bytes or buffer, not str
```

(TypeError: данные должны быть типа bytes или buffer, а не str)

```
C:\...\PP4E\Internet\Sockets> test-streams-binary.py
"""
b'"""\r'
99 spam

Traceback (most recent call last):
  File "C:\...\PP4E\Internet\Sockets\test-stream-modes.py", line 27,
        in <module>
    writer( open('temp', 'w', 0) )  # ОШИБКА в open: буферизация
                                    # в текстовом...
ValueError: can't have unbuffered text I/O
```

(ValueError: не поддерживается небуферизованный ввод-вывод текста)

То же самое относится к объектам файлов-оберток для сокетов, создаваемых с помощью метода `makefile` сокетов – они должны открываться в текстовом режиме, чтобы обеспечить поддержку `print`, и также должны открываться в текстовом режиме, если желательно, чтобы функция `input` принимала текстовые строки, но текстовый режим препятствует использованию небуферизованного режима для файлов:

```
>>> from socket import *
>>> s = socket()            # по умолчанию для tcp/ip (AF_INET, SOCK_STREAM)
>>> s.makefile('w', 0)      # эта инструкция работала в Python 2.X
Traceback (most recent call last):
  File "C:\Python31\lib\socket.py", line 151, in makefile
ValueError: unbuffered streams must be binary
```

(ValueError: небуферизованные потоки ввода-вывода должны быть двоичными)

Построчная буферизация

Файлы-обертки сокетов в текстов ом режиме принимают также значение 1 в аргументе, определяющем режим буферизации, что позволяет определить режим *построчной буферизации* вместо режима полной буферизации:

```
>>> from socket import *
>>> s = socket()
>>> f = s.makefile('w', 1)  # то же, что и buffering=1, но действует
                            # как режим полной буферизации!
```

Похоже, что этот режим ничем не отличается от режима полной буферизации и по-прежнему требует вручную выталкивать выходной буфер, чтобы обеспечить передачу строк по мере их вывода. Рассмотрим простые сценарии сервера и клиента, представленные в примерах 12.13 и 12.14. Сервер просто читает три сообщения, используя интерфейс сокетов непосредственно.

Пример 12.13. PP4E\Internet\Sockets\socket-unbuff-server.py

```
from socket import *       # читает три сообщения непосредственно из сокета
sock = socket()
sock.bind(('', 60000))
sock.listen(5)
print('accepting...')
conn, id = sock.accept()   # блокируется, пока не подключится клиент

for i in range(3):
    print('receiving...')
    msg = conn.recv(1024)  # блокируется, пока не поступят данные
    print(msg)             # выведет все строки сразу, если не выталкивать
                           # буфер вручную
```

Клиент, представленный в примере 12.14, отправляет три сообщения. Первые два отправляются через файл-обертку сокета, а последнее – прямым обращением к сокету. Вызовы метода flush здесь закомментированы, но оставлены, чтобы вы могли поэкспериментировать с ними, а вызовы функции sleep заставляют сервер ждать поступления данных.

Пример 12.14. PP4\Internet\Sockets\socket-unbuff-client.py

```
import time      # отправляет три сообщения через файл-обертку и сокет
from socket import *
sock = socket()  # по умолчанию=AF_INET, SOCK_STREAM (tcp/ip)
sock.connect(('localhost', 60000))
file = sock.makefile('w', buffering=1)  # по умолчанию=полная буферизация,
                                        # 0=ошибка, 1 не включает построчную
                                        # буферизацию!
print('sending data1')
file.write('spam\n')
time.sleep(5)  # следующий вызов flush() должен вызвать немедленную передачу
#file.flush()  # раскомментируйте вызовы flush(), чтобы увидеть разницу
```

```
print('sending data2')     # дополнительный вывод в файл не приводит
print('eggs', file=file)   # к выталкиванию буфера
time.sleep(5) # вывод будет принят сервером только после выталкивания буфера
#file.flush() # или после завершения

print('sending data3')     # низкоуровневый двоичный интерфейс выполняет
передачу
sock.send(b'ham\n')   # немедленно, эта строка будет принята первой, если
time.sleep(5)         # в первых двух случаях не выталкивать буферы вручную!
```

Запустите сначала сервер в одном окне, а затем клиента – в другом (или, в Unix-подобных системах, запустите сначала сервер в фоновом режиме). Ниже показан вывод в окне сервера – передача сообщений, отправленных через файл-обертку сокета, откладывается до завершения программы-клиента, а передача данных, отправляемых через низкоуровневый интерфейс сокета, выполняется немедленно:

```
C:\...\PP4E\Internet\Sockets> socket-unbuff-server.py
accepting...
receiving...
b'ham\n'
receiving...
b'spam\r\neggs\r\n'
receiving...
b''
```

В окне клиента строки «sending» появляются через каждые 5 секунд. Третье сообщение появится в окне сервера через 10 секунд, а передача первого и второго сообщений, отправленных через файл-обертку, будет отложена до завершения клиента (на 15 секунд), потому что файл-обертка действует в режиме полной буферизации. Если в клиенте раскомментировать вызовы метода `flush`, все три сообщения по очереди будут появляться в окне сервера с интервалом 5 секунд (третье сообщение появится после второго):

```
C:\...\PP4E\Internet\Sockets> socket-unbuff-server.py
accepting...
receiving...
b'spam\r\n'
receiving...
b'eggs\r\n'
receiving...
b'ham\n'
```

Иными словами, даже когда запрошена построчная буферизация, вывод в файл-обертку сокета (и, соответственно, вывод функции `print`) будет сохраняться в буфере, пока программа не завершит работу, или пока выходной буфер не будет вытолкнут вручную, или пока буфер не переполнится.

Решения

Чтобы избежать задержки вывода или взаимоблокировки, сценарии, которые должны отправлять данные ожидающим программам за счет вывода в файлы-обертки сокетов (то есть с помощью `print` или `sys.stdout.write`), должны предусматривать выполнение одного из следующих пунктов:

- Периодически вызывать `sys.stdout.flush`, чтобы вытолкнуть содержимое буфера и обеспечить его отправку по мере вывода, как показано в примере 12.11.
- Запускаться с ключом `-u` интерпретатора Python, если это возможно, чтобы принудительно отключить буферизацию потоков вывода. Этот прием может применяться к немодифицированным программам, порожденным с помощью инструментов для работы с каналами, таких как `os.popen`. Но он *не* поможет в данном случае, потому что мы вручную переустанавливаем файлы потоков ввода-вывода, назначая им буферизованные текстовые файлы-обертки сокетов уже после того, как процесс будет запущен. Чтобы убедиться в этом, раскомментируйте вызовы `flush` в примере 12.11 и вызов `sleep` в конце и запустите его с ключом `-u`: вывод первого теста по-прежнему появится с задержкой в 5 секунд.
- Использовать *потоки выполнения*, чтобы избежать блокирования при чтении из сокетов, что особенно важно, если принимающей программой является графический интерфейс, который не должен зависеть от вызова метода `flush` на стороне клиента. Дополнительные указания вы найдете в главе 10. В действительности этот подход не решает проблему – порожденный поток выполнения, производящий чтение, также может оказаться заблокированным или взаимно заблокировать программу, выполняющую запись, однако в такой ситуации графический интерфейс хотя бы останется активным.
- Реализовать собственные, *нестандартные* объекты-обертки сокетов, которые будут перехватывать операции записи текста, кодировать его в двоичное представление и передавать методу `send` сокета. Метод `socket.makefile` – это, в действительности, всего лишь инструмент для удобства, и мы всегда можем реализовать собственную обертку с более специализированными возможностями. Для подсказки смотрите реализацию класса `GuiOutput` в главе 10, класс перенаправления потоков ввода-вывода в главе 3 и классы в стандартном модуле `io` (на которых основаны инструменты ввода-вывода языка Python и которые можно подмешивать в собственные классы).
- Вообще не использовать `print` и выполнять обмен данными с применением «родных» интерфейсов инструментов IPC, таких как низкоуровневые методы сокетов `send` и `recv`, – они выполняют передачу данных немедленно и не предусматривают их буферизацию, как методы файлов. Таким способом мы можем непосредственно передавать

простые строки байтов или использовать инструменты dumps и loads модуля pickle для преобразования объектов Python в строки байтов и обратно при передаче их непосредственно через сокеты (подробнее о модуле pickle рассказывается в главе 17).

Последний вариант является наиболее прямым (кроме того, функции перенаправления из вспомогательного модуля возвращают простые сокеты как раз для поддержки подобного варианта), но он может использоваться далеко не во всех случаях; особенно сложно его будет применить к существующим или многорежимным сценариям. Во многих случаях гораздо проще может оказаться добавить вызовы flush в программы командной строки, потоки ввода-вывода которых могут быть связаны с другими программами через сокеты.

Буферизация в других контекстах: еще раз о каналах

Имейте также в виду, что буферизация потоков ввода-вывода и взаимоблокировки являются более общими проблемами, которые затрагивают не только файлы-обертки сокетов. Мы уже исследовали эту тему в главе 5. Однако, в качестве краткого напоминания, в примере 12.15 приводится сценарий, не использующий сокеты, в котором режим полной буферизации не действует, когда его стандартный поток вывода подключен к терминалу (когда сценарий запускается из командной строки, его вывод буферизуется построчно), и действует, когда он подключен к чему-то другому (включая сокет или канал).

Пример 12.15. PP4E\Internet\Sockets\pipe-unbuff-writer.py

```
# вывод с построчной буферизацией (небуферизованный), когда stdout подключен
# к терминалу; при подключении к другим устройствам по умолчанию выполняется
# полная буферизация: используйте -u или sys.stdout.flush(), чтобы избежать
# задержки вывода в канал/сокет

import time, sys
for i in range(5):              # режим буферизации потока влияет на print
    print(time.asctime())       # и на прямые операции доступа к файлу потока
    sys.stdout.write('spam\n') # если sys.stdout не был переустановлен
    time.sleep(2)               # в другой файл
```

Несмотря на то, что в Python 3.X функция print требует, чтобы файл был открыт в текстовом режиме, в версии 3.X по-прежнему можно подавить полную буферизацию потока вывода с помощью флага -u. Использование этого флага в примере 12.16 заставляет сообщения, которые порождаемый сценарий печатает каждые 2 секунды, появляться по мере того, как они отправляются в поток вывода. В случае отсутствия этого флага выводимые данные появляются все сразу через 10 секунд, когда дочерний сценарий завершит работу, если только он не вызывает sys.stdout.flush в каждой итерации.

Пример 12.16. PP4E\Internet\Sockets\pipe-unbuff-reader.py

```
# вывод появится только через 10 секунд, если не использовать флаг Python -u
# или sys.stdout.flush(); однако вывод будет появляться каждые 2 секунды,
# если использовать любой из этих двух вариантов

import os                                                # итератор читает
for line in os.popen('python -u pipe-unbuff-writer.py'): # строки
    print(line, end='')                              # блокируется без -u!
```

Ниже приводится вывод этого сценария. В отличие от примеров с сокетами, он автоматически запускает пишущий сценарий, поэтому нам не требуется открывать отдельное окно консоли для тестирования. В главе 5 говорилось, что функция os.popen также принимает аргумент buffering, как и метод socket.makefile, но он не оказывает влияния на буферизацию потоков ввода-вывода порождаемых программ и поэтому не может предотвратить буферизацию потока вывода в подобных ситуациях.

```
C:\...\PP4E\Internet\Sockets> pipe-unbuff-reader.py
Wed Apr 07 09:32:28 2010
spam
Wed Apr 07 09:32:30 2010
spam
Wed Apr 07 09:32:32 2010
spam
Wed Apr 07 09:32:34 2010
spam
Wed Apr 07 09:32:36 2010
spam
```

Таким образом, ключ -u по-прежнему можно использовать в версии 3.X для решения проблемы буферизации при соединении программ, если только стандартным потокам в порождаемых программах не назначаются другие объекты, как это делается в примере 12.11 для перенаправления потоков ввода-вывода в сокеты. В случае перенаправления в сокеты может потребоваться вручную вызывать метод flush или заменить обертки сокетов.

Сокеты и каналы

Итак, зачем вообще использовать сокеты для перенаправления? Проще говоря, для обеспечения независимости сервера и возможности использования в сети. Обратите внимание, что при использовании каналов не очевидно, кто должен называться «сервером», а кто «клиентом», потому что ни один из сценариев не выполняется непрерывно. Фактически, это один из основных недостатков использования каналов вместо сокетов в подобных ситуациях. Поскольку для использования каналов необходимо, чтобы *одна программа порождала другую*, каналы не могут использоваться для реализации взаимодействий с долгоживущими или удаленными серверами, как сокеты.

При использовании сокетов мы можем запускать клиенты и серверы независимо друг от друга, при этом серверы могут выполняться постоянно и обслуживать множество клиентов (хотя для этого придется внести некоторые изменения в функцию инициализации сервера в нашем вспомогательном модуле). Кроме того, возможность передачи имен удаленных компьютеров инструментам перенаправления в сокеты позволяет клиентам соединяться с серверами, выполняющимися на совершенно разных компьютерах. Как мы узнали в главе 5, именованные каналы (fifo) также обеспечивают независимость клиентов и серверов, но в отличие от сокетов они обычно могут применяться только в пределах локального компьютера и поддерживаются не всеми платформами.

Поэкспериментируйте с этим программным кодом, чтобы лучше вникнуть в его суть. Попробуйте также изменить пример 12.11, чтобы вместо сервера или параллельно с сервером он запускал клиентские функции в дочернем процессе, с вызовами и без вызовов метода `flush` и с вызовом функции `time.sleep` в конце, чтобы отложить завершение. Применение операции запуска дочернего процесса может также повлиять на надежность данной реализации диалога через сокеты, но мы не будем углубляться в обсуждение этого вопроса в интересах экономии места.

Несмотря на необходимость заботиться о кодировании текста и решать проблему буферизации потоков ввода-вывода, вспомогательный модуль в примере 12.10 все равно представляет весьма интересное решение – операции ввода-вывода автоматически выполняются через сетевые или локальные соединения сокетов, а чтобы задействовать этот модуль, в обычные сценарии требуется внести минимум изменений. Во многих случаях этот прием может существенно расширить область применения сценариев.

В следующем разделе мы снова будем использовать метод `makefile` для обертывания сокета объектом, похожим на файл, чтобы обеспечить возможность построчного чтения с применением обычных приемов и методов текстовых файлов. Строго говоря, в этом нет большой необходимости – мы могли бы читать строки, как строки байтов с помощью метода `recv` сокета. Однако в целом метод `makefile` удобно использовать, когда желательно работать с сокетами, как с простыми файлами. Двинемся дальше, чтобы увидеть действующий пример.

Простой файловый сервер на Python

Настало время для реализации более реалистичного примера. Завершим эту главу, применив некоторые из рассмотренных нами идей, относящихся к сокетам, для решения более полезной задачи, чем простая пересылка текста туда-обратно. В примере 12.17 представлена логика реализации сервера и клиента, необходимая для передачи файла с сервера на компьютер клиента через сокеты.

Этот сценарий реализует простую систему *загрузки файлов* с сервера. Один ее экземпляр выполняется на компьютере, где находятся загружаемые файлы (на сервере), а другой – на компьютере, куда должны копироваться файлы (на клиенте). Аргументы командной строки определяют, в каком качестве используется сценарий, а также могут использоваться для определения имени компьютера сервера и номера порта, через который должна производиться передача. Экземпляр сервера может отвечать на любое количество запросов файлов клиентами на порту, который он слушает, так как каждый запрос обслуживается в отдельном потоке.

Пример 12.17. PP4E\Internet\Sockets\getfile.py

```
"""
##############################################################################
реализует логику работы клиента и сервера для передачи произвольного файла
от сервера клиенту через сокет; использует простой протокол с управляющей
информацией вместо отдельных сокетов для передачи управляющих воздействий
и данных (как в ftp), обработка каждого клиентского запроса выполняется
в отдельном потоке, где организован цикл поблочной передачи содержимого
файла; более высокоуровневую схему организации транспортировки вы найдете
в примерах ftplib;
##############################################################################
"""

import sys, os, time, _thread as thread
from socket import *

blksz = 1024
defaultHost = 'localhost'
defaultPort = 50001

helptext = """
Usage...
server=> getfile.py -mode server           [-port nnn] [-host hhh|localhost]
client=> getfile.py [-mode client] -file fff [-port nnn] [-host
hhh|localhost]
"""

def now():
    return time.asctime()

def parsecommandline():
    dict = {}             # поместить в словарь для упрощения поиска
    args = sys.argv[1:]   # пропустить имя программы в начале аргументов
    while len(args) >= 2: # пример: dict['-mode'] = 'server'
        dict[args[0]] = args[1]
        args = args[2:]
    return dict

def client(host, port, filename):
```

```
    sock = socket(AF_INET, SOCK_STREAM)
    sock.connect((host, port))
    sock.send((filename + '\n').encode()) # имя файла с каталогом: bytes
    dropdir = os.path.split(filename)[1]  # имя файла в конце пути
    file = open(dropdir, 'wb')            # создать локальный файл в cwd
    while True:
        data = sock.recv(blksz)       # получать единовременно до 1 Кбайта
        if not data: break            # до закрытия сервером
        file.write(data)              # сохранить данные в локальном файле
    sock.close()
    file.close()
    print('Client got', filename, 'at', now())

def serverthread(clientsock):
    sockfile = clientsock.makefile('r')     # обернуть сокет объектом файла
    filename = sockfile.readline()[:-1]     # получить имя файла
                                            # без конца строки
    try:
        file = open(filename, 'rb')
        while True:
            bytes = file.read(blksz)        # читать/отправлять по 1 Кбайту
            if not bytes: break             # до полной передачи файла
            sent = clientsock.send(bytes)
            assert sent == len(bytes)
    except:
        print('Error downloading file on server:', filename)
    clientsock.close()

def server(host, port):
    serversock = socket(AF_INET, SOCK_STREAM)  # слушать на сокете TCP/IP
    serversock.bind((host, port))           # обслуживать клиентов в потоках
    serversock.listen(5)
    while True:
        clientsock, clientaddr = serversock.accept()
        print('Server connected by', clientaddr, 'at', now())
        thread.start_new_thread(serverthread, (clientsock,))

def main(args):
    host = args.get('-host', defaultHost)       # аргументы или умолчания
    port = int(args.get('-port', defaultPort))  # строка в argv
    if args.get('-mode') == 'server':       # None, если нет -mode: клиент
        if host == 'localhost': host = ''   # иначе потерпит неудачу
        server(host, port)                  # при удаленной работе
    elif args.get('-file'):                 # в режиме клиента нужен -file
        client(host, port, args['-file'])
    else:
        print(helptext)

if __name__ == '__main__':
    args = parsecommandline()
    main(args)
```

В этом сценарии нет ничего особенного в сравнении с уже встречавшимися примерами. В зависимости от аргументов командной строки он вызывает одну из двух функций:

- Функция `server` направляет все поступающие клиентские запросы в потоки, отправляющие байты запрошенного файла.
- Функция `client` посылает серверу имя файла и сохраняет полученные от него байты в локальном файле с таким же именем.

Наибольшая новизна заключается в протоколе между клиентом и сервером: клиент начинает диалог с сервером путем отправки ему строки с именем файла, оканчивающейся символом конца строки и содержащей путь к файлу на сервере. На сервере порожденный поток извлекает имя запрошенного файла, читая данные из сокета клиента, открывает запрошенный файл и отправляет его клиенту по частям.

Запуск сервера файлов и клиентов

Так как обслуживание клиентов на сервере выполняется в отдельных потоках, опробовать сервер и клиент можно на одном и том же компьютере с Windows. Сначала запустим экземпляр сервера, и пока он работает, запустим на том же компьютере два экземпляра клиента:

```
[окно сервера, localhost]
C:\...\Internet\Sockets> python getfile.py -mode server
Server connected by ('127.0.0.1', 59134) at Sun Apr 25 16:26:50 2010
Server connected by ('127.0.0.1', 59135) at Sun Apr 25 16:27:21 2010

[окно клиента, localhost]
C:\...\Internet\Sockets> dir /B *.gif *.txt
File Not Found

C:\...\Internet\Sockets> python getfile.py -file testdir\ora-lp4e.gif
Client got testdir\ora-lp4e.gif at Sun Apr 25 16:26:50 2010

C:\...\Internet\Sockets> python getfile.py -file testdir\textfile.txt
                             -port 50001
Client got testdir\textfile.txt at Sun Apr 25 16:27:21 2010
```

Клиенты запускаются в каталоге, куда нужно поместить загружаемые файлы – реализация экземпляра клиента отбрасывает пути на сервере при создании локального файла. Здесь «загрузка» просто копирует запрошенные файлы в локальный родительский каталог (команда DOS `fc` сравнивает содержимое файлов):

```
C:\...\Internet\Sockets> dir /B *.gif *.txt
ora-lp4e.gif
textfile.txt

C:\...\Internet\Sockets> fc /B ora-lp4e.gif testdir/ora-lp4e.gif
FC: no differences encountered
```

```
C:\...\Internet\Sockets> fc textfile.txt testdir\textfile.txt
FC: no differences encountered
```

Как обычно, сервер и клиенты можно запускать на разных компьютерах. Например, ниже показаны команды, которые можно было бы использовать для запуска сценария как сервера на удаленном компьютере и получения файлов на локальном компьютере. Попробуйте выполнить их у себя, чтобы увидеть вывод сервера и клиентов.

[окно удаленного сервера]

```
[...]$ python getfile.py -mode server
```

[окно клиента: обслуживание клиента выполняется в отдельном потоке на сервере]

```
C:\...\Internet\Sockets> python getfile.py -mode client
                             -host learning-python.com
                             -port 50001 -file python.exe

C:\...\Internet\Sockets> python getfile.py
                             -host learning-python.com -file index.html
```

Замечание, касающееся безопасности: реализация сервера готова отправить любой файл, находящийся на сервере, имя которого получено от клиента, если сервер выполняется под именем пользователя, имеющего право чтения запрошенного файла. Если вас волнует проблема защиты некоторых своих файлов на сервере, следует добавить логику, запрещающую загрузку защищенных файлов. Я оставляю ее реализацию читателю в качестве упражнения, но такие проверки имен файлов будут реализованы в утилите загрузки `getfile`, далее в книге.[1]

Добавляем графический интерфейс пользователя

После суеты вокруг графических интерфейсов в предыдущей части книги вы могли заметить, что на протяжении всей этой главы мы жили в мире командной строки – наши клиенты и серверы сокетов запускались из простых командных оболочек DOS или Linux. Однако ничто не мешает нам добавить в некоторые из этих сценариев красивый пользовательский интерфейс, «point-and-click» («укажи-и-щелкни»), – графические интерфейсы и сетевые сценарии не являются взаимно исклю-

[1] Там мы встретимся еще с тремя программами `getfile`, прежде чем распрощаться со сценариями для Интернета. В следующей главе сценарий *getfile.py* будет использовать не прямое обращение к сокетам, а получать файлы с помощью интерфейса FTP более высокого уровня, а сценарии *http-getfile* будут получать файлы по протоколу HTTP. В главе 15 представлен CGI-сценарий *getfile.py*, передающий содержимое файла через порт HTTP в ответ на запрос, производимый веб-броузером (файлы отправляются как вывод CGI-сценария). Все четыре схемы загрузки, представленные в этой книге, в конечном счете, используют сокеты, но только в этой версии они используются явно.

чающими технологиями. На самом деле их правильное совместное использование может оказаться достаточно привлекательным.

Например, нетрудно реализовать на основе tkinter графический интерфейс для клиентской части сценария getfile, с которым мы только что познакомились. Такой инструмент, выполняясь на клиентском компьютере, может просто выводить всплывающее окно с виджетами Entry для ввода имени файла, сервера и так далее. После ввода параметров загрузки интерфейс пользователя может импортировать и вызвать функцию getfile.client с соответствующими аргументами либо сконструировать и выполнить команду запуска getfile.py с помощью таких инструментов, как os.system, os.popen, subprocess и так далее.

Использование фреймов рядов и команд

Для большей конкретности рассмотрим очень бегло несколько простых сценариев, добавляющих интерфейс на основе tkinter к клиентской стороне программы getfile. Все эти примеры предполагают, что серверная часть getfile уже запущена, – они просто добавляют графический интерфейс к клиентской стороне программы, облегчающий загрузку файла с сервера. Первый из них, представленный в примере 12.18, создает диалоговое окно для ввода данных о сервере, порте и имени файла, используя приемы конструирования форм, с которыми мы встречались в главах 8 и 9, а потом строит соответствующую команду getfile и выполняет ее с помощью функции os.system, рассматривавшейся во второй части книги.

Пример 12.18. PP4E\Internet\Sockets\getfilegui-1.py

```
"""
запускает сценарий getfile в режиме клиента из простого
графического интерфейса на основе tkinter;
точно так же можно было бы использовать os.fork+exec, os.spawnv
(смотрите модуль Launcher);
в windows: замените 'python' на 'start', если каталог
с интерпретатором не перечислен в переменной окружения PATH;
"""

import sys, os
from tkinter import *
from tkinter.messagebox import showinfo

def onReturnKey():
    cmdline = ('python getfile.py -mode client -file %s -port %s -host %s' %
                      (content['File'].get(),
                       content['Port'].get(),
                       content['Server'].get()))
    os.system(cmdline)
    showinfo('getfilegui-1', 'Download complete')
```

```
box = Tk()
labels = ['Server', 'Port', 'File']
content = {}
for label in labels:
    row = Frame(box)
    row.pack(fill=X)
    Label(row, text=label, width=6).pack(side=LEFT)
    entry = Entry(row)
    entry.pack(side=RIGHT, expand=YES, fill=X)
    content[label] = entry

box.title('getfilegui-1')
box.bind('<Return>', (lambda event: onReturnKey()))
mainloop()
```

Если запустить этот сценарий, он создаст форму, изображенную на рис. 12.1. Нажатие клавиши Enter запускает экземпляр программы getfile в режиме клиента. Когда сгенерированная команда вызова getfile завершается, появляется окно подтверждения, изображенное на рис. 12.2.

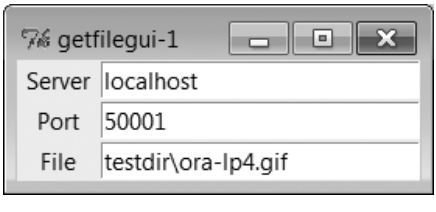


Рис. 12.1. *Сценарий getfilegui-1 в действии*

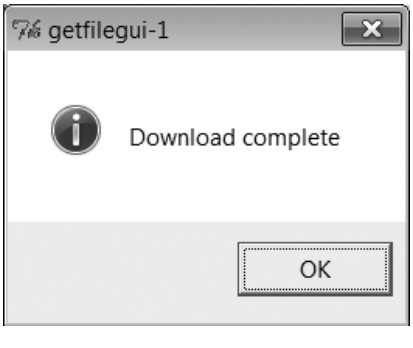


Рис. 12.2. *Окно подтверждения getfilegui-1*

Использование сеток и вызовов функций

В первом сценарии пользовательского интерфейса (пример 12.18) для создания формы ввода использован менеджер компоновки pack и фреймы рядов с метками фиксированной длины, а клиент getfile выполняется как самостоятельная программа. Как мы узнали в главе 9, для расположения элементов на форме с тем же успехом можно просто использовать менеджер grid, а также импортировать и вызвать функцию, реализующую логику клиента, а не запускать самостоятельную программу. Это решение демонстрируется в примере 12.19.

Пример 12.19. PP4E\Internet\Sockets\getfilegui-2.py

```
"""
то же самое, но с компоновкой по сетке и импортом с вызовом вместо
компоновки менеджером pack и командной строки; непосредственные вызовы
функций обычно выполняются быстрее, чем запуск файлов;
"""

import getfile
from tkinter import *
from tkinter.messagebox import showinfo

def onSubmit():
    getfile.client(content['Server'].get(),
                   int(content['Port'].get()),
                   content['File'].get())
    showinfo('getfilegui-2', 'Download complete')

box = Tk()
labels = ['Server', 'Port', 'File']
rownum = 0
content = {}
for label in labels:
    Label(box, text=label).grid(column=0, row=rownum)
    entry = Entry(box)
    entry.grid(column=1, row=rownum, sticky=E+W)
    content[label] = entry
    rownum += 1

box.columnconfigure(0, weight=0)      # сделать растягиваемым
box.columnconfigure(1, weight=1)
Button(text='Submit', command=onSubmit).grid(row=rownum, column=0,
                                             columnspan=2)

box.title('getfilegui-2')
box.bind('<Return>', (lambda event: onSubmit()))
mainloop()
```

Эта версия создает похожее окно (рис. 12.3), в нижнюю часть которого добавлена кнопка, выполняющая то же действие, что нажатие клавиши Enter, – она запускает процедуру клиента getfile. Вообще говоря,

импорт и вызов функций (как в этом примере) происходит быстрее, чем выполнение команд, особенно при многократном выполнении. Сценарий getfile позволяет использовать его любым способом – как программу или как библиотеку функций.

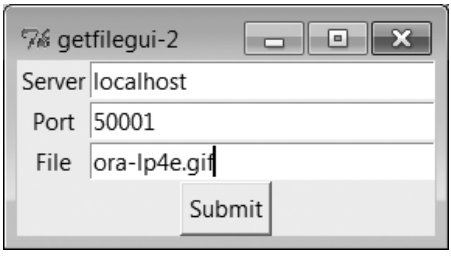


Рис. 12.3. *Сценарий getfilegui-2 в действии*

Многократно используемый класс формы

Если вы похожи на меня, то писать всю реализацию компоновки формы для этих двух сценариев покажется вам утомительным, какой бы менеджер компоновки, pack или grid, вы ни использовали. Мне это показалось настолько скучным, что я решил написать класс структуры формы общего назначения, представленный в примере 12.20, который выполняет большую часть черновой работы по компоновке элементов графического интерфейса.

Пример 12.20. PP4E\Internet\Sockets\form.py

```
"""
##################################################################
многократно используемый класс формы, задействованный
в сценарии getfilegui (и в других)
##################################################################
"""

from tkinter import *
entrysize = 40

class Form:                                     # немодальное окно формы
    def __init__(self, labels, parent=None):    # передать список меток полей
        labelsize = max(len(x) for x in labels) + 2
        box = Frame(parent)                       # в окне есть ряды, кнопка
        box.pack(expand=YES, fill=X)              # ряды оформлены как фреймы
        rows = Frame(box, bd=2, relief=GROOVE)    # нажатие кнопки или Enter
        rows.pack(side=TOP, expand=YES, fill=X)   # вызывают метод onSubmit
        self.content = {}
        for label in labels:
            row = Frame(rows)
            row.pack(fill=X)
```

```
            Label(row, text=label, width=labelsize).pack(side=LEFT)
            entry = Entry(row, width=entrysize)
            entry.pack(side=RIGHT, expand=YES, fill=X)
            self.content[label] = entry
        Button(box, text='Cancel', command=self.onCancel).pack(side=RIGHT)
        Button(box, text='Submit', command=self.onSubmit).pack(side=RIGHT)
        box.master.bind('<Return>', (lambda event: self.onSubmit()))

    def onSubmit(self):                # переопределить этот метод
        for key in self.content:       # ввод пользователя
            print(key, '\t=>\t', self.content[key].get()) # в self.content[k]

    def onCancel(self):                # переопределить при необходимости
        Tk().quit()                    # по умолчанию осуществляет выход

class DynamicForm(Form):
    def __init__(self, labels=None):
        labels = input('Enter field names: ').split()
        Form.__init__(self, labels)
    def onSubmit(self):
        print('Field values...')
        Form.onSubmit(self)
        self.onCancel()

if __name__ == '__main__':
    import sys
    if len(sys.argv) == 1:
        Form(['Name', 'Age', 'Job'])  # предопределенные поля остаются
    else:                             # после передачи
        DynamicForm()                 # динамически созданные поля
    mainloop()                        # исчезают после передачи
```

Сравните этот подход с тем, что был реализован в функции конструирования рядов форм, которую мы написали в главе 10, в примере 10.9. В то время как этот пример заметно уменьшает объем программного кода, необходимого для его использования, он реализует более полную и автоматизированную схему – модуль конструирует форму целиком, исходя из заданного набора имен меток, и предоставляет словарь со всеми виджетами полей ввода, готовыми для извлечения информации.

Если запустить этот модуль как самостоятельный сценарий, выполняется программный код самотестирования, находящийся в конце. При запуске без аргументов (или двойным щелчком в проводнике файлов Windows) программный код самопроверки генерирует форму с готовыми полями, как показано на рис. 12.4, и выводит значения полей при нажатии клавиши Enter или щелчке на кнопке Submit:

```
C:\...\PP4E\Internet\Sockets> python form.py
Age     =>      40
Name    =>      Bob
Job     =>      Educator, Entertainer
```

Рис. 12.4. *Тест формы, предопределенные поля ввода*

При запуске с аргументами командной строки программный код самопроверки в модуле класса формы предлагает ввести произвольную группу имен полей формы. При желании поля могут создаваться динамически. На рис. 12.5 показана форма для ввода, сконструированная в результате приведенного ниже диалога в консоли. Имена полей могут быть взяты из командной строки, но в таких простых проверках столь же хорошо действует и встроенная функция `input`. В этом режиме графический интерфейс исчезает после первой передачи данных, потому что так определено в методе `DynamicForm.onSubmit`:

```
C:\...\PP4E\Internet\Sockets> python form.py -
Enter field names: Name Email Web Locale
Field values...
Locale  =>      Florida
Web     =>      http://learning-python.com
Name    =>      Book
Email   =>      pp4e@learning-python.com
```

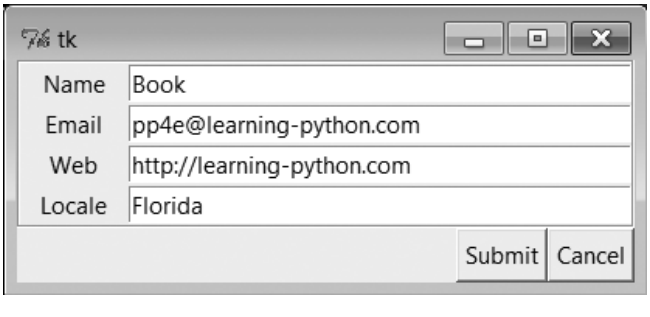


Рис. 12.5. *Тест формы, динамические поля ввода*

И последнее, но немаловажное замечание. В примере 12.21 приводится еще одна реализация интерфейса пользователя для `getfile`, на этот раз построенного с помощью многократно используемого класса компоновки формы. Необходимо лишь заполнить список меток и предоставить свой метод обратного вызова `onSubmit`. Все действия по созданию формы

совершаются «бесплатно» в результате импорта многократно используемого суперкласса `Form`.

Пример 12.21. PP4E\Internet\Sockets\getfilegui.py

```
"""
запускает функцию client из модуля getfile и реализует графический
интерфейс на основе многократно используемого класса формы;
с помощью os.chdir выполняет переход в требуемый локальный каталог,
если указан (getfile сохраняет файл в cwd);
что сделать: использовать потоки выполнения, вывести индикатор
хода выполнения операции и отобразить вывод getfile;
"""


from form import Form
from tkinter import Tk, mainloop
from tkinter.messagebox import showinfo
import getfile, os

class GetfileForm(Form):
    def __init__(self, oneshot=False):
        root = Tk()
        root.title('getfilegui')
        labels = ['Server Name', 'Port Number', 'File Name', 'Local Dir?']
        Form.__init__(self, labels, root)
        self.oneshot = oneshot

    def onSubmit(self):
        Form.onSubmit(self)
        localdir = self.content['Local Dir?'].get()
        portnumber = self.content['Port Number'].get()
        servername = self.content['Server Name'].get()
        filename = self.content['File Name'].get()
        if localdir:
            os.chdir(localdir)
        portnumber = int(portnumber)
        getfile.client(servername, portnumber, filename)
        showinfo('getfilegui', 'Download complete')
        if self.oneshot: Tk().quit()  # иначе останется в последнем localdir

if __name__ == '__main__':
    GetfileForm()
    mainloop()
```

Импортированный здесь класс компоновки формы может быть использован любой программой, где требуется организовать ввод данных в виде формы. При использовании в данном сценарии в Windows 7 получается интерфейс пользователя, как показано на рис. 12.6 (и похожий на других платформах).

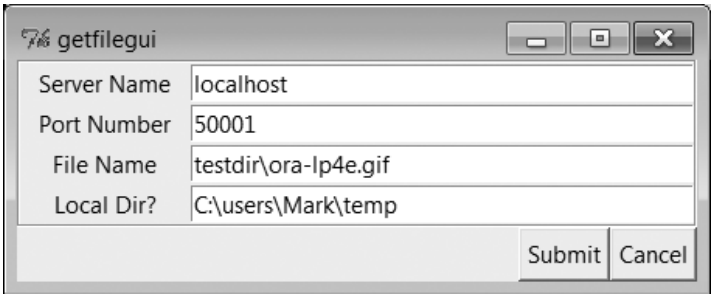


Рис. 12.6. *Сценарий getfilegui в действии*

Щелчок на кнопке Submit или нажатие клавиши Enter в этой форме, как и раньше, заставляет сценарий getfilegui вызвать импортированную функцию клиентской части getfile.client. Однако на сей раз сначала производится переход в указанный в форме локальный каталог, куда следует сохранить полученный файл (getfile сохраняет файл в текущем рабочем каталоге, каким бы он ни был при вызове сценария). Ниже приводятся сообщения, которые выводятся в консоли клиента, а также результат проверки переданного файла – сервер все так же действует в каталоге выше testdir, а клиент сохраняет файл в каком-то другом месте после извлечения его из сокета:

```
C:\...\Internet\Sockets> getfilegui.py
Local Dir?      =>      C:\users\Mark\temp
File Name       =>      testdir\ora-lp4e.gif
Server Name     =>      localhost
Port Number     =>      50001
Client got testdir\ora-lp4e.gif at Sun Apr 25 17:22:39 2010

C:\...\Internet\Sockets> fc /B C:\Users\mark\temp\ora-lp4e.gif
                                testdir\ora-lp4e.gif
FC: no differences encountered
```

Как обычно, с помощью этого интерфейса можно соединяться с серверами, которые выполняются локально на том же компьютере (как здесь) или удаленно. Если у вас сервер выполняется удаленно, укажите другое имя компьютера сервера и путь к файлу – волшебная сила сокетов «просто действует», независимо от того, где выполняется сервер, локально или удаленно.

Здесь стоит сделать одно предупреждение: графический интерфейс фактически замирает, пока происходит загрузка (даже перерисовка экрана не выполняется – попробуйте заслонить окно и снова открыть его, и вы поймете, что я имею в виду). Положение можно улучшить, запустив загрузку в отдельном потоке выполнения, но пока мы не увидим,

как это делается, – в следующей главе, где будем исследовать протокол FTP, – следует считать это замечание предварительным знакомством с проблемой.

В завершение несколько последних замечаний. Во-первых, я должен отметить, что сценарии, представленные в этой главе, применяют приемы использования tkinter, которые мы уже видели раньше и здесь не станем подробно рассматривать их в интересах экономии места. Советы по реализации можно найти в главах этой книги, посвященных графическому интерфейсу.

Имейте также в виду, что все эти интерфейсы добавляются к уже существующим сценариям, повторно используя их реализацию, – таким способом можно снабдить графическим интерфейсом любой инструмент командной строки, сделав его более привлекательным и дружественным пользователю. Например, в главе 14 мы познакомимся с более удобным клиентским интерфейсом пользователя на основе tkinter, предназначенным для чтения и отправки электронной почты через сокеты (PyMailGui), который в общем-то лишь добавляет графический интерфейс к средствам обработки электронной почты. Вообще говоря, графические интерфейсы часто могут быть добавлены к программам почти что задним числом. Хотя степень разделения интерфейса пользователя и базовой логики может быть различной в каждой программе, отделение одного от другого облегчает возможность сосредоточиться на каждом из них в отдельности.

И наконец, теперь, когда я показал, как создавать интерфейсы пользователя поверх сценария `getfile` из этой главы, должен также сказать, что в действительности они не столь полезны, как может показаться. В частности, клиенты `getfile` могут общаться только с теми компьютерами, на которых выполняется сервер `getfile`. В следующей главе мы откроем для себя еще один способ загрузки файлов с сервера, протокол FTP, который также основывается на сокетах, но предоставляет интерфейс более высокого уровня и доступен в качестве стандартной службы на многих компьютерах в Сети. Обычно не требуется запускать индивидуально разработанный сервер для передачи файлов через FTP, как мы это делали с `getfile`. Сценарии с графическим интерфейсом пользователя, представленные в этой главе, можно легко изменить, чтобы получить нужный файл с помощью инструментов FTP, имеющихся в Python, а не модуля `getfile`. Но я не стану сейчас все рассказывать, а просто предложу продолжить чтение.

Использование последовательных портов

Сокеты, главный предмет этой главы, служат в сценариях Python программным интерфейсом к сетевым соединениям. Как было показано выше, они позволяют писать сценарии, обменивающиеся данными с компьютерами, расположенными в произвольном месте сети, и образуют становой хребет Интернета и Веб.

Однако если вы ищете более низкоуровневые средства для связи с устройствами в целом, вас может заинтересовать тема интерфейсов Python к последовательным портам. Эта тема напрямую не связана со сценариями для Интернета, однако она достаточно близка по духу и достаточно часто обсуждается в Сети, чтобы быть кратко рассмотренной здесь.

Используя интерфейсы к последовательным портам, сценарии могут взаимодействовать с такими устройствами, как мышь, модем и целым рядом других последовательных устройств. Интерфейсы последовательных портов применяются также для связи с устройствами, подключаемыми через инфракрасные порты (например, с карманными компьютерами и удаленными модемами). Такие интерфейсы позволяют сценариям вмешиваться в потоки необработанных данных и реализовывать собственные протоколы взаимодействий с устройствами. Для создания и извлечения упакованных двоичных данных, передаваемых через эти порты, можно также использовать дополнительные инструменты, предоставляемые стандартными модулями Python `ctypes` и `struct`.

В настоящее время существует не менее трех способов реализовать на языке Python прием и передачу данных через последовательные порты. Наибольшего внимания заслуживает расширение *PySerial*, распространяемое с открытыми исходными текстами, которое позволяет реализовать управление последовательными портами на языке Python в Windows и Linux, а также в BSD Unix, Jython (для Java) и IronPython (для .Net и Mono). К сожалению, здесь не так много места, чтобы обсудить эти или любые другие инструменты для работы с последовательными портами с той или иной степенью подробности. Как обычно, чтобы получить самые свежие сведения по этой теме, следует использовать поисковые системы в Интернете.

13

Сценарии на стороне клиента

«Свяжись со мной!»

В предыдущей главе мы познакомились с основами Интернета и исследовали сокеты – механизм взаимодействий, посредством которого осуществляется передача потоков байтов через Сеть. В данной главе мы поднимемся на один уровень выше в иерархии инкапсуляции и направим внимание на инструменты Python, которые обеспечивают поддержку стандартных протоколов Интернета на стороне клиента.

В начале предыдущей главы были кратко описаны протоколы Интернета верхнего уровня, и если вы пропустили этот материал при первом чтении, к нему, вероятно, стоит вернуться. Вкратце, протоколы определяют порядок обмена информацией, происходящего при выполнении большинства знакомых нам задач Интернета – чтении электронной почты, передаче файлов по FTP, загрузке веб-страниц и так далее.

В основе своей все эти диалоги протоколов осуществляются через сокеты с использованием фиксированных и стандартных структур сообщений и номеров портов, поэтому в некотором смысле данная глава основана на предыдущей. Но, как будет показано далее, модули протоколов Python скрывают большую часть деталей – сценариям обычно приходится иметь дело только с простыми объектами и методами, в то время как Python автоматизирует логику сокетов и сообщений, требуемую протоколом.

В этой главе мы сосредоточимся на модулях Python протоколов FTP и электронной почты, но попутно взглянем и на некоторые другие (новостей NNTP, веб-страниц HTTP и так далее). Из-за важного положения, занимаемого в Интернете электронной почтой, мы уделим ей много внимания в этой главе, так же как и в последующих двух – инструменты

и приемы, представленные здесь, мы будем использовать для создания крупных примеров клиентских и серверных программ PyMailGUI и PyMailCGI в главах 14 и 16.

Все инструменты, используемые в примерах этой главы, присутствуют в стандартной библиотеке Python и поставляются вместе с системой Python. Все примеры, представленные здесь, предназначены для выполнения на клиентской стороне сетевого соединения – эти сценарии соединяются с уже действующим сервером, которому они передают запросы, и могут выполняться на обычном ПК или другом клиентском устройстве (они требуют только возможности соединения с сервером). И как обычно, весь программный код, представленный здесь, разрабатывался также с целью показать приемы программирования на языке Python в целом – мы будем реорганизовывать примеры использования FTP и перепаковывать примеры работы с электронной почтой, демонстрируя объектно-ориентированное программирование (ООП) в действии.

В следующей главе мы рассмотрим законченный пример клиентской программы, после чего перейдем к изучению сценариев, которые, напротив, предназначены для выполнения на стороне сервера. Программы на языке Python способны также генерировать страницы на веб-сервере, и в мире Python имеется вся необходимая поддержка для создания серверов HTTP, электронной почты и FTP. А пока займемся клиентом.[1]

FTP: передача файлов по сети

Как было показано в предыдущей главе, сокеты используются для выполнения самых разных действий в Сети. В частности, пример `getfile` из предыдущей главы обеспечивал передачу между машинами файлов целиком. Однако на практике многое из происходящего в Сети обеспечивается протоколами более высокого уровня. Протоколы действуют

[1] В языке Python имеется поддержка и других технологий, которые также можно отнести к разряду «клиентских», таких как апплеты Jython/Java, веб-службы XML-RPC и SOAP и инструменты создания полнофункциональных интернет-приложений, таких как Flex, Silverlight, pyjamas и AJAX. Они уже были представлены ранее в главе 12. Такие инструменты тесно связаны с понятием веб-взаимодействий – они либо расширяют возможности веб-броузера, выполняющегося на клиентском компьютере, либо упрощают доступ к веб-серверу со стороны клиента. С приемами расширения возможностей броузера мы познакомимся в главах 15 и 16; здесь же под клиентскими сценариями мы будем подразумевать клиентскую сторону протоколов, часто используемых в Интернете, таких как FTP и электронная почта, не зависящих от Веб или веб-броузеров. В своей основе веб-броузеры являются всего лишь обычными приложениями с графическим интерфейсом, которые используют поддержку протоколов на стороне клиента, включая и те, что мы будем изучать здесь, такие как HTTP и FTP. Дополнительные сведения о приемах, применяемых на стороне клиента, вы найдете в главе 12, а также в конце этой главы.

поверх сокетов и скрывают значительную часть сложностей сетевых сценариев, которые мы видели в примерах в предыдущей главе.

FTP (File Transfer Protocol, протокол передачи файлов) – один из наиболее часто используемых протоколов Интернета. Он определяет модель взаимодействия более высокого уровня, в основе которой лежит обмен строками команд и содержимым файлов через сокеты. Протокол FTP позволяет решать те же задачи, что и сценарий `getfile` из предыдущей главы, но использует более простой, стандартный и универсальный интерфейс – FTP позволяет запрашивать файлы любой машине-серверу, которая поддерживает FTP, не требуя, чтобы на ней выполнялся наш специализированный сценарий `getfile`. FTP позволяет также выполнять более сложные операции, такие как выгрузка файлов на сервер, получение содержимого удаленного каталога и многое другое.

В действительности FTP выполняется поверх двух сокетов: один из них служит для передачи управляющих команд между клиентом и сервером (порт 21), а другой – для передачи байтов. Благодаря использованию модели с двумя сокетами FTP устраняет возможность взаимной блокировки (то есть передача в сокетах данных не блокирует диалога в управляющих сокетах). И наконец, существует вспомогательный модуль Python `ftplib`, который позволяет выгружать файлы на удаленный сервер и загружать с него посредством FTP, не имея дело ни с низкоуровневыми вызовами сокетов, ни с деталями протокола FTP.

Передача файлов с помощью ftplib

Поскольку интерфейс Python к протоколу FTP очень прост, перейдем сразу к практическому примеру. Сценарий, представленный в примере 13.1, автоматически загружает и открывает удаленный файл с помощью Python. Если быть более точными, этот сценарий Python выполняет следующие действия:

1. Загружает файл изображения (по умолчанию) с удаленного сайта FTP.
2. Открывает загруженный файл с помощью утилиты, реализованной нами в главе 6 (пример 6.23).

Часть, которая выполняет загрузку, будет работать на любом компьютере, где есть Python и соединение с Интернетом. Однако вам, вероятно, придется изменить настройки в сценарии таким образом, чтобы он обращался к вашему серверу FTP и загружал ваш файл. Часть сценария, которая открывает файл, будет работать, если *playfile.py* поддерживает вашу платформу – смотрите подробности в главе 6 и внесите соответствующие изменения, если это необходимо.

Пример 13.1. PP4E\Internet\Ftp\getone.py

```
#!/usr/local/bin/python
"""
Сценарий на языке Python для загрузки медиафайла по FTP и его проигрывания.
```

```
Использует модуль ftplib, реализующий поддержку протокола ftp на основе
сокетов. Протокол FTP использует 2 сокета (один для данных и один
для управления - на портах 20 и 21) и определяет форматы текстовых
сообщений, однако модуль ftplib скрывает большую часть деталей этого
протокола. Измените настройки в соответствии со своим сайтом/файлом.
"""

import os, sys
from getpass import getpass  # инструмент скрытого ввода пароля
from ftplib import FTP       # инструменты FTP на основе сокетов

nonpassive = False           # использовать активный режим FTP?
filename   = 'monkeys.jpg'   # загружаемый файл
dirname    = '.'             # удаленный каталог, откуда загружается файл
sitename   = 'ftp.rmi.net'   # FTP-сайт, к которому выполняется подключение
userinfo   = ('lutz', getpass('Pswd?'))      # () - для анонимного доступа
if len(sys.argv) > 1: filename = sys.argv[1] # имя файла в командной строке?

print('Connecting...')
connection = FTP(sitename)        # соединиться с FTP-сайтом
connection.login(*userinfo)       # по умолчанию анонимный доступ
connection.cwd(dirname)           # передача порциями по 1 Кбайту
if nonpassive:                    # использовать активный режим FTP,
    connection.set_pasv(False)    # если этого требует сервер

print('Downloading...')
localfile = open(filename, 'wb')  # локальный файл, куда сохраняются данные
connection.retrbinary('RETR ' + filename, localfile.write, 1024)
connection.quit()
localfile.close()

if input('Open file?') in ['Y', 'y']:
    from PP4E.System.Media.playfile import playfile
    playfile(filename)
```

Большинство деталей реализации протокола FTP инкапсулировано в импортируемом модуле Python `ftplib`. Данный сценарий использует самые простые интерфейсы `ftplib` (остальные мы увидим чуть позже, в этой же главе), но они являются достаточно представительными для модуля в целом.

Чтобы открыть соединение с удаленным (или локальным) сервером FTP, нужно создать экземпляр класса `ftplib.FTP`, передав ему имя (доменное или IP-адрес) компьютера, с которым нужно соединиться:

```
connection = FTP(sitename)      # соединиться с FTP-сайтом
```

Если при этом вызове не возбуждается исключение, полученный объект FTP экспортирует методы, соответствующие обычным операциям FTP. Сценарии Python действуют подобно типичным программам FTP-клиентов – нужно просто заменить обычные вводимые или выбираемые команды вызовами методов:

```
connection.login(*userinfo)     # по умолчанию анонимный доступ
connection.cwd(dirname)         # передача порциями по 1 Кбайту
```

После подключения производится регистрация и переход в удаленный каталог, где находится требуемый файл. Метод `login` позволяет передавать дополнительные необязательные аргументы, определяющие имя пользователя и пароль. По умолчанию выполняется анонимная регистрация FTP. Обратите внимание на флаг `nonpassive`, используемый в этом сценарии:

```
if nonpassive:                  # использовать активный режим FTP,
    connection.set_pasv(False)  # если этого требует сервер
```

Если этот флаг установлен в значение `True`, сценарий будет осуществлять передачу файла не в пассивном режиме FTP, используемом по умолчанию, а в активном. Мы не будем углубляться здесь в детали отличий между режимами (режим определяет, с какой стороны соединения производится выбор номера порта для передачи). Но если у вас возникнут проблемы с передачей файлов с помощью какого-либо сценария FTP из этой главы, попробуйте сначала использовать активный режим. В Python 2.1 и более поздних версиях, по умолчанию используется пассивный режим FTP. Теперь откроем локальный файл, куда будет сохраняться содержимое принимаемого файла, и выполним загрузку:

```
localfile = open(filename, 'wb')
connection.retrbinary('RETR ' + filename, localfile.write, 1024)
```

После перехода в целевой каталог вызывается метод `retrbinary` для загрузки целевого файла с сервера в двоичном режиме. Для завершения вызова `retrbinary` требуется некоторое время, поскольку должен быть загружен большой файл. Метод принимает три аргумента:

- Строка команды FTP, в данном случае строка `RETR` *`имя_файла`*, являющаяся стандартным форматом загрузки по FTP.
- Функция или метод, которым Python передает каждый блок загруженных байтов файла, – в данном случае метод `write` вновь созданного и открытого локального файла.
- Размер этих блоков байтов. В данном случае каждый раз загружается 1024 байта, но если этот аргумент опущен, используется значение по умолчанию.

Так как этот сценарий создает локальный файл с именем `localfile`, таким же, как у загружаемого удаленного файла, и передает его метод `write` методу получения FTP, содержимое удаленного файла автоматически окажется в локальном файле на стороне клиента после завершения загрузки.

Обратите внимание, что этот файл открывается в двоичном режиме `wb`: если этот сценарий выполняется в Windows, нужно избежать автоматического преобразования байтов `\n` в последовательности байтов `\r\n` – как мы видели в главе 4, эта операция автоматически выполня-

ется в Windows при записи в файлы, открытые в текстовом режиме `w`. Нам также необходимо избежать проблем с кодировкой Юникода в Python 3.X – как мы знаем из той же главы 4, при записи в текстовом режиме выполняется кодирование строк, что является излишним для двоичных файлов, таких как изображения. А кроме того, текстовый режим не позволил бы библиотечному методу `retrbinary` передавать строки `bytes` методу `write` текстового файла, поэтому режим `wb` фактически является здесь единственно допустимым (подробнее о режимах открытия файлов для записи мы поговорим ниже).

Наконец, вызывается метод FTP `quit`, чтобы разорвать соединение с сервером, и с помощью метода `close` вручную закрывается локальный файл, чтобы вытолкнуть выходные буферы на диск и обеспечить возможность дальнейшей обработки файла (без вызова `close` части файла могут остаться в выходных буферах):

```
connection.quit()
localfile.close()
```

Вот и все, что нужно сделать. Все детали протокола FTP, сокетов и работы в сети скрыты за интерфейсом модуля `ftplib`. Ниже приводятся результаты работы этого сценария в Windows 7 – после загрузки файл изображения появляется на экране моего ноутбука, в окне программы просмотра, как показано на рис. 13.1. Измените имя сервера и файла в этом сценарии, чтобы опробовать его со своим сервером и своим файлом, и обязательно проверьте, чтобы переменная окружения `PYTHONPATH` включала путь к корневому каталогу примеров *PP4E*, так как здесь выполняется импортирование модулей из дерева каталогов с примерами:

```
C:\...\PP4E\Internet\Ftp> python getone.py
Pswd?
Connecting...
Downloading...
Open file?y
```

Обратите внимание, что здесь для запроса пароля FTP используется стандартная функция Python `getpass.getpass`. Подобно встроенной функции `input`, она выводит приглашение к вводу и читает строку, вводимую пользователем в консоли. В отличие от `input`, функция `getpass` не выводит вводимые символы на экран (смотрите пример `moreplus` перенаправления потоков ввода-вывода в главе 3, где демонстрируются похожие инструменты). Этот прием удобно использовать для сокрытия паролей от постороннего глаза. Но будьте внимательны – графический интерфейс IDLE после предупреждения выводит все символы пароля!

Обратите особое внимание, что этот обычный в остальных отношениях сценарий Python способен получать данные с произвольных удаленных сайтов FTP и компьютеров. При наличии ссылки с помощью подобных интерфейсов сценариями Python может быть получена любая информация, опубликованная на сервере FTP в Сети.

Рис. 13.1. *Файл изображения, загруженный по FTP и открытый на локальном компьютере*

Использование пакета urllib для загрузки файлов

FTP является лишь одним из способов передачи информации через Сеть, и в библиотеке Python есть более универсальные средства для выполнения такой загрузки, как в предыдущем сценарии. Пожалуй, наиболее простым в этом отношении является модуль `urllib.request`: получив строку с адресом в Интернете – адрес URL, или унифицированный указатель ресурса (Uniform Resource Locator) – этот модуль открывает соединение с указанным сервером и возвращает объект, похожий на файл, который можно читать с помощью обычных вызовов методов объекта файла (например, `read`, `readline`).

Такой высокоуровневый интерфейс может быть применен для загрузки всего, что имеет адрес в Сети, – файлов, опубликованных на FTP-сайтах (используя адреса URL, начинающиеся с *ftp://*), веб-страниц и вывода сценариев, расположенных на удаленных серверах (используя адреса URL, начинающиеся с *http://*), и даже локальных файлов (используя адреса URL, начинающиеся с *file://*). В частности, сценарий в при-

мере 13.2 делает то же, что и сценарий в примере 13.1, но для получения файла дистрибутива с исходными текстами вместо модуля конкретного протокола ftplib использует более универсальный модуль urllib.request.

Пример 13.2. PP4E\Internet\Ftp\getone-urllib.py

```
#!/usr/local/bin/python
"""
Сценарий на языке Python для загрузки файла по строке адреса URL;
вместо ftplib использует более высокоуровневый модуль urllib;
urllib поддерживает протоколы FTP, HTTP, HTTPS на стороне клиента,
локальные файлы, может работать с прокси-серверами, выполнять инструкции
перенаправления, принимать cookies и многое другое; urllib также
позволяет загружать страницы html, изображения, текст и так далее;
смотрите также парсеры Python разметки html/xml веб-страниц,
получаемых с помощью urllib, в главе 19;
"""

import os, getpass
from urllib.request import urlopen   # веб-инструменты на основе сокетов

filename = 'monkeys.jpg'             # имя удаленного/локального файла
password = getpass.getpass('Pswd?')

remoteaddr = 'ftp://lutz:%s@ftp.rmi.net/%s;type=i' % (password, filename)
print('Downloading', remoteaddr)

# такой способ тоже работает:
# urllib.request.urlretrieve(remoteaddr, filename)

remotefile = urlopen(remoteaddr)   # возвращает объект типа файла для ввода
localfile = open(filename, 'wb')   # локальный файл для сохранения данных
localfile.write(remotefile.read())
localfile.close()
remotefile.close()
```

Обратите внимание, что здесь выходной файл снова открывается в двоичном режиме – данные, получаемые модулем urllib, возвращаются в виде строк байтов, даже веб-страницы HTTP. Не ломайте голову над устройством строки URL, использованной здесь, – она, безусловно, сложна, но мы подробно будем рассматривать структуру адресов URL в целом в главе 15. Мы также вновь будем обращаться к urllib в этой и последующих главах для получения веб-страниц, форматирования сгенерированных строк URL и получения вывода удаленных сценариев в Сети.

С технической точки зрения urllib.request поддерживает целый ряд протоколов Интернета (HTTP, FTP и локальные файлы). В отличие от ftplib, модуль urllib.request используется в целом для чтения удаленных объектов, но не для записи или выгрузки их на сервер (хотя прото-

колы HTTP и FTP поддерживают такую возможность). Как и при использовании ftplib, получение данных обычно должно осуществляться в отдельных потоках выполнения, если блокировка составляет предмет для беспокойства. Однако базовый интерфейс, показанный в этом сценарии, прост. Вызов:

```
remotefile = urllib.request.urlopen(remoteaddr) # возвращает объект
                                                # типа файла для ввода
```

соединяется с сервером, указанным в строке URL remoteaddr, и возвращает объект типа файла, подключенный к потоку загрузки (здесь – сокет FTP). Вызов метода read извлекает содержимое файла, которое записывается в локальный файл на стороне клиента. Еще более простой интерфейс:

```
urllib.request.urlretrieve(remoteaddr, filename)
```

также открывает локальный файл и записывает в него загружаемые байты, что в данном сценарии выполняется вручную. Такой интерфейс удобен, если нужно загрузить файл, но менее полезен, если требуется сразу же обрабатывать его данные.

В любом случае конечный результат один и тот же: требуемый файл, находящийся на сервере, оказывается на компьютере клиента. Вывод этого сценария такой же, как в первоначальной версии, но на этот раз мы не пытаемся автоматически открыть загруженный файл (я изменил пароль в адресе URL, чтобы не искушать судьбу):

```
C:\...\PP4E\Internet\Ftp> getone-urllib.py
Pswd?
Downloading ftp://lutz:xxxxxx@ftp.rmi.net/monkeys.jpg;type=i

C:\...\PP4E\Internet\Ftp> fc monkeys.jpg test\monkeys.jpg
FC: no differences encountered

C:\...\PP4E\Internet\Ftp> start monkeys.jpg
```

Дополнительные примеры загрузки файлов с использованием модуля urllib вы найдете в разделе с описанием протокола HTTP, далее в этой главе, а примеры серверных сценариев – в главе 15. Как будет показано в главе 15, такие инструменты, как функция urlopen из модуля urllib.request, позволяют сценариям загружать удаленные файлы и вызывать программы, находящиеся на удаленных серверах, благодаря чему они могут служить удобными инструментами тестирования и использования веб-сайтов. В главе 15 мы также увидим, что модуль urllib.parse включает инструменты форматирования (экранирования) строк URL для обеспечения безопасной передачи.

Утилиты FTP get и put

Почти всегда, когда я рассказываю об интерфейсах ftplib на занятиях по Python, учащиеся интересуются, для чего программист должен ука-

зывать строку RETR в методе загрузки. Это хороший вопрос: строка RETR является именем команды загрузки в протоколе FTP, но, как уже говорилось, модуль ftplib призван инкапсулировать этот протокол. Как мы увидим чуть ниже, при выгрузке на сервер также требуется указывать странную строку STOR. Это шаблонный программный код, который для начала приходится принимать на веру, но который напрашивается на этот вопрос. Вы, конечно, можете предложить свою заплатку для ftplib, но это не самый хороший совет начинающим изучать Python, а кроме того, такая заплата может нарушить работоспособность существующих сценариев (имеется причина, по которой интерфейс должен иметь такой вид).

Лучше будет ответить так: Python упрощает возможность расширения стандартных библиотечных модулей собственными интерфейсами более высокого уровня – с помощью всего лишь нескольких строк многократно используемого программного кода можно заставить интерфейс FTP в Python выглядеть так, как вы захотите. Например, можно взять и написать вспомогательные модули, обертывающие интерфейсы ftplib и скрывающие строку RETR. Если поместить эти модули в каталог, включенный в переменную окружения PYTHONPATH, они станут столь же доступными, как сам модуль ftplib, и будут автоматически использоваться в любом сценарии Python, который может быть написан в будущем. Помимо устранения необходимости в строке RETR модуль-оболочка может использовать допущения, которые упрощают операции FTP до единственного вызова функции.

Например, при наличии модуля, который инкапсулирует и упрощает ftplib, наш сценарий для загрузки и запуска файлов можно было бы сократить еще больше, что иллюстрирует сценарий в примере 11.3, в сущности состоящий из двух вызовов функций и ввода пароля, но дающий тот же результат, что и сценарий в примере 13.1.

Пример 13.3. PP4E\Internet\Ftp\getone-modular.py

```
#!/usr/local/bin/python
"""
Сценарий на языке Python для загрузки медиафайла по FTP и его проигрывания.
Использует getfile.py, вспомогательный модуль, инкапсулирующий
этап загрузки по FTP.
"""

import getfile
from getpass import getpass
filename = 'monkeys.jpg'

# получить файл с помощью вспомогательного модуля
getfile.getfile(file=filename,
                site='ftp.rmi.net',
                dir ='.',
                user=('lutz', getpass('Pswd?')),
                refetch=True)
```

```
# остальная часть сценария осталась без изменений
if input('Open file?') in ['Y', 'y']:
    from PP4E.System.Media.playfile import playfile
    playfile(filename)
```

Помимо того что в этом варианте существенно уменьшилось количество строк, основное тело этого сценария разбито на отдельные файлы, которые можно повторно использовать в других ситуациях. Если когда-либо вновь потребуется загрузить файл, достаточно импортировать существующую функцию, а не заниматься редактированием путем копирования и вставки. Операцию загрузки потребуется изменить только в одном файле, а не во всех местах, куда был скопирован шаблонный программный код; можно даже сделать так, чтобы функция `getfile.getfile` использовала `urllib` вместо `ftplib`, никак не затронув при этом его клиентов. Это хорошая конструкция.

Утилита загрузки

И как же можно было бы написать такую обертку интерфейса FTP (задаст читатель риторический вопрос)? При наличии библиотечного модуля `ftplib` создать обертку для загрузки конкретного файла из конкретного каталога достаточно просто. Объекты соединений FTP поддерживают два метода загрузки:

`retrbinary`

Этот метод загружает запрашиваемый файл в двоичном режиме, блоками посылая его байты указанной функции, без преобразования символов конца строки. Обычно в качестве функции указывается метод `write` объекта открытого локального файла, благодаря которому байты помещаются в локальный файл на стороне клиента.

`retrlines`

Этот метод загружает запрашиваемый файл в режиме текста ASCII, посылая заданной функции строки текста с удаленными символами конца строки. Обычно указанная функция добавляет символ новой строки `\n` (преобразуемый в зависимости от платформы клиента) и записывает строку в локальный файл.

Позднее мы встретимся с примером использования метода `retrlines` – вспомогательный модуль `getfile` в примере 13.4 всегда осуществляет передачу в двоичном режиме с помощью метода `retrbinary`. Это означает, что файлы загружаются в точности в том виде, в каком они находятся на сервере, байт в байт, сохраняя для текстовых файлов те символы конца строки, которые приняты на сервере (если эти символы выглядят необычно в вашем текстовом редакторе, может потребоваться преобразовать их после загрузки – указания смотрите в справке к своему текстовому редактору или к командной оболочке или напишите сценарий Python, который открывал бы и записывал текст так, как необходимо).

Пример 13.4. PP4E\Internet\Ftp\getfile.py

```
#!/usr/local/bin/python
"""
Загружает произвольный файл по FTP. Используется анонимный доступ к FTP,
если не указан кортеж user=(имя, пароль). В разделе самопроверки
используются тестовый FTP-сайт и файл.
"""

from ftplib import FTP      # инструменты FTP на основе сокетов
from os.path import exists  # проверка наличия файла

def getfile(file, site, dir, user=(), *, verbose=True, refetch=False):
    """
    загружает файл по ftp с сайта/каталога, используя анонимный доступ
    или действительную учетную запись, двоичный режим передачи
    """
    if exists(file) and not refetch:
        if verbose: print(file, 'already fetched')
    else:
        if verbose: print('Downloading', file)
        local = open(file, 'wb')  # локальный файл с тем же именем
        try:
            remote = FTP(site)    # соединиться с FTP-сайтом
            remote.login(*user)   # для анонимного =() или (имя, пароль)
            remote.cwd(dir)
            remote.retrbinary('RETR ' + file, local.write, 1024)
            remote.quit()
        finally:
            local.close()                     # закрыть файл в любом случае
        if verbose: print('Download done.')  # исключения обрабатывает
                                              # вызывающая программа

if __name__ == '__main__':
    from getpass import getpass
    file = 'monkeys.jpg'
    dir  = '.'
    site = 'ftp.rmi.net'
    user = ('lutz', getpass('Pswd?'))
    getfile(file, site, dir, user)
```

Этот модуль, по сути, просто придает иную форму программному коду FTP, использовавшемуся выше для получения файла изображения, с целью сделать его более простым и многократно используемым. Так как экспортируемая здесь функция `getfile.getfile` является вызываемой, она стремится быть максимально надежной и широко используемой, но даже такая маленькая функция требует некоторых конструктивных решений. Вот несколько замечаний по использованию:

Режим FTP

Функция `getfile` в этом сценарии по умолчанию использует анонимный режим доступа по FTP, однако имеется возможность передать

в аргументе user кортеж из двух элементов с именем пользователя и паролем, чтобы зарегистрироваться на удаленном сервере в неанонимном режиме. Для работы по FTP в анонимном режиме не передавайте этот аргумент или передайте в нем пустой кортеж (). Метод login объекта FTP принимает два необязательных аргумента, обозначающие имя пользователя и пароль, а синтаксис вызова function(*args), используемый в примере 13.4, отправляет ему тот кортеж, который был передан в аргументе user, в виде отдельных аргументов.

Режимы обработки

Последние два аргумента (verbose, refetch) позволяют отключить сообщения о состоянии, выводимые в поток stdout (возможно, нежелательные в контексте графического интерфейса), и принудительно выполнить загрузку, даже если локальный файл уже существует (загрузка перезаписывает существующий файл).

Эти два аргумента оформлены как аргументы со значениями по умолчанию, которые могут передаваться *только как именованные аргументы* в Python 3.X, поэтому при использовании они должны передаваться по имени, а не по позиции. Аргумент user, напротив, может передаваться любым способом, если он вообще передается. Передача этих аргументов только в виде именованных предотвращает ошибочное сопоставление значения verbose или refetch с аргументом user, если он отсутствует в вызове функции.

Протокол обработки исключений

Предполагается, что исключения будут обрабатываться вызывающей программой. Данная функция заключает загрузку в оператор try/finally, чтобы гарантировать закрытие локального выходного файла, но разрешает дальнейшее распространение исключения. Например, при использовании в графическом интерфейсе или при вызове в отдельном потоке выполнения исключения могут потребовать особой обработки, о которой этот модуль ничего не знает.

Самотестирование

Когда этот модуль запускается как самостоятельный сценарий, он загружает с целью самопроверки файл изображения с моего веб-сайта (укажите здесь свой сервер и файл), но обычно этой функции передаются определенные имена файлов, сайтов и каталогов для FTP.

Режим открытия файла

Как и в предыдущих примерах, этот сценарий открывает локальный выходной файл в двоичном режиме wb, чтобы подавить преобразование символов конца строки и обеспечить соответствие модели строк Юникода в Python 3.X. Как мы узнали в главе 4, файлы с действительно двоичными данными могут содержать байты со значением \n, соответствующим символу конца строки. Открытие их в текстовом режиме w приведет к автоматическому преобразованию этих байтов в последовательность \r\n при записи в Windows локально.

Эта проблема наблюдается только в Windows – режим `w` изменяет символы конца не во всех системах.

Однако, как мы узнали в той же главе 4, двоичный режим необходим также для подавления автоматического кодирования символов Юникода, выполняемого в Python 3.X при работе с текстовыми файлами. Если бы мы использовали текстовый файл, Python попытался бы выполнить кодирование полученных данных при записи, используя кодировку по умолчанию или указанную явно, что может приводить к ошибкам при работе с некоторыми текстовыми данными, и обычно приводит к ошибкам при работе с действительно двоичными данными, такими как изображения и аудиоданные.

Поскольку метод `retrbinary` в версии 3.X будет пытаться записывать строки `bytes`, мы в действительности просто не сможем открыть выходной файл в текстовом режиме. В противном случае метод `write` будет возбуждать исключение. Напомню, что текстовые файлы в Python 3.X требуют при записи передавать строки типа `str`, а двоичные файлы ожидают получить строки `bytes`. Поскольку метод `retrbinary` записывает строки `bytes`, а метод `retrlines` – строки `str`, они неявно требуют открывать выходной файл в двоичном или в текстовом режиме соответственно. Данное ограничение действует независимо от проблемы преобразования символов конца строки и кодирования символов Юникода, но его удовлетворение фактически решает и эти проблемы.

Как мы увидим в последующих примерах, операции загрузки файлов в текстовом режиме накладывают дополнительные требования, касающиеся кодировки. В действительности, модуль `ftplib` может служить отличным примером влияния модели строк Юникода в Python 3.X на практическую реализацию. Постоянно используя двоичный режим в этом сценарии, мы полностью уходим от этой проблемы.

Модель каталогов

Данная функция использует одно и то же имя для идентификации удаленного файла и локального файла, в котором должно быть сохранено загруженное содержимое. Поэтому ее следует выполнять в том каталоге, где должен оказаться загруженный файл. При необходимости переместиться в нужный каталог используйте `os.chdir`. (Можно было бы сделать так, чтобы аргумент `file` представлял имя локального файла, и убирать из него локальный каталог с помощью `os.path.split` или принимать два аргумента с именами файлов – локального и удаленного.)

Обратите также внимание, что несмотря на свое название, этот модуль значительно отличается от сценария *getfile.py*, рассматривавшегося в конце материала по сокетам в предыдущей главе. Основанный на сокетах модуль `getfile` реализовывал логику клиента и сервера для за-

грузки файла с сервера на компьютер клиента непосредственно через сокеты.

Этот новый модуль getfile является исключительно инструментом клиента. Для запроса файла с сервера вместо непосредственного использования сокетов в нем применяется стандартный протокол FTP. Все детали работы с сокетами скрыты в реализации протокола FTP для клиента внутри модуля ftplib. Кроме того, сервер здесь является программой, постоянно выполняемой на компьютере сервера, которая ждет запросов FTP на сокете и отвечает на них, используя выделенный порт FTP (номер 21). Таким образом, этому сценарию требуется, чтобы на компьютере, где находится нужный файл, работал сервер FTP, и весьма вероятно, такой сервер там есть.

Утилита выгрузки

Если уж мы ввязались в это дело, напишем сценарий для выгрузки (upload) по FTP одиночного файла на удаленный компьютер. Интерфейсы выгрузки на сервер в модуле реализации протокола FTP симметричны интерфейсам загрузки с сервера. Если есть подключенный объект FTP:

- Посредством его метода storbinary можно выгружать на сервер байты из открытого объекта локального файла.
- Посредством его метода storlines можно выгружать на сервер текст в режиме ASCII из открытого объекта локального файла.

В отличие от интерфейсов загрузки с сервера, обоим этим методам передается объект файла целиком, а не метод этого объекта (или другая функция). С методом storlines мы еще встретимся в более позднем примере. Вспомогательный модуль, представленный в примере 13.5, использует метод storbinary, таким образом, файл, имя которого передается методу, всегда передается дословно – в двоичном режиме, без кодирования символов Юникода или преобразования символов конца строки соответственно соглашениям, принятым на целевой платформе. Если этот сценарий выгрузит текстовый файл, он будет получен точно в том виде, в каком хранился на компьютере, откуда поступил, со всеми символами конца строки и в кодировке, используемой на стороне клиента.

Пример 13.5. PP4E\Internet\Ftp\putfile.py

```
#!/usr/local/bin/python
"""
Выгружает произвольный файл по FTP в двоичном режиме.
Использует анонимный доступ к ftp, если функции не был передан
кортеж user=(имя, пароль) аргументов.
"""

import ftplib                # инструменты FTP на основе сокетов
```

```
def putfile(file, site, dir, user=(), *, verbose=True):
    """
    выгружает произвольный файл по FTP на сайт/каталог, используя анонимный
    доступ или действительную учетную запись, двоичный режим передачи
    """
    if verbose: print('Uploading', file)
    local  = open(file, 'rb')  # локальный файл с тем же именем
    remote = ftplib.FTP(site)  # соединиться с FTP-сайтом
    remote.login(*user)        # анонимная или действительная учетная запись
    remote.cwd(dir)
    remote.storbinary('STOR ' + file, local, 1024)
    remote.quit()
    local.close()
    if verbose: print('Upload done.')

if __name__ == '__main__':
    site = 'ftp.rmi.net'
    dir  = '.'
    import sys, getpass
    pswd = getpass.getpass(site+' pswd?')     # имя файла в командной строке
    putfile(sys.argv[1], site, dir, user=('lutz', pswd))   # действительная
                                                           # учетная запись
```

Обратите внимание, что для переносимости локальный файл на этот раз открывается в двоичном режиме `rb`, чтобы предотвратить автоматическое преобразование символа конца строки. Если файл действительно является двоичным, было бы нежелательно, чтобы из него таинственным образом исчезли байты, значением которых окажется символ возврата каретки `\r`, когда он пересылается клиентом, выполняющимся в Windows. Нам также требуется подавить кодирование символов Юникода при передаче нетекстовых файлов и требуется читать из исходного файла строки `bytes`, которые ожидает получить метод `storbinary`, выполняющий операцию выгрузки (подробнее о режимах открытия входного файла рассказывается ниже).

Программный код самотестирования этого сценария выгружает файл, имя которого указано в командной строке, но обычно вы будете передавать ему строки действительных имен файла, сайта и каталога. Кроме того, как и в утилите загрузки файла, можно передать кортеж (*имя_пользователя*, *пароль*) в качестве аргумента `user` для работы в режиме неанонимного доступа (по умолчанию используется анонимный доступ).

Воспроизведение музыкальной темы Monty Python

Пришло время немного поразвлечься. Воспользуемся этими сценариями для передачи и воспроизведения звукового файла с музыкальной темой Monty Python, находящегося на моем веб-сайте. Прежде всего, напишем модуль, представленный в примере 13.6, который загружает и воспроизводит файл.

Пример 13.6. PP4E\Internet\Ftp\sousa.py

```
#!/usr/local/bin/python
"""
Порядок использования: sousa.py. Загружает и проигрывает музыкальную
тему Monty Python. В текущем виде может не работать в вашей системе:
он требует, чтобы компьютер был подключен к Интернету, имелась
учетная запись на сервере FTP, и использует аудиофильтры в Unix и плеер
файлов .au в Windows. Настройте этот файл и файл playfile.py, как требуется.
"""


from getpass import getpass
from PP4E.Internet.Ftp.getfile import getfile
from PP4E.System.Media.playfile import playfile

file = 'sousa.au'      # координаты по умолчанию файла
site = 'ftp.rmi.net'   # с музыкальной темой Monty Python
dir  = '.'
user = ('lutz', getpass('Pswd?'))

getfile(file, site, dir, user)     # загрузить аудиофайл по FTP
playfile(file)                     # передать его аудиоплееру

# import os
# os.system('getone.py sousa.au')  # эквивалент командной строки
```

В этом сценарии нет ничего нового, потому что он просто объединяет два инструмента, уже созданных нами. Мы повторно использовали здесь функцию `getfile` из примера 13.4 для загрузки файла и модуль `playfile` из главы 6 (пример 6.23) для проигрывания аудиофайла после его загрузки (вернитесь к тому примеру за дополнительными подробностями о том, как выполняется проигрывание). Обратите также внимание на две последние строки в этом файле – мы могли бы добиться того же эффекта, передав имя аудиофайла как аргумент командной строки нашему первоначальному сценарию, но этот путь менее прямой.

В текущем виде сценарий предполагает использование моей учетной записи на сервере FTP. Настройте сценарий на использование своей учетной записи (ранее этот файл можно было загрузить с анонимного FTP-сайта *ftp.python.org*, но он был закрыт из-за проблем с безопасностью между изданиями этой книги). После настройки этот сценарий будет работать на любом компьютере с Python, выходом в Интернет и зарегистрированным в системе аудиоплеером; он действует на моем ноутбуке с Windows и широкополосным соединением с Интернетом (если бы это было возможно, я вставил бы сюда гиперссылку на звуковой файл, чтобы показать, как он звучит):

```
C:\...\PP4E\Internet\Ftp> sousa.py
Pswd?
Downloading sousa.au
Download done.
```

```
C:\...\PP4E\Internet\Ftp> sousa.py
Pswd?
sousa.au already fetched
```

Модули getfile и putfile также могут использоваться для перемещения образца звукового файла. Оба они могут быть импортированы клиентами, желающими использовать их функции, или запущены как программы верхнего уровня, выполняющие самотестирование. Запустим их из командной строки и интерактивной оболочки и посмотрим, как они работают. При автономном выполнении в командной строке передаются параметры и используются настройки файла по умолчанию:

```
C:\...\PP4E\Internet\Ftp> putfile.py sousa.py
ftp.rmi.net pswd?
Uploading sousa.py
Upload done.
```

При импортировании параметры явно передаются функциям:

```
C:\...\PP4E\Internet\Ftp> python
>>> from getfile import getfile
>>> getfile(file='sousa.au',site='ftp.rmi.net', dir='.', user=('lutz',
                                                                 'XXX'))
sousa.au already fetched

C:\...\PP4E\Internet\Ftp> del sousa.au

C:\...\PP4E\Internet\Ftp> python
>>> from getfile import getfile
>>> getfile(file='sousa.au',site='ftp.rmi.net', dir='.', user=('lutz',
                                                                 'XXX'))
Downloading sousa.au
Download done.
>>> from PP4E.System.Media.playfile import playfile
>>> playfile('sousa.au')
```

Хотя модуль Python ftplib сам автоматизирует работу с сокетами и форматирование сообщений FTP, тем не менее, наши собственные инструменты, подобные этим, могут упростить процесс еще больше.

Добавляем пользовательский интерфейс

Если вы читали предыдущую главу, то должны помнить, что она завершалась кратким обзором сценариев, добавляющих интерфейс пользователя к основанному на сокетах сценарию getfile – он передавал файлы через сокеты, используя специфические правила, а не FTP. В конце обзора было отмечено, что FTP предоставляет гораздо более универсальный способ перемещения файлов, потому что серверы FTP широко распространены в Сети. В иллюстративных целях в примере 13.7 приводится видоизмененная версия пользовательского интерфейса из предыдущей главы, реализованная как новый подкласс универсально-

го построителя форм из предыдущей главы, представленного в примере 12.20.

Пример 13.7. PP4E\Internet\Ftp\getfilegui.py

```
"""
###############################################################################
вызывает функцию FTP getfile из многократно используемого класса формы
графического интерфейса; использует os.chdir для перехода в целевой
локальный каталог (getfile в настоящее время предполагает,
что в имени файла отсутствует префикс пути к локальному каталогу);
вызывает getfile.getfile в отдельном потоке выполнения, что позволяет
выполнять несколько запросов одновременно и избежать блокировки
графического интерфейса на время загрузки; отличается от основанного
на сокетах getfilegui, но повторно использует класс Form построителя
графического интерфейса; в данном виде поддерживает как анонимный доступ
к FTP, так и с указанием имени пользователя;

предостережение: содержимое поля ввода пароля здесь не скрывается
за звездочками, ошибки выводятся в консоль, а не в графический интерфейс
(потоки выполнения не могут обращаться к графическому интерфейсу в Windows),
поддержка многопоточной модели выполнения реализована не на все 100%
(существует небольшая задержка между os.chdir и открытием локального
выходного файла в getfile) и можно было бы выводить диалог "сохранить как",
для выбора локального каталога, и диалог с содержимым удаленного каталога,
для выбора загружаемого файла; читателям предлагается самостоятельно
добавить эти улучшения;
###############################################################################
"""

from tkinter import Tk, mainloop
from tkinter.messagebox import showinfo
import getfile, os, sys, _thread              # здесь FTP-версия getfile
from PP4E.Internet.Sockets.form import Form   # использовать инструмент форм

class FtpForm(Form):
    def __init__(self):
        root = Tk()
        root.title(self.title)
        labels = ['Server Name', 'Remote Dir', 'File Name',
                  'Local Dir', 'User Name?', 'Password?']
        Form.__init__(self, labels, root)
        self.mutex = _thread.allocate_lock()
        self.threads = 0

    def transfer(self, filename, servername, remotedir, userinfo):
        try:
            self.do_transfer(filename, servername, remotedir, userinfo)
            print('%s of "%s" successful' % (self.mode, filename))
        except:
            print('%s of "%s" has failed:' % (self.mode, filename), end=' ')
            print(sys.exc_info()[0], sys.exc_info()[1])
```

```
        self.mutex.acquire()
        self.threads -= 1
        self.mutex.release()

    def onSubmit(self):
        Form.onSubmit(self)
        localdir   = self.content['Local Dir'].get()
        remotedir  = self.content['Remote Dir'].get()
        servername = self.content['Server Name'].get()
        filename   = self.content['File Name'].get()
        username   = self.content['User Name?'].get()
        password   = self.content['Password?'].get()
        userinfo   = ()
        if username and password:
            userinfo = (username, password)
        if localdir:
            os.chdir(localdir)
        self.mutex.acquire()
        self.threads += 1
        self.mutex.release()
        ftpargs = (filename, servername, remotedir, userinfo)
        _thread.start_new_thread(self.transfer, ftpargs)
        showinfo(self.title, '%s of "%s" started' % (self.mode, filename))

    def onCancel(self):
        if self.threads == 0:
            Tk().quit()
        else:
            showinfo(self.title,
                     'Cannot exit: %d threads running' % self.threads)

class FtpGetfileForm(FtpForm):
    title = 'FtpGetfileGui'
    mode = 'Download'
    def do_transfer(self, filename, servername, remotedir, userinfo):
        getfile.getfile(filename, servername, remotedir,
                        userinfo, verbose=False, refetch=True)

if __name__ == '__main__':
    FtpGetfileForm()
    mainloop()
```

Если вернуться в конец предыдущей главы, можно обнаружить, что эта версия по структуре аналогична приведенной там. В действительности они даже названы одинаково (и отличаются только тем, что находятся в разных каталогах). Однако в данном примере класс умеет пользоваться модулем `getfile`, использующим протокол FTP, приведенным в начале данной главы, а не основанным на сокетах модулем `getfile`, с которым мы познакомились в предыдущей главе. Кроме того, эта версия создает большее количество полей ввода, как видно на рис. 13.2.

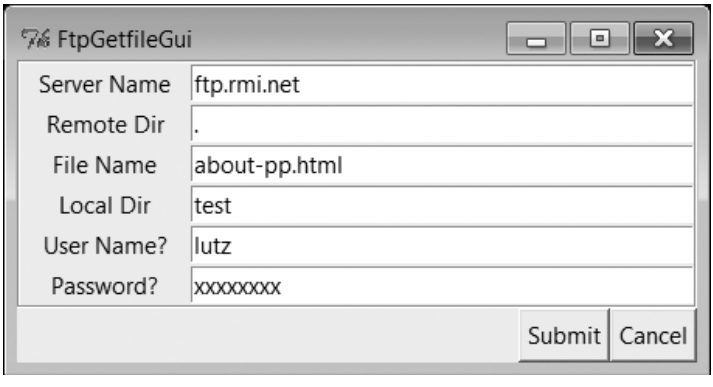


Рис. 13.2. *Форма ввода для FTP-версии модуля getfile*

Обратите внимание, что здесь в поле ввода имени файла можно ввести абсолютный путь к файлу. Если его не указать, сценарий будет искать файл в текущем рабочем каталоге, который изменяется после каждой загрузки с сервера и может зависеть от того, где запускается графический интерфейс (то есть текущий каталог будет иным, когда сценарий запускается из программы PyDemos, находящейся в корневом каталоге дерева примеров). При щелчке на кнопке Submit в этом графическом интерфейсе (или нажатии клавиши Enter) сценарий просто передает значения полей ввода формы как аргументы функции FTP `getfile.getfile`, показанной выше в этом разделе в примере 13.4. Кроме того, он выводит окно, сообщающее о начале загрузки (рис. 13.3).

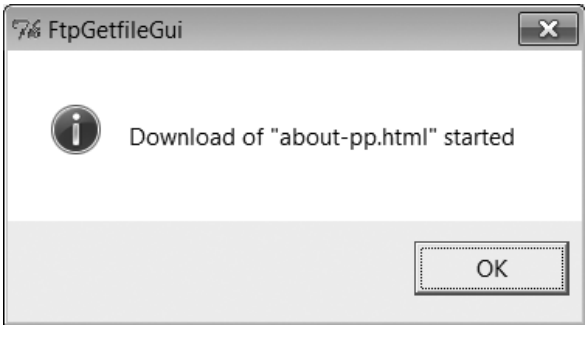


Рис. 13.3. *Информационное окно для FTP-версии модуля getfile*

Все сообщения о состоянии загрузки, включая сообщения об ошибках FTP, данная версия выводит в окно консоли. Ниже показаны сообщения, которые выводятся в случае успешной загрузки двух файлов и одной неудачной попытки (для удобочитаемости я добавил пустые строки):

```
C:\...\PP4E\Internet\Ftp> getfilegui.py
Server Name     =>      ftp.rmi.net
User Name?      =>      lutz
Local Dir       =>      test
File Name       =>      about-pp.html
Password?       =>      xxxxxxxx
Remote Dir      =>      .
Download of "about-pp.html" successful

Server Name     =>      ftp.rmi.net
User Name?      =>      lutz
Local Dir       =>      C:\temp
File Name       =>      ora-lp4e-big.jpg
Password?       =>      xxxxxxxx
Remote Dir      =>      .
Download of "ora-lp4e-big.jpg" successful

Server Name     =>      ftp.rmi.net
User Name?      =>      lutz
Local Dir       =>      C:\temp
File Name       =>      ora-lp4e.jpg
Password?       =>      xxxxxxxx
Remote Dir      =>      .
Download of "ora-lp4e.jpg" has failed: <class 'ftplib.error_perm'>
550 ora-lp4e.jpg: No such file or directory
```

При наличии имени пользователя и пароля загрузчик производит регистрацию под определенной учетной записью. Для анонимного доступа к FTP нужно оставить поля имени пользователя и пароля пустыми.

Теперь, чтобы проиллюстрировать поддержку многопоточной модели выполнения в данном графическом интерфейсе, запустим загрузку большого файла, а затем, пока продолжается загрузка этого файла, попробуем запустить загрузку другого файла. Графический интерфейс остается активным во время загрузки, поэтому просто изменим значения полей ввода и еще раз нажмем кнопку Submit.

Загрузка второго файла начнется и будет выполняться параллельно той, что была запущена первой, так как каждая загрузка выполняется в отдельном потоке выполнения и одновременно могут быть активными несколько соединений с Интернетом. Сам графический интерфейс остается активным во время загрузки только потому, что загрузка происходит в отдельном потоке выполнения – если бы это было не так, даже перерисовка экрана не происходила бы до завершения загрузки.

Мы уже обсуждали потоки выполнения в главе 5, а особенности их использования в графических интерфейсах – в главах 9 и 10, но данный сценарий иллюстрирует некоторые практические аспекты использования потоков выполнения:

- Эта программа избегает производить какие-либо действия с графическим интерфейсом в потоках, выполняющих загрузку. Как мы уже

знаем, работать с графическим интерфейсом может только тот поток выполнения, который его создает.

- Чтобы избежать остановки порожденных потоков загрузки на некоторых платформах, графический интерфейс не должен завершаться, пока продолжается хотя бы одна загрузка. Он следит за количеством потоков, в которых производится загрузка, и выводит окно, если попытаться закрыть графический интерфейс нажатием кнопки Cancel во время загрузки.

В главе 10 мы познакомились со способами, позволяющими обойти правило, запрещающее трогать графический интерфейс из потоков выполнения, и мы будем применять их в примере PyMailGui, в следующей главе. Для обеспечения переносимости действительно нельзя закрывать графический интерфейс, пока счетчик активных потоков не уменьшится до нуля. Для достижения того же эффекта можно было бы использовать модель выхода из модуля threading, представленного в главе 5. Ниже показан вывод, который появляется в окне консоли, когда одновременно выполняется загрузка двух файлов:

```
C:\...\PP4E\Internet\Ftp> python getfilegui.py
Server Name     =>      ftp.rmi.net
User Name?      =>      lutz
Local Dir       =>      C:\temp
File Name       =>      spain08.JPG
Password?       =>      xxxxxxxx
Remote Dir      =>      .

Server Name     =>      ftp.rmi.net
User Name?      =>      lutz
Local Dir       =>      C:\temp
File Name       =>      index.html
Password?       =>      xxxxxxxx
Remote Dir      =>      .

Download of "index.html" successful
Download of "spain08.JPG" successful
```

Конечно, этот сценарий ненамного полезнее, чем инструмент командной строки, но его можно легко модифицировать, изменив программный код на языке Python, и его вполне достаточно, чтобы считать его простым наброском интерфейса пользователя FTP. Кроме того, поскольку этот графический интерфейс выполняет загрузку в отдельных потоках выполнения, из него можно одновременно загружать несколько файлов, не запуская другие программы клиентов FTP.

Пока мы озабочены графическим интерфейсом, добавим также простой интерфейс и к утилите putfile. Сценарий в примере 13.8 создает диалог для запуска выгрузки файлов на сервер в отдельных потоках выполнения и использует базовую логику работы с протоколом FTP, импортированную из примера 13.5. Он почти не отличается от только что написан-

ного графического интерфейса для getfile, поэтому ничего нового о нем сказать нельзя. В действительности, поскольку операции приема и отправки столь схожи с точки зрения интерфейса, значительная часть логики формы для получения была умышленно выделена в один родовой класс (FtpForm), благодаря чему изменения требуется производить только в одном месте. Таким образом, графический интерфейс к сценарию отправки по большей части повторно использует графический интерфейс к сценарию получения, с измененными метками и методом передачи. Он находится в отдельном файле, что облегчает его запуск как самостоятельной программы.

Пример 13.8. PP4E\Internet\Ftp\putfilegui.py

```
"""
###############################################################################
запускает функцию FTP putfile из многократно используемого класса формы
графического интерфейса; см. примечания в getfilegui: справедливыми
остаются большинство тех же предупреждений; формы для получения
и отправки выделены в единый класс, чтобы производить изменения
лишь в одном месте;
###############################################################################
"""


from tkinter import mainloop
import putfile, getfilegui

class FtpPutfileForm(getfilegui.FtpForm):
    title = 'FtpPutfileGui'
    mode = 'Upload'
    def do_transfer(self, filename, servername, remotedir, userinfo):
        putfile.putfile(filename, servername, remotedir,
                        userinfo, verbose=False)

if __name__ == '__main__':
    FtpPutfileForm()
    mainloop()
```

Графический интерфейс этого сценария весьма похож на графический интерфейс сценария загрузки файлов, поскольку действует практически тот же программный код. Отправим несколько файлов с клиентского компьютера на сервер. На рис. 13.4 показано состояние графического интерфейса во время отправки одного из файлов.

А ниже приводится вывод в окне консоли при последовательной отправке двух файлов. Здесь выгрузка файлов также выполняется в отдельных потоках, поэтому, если попытаться запустить выгрузку нового файле прежде, чем закончится выгрузка текущего, они перекроются по времени:

```
C:\...\PP4E\Internet\Ftp\test> ..\putfilegui.py
Server Name     =>      ftp.rmi.net
User Name?      =>      lutz
```

```
Local Dir       =>      .
File Name       =>      sousa.au
Password?       =>      xxxxxxxx
Remote Dir      =>      .
Upload of "sousa.au" successful

Server Name     =>      ftp.rmi.net
User Name?      =>      lutz
Local Dir       =>      .
File Name       =>      about-pp.html
Password?       =>      xxxxxxxx
Remote Dir      =>      .
Upload of "about-pp.html" successful
```

Рис. 13.4. *Форма ввода для FTP-версии модуля putfile*

Наконец, можно увязать оба графических интерфейса в единый запускающий сценарий, который умеет запускать интерфейсы получения и отправки независимо от того, в каком каталоге мы находимся при запуске сценария и от платформы, на которой он выполняется. Этот процесс приводится в примере 13.9.

Пример 13.9. PP4E\Internet\Ftp\PyFtpGui.pyw

```
"""
###############################################################################
запускает графические интерфейсы получения и отправки по ftp независимо
от каталога, в котором находится сценарий; сценарий не обязательно
должен находиться в os.getcwd; можно также жестко определить путь
в $PP4EHOME или guessLocation; можно было бы также так: [from PP4E.
launchmodes import PortableLauncher, PortableLauncher('getfilegui', '%s/
getfilegui.py' % mydir)()], но в Windows понадобилось бы всплывающее
окно DOS для вывода сообщений о состоянии, описывающих выполняемые операции;
###############################################################################
"""
```

```
import os, sys
print('Running in: ', os.getcwd())

# PP3E
# from PP4E.Launcher import findFirst
# mydir = os.path.split(findFirst(os.curdir, 'PyFtpGui.pyw'))[0]

# PP4E
from PP4E.Tools.find import findlist
mydir = os.path.dirname(findlist('PyFtpGui.pyw', startdir=os.curdir)[0])

if sys.platform[:3] == 'win':
    os.system('start %s\getfilegui.py' % mydir)
    os.system('start %s\putfilegui.py' % mydir)
else:
    os.system('python %s/getfilegui.py &' % mydir)
    os.system('python %s/putfilegui.py &' % mydir)
```

Обратите внимание, что здесь мы повторно используем утилиту `find` из примера 6.13 в главе 6 – на сей раз чтобы определить путь к домашнему каталогу сценария, необходимый для конструирования командных строк. При использовании программ запуска в корневом каталоге дерева примеров или при вызове сценария из командной строки в любом другом каталоге текущий рабочий каталог может не совпадать с каталогом, где находится сценарий. В предыдущем издании вместо поиска своего собственного каталога этот сценарий использовал инструмент из модуля `Launcher` (аналогичные решения вы найдете в пакете с примерами).

Если запустить этот сценарий, на экране появятся оба графических интерфейса – для получения и отправки – как отдельные независимо выполняемые программы. Альтернативой может быть прикрепление обеих форм к единому интерфейсу. Конечно, можно создать значительно более замысловатые интерфейсы. Например, можно воспользоваться всплывающими диалогами для выбора локальных файлов и отобразить виджеты, сообщающие о ходе выполнения текущей операции загрузки или выгрузки. Можно даже отобразить доступные на удаленном сервере файлы в окне списка, запросив содержимое удаленного каталога через соединение FTP. Однако, чтобы научиться добавлять такие функции, нужно перейти к следующему разделу.

Передача каталогов с помощью ftplib

В былые времена для управления своим веб-сайтом, находящимся на сервере провайдера Интернета (ISP), я использовал Telnet. Я регистрировался на веб-сервере в окне командной оболочки и редактировал свои файлы непосредственно на удаленном компьютере. Все файлы моего сайта существовали в единственном экземпляре и находились на сервере провайдера. Кроме того, изменение содержимого сайта можно было

выполнять с любого компьютера, где имелся клиент Telnet, – идеальное решение для тех, чья работа связана с частыми поездками.[1]

Но времена изменились. Подобно большинству персональных веб-сайтов в настоящее время все мои файлы хранятся и редактируются на моем ноутбуке, и я передаю их на сервер провайдера и обратно по мере необходимости. Часто требуется передать всего один-два файла и для этого можно воспользоваться клиентом FTP командной строки. Однако иногда мне требуется простой способ передачи всех файлов сайта. Загрузка каталога может потребоваться, например, чтобы определить, какие файлы изменились. Иногда изменения столь обширны, что проще выгрузить весь сайт за один прием.

Существуют различные решения этой задачи (включая возможности инструментов конструирования сайтов), однако Python тоже может оказать помощь: написав сценарии Python, автоматизирующие задачи выгрузки и загрузки файлов для обеспечения возможности сопровождения сайта на моем ноутбуке, я получил переносимое и мобильное решение. Так как сценарии Python, использующие протокол FTP, могут работать на любом компьютере, поддерживающем сокеты, они могут выполняться на моем ноутбуке и практически на любом другом компьютере, где установлен Python. Более того, те же сценарии, с помощью которых файлы страниц перемещаются на мой компьютер и обратно, можно использовать для копирования моего сайта на другой веб-сервер с целью создания резервной копии на случай, если возникнет остановка в работе моего провайдера. Этот прием иногда называется созданием *зеркала* – копии удаленного сайта.

Загрузка каталогов сайта

Следующие два сценария призваны удовлетворить эту потребность. Первый из них, *downloadflat.py*, автоматически загружает (то есть копирует) по FTP все файлы из каталога удаленного сайта в каталог на локальном компьютере. В настоящее время я храню основные копии файлов моего сайта на своем ПК, но в действительности использую этот сценарий двумя способами:

---

[1] На самом деле это не совсем так. Во втором издании книги в этом месте приводился горестный рассказ о том, как мой провайдер вынудил своих пользователей отучиться от доступа по Telnet. В настоящее время это уже не кажется такой большой проблемой. Обычная практика в Интернете, получившая широкое распространение в короткие сроки. Один из моих сайтов даже вырос настолько, что его стало слишком сложно редактировать вручную (конечно, кроме случаев, когда это обусловлено необходимостью обхода ошибок в инструменте конструирования сайтов). Только вдумайтесь, что значит присутствие Python в Веб. А ведь когда я впервые столкнулся с Python в 1992 году, это был набор кодированных сообщений электронной почты, которые пользователи декодировали, объединяли и уповали, что полученный результат будет работать. «Да, да, понимаю – давай, дед, расскажи еще...»

- Чтобы загрузить мой веб-сайт на клиентский компьютер, где я хочу заняться редактированием, я загружаю содержимое веб-каталога своей учетной записи на сервере моего интернет-провайдера.
- Чтобы сделать зеркальную копию моего сайта на другом сервере, я периодически выполняю этот сценарий на целевом компьютере, если он поддерживает Telnet или безопасную оболочку SSH – в противном случае я просто загружаю файлы веб-сервера на один компьютер и выгружаю их оттуда на требуемый сервер.

Вообще говоря, этот сценарий (представленный в примере 13.10) загрузит полный каталог файлов на любой компьютер с Python и сокетами с любого компьютера, на котором работает сервер FTP.

Пример 13.10. PP4E\Internet\Ftp\Mirror\downloadflat.py

```
#!/bin/env python
"""
###############################################################################
использует протокол FTP для копирования (загрузки) всех файлов
из единственного каталога на удаленном сайте в каталог на локальном
компьютере; запускайте этот сценарий периодически для создания зеркала
плоского каталога FTP-сайта, находящегося на сервере вашего провайдера;
для анонимного доступа установите переменную remoteuser в значение
'anonymous'; чтобы пропускать ошибки загрузки файлов, можно было бы
использовать инструкцию try, но в случае таких ошибок FTP-соединение скорее
всего все равно будет закрыто автоматически; можно было бы перед передачей
каждого нового файла переустанавливать соединение, создавая новый экземпляр
класса FTP: сейчас устанавливается всего одно соединение; в случае неудачи
попробуйте записать в переменную nonpassive значение True, чтобы
использовать активный режим FTP, или отключите брандмауэр; кроме того,
работоспособность этого сценария зависит от настроек сервера FTP
и возможных ограничений на загрузку.
###############################################################################
"""


import os, sys, ftplib
from getpass import getpass
from mimetypes import guess_type

nonpassive = False          # в 2.1+ по умолчанию пассивный режим FTP
remotesite = 'home.rmi.net' # загрузить с этого сайта
remotedir  = '.'            # и из этого каталога (например, public_html)
remoteuser = 'lutz'
remotepass = getpass('Password for %s on %s: ' % (remoteuser, remotesite))
localdir   = (len(sys.argv) > 1 and sys.argv[1]) or '.'
cleanall   = input('Clean local directory first? ')[:1] in ['y', 'Y']

print('connecting...')
connection = ftplib.FTP(remotesite)       # соединиться с FTP-сайтом
connection.login(remoteuser, remotepass)  # зарегистрироваться
                                          # с именем/паролем
```

```
connection.cwd(remotedir)                  # перейти в копируемый каталог
if nonpassive:                             # принудительный переход
                                           # в активный режим FTP
    connection.set_pasv(False)             # большинство серверов работают
                                           # в пассивном режиме

if cleanall:                               # сначала удалить все локальные
    for localname in os.listdir(localdir):     # файлы, чтобы избавиться от
        try:                                   # устаревших копий os.listdir
            print('deleting local', localname) # пропускает . и ..
            os.remove(os.path.join(localdir, localname))
        except:
            print('cannot delete local', localname)

count = 0                          # загрузить все файлы из удаленного каталога
remotefiles = connection.nlst()    # nlst() возвращает список файлов
                                   # dir() возвращает полный список
for remotename in remotefiles:
    if remotename in ('.', '..'): continue      # некоторые серверы
                                                # включают . и ..
    mimetype, encoding = guess_type(remotename) # например,
                                                # ('text/plain','gzip')
    mimetype = mimetype or '?/?'                # допускается (None, None)
    maintype = mimetype.split('/')[0]           # .jpg ('image/jpeg', None')

    localpath = os.path.join(localdir, remotename)
    print('downloading', remotename, 'to', localpath, end=' ')
    print('as', maintype, encoding or '')

    if maintype == 'text' and encoding == None:
        # использовать текстовый файл и режим передачи ascii
        # использовать кодировку, совместимую с ftplib
        localfile = open(localpath, 'w', encoding=connection.encoding)
        callback = lambda line: localfile.write(line + '\n')
        connection.retrlines('RETR ' + remotename, callback)

    else:
        # использовать двоичный файл и двоичный режим предачи
        localfile = open(localpath, 'wb')
        connection.retrbinary('RETR ' + remotename, localfile.write)

    localfile.close()
    count += 1

connection.quit()
print('Done:', count, 'files downloaded.')
```

В этом примере не так много нового для обсуждения в сравнении с примерами использования протокола FTP, которые мы видели выше. Здесь мы открываем соединение с удаленным сервером FTP, регистрируемся с необходимыми именем пользователя и паролем (в этом сценарии не

используется анонимный доступ к FTP) и переходим в требуемый удаленный каталог. Новыми здесь, однако, являются циклы обхода всех файлов в локальном и удаленном каталогах, загрузка в текстовом режиме и удаление файлов:

Удаление всех локальных файлов

У этого сценария есть параметр `cleanall`, значение которого запрашивается у пользователя. Если он включен, сценарий перед загрузкой удаляет все файлы из локального каталога, чтобы гарантировать отсутствие файлов, которых нет на сервере (они могут сохраниться от предыдущей загрузки). Чтобы удалить локальные файлы, сценарий вызывает `os.listdir` и получает список имен файлов в каталоге, затем удаляет каждый из них с помощью `os.remove`. Если вы забыли, как действуют эти функции, смотрите их описание в главе 4 (или руководство по библиотеке Python).

Обратите внимание, что для объединения пути к каталогу с именем файла в соответствии с соглашениями, принятыми на текущей платформе, используется функция `os.path.join`; `os.listdir` возвращает имена файлов без путей, а этот сценарий не обязательно выполняется в локальном каталоге, куда будут помещены загружаемые файлы. Локальным каталогом по умолчанию станет текущий каталог («.»), но его можно изменить с помощью аргумента командной строки, передаваемого сценарию.

Загрузка всех удаленных файлов

Чтобы вытащить из удаленного каталога все файлы, сначала нужно получить список их имен. Объект класса `FTP` обладает методом `nlst`, который является удаленным эквивалентом функции `os.listdir`: `nlst` возвращает список строк – имен всех файлов в текущем удаленном каталоге. Получив такой список, мы просто обходим его в цикле и выполняем FTP-команды получения файлов поочередно для каждого имени файла (подробнее об этом чуть ниже).

Метод `nlst` в некоторой мере похож на запрос списка содержимого каталога командой `ls` в обычных интерактивных программах FTP, но Python автоматически преобразует текст листинга в список имен файлов. Методу можно передать имя удаленного каталога, содержимое которого нужно получить. По умолчанию он возвращает список файлов для текущего каталога сервера. Родственный метод класса `FTP`, `dir`, возвращает список строк, порожденных командой FTP `LIST`. Результат ее похож на ввод команды `dir` в сеансе FTP, а строки результата, в отличие от `nlst`, содержат полные сведения о файлах. Если потребуется получить более подробную информацию обо всех удаленных файлах, следует вызвать метод `dir` и проанализировать его результаты (как это делается, будет показано в одном из последующих примеров).

Обратите внимание, что здесь выполняется пропуск «.» и «..» – текущего и родительского каталогов, если они присутствуют в списке

элементов удаленного каталога. В отличие от os.listdir, некоторые (но не все) серверы включают эти элементы в список, поэтому нам необходимо либо пропускать их, либо обрабатывать исключения, к которым они могут привести (подробнее об этом будет рассказываться ниже, когда мы начнем использовать dir).

Выбор режима передачи в зависимости от типа файла

Выше мы уже обсуждали режимы открытия выходных файлов для FTP, но теперь, когда мы начинаем рассматривать возможность передачи текстовых файлов, я могу довести это обсуждение до конца. Для обработки кодировок и сохранения символов конца строки в соответствии с особенностями платформы, где находятся мои веб-файлы, этот сценарий по-разному выполняет передачу текстовых и двоичных файлов. Для выбора между текстовым и двоичным режимом передачи каждого файла он использует стандартный модуль mimetypes.

Мы уже встречались с модулем mimetypes в главе 6, в примере 6.23, где он использовался для реализации проигрывания медиафайлов (за пояснениями обращайтесь к тексту примера и описанию к нему). Здесь модуль mimetypes используется для выбора между текстовым и двоичным режимами передачи файла, исходя из расширения его имени. Например, веб-страницы HTML и простые текстовые файлы передаются в текстовом режиме с автоматическим отображением символов конца строки, а изображения и архивы передаются в двоичном режиме.

Загрузка файлов: двоичный и текстовый режимы

Двоичные файлы загружаются с помощью метода retrbinary, с которым мы познакомились ранее, и локального режима открытия wb. Этот режим необходим, чтобы обеспечить возможность передачи строк bytes методу write, вызываемому методом retrbinary, а также подавить преобразование байтов конца строки и кодирование символов Юникода. Опять же текстовый режим в Python 3.X требует, чтобы в этом режиме в файлы записывался текст, который можно кодировать в указанную кодировку, а попытка записи двоичных данных, таких как изображения, может вызывать ошибки кодирования. Этот сценарий может выполняться в Windows или в Unix-подобных системах, и было бы нежелательно, чтобы в Windows байты \n в изображениях замещались парой байтов \r\n. Здесь не используется третий аргумент, определяющий размер блока, – по умолчанию он принимает достаточно разумное значение.

Для загрузки *текстовых файлов* этот сценарий использует другой метод – retrlines, передавая ему функцию, которая должна вызываться для каждой загруженной строки текстового файла. Функция-обработчик строки текста обычно просто принимает строку str и записывает ее в локальный текстовый файл. Но обратите внимание, что функция-обработчик, создаваемая здесь lambda-выражением, добавляет также в конец переданной ей строки символ перевода

строки \n. Метод Python retrlines удаляет из строк все символы перевода строки, чтобы обойти различия между платформами. Добавляя \n, сценарий обеспечивает добавление правильной последовательности символов перевода строки для той платформы, на которой выполняется сценарий (\n или \r\n).

Конечно, чтобы такое автоматическое отображение символов \n действовало в сценарии, необходимо также открывать выходные текстовые файлы в текстовом режиме w, а не в режиме wb – при записи данных в файл в Windows происходит отображение \n в \r\n. Как уже обсуждалось ранее, текстовый режим также подразумевает, что при вызове метода write файла из метода retrlines ему будет передаваться строка str, и при записи текст будет кодироваться с применением указанной кодировки.

Обратите внимание, что в вызове функции open, открывающей выходной текстовый файл, вместо кодировки по умолчанию мы явно используем *схему кодирования*, извлекая ее из объекта connection класса FTP. Без этого сценарий завершался с ошибкой UnicodeEncodeError при попытке загрузить некоторые файлы с моего сайта. В методе retrlines объект FTP читает данные из удаленного файла через сокет, используя файл-обертку в текстовом режиме, и явно указывает схему кодирования для декодирования байтов. Поскольку в любом случае объект класса FTP не может предложить ничего лучше, мы используем эту же кодировку и при открытии выходного файла.

Для декодирования получаемого текста (а также для кодирования текста при передаче) объекты FTP по умолчанию используют кодировку latin1, но ее можно изменить, записав требуемое имя кодировки в атрибут encoding. Локальные текстовые файлы сохраняются в кодировке, используемой модулем ftplib, поэтому они совместимы с кодировками, применяемыми к текстовым данным, которые производятся и передаются этим модулем.

Мы могли бы также попробовать перехватывать исключения кодирования символов Юникода для файлов, несовместимых с кодировкой, используемой объектом FTP, но при тестировании в Python 3.1 выяснилось, что исключения оставляют объект FTP в неисправном состоянии. Как вариант, мы могли бы также использовать двоичный режим wb для выходных локальных текстовых файлов и вручную кодировать строки с помощью метода line.encode, или просто использовать метод retrbinary и двоичный режим выходных файлов во всех случаях, но в обоих ситуациях мы лишились бы возможности отображать символы конца строки переносимым способом – потеряется весь смысл различения типов файлов в этом контексте.

Лучше один раз увидеть, чем сто раз услышать. Ниже приводится команда, которую я использую для загрузки всего моего веб-сайта поддержки книги с сервера моего провайдера на ноутбук с Windows за один шаг:

```
C:\...\PP4E\Internet\Ftp\Mirror> downloadflat.py test
Password for lutz on home.rmi.net:
Clean local directory first? y
connecting...
deleting local 2004-longmont-classes.html
deleting local 2005-longmont-classes.html
deleting local 2006-longmont-classes.html
deleting local about-hopl.html
deleting local about-lp.html
deleting local about-lp2e.html
deleting local about-pp-japan.html
```

...часть строк опущена...

```
downloading 2004-longmont-classes.html to test\2004-longmont-classes.html
                 as text
downloading 2005-longmont-classes.html to test\2005-longmont-classes.html
                as text
downloading 2006-longmont-classes.html to test\2006-longmont-classes.html
                as text
downloading about-hopl.html to test\about-hopl.html as text
downloading about-lp.html to test\about-lp.html as text
downloading about-lp2e.html to test\about-lp2e.html as text
downloading about-pp-japan.html to test\about-pp-japan.html as text
```

...часть строк опущена...

```
downloading ora-pyref4e.gif to test\ora-pyref4e.gif as image
downloading ora-lp4e-big.jpg to test\ora-lp4e-big.jpg as image
downloading ora-lp4e.gif to test\ora-lp4e.gif as image
downloading pyref4e-updates.html to test\pyref4e-updates.html as text
downloading lp4e-updates.html to test\lp4e-updates.html as text
downloading lp4e-examples.html to test\lp4e-examples.html as text
downloading LP4E-examples.zip to test\LP4E-examples.zip as application
Done: 297 files downloaded.
```

Эта процедура может занять несколько минут, в зависимости от размера вашего сайта и скорости вашего соединения (это связано с ограничениями скорости передачи в сети, и для передачи моего веб-сайта на мой ноутбук с беспроводным подключением обычно требуется примерно две-три минуты). Однако такая процедура значительно точнее и проще, чем загрузка файлов вручную. Этот сценарий обходит весь список файлов на веб-сервере, возвращаемый методом `nlst`, и поочередно загружает каждый из них по протоколу FTP (то есть через сокеты). Для имен, очевидно указывающих на текстовые данные, используется текстовый режим передачи, а для остальных – двоичный режим.

Прежде чем запускать сценарий таким способом, я проверяю, чтобы начальные операции присваивания в нем отражали участвующие в обмене компьютеры, а затем запускаю его из локального каталога, в который хочу поместить копию сайта. Поскольку каталог для загрузки обычно

отличается от каталога, где находится сам сценарий, интерпретатору необходимо указать полный путь к файлу сценария. Например, когда я выполняю этот сценарий в сеансе Telnet или SSH, каталог выполнения и каталог, где находится сам сценарий, отличаются, но сценарий работает одинаково.

Если вы решите удалить локальные файлы из каталога загрузки, то можете получить на экране серию сообщений «deleting local...» и только потом строки «downloading...»: это автоматически удалит весь лишний мусор, оставшийся от предыдущих операций загрузки. А если вы ошибетесь при вводе пароля, Python возбудит исключение – мне иногда приходится запускать сценарий повторно (и вводить пароль медленнее):

```
C:\...\PP4E\Internet\Ftp\Mirror> downloadflat.py test
Password for lutz on home.rmi.net:
Clean local directory first?
connecting...
Traceback (most recent call last):
  File "C:\...\PP4E\Internet\Ftp\Mirror\downloadflat.py", line 29,
        in <module>
    connection.login(remoteuser, remotepass)   # зарегистрироваться
                                               # с именем/паролем
  File "C:\Python31\lib\ftplib.py", line 375, in login
    if resp[0] == '3': resp = self.sendcmd('PASS ' + passwd)
  File "C:\Python31\lib\ftplib.py", line 245, in sendcmd
    return self.getresp()
  File "C:\Python31\lib\ftplib.py", line 220, in getresp
    raise error_perm(resp)
ftplib.error_perm: 530 Login incorrect.
```

Следует отметить, что этот сценарий частично настраивается операциями присваивания, производимыми в начале файла. Кроме того, пароль и признак необходимости удаления файлов указываются при интерактивном вводе, и разрешается указывать один аргумент командной строки – имя локального каталога для загрузки файлов (по умолчанию используется «.», то есть каталог, в котором выполняется сценарий). Можно было бы использовать аргументы командной строки для настройки других параметров загрузки, но благодаря простоте Python и отсутствию этапов компиляции/сборки изменять настройки в тексте сценариев Python обычно не труднее, чем вводить слова в командной строке.

Для проверки возможных различий между версиями файлов после такой пакетной загрузки или выгрузки можно воспользоваться сценарием `diffall`, представленным в главе 6, в примере 6.12. Например, я обнаруживаю файлы, изменившиеся с течением времени из-за обновлений на нескольких платформах, сравнивая загруженную локальную копию моего веб-сайта с помощью такой команды: `C:\...\PP4E\Internet\Ftp> ..\..\System\Filetools\diffall.py Mirror\test C:\...\Websites\public_html`. Подробности об этом инструменте смотрите в главе 6, а пример результатов, получаемых с его помощью, смот-

рите и в файле *diffall.out.txt*, находящемся в каталоге *diffs* в дереве примеров. Различия между текстовыми файлами в основном обусловлены либо наличием завершающего символа перевода строки в конце файла, либо различиями между символами перевода строки, возникшими в результате передачи в двоичном режиме, которые команда `fc` в Windows и FTP-серверы не замечают.

Выгрузка каталогов сайтов

Выгрузка на сервер целого каталога симметрична загрузке с сервера: в основном требуется поменять местами локальные и удаленные компьютеры и операции в только что рассмотренной программе. Сценарий в примере 13.11, используя протокол FTP, копирует все файлы из каталога на локальном компьютере, в котором он был запущен, в каталог на удаленном компьютере.

Я действительно пользуюсь этим сценарием, чаще всего для выгрузки одним махом всех файлов, которые я поддерживаю на своем ноутбуке, на сервер моего интернет-провайдера. Иногда я также с его помощью копирую свой сайт с моего ПК на сервер зеркала или с сервера зеркала обратно на сервер провайдера. Этот сценарий выполняется на любом компьютере, где есть Python и сокеты, благодаря чему он может копировать каталог с любого компьютера, подключенного к Сети, на любой компьютер, где работает FTP-сервер. Чтобы осуществить требуемую пересылку, достаточно лишь поменять соответствующим образом начальные настройки в этом модуле.

Пример 13.11. PP4E\Internet\Ftp\Mirror\uploadflat.py

```
#!/bin/env python
"""
##############################################################################
использует FTP для выгрузки всех файлов из локального каталога на удаленный
сайт/каталог; например, сценарий можно использовать для копирования
файлов веб/FTP сайта с вашего ПК на сервер провайдера; выполняет выгрузку
плоского каталога: вложенные каталоги можно копировать с помощью
сценария uploadall.py. дополнительные примечания смотрите в комментариях
в downloadflat.py: этот сценарий является его зеркальным отражением.
##############################################################################
"""

import os, sys, ftplib
from getpass import getpass
from mimetypes import guess_type

nonpassive = False                 # пассивный режим FTP по умолчанию
remotesite = 'learning-python.com' # выгрузить на этот сайт
remotedir  = 'books'               # с компьютера, где выполняется сценарий
remoteuser = 'lutz'
remotepass = getpass('Password for %s on %s: ' % (remoteuser, remotesite))
localdir   = (len(sys.argv) > 1 and sys.argv[1]) or '.'
```

```
cleanall   = input('Clean remote directory first? ')[:1] in ['y', 'Y']

print('connecting...')
connection = ftplib.FTP(remotesite)      # соединиться с FTP-сайтом
connection.login(remoteuser, remotepass) # зарегистрироваться
                                         # с именем/паролем
connection.cwd(remotedir)                # перейти в каталог копирования
if nonpassive:                 # принудительный переход в активный режим FTP
    connection.set_pasv(False) # большинство серверов работают
                               # в пассивном режиме

if cleanall:
    for remotename in connection.nlst():          # уничтожить все удаленные
        try:                                      # файлы, чтобы избавиться
            print('deleting remote', remotename)  # от устаревших копий
            connection.delete(remotename)         # пропустить . и ..
        except:
            print('cannot delete remote', remotename)

count = 0                         # выгрузить все локальные файлы
localfiles = os.listdir(localdir) # listdir() отбрасывает путь к каталогу
                                  # любая ошибка завершит сценарий
for localname in localfiles:
    mimetype, encoding = guess_type(localname)  # например,
                                                # ('text/plain','gzip')
    mimetype = mimetype or '?/?'                # допускается (None, None)
    maintype = mimetype.split('/')[0]           # .jpg ('image/jpeg', None')

    localpath = os.path.join(localdir, localname)
    print('uploading', localpath, 'to', localname, end=' ')
    print('as', maintype, encoding or '')

    if maintype == 'text' and encoding == None:
        # использовать двоичный файл и режим передачи ascii
        # для поддержки логики обработки символов конца строки
        # требуется использовать режим rb
        localfile = open(localpath, 'rb')
        connection.storlines('STOR ' + localname, localfile)

    else:
        # использовать двоичный файл и двоичный режим предачи
        localfile = open(localpath, 'rb')
        connection.storbinary('STOR ' + localname, localfile)

    localfile.close()
    count += 1

connection.quit()
print('Done:', count, 'files uploaded.')
```

Как и сценарий загрузки каталога с сервера, эта программа иллюстрирует ряд новых интерфейсов и приемов написания сценариев FTP:

Удаление всех файлов в копии на сервере

Как и операция загрузки каталога, выгрузка на сервер начинается с запроса необходимости удаления всех файлов в целевом каталоге сервера перед копированием туда новых файлов. Параметр `cleanall` полезно использовать, если какие-то файлы были удалены в локальной копии каталога у клиента – отсутствующие файлы останутся в копии на сервере, если сначала не удалить там все файлы.

Чтобы реализовать зачистку сервера, этот сценарий просто получает список всех файлов в каталоге на сервере с помощью метода `nlst` объекта `FTP` и поочередно удаляет их с помощью метода `delete` объекта `FTP`. При наличии права на удаление каталог будет очищен (права доступа к файлам зависят от учетной записи, под которой производится регистрация при подключении к серверу). К моменту выполнения операции удаления файлов мы уже находимся в целевом каталоге на сервере, поэтому в именах файлов не требуется указывать путь к каталогу. Обратите внимание, что на некоторых серверах метод `nlst` может возбуждать исключение, если удаленный каталог пуст – мы не перехватываем это исключение здесь, но можем просто не указывать необходимость очистки при второй попытке, если возникла такая ошибка. Мы обрабатываем исключения, связанные с удалением файлов, потому что некоторые серверы могут возвращать в списке файлов имена каталогов «.» и «..».

Запись всех локальных файлов

Чтобы применить операцию выгрузки на сервер к каждому файлу в локальном каталоге, мы сначала получаем список локальных файлов вызовом стандартной функции `os.listdir` и добавляем путь к локальному каталогу в начало имени каждого файла с помощью функции `os.path.join`. Напомню, что `os.listdir` возвращает имена файлов без пути к каталогу, а исходный каталог может отличаться от каталога исполнения сценария, если он передан в командной строке.

Выгрузка на сервер: двоичный и текстовый режимы

Этот сценарий может выполняться в Windows и Unix-подобных системах, поэтому текстовые файлы должны обрабатываться особым способом. Как и при загрузке каталога с сервера, этот сценарий выбирает текстовый или двоичный режим передачи, определяя тип файла с помощью модуля `mimetypes` по его расширению в имени – файлы HTML и текстовые файлы перемещаются в текстовом режиме FTP. Мы уже знакомы с методом `storbinary` объекта `FTP`, применяемым для выгрузки файлов в двоичном режиме, – на удаленном сайте оказывается точная, байт в байт, копия файла.

Передача в текстовом режиме действует почти так же: метод `storlines` принимает строку команды FTP и объект локального файла (или похожего на файл) и просто копирует каждую строку локального файла в одноименный файл на удаленном компьютере.

Однако обратите внимание, что в Python 3.X локальный текстовый файл должен открываться в двоичном режиме rb. Входные текстовые файлы обычно открываются в текстовом режиме r, что обеспечивает декодирование в символы Юникода и преобразование в Windows любых последовательностей \r\n конца строки в платформонезависимые символы \n в процессе чтения строк. Однако модуль ftplib в Python 3.1 требует, чтобы текстовые файлы открывались в двоичном режиме rb, потому что он сам преобразует все символы конца строки в последовательности \r\n при передаче, а для этого он должен читать строки как простые последовательности байтов с помощью readlines и выполнять обработку строк bytes, которые возвращаются в двоичном режиме.

Модуль ftplib мог работать в Python 2.X со строками, читаемыми из файлов в текстовом режиме, но лишь по той простой причине, что в этой версии отсутствовал отдельный тип bytes – символы \n просто замещались последовательностями \r\n. Открытие локального файла в двоичном режиме также означает, что при чтении его модулем ftplib не будет выполняться декодирование в символы Юникода: текст передается через сокеты как строка байтов в кодированной форме. Все это, конечно, является наглядным примером влияния особенностей кодирования Юникода – за подробностями обращайтесь к файлу модуля *ftplib.py* в каталоге с исходными текстами библиотеки Python.

В двоичном режиме передачи все обстоит еще проще – двоичные файлы открываются в режиме rb, чтобы подавить декодирование в символы Юникода и другие автоматические преобразования, и возвращают строки bytes, ожидаемые модулем ftplib при чтении. Двоичные данные не являются текстом Юникода, и нам не нужно, чтобы байты в аудиофайле, значение которых случайно совпадает с \r, таинственным образом исчезали при чтении в Windows.

Как и в сценарии загрузки каталога с сервера, эта программа обходит все файлы, которые должны быть переданы (в данном случае в локальном каталоге), и поочередно передает их – в текстовом или двоичном режиме в зависимости от имени файла. Ниже приводится команда, которой я пользуюсь для выгрузки целиком моего сайта с моего ноутбука с Windows на удаленный Linux-сервер моего интернет-провайдера за один шаг:

```
C:\...\PP4E\Internet\Ftp\Mirror> uploadflat.py test
Password for lutz on learning-python.com:
Clean remote directory first? y
connecting...
deleting remote .
cannot delete remote .
deleting remote ..
cannot delete remote ..
deleting remote 2004-longmont-classes.html
deleting remote 2005-longmont-classes.html
```

```
deleting remote 2006-longmont-classes.html
deleting remote about-lp1e.html
deleting remote about-lp2e.html
deleting remote about-lp3e.html
deleting remote about-lp4e.html
```

...часть строк опущена...

```
uploading test\2004-longmont-classes.html to 2004-longmont-classes.html
               as text
uploading test\2005-longmont-classes.html to 2005-longmont-classes.html
              as text
uploading test\2006-longmont-classes.html to 2006-longmont-classes.html
              as text
uploading test\about-lp1e.html to about-lp1e.html as text
uploading test\about-lp2e.html to about-lp2e.html as text
uploading test\about-lp3e.html to about-lp3e.html as text
uploading test\about-lp4e.html to about-lp4e.html as text
uploading test\about-pp-japan.html to about-pp-japan.html as text
```

...часть строк опущена...

```
uploading test\whatsnew.html to whatsnew.html as text
uploading test\whatsold.html to whatsold.html as text
uploading test\wxPython.doc.tgz to wxPython.doc.tgz as application gzip
uploading test\xlate-lp.html to xlate-lp.html as text
uploading test\zaurus0.jpg to zaurus0.jpg as image
uploading test\zaurus1.jpg to zaurus1.jpg as image
uploading test\zaurus2.jpg to zaurus2.jpg as image
uploading test\zoo-jan-03.jpg to zoo-jan-03.jpg as image
uploading test\zopeoutline.htm to zopeoutline.htm as text
Done: 297 files uploaded.
```

Для моего сайта на моем текущем ноутбуке с беспроводным подключением весь процесс обычно занимает около шести минут, в зависимости от загруженности сервера. Как и в примере с загрузкой каталога, я часто выполняю эту команду из локального каталога, где хранятся мои веб-файлы, и передаю интерпретатору Python полный путь к сценарию. Когда я запускаю этот сценарий на Linux-сервере, он действует так же, но я указываю иные пути к сценарию и к каталогу с моими веб-файлами.[1]

[1] Примечания к использованию: эти сценарии сильно зависят от корректной работы FTP-сервера. В течение некоторого времени в сценарии выгрузки каталога на сервер при работе через мое текущее широкополосное соединение иногда возникали ошибки, связанные с превышением предельного времени ожидания. Позднее эти ошибки пропали – когда мой интернет-провайдер исправил настройки своего сервера. Если у вас возникнут проблемы, попробуйте запустить этот сценарий с другим сервером; иногда может помочь разъединение и соединение перед передачей каждого файла (некоторые серверы могут ограничивать количество соединений).

Реорганизация сценариев выгрузки и загрузки для многократного использования

Сценарии выгрузки и загрузки каталога, представленные в двух предыдущих разделах, вполне справляются со своей работой и были единственными примерами использования протокола FTP, исключая логику использования `mimetypes`, вошедшими во второе издание этой книги. Однако, если внимательно изучить эти два сценария, можно заметить общие черты, объединяющие их. Фактически они в значительной степени совпадают – в них используется идентичный программный код настройки параметров, соединения с FTP-сервером и определения типа файла. Конечно, со временем появились некоторые отличия в деталях, но часть программного кода определенно была просто скопирована из одного файла в другой.

Хотя такая избыточность и не является поводом для тревоги, особенно если в будущем не планируется изменять эти сценарии, тем не менее, в программных проектах в целом это обстоятельство вызывает неудобства. Когда имеется две копии идентичного программного кода, это не только опасно тем, что со временем они потеряют свою идентичность (при этом может быть утрачено единообразие пользовательского интерфейса и поведения сценариев), но и вынудит вас удваивать свои усилия, когда потребуется изменить программный код сразу в двух местах. Если только вы не любитель лишней работы, есть смысл приложить усилия, чтобы избежать такой избыточности.

Избыточность становится особенно явной, стоит только посмотреть на сложный программный код, использующий `mimetypes` для определения типов файлов. Повторение ключевого программного кода в нескольких местах практически всегда относится к неудачным идеям – не только потому, что нам придется вспоминать, как он действует, когда нам потребуется повторно писать ту же самую утилиту, но и потому, что такая организация способствует появлению ошибок.

Версия на основе функций

Оригинальные версии сценариев загрузки и выгрузки включают программный код верхнего уровня, который опирается на глобальные переменные. Такая структура плохо поддается повторному использованию – программный код выполняется немедленно, на этапе импортирования, и его сложно обобщить для применения в различных контекстах. Хуже того, его сложно сопровождать – всякий раз, щелкая на кнопке Paste (Копировать), чтобы скопировать фрагмент существующего программного кода, вы увеличиваете сложность внесения изменений в будущем.

Чтобы показать более удачное решение, в примере 13.12 демонстрируется один из способов *рефакторинга* (реорганизации) сценария загрузки. Благодаря обертыванию различных частей в функции они становятся

доступными для повторного использования в других модулях, включая и программу выгрузки.

Пример 13.12. PP4E\Internet\Ftp\Mirror\downloadflat_modular.py

```
#!/bin/env python
"""
##############################################################################
использует протокол FTP для копирования (загрузки) всех файлов из каталога
на удаленном сайте в каталог на локальном компьютере; эта версия действует
точно так же, но была реорганизована с целью завернуть фрагменты
программного кода в функции, чтобы их можно было повторно использовать
в сценарии выгрузки каталога и, возможно, в других программах в будущем –
в противном случае избыточность программного кода может с течением времени
привести к появлению различий в изначально одинаковых фрагментах и усложнит
сопровождение.
##############################################################################
"""

import os, sys, ftplib
from getpass import getpass
from mimetypes import guess_type, add_type

defaultSite = 'home.rmi.net'
defaultRdir = '.'
defaultUser = 'lutz'

def configTransfer(site=defaultSite, rdir=defaultRdir, user=defaultUser):
    """
    принимает параметры выгрузки или загрузки
    из-за большого количества параметров использует класс
    """
    class cf: pass
    cf.nonpassive = False  # пассивный режим FTP, по умолчанию в 2.1+
    cf.remotesite = site   # удаленный сайт куда/откуда выполняется передача
    cf.remotedir  = rdir   # и каталог ('.' означает корень учетной записи)
    cf.remoteuser = user
    cf.localdir   = (len(sys.argv) > 1 and sys.argv[1]) or '.'
    cf.cleanall   = input('Clean target directory first? ')[:1] in ['y','Y']
    cf.remotepass = getpass(
        'Password for %s on %s:' % (cf.remoteuser, cf.remotesite))
    return cf

def isTextKind(remotename, trace=True):
    """
    использует mimetype для определения принадлежности файла
    к текстовому или двоичному типу
    'f.html' определяется как ('text/html', None): текст
    'f.jpeg' определяется как ('image/jpeg', None): двоичный
    'f.txt.gz' определяется как ('text/plain', 'gzip'): двоичный
    файлы с неизвестными расширениями определяются как (None, None): двоичные
```

```
    модуль mimetype способен также строить предположения об именах
    исходя из типа: смотрите пример PyMailGUI
    """
    add_type('text/x-python-win', '.pyw')        # отсутствует в таблицах
    mimetype,encoding = guess_type(remotename,strict=False) # разрешить
                                                 # расширенную интерпретацию
    mimetype = mimetype or '?/?'                 # тип неизвестен?
    maintype = mimetype.split('/')[0]            # получить первый элемент
    if trace: print(maintype, encoding or '')
    return maintype == 'text' and encoding == None   # не сжатый

def connectFtp(cf):
    print('connecting...')
    connection = ftplib.FTP(cf.remotesite)       # соединиться с FTP-сайтом
    connection.login(cf.remoteuser, cf.remotepass)   # зарегистрироваться
    connection.cwd(cf.remotedir)                     # перейти в каталог
    if cf.nonpassive:       # переход в активный режим FTP при необходимости
        connection.set_pasv(False) # большинство работают в пассивном режиме
    return connection

def cleanLocals(cf):
    """
    пытается удалить все локальные файлы, чтобы убрать устаревшие копии
    """
    if cf.cleanall:
        for localname in os.listdir(cf.localdir): # список локальных файлов
            try:                                  # удаление локального файла
                print('deleting local', localname)
                os.remove(os.path.join(cf.localdir, localname))
            except:
                print('cannot delete local', localname)

def downloadAll(cf, connection):
    """
    загружает все файлы из удаленного каталога в соответствии с настройками
    в cf; метод nlst() возвращает список файлов, dir() - полный список
    с дополнительными подробностями
    """
    remotefiles = connection.nlst()        # nlst - список файлов на сервере
    for remotename in remotefiles:
        if remotename in ('.', '..'): continue
        localpath = os.path.join(cf.localdir, remotename)
        print('downloading', remotename, 'to', localpath, 'as', end=' ')
        if isTextKind(remotename):
            # использовать текстовый режим передачи
            localfile = open(localpath, 'w', encoding=connection.encoding)
            def callback(line): localfile.write(line + '\n')
            connection.retrlines('RETR ' + remotename, callback)
        else:
            # использовать двоичный режим передачи
            localfile = open(localpath, 'wb')
```

```
            connection.retrbinary('RETR ' + remotename, localfile.write)
        localfile.close()
    connection.quit()
    print('Done:', len(remotefiles), 'files downloaded.')

if __name__ == '__main__':
    cf = configTransfer()
    conn = connectFtp(cf)
    cleanLocals(cf) # не удалять файлы, если соединение не было установлено
    downloadAll(cf, conn)
```

Сравните эту версию с оригиналом. Этот сценарий и все остальные в этом разделе действуют точно так же, как и оригинальные программы загрузки и выгрузки плоского каталога. Хотя мы и не изменили поведение сценария, тем не менее, мы радикально изменили его структуру – его реализация теперь представляет собой комплект инструментов, которые можно импортировать и повторно использовать в других программах.

В примере 13.13 приводится реорганизованная версия программы выгрузки, которая теперь стала значительно проще и использует один и тот же программный код совместно со сценарием загрузки, благодаря чему изменения при необходимости потребуется вносить только в одном месте.

Пример 13.13. PP4E\Internet\Ftp\Mirror\uploadflat_modular.py

```
#!/bin/env python
"""
##############################################################################
использует FTP для выгрузки всех файлов из локального каталога на удаленный
сайт/каталог; эта версия повторно использует функции из сценария загрузки,
чтобы избежать избыточности программного кода;
##############################################################################
"""

import os
from downloadflat_modular import configTransfer, connectFtp, isTextKind

def cleanRemotes(cf, connection):
    """
    пытается сначала удалить все файлы в каталоге на сервере,
    чтобы ликвидировать устаревшие копии
    """
    if cf.cleanall:
        for remotename in connection.nlst():     # список файлов на сервере
            try:                                 # удаление файла на сервере
                print('deleting remote',remotename)  # пропустить . и ..
                connection.delete(remotename)
            except:
                print('cannot delete remote', remotename)
```

```
def uploadAll(cf, connection):
    """
    выгружает все файлы в каталог на сервере в соответствии с настройками cf
    listdir() отбрасывает пути к каталогам, любые ошибки завершают сценарий
    """
    localfiles = os.listdir(cf.localdir) # listdir - список локальных файлов
    for localname in localfiles:
        localpath = os.path.join(cf.localdir, localname)
        print('uploading', localpath, 'to', localname, 'as', end=' ')
        if isTextKind(localname):
            # использовать текстовый режим передачи
            localfile = open(localpath, 'rb')
            connection.storlines('STOR ' + localname, localfile)
        else:
            # использовать двоичный режим передачи
            localfile = open(localpath, 'rb')
            connection.storbinary('STOR ' + localname, localfile)
        localfile.close()
    connection.quit()
    print('Done:', len(localfiles), 'files uploaded.')

if __name__ == '__main__':
    cf = configTransfer(site='learning-python.com', rdir='books',
                        user='lutz')
    conn = connectFtp(cf)
    cleanRemotes(cf, conn)
    uploadAll(cf, conn)
```

Благодаря повторному использованию программного кода сценарий выгрузки не только стал заметно проще, но он также будет наследовать все изменения, которые будут выполняться в модуле загрузки. Например, функция `isTextKind` позднее была дополнена программным кодом, который добавляет расширение *.pyw* в таблицы `mimetypes` (по умолчанию тип этих файлов не распознается), – поскольку теперь это совместно используемая функция, изменения автоматически будут унаследованы и программой выгрузки.

Этот сценарий и программы, которые будут его импортировать, достигает тех же целей, что и оригинал, но простота обслуживания, которую они приобретают, имеет большое значение в мире разработки программного обеспечения. Ниже приводится пример загрузки сайта с одного сервера и выгрузки его на другой сервер:

```
C:\...\PP4E\Internet\Ftp\Mirror> python downloadflat_modular.py test
Clean target directory first?
Password for lutz on home.rmi.net:
connecting...
downloading 2004-longmont-classes.html to test\2004-longmont-classes.html
            as text
...часть строк опущена...
downloading relo-feb010-index.html to test\relo-feb010-index.html as text
```

```
Done: 297 files downloaded.

C:\...\PP4E\Internet\Ftp\Mirror> python uploadflat_modular.py test
Clean target directory first?
Password for lutz on learning-python.com:
connecting...
uploading test\2004-longmont-classes.html to 2004-longmont-classes.html
          as text
...часть строк опущена...
uploading test\zopeoutline.htm to zopeoutline.htm as text
Done: 297 files uploaded.
```

Версия на основе классов

Подход на основе функций, использованный в последних двух примерах, решает проблему избыточности, но эти реализации местами выглядят несколько неуклюжими. Например, объект `cf` с параметрами настройки, образующий собственное пространство имен, замещает глобальные переменные и разрывает зависимости между файлами. При этом, начав создавать объекты, моделирующие пространства имен, можно заметить, что поддержка ООП в языке Python обеспечивает более естественный способ организации программного кода. В качестве последнего витка развития в примере 13.14 приводится еще одна попытка реорганизовать программный код для работы с FTP, использующий возможности классов в языке Python.

Пример 13.14. PP4E\Internet\Ftp\Mirror\ftptools.py

```
#!/bin/env python
"""
##############################################################################
использует протокол FTP для загрузки из удаленного каталога или выгрузки
в удаленный каталог всех файлов сайта; для организации пространства имен
и обеспечения более естественной структуры программного кода в этой версии
используются классы и приемы ООП; мы могли бы также организовать сценарий
как суперкласс, выполняющий загрузку, и подкласс, выполняющий выгрузку,
который переопределяет методы очистки каталога и передачи файла, но это
усложнило бы в других клиентах возможность выполнения обеих операций,
загрузки и выгрузки; для сценария uploadall и, возможно, для других также
предусмотрены методы, выполняющие выгрузку/загрузку единственного файла,
которые используются в цикле в оригинальных методах;
##############################################################################
"""


import os, sys, ftplib
from getpass import getpass
from mimetypes import guess_type, add_type

# значения по умолчанию для всех клиентов
dfltSite = 'home.rmi.net'
dfltRdir = '.'
```

```
dfltUser = 'lutz'

class FtpTools:

   # следующие три метода допускается переопределять
    def getlocaldir(self):
        return (len(sys.argv) > 1 and sys.argv[1]) or '.'

    def getcleanall(self):
        return input('Clean target dir first?')[:1] in ['y','Y']

    def getpassword(self):
        return getpass(
            'Password for %s on %s:' % (self.remoteuser, self.remotesite))

    def configTransfer(self, site=dfltSite, rdir=dfltRdir, user=dfltUser):
        """
        принимает параметры операции выгрузки или загрузки
        из значений по умолчанию в модуле, из аргументов,
        из командной строки, из ввода пользователя
        анонимный доступ к ftp: user='anonymous' pass=emailaddr
        """
        self.nonpassive = False # пассивный режим FTP по умолчанию в 2.1+
        self.remotesite = site  # удаленный сайт куда/откуда вып-ся передача
        self.remotedir  = rdir  # и каталог ('.' - корень учетной записи)
        self.remoteuser = user
        self.localdir   = self.getlocaldir()
        self.cleanall   = self.getcleanall()
        self.remotepass = self.getpassword()

    def isTextKind(self, remotename, trace=True):
        """
        использует mimetype для определения принадлежности файла
        к текстовому или двоичному типу
        'f.html' определяется как ('text/html', None): текст
        'f.jpeg' определяется как ('image/jpeg', None): двоичный
        'f.txt.gz' определяется как ('text/plain', 'gzip'): двоичный
        неизвестные расширения определяются как (None, None): двоичные
        модуль mimetype способен также строить предположения об именах
        исходя из типа: смотрите пример PyMailGUI
        """
        add_type('text/x-python-win', '.pyw') # отсутствует в таблицах
        mimetype, encoding = guess_type(remotename, strict=False) # разрешить
                                              # расширенную интерпретацию
        mimetype = mimetype or '?/?'          # тип неизвестен?
        maintype = mimetype.split('/')[0]     # получить первый элемент
        if trace: print(maintype, encoding or '')
        return maintype == 'text' and encoding == None   # не сжатый

    def connectFtp(self):
        print('connecting...')
        connection = ftplib.FTP(self.remotesite)  # соединиться с FTP-сайтом
```

```
        connection.login(self.remoteuser, self.remotepass) # зарегистрироваться
        connection.cwd(self.remotedir)                     # перейти в каталог
        if self.nonpassive:                     # переход в активный режим FTP
                                                # при необходимости
            connection.set_pasv(False) # большинство - в пассивном режиме
        self.connection = connection

    def cleanLocals(self):
        """
        пытается удалить все локальные файлы, чтобы убрать устаревшие копии
        """
        if self.cleanall:
            for localname in os.listdir(self.localdir):  # локальные файлы
                try:                                     # удаление файла
                    print('deleting local', localname)
                    os.remove(os.path.join(self.localdir, localname))
                except:
                    print('cannot delete local', localname)

    def cleanRemotes(self):
        """
        пытается сначала удалить все файлы в каталоге на сервере,
        чтобы ликвидировать устаревшие копии
        """
        if self.cleanall:
            for remotename in self.connection.nlst(): # список файлов
                try:                                  # удаление файла
                    print('deleting remote', remotename)
                    self.connection.delete(remotename)
                except:
                    print('cannot delete remote', remotename)

    def downloadOne(self, remotename, localpath):
        """
        загружает один файл по FTP в текстовом или двоичном режиме
        имя локального файла не обязательно должно соответствовать имени
        удаленного файла
        """
        if self.isTextKind(remotename):
            localfile = open(localpath, 'w',
                             encoding=self.connection.encoding)
            def callback(line): localfile.write(line + '\n')
            self.connection.retrlines('RETR ' + remotename, callback)
        else:
            localfile = open(localpath, 'wb')
            self.connection.retrbinary('RETR ' + remotename, localfile.write)
        localfile.close()

    def uploadOne(self, localname, localpath, remotename):
        """
        выгружает один файл по FTP в текстовом или двоичном режиме
```

```
        имя удаленного файла не обязательно должно соответствовать
        имени локального файла
        """
        if self.isTextKind(localname):
            localfile = open(localpath, 'rb')
            self.connection.storlines('STOR ' + remotename, localfile)
        else:
            localfile = open(localpath, 'rb')
            self.connection.storbinary('STOR ' + remotename, localfile)
        localfile.close()

    def downloadDir(self):
        """
        загружает все файлы из удаленного каталога в соответствии
        с настройками; метод nlst() возвращает список файлов, dir() -
        полный список с дополнительными подробностями
        """
        remotefiles = self.connection.nlst()    # nlst - список файлов
                                                # на сервере
        for remotename in remotefiles:
            if remotename in ('.', '..'): continue
            localpath = os.path.join(self.localdir, remotename)
            print('downloading', remotename, 'to', localpath, 'as', end=' ')
            self.downloadOne(remotename, localpath)
        print('Done:', len(remotefiles), 'files downloaded.')

    def uploadDir(self):
        """
        выгружает все файлы в каталог на сервере в соответствии
        с настройками listdir() отбрасывает пути к каталогам,
        любые ошибки завершают сценарий
        """
        localfiles = os.listdir(self.localdir) # listdir - локальные файлы
        for localname in localfiles:
            localpath = os.path.join(self.localdir, localname)
            print('uploading', localpath, 'to', localname, 'as', end=' ')
            self.uploadOne(localname, localpath, localname)
        print('Done:', len(localfiles), 'files uploaded.')

    def run(self, cleanTarget=lambda:None, transferAct=lambda:None):
        """
        выполняет весь сеанс FTP
        по умолчанию очистка каталога и передача не выполняются
        не удаляет файлы, если соединение с сервером установить не удалось
        """
        self.connectFtp()
        cleanTarget()
        transferAct()
        self.connection.quit()

if __name__ == '__main__':
```

```
ftp = FtpTools()
xfermode = 'download'
if len(sys.argv) > 1:
    xfermode = sys.argv.pop(1) # получить и удалить второй аргумент
if xfermode == 'download':
    ftp.configTransfer()
    ftp.run(cleanTarget=ftp.cleanLocals, transferAct=ftp.downloadDir)
elif xfermode == 'upload':
    ftp.configTransfer(site='learning-python.com',
                       rdir='books', user='lutz')
    ftp.run(cleanTarget=ftp.cleanRemotes, transferAct=ftp.uploadDir)
else:
    print('Usage: ftptools.py ["download" | "upload"] [localdir]')
```

Фактически этот последний вариант объединяет в одном файле реализацию операций выгрузки и загрузки, потому что они достаточно тесно связаны между собой. Как и прежде, общий программный код был выделен в отдельные методы, чтобы избежать избыточности. Новым здесь является то обстоятельство, что для хранения параметров используется сам экземпляр класса (они превратились в атрибуты объекта `self`). Изучите программный код этого примера, чтобы лучше понять, как была выполнена реорганизация.

Эта версия действует, как и оригинальные сценарии загрузки и выгрузки сайта. Подробности, касающиеся использования, вы найдете в конце модуля, в программном коде самотестирования, а выполняемая операция определяется аргументом командной строки со значением «download» (загрузка) или «upload» (выгрузка). Мы не изменили принцип действия, а реорганизовали программный код, чтобы его проще было сопровождать и использовать в других программах:

```
C:\...\PP4E\Internet\Ftp\Mirror> ftptools.py download test
Clean target dir first?
Password for lutz on home.rmi.net:
connecting...
downloading 2004-longmont-classes.html to test\2004-longmont-classes.html
            as text
...часть строк опущена...
downloading relo-feb010-index.html to test\relo-feb010-index.html as text
Done: 297 files downloaded.

C:\...\PP4E\Internet\Ftp\Mirror> ftptools.py upload test
Clean target dir first?
Password for lutz on learning-python.com:
connecting...
uploading test\2004-longmont-classes.html to 2004-longmont-classes.html
            as text
...часть строк опущена...
uploading test\zopeoutline.htm to zopeoutline.htm as text
Done: 297 files uploaded.
```

Хотя этот файл по-прежнему можно запускать как сценарий командной строки, в действительности его класс теперь представляет собой пакет инструментов для работы с протоколом FTP, который можно подмешивать к другим классам и повторно использовать в других программах. Программный код, обернутый в класс, проще поддается адаптации под более специализированные требования путем переопределения его методов – его методы, возвращающие параметры настройки, такие как `getlocaldir`, например, можно переопределять в подклассах для решения специфических задач.

Но самое важное, пожалуй, заключается в том, что классы упрощают многократное использование программного кода. Клиенты этого файла смогут выгружать и загружать каталоги за счет создания подклассов или встраивания экземпляров этого класса и вызова их методов. Чтобы увидеть один из примеров такого применения, перейдем к следующему разделу.

Передача деревьев каталогов с помощью ftplib

Возможно, самым большим ограничением рассмотренных сценариев загрузки веб-сайта с сервера и выгрузки его на сервер является допущение, что каталог сайта является плоским (отсюда их названия)[1], то есть оба сценария передают только простые файлы и не обрабатывают вложенные подкаталоги внутри пересылаемого веб-каталога.

Для моих целей это часто разумное ограничение. Для простоты я избегаю вложенных подкаталогов и храню свой домашний сайт, созданный для поддержки книги, как простой каталог файлов. Однако для других сайтов, включая тот, который я держу на другом компьютере, сценариями пересылки файлов проще пользоваться, если они попутно также автоматически пересылают подкаталоги.

Выгрузка локального дерева каталогов

Оказывается, поддержка выгрузки подкаталогов осуществляется достаточно просто – требуется добавить только немного рекурсии и вызовы методов для создания удаленных каталогов. Сценарий выгрузки на сервер, представленный в примере 13.15, расширяет версию на основе классов, которую мы только что видели в примере 13.14, и осуществляет выгрузку всех подкаталогов, вложенных в пересылаемый каталог. Более того, он рекурсивно пересылает каталоги, находящиеся внутри подкаталогов, – все дерево каталогов, содержащееся внутри передаваемого каталога верхнего уровня, выгружается в целевой каталог на удаленном сервере.

С точки зрения организации программного кода пример 13.15 является простой адаптацией класса `FtpTools` из предыдущего раздела – в дейст-

[1] flat (англ.) – плоский. – *Прим. ред.*

вительности мы просто реализовали подкласс с методом, выполняющим рекурсивную выгрузку. Как одно из следствий, мы бесплатно получаем инструменты, хранящие параметры настройки, проверяющие тип содержимого, выполняющие соединение с сервером и выгрузку – благодаря применению ООП часть работы оказалась выполненной еще до того, как мы к ней приступили.

Пример 13.15. PP4E\Internet\Ftp\Mirror\uploadall.py

```
#!/bin/env python
"""
##############################################################################
расширяет класс FtpTools, обеспечивая возможность выгрузки всех файлов
и подкаталогов из локального дерева каталогов в удаленный каталог
на сервере; поддерживает вложенные подкаталоги, но не поддерживает
операцию cleanall (для этого необходимо анализировать листинги FTP,
чтобы определять удаленные каталоги: смотрите сценарий cleanall.py);
для выгрузки подкаталогов используется os.path.isdir(path),
которая проверяет, является ли в действительности локальный файл каталогом,
метод FTP().mkd(path) - для создания каталогов на сервере (вызов обернут
инструкцией try, на случай если каталог уже существует на сервере),
и рекурсия - для выгрузки всех файлов/каталогов внутри
вложенного подкаталога.
##############################################################################
"""

import os, ftptools

class UploadAll(ftptools.FtpTools):
    """
    выгружает дерево подкаталогов целиком
    предполагается, что каталог верхнего уровня уже существует на сервере
    """
    def __init__(self):
        self.fcount = self.dcount = 0

    def getcleanall(self):
        return False                            # даже не спрашивать

    def uploadDir(self, localdir):
        """
        для каждого каталога в дереве выгружает простые файлы,
        производит рекурсивный вызов для подкаталогов
        """
        localfiles = os.listdir(localdir)
        for localname in localfiles:
            localpath = os.path.join(localdir, localname)
            print('uploading', localpath, 'to', localname, end=' ')
            if not os.path.isdir(localpath):
                self.uploadOne(localname, localpath, localname)
                self.fcount += 1
```

```
            else:
                try:
                    self.connection.mkd(localname)
                    print('directory created')
                except:
                    print('directory not created')
                self.connection.cwd(localname) # изменить удаленный каталог
                self.uploadDir(localpath)      # выгрузить локальный каталог
                self.connection.cwd('..')      # вернуться обратно
                self.dcount += 1
                print('directory exited')

if __name__ == '__main__':
    ftp = UploadAll()
    ftp.configTransfer(site='learning-python.com',
                       rdir='training', user='lutz')
    ftp.run(transferAct = lambda: ftp.uploadDir(ftp.localdir))
    print('Done:', ftp.fcount, 'files and',
          ftp.dcount, 'directories uploaded.')
```

Как и сценарий выгрузки плоских каталогов, этот сценарий может выполняться на любом компьютере с Python и сокетами и осуществлять выгрузку на любой компьютер, где выполняется сервер FTP. Я запускаю его на своем ноутбуке и на других серверах через Telnet или SSH, чтобы выгрузить сайты на сервер моего интернет-провайдера.

Центральным пунктом в этом сценарии является вызов функции `os.path.isdir` в самом начале. Если эта проверка обнаруживает подкаталог в текущем локальном каталоге, с помощью метода `connection.mkd` на удаленном сервере создается одноименный каталог, с помощью `connection.cwd` выполняется вход в него и рекурсивный спуск в подкаталог на локальном компьютере (рекурсия здесь совершенно необходима, так как дерево каталогов может иметь произвольную структуру и глубину). Как и все методы объекта `FTP`, методы `mkd` и `cwd` посылают команды FTP удаленному серверу. Выйдя из локального подкаталога, выполняется выход из удаленного подкаталога `cwd('..')`, чтобы подняться в родительский каталог на сервере, и обход файлов продолжается – возврат из рекурсивного вызова автоматически восстанавливает местоположение в текущем каталоге. Остальная часть сценария примерно совпадает с первоначальным вариантом.

В целях экономии места я оставляю более глубокое изучение этого варианта в качестве упражнения для читателя. Чтобы получить более полное представление, попробуйте изменить этот сценарий, чтобы он не полагался больше на наличие на сервере каталога верхнего уровня. Как обычно в мире программного обеспечения, данную особенность можно реализовать различными способами.

Ниже приводится пример вывода сценария `uploadall`, запущенного для выгрузки сайта, содержащего несколько уровней вложенности подкаталогов, который я сопровождаю с помощью инструментов конструиро-

вания сайтов. Он напоминает вывод сценария выгрузки плоского каталога (чего вполне можно было ожидать, учитывая, что данный сценарий использует большую часть того же программного кода через механизм наследования), но обратите внимание, что попутно он выгружает вложенные подкаталоги:

```
C:\...\PP4E\Internet\Ftp\Mirror> uploadall.py Website-Training
Password for lutz on learning-python.com:
connecting...
uploading Website-Training\2009-public-classes.htm
                                        to 2009-public-classes.htm text
uploading Website-Training\2010-public-classes.html
                                        to 2010-public-classes.html text
uploading Website-Training\about.html to about.html text
uploading Website-Training\books to books directory created
uploading Website-Training\books\index.htm to index.htm text
uploading Website-Training\books\index.html to index.html text
uploading Website-Training\books\_vti_cnf to _vti_cnf directory created
uploading Website-Training\books\_vti_cnf\index.htm to index.htm text
uploading Website-Training\books\_vti_cnf\index.html to index.html text
directory exited
directory exited
uploading Website-Training\calendar.html to calendar.html text
uploading Website-Training\contacts.html to contacts.html text
uploading Website-Training\estes-nov06.htm to estes-nov06.htm text
uploading Website-Training\formalbio.html to formalbio.html text
uploading Website-Training\fulloutline.html to fulloutline.html text
```

...часть строк опущена...

```
uploading Website-Training\_vti_pvt\writeto.cnf to writeto.cnf ?
uploading Website-Training\_vti_pvt\_vti_cnf to _vti_cnf directory created
uploading Website-Training\_vti_pvt\_vti_cnf\_x_todo.htm to _x_todo.htm text
uploading Website-Training\_vti_pvt\_vti_cnf\_x_todoh.htm
                                            to _x_todoh.htm text
directory exited
uploading Website-Training\_vti_pvt\_x_todo.htm to _x_todo.htm text
uploading Website-Training\_vti_pvt\_x_todoh.htm to _x_todoh.htm text
directory exited
Done: 366 files and 18 directories uploaded.
```

В текущей реализации сценарий в примере 13.15 обрабатывает только выгрузку деревьев каталогов – для тех, кто сопровождает веб-сайты на своем локальном компьютере и периодически выгружает изменения на сервер, рекурсивная выгрузка обычно имеет большую практическую ценность, чем рекурсивная загрузка. Чтобы также *загружать* (создавать зеркальные копии) веб-сайты, содержащие вложенные подкаталоги, сценарий должен проанализировать вывод команды получения списка файлов в удаленном каталоге. По той же самой причине сценарий рекурсивной выгрузки не поддерживает очистку удаленного дерева каталогов – для реализации данной особенности также потребовалось бы

организовать анализ списков удаленных файлов. Как это сделать, демонстрируется в следующем разделе.

Удаление деревьев каталогов на сервере

И последний пример многократного использования программного кода: когда я приступил к тестированию сценария uploadall из предыдущего раздела, он содержал ошибку, которая вызывала бесконечный рекурсивный цикл, снова и снова копируя весь сайт в новые подкаталоги, пока FTP-сервер не закрывал соединение (непредусмотренная особенность программы!). Фактически выгрузка продолжалась до достижения 13 уровня вложенности, прежде чем сервер закрывал соединение – это приводило к блокированию моего сайта, заставляя исправить ошибку.

Чтобы избавиться от всех файлов, выгруженных по ошибке, я быстро написал сценарий, представленный в примере 13.16, в экстремальном (даже в паническом) режиме. Он удаляет все файлы и вложенные подкаталоги в дереве на удаленном сервере. К счастью, это оказалось очень просто, учитывая, что пример 13.16 унаследовал повторно используемые инструменты от суперкласса FtpTools. В данном примере просто определяется расширение, рекурсивно удаляющее файлы и каталоги на сервере. Даже в таком вынужденном режиме, в каком я оказался, можно с успехом использовать преимущества ООП.

Пример 13.16. PP4E\Internet\Ftp\Mirror\cleanall.py

```
#!/bin/env python
"""
##############################################################################
расширяет класс FtpTools возможностью удаления файлов и подкаталогов
в дереве каталогов на сервере; поддерживает удаление вложенных подкаталогов;
зависит от формата вывода команды dir(), который может отличаться
на некоторых серверах! - смотрите подсказки в файле
Tools\Scripts\ftpmirror.py, в каталоге установки Python;
добавьте возможность загрузки дерева каталогов с сервера;
##############################################################################
"""

from ftptools import FtpTools

class CleanAll(FtpTools):
    """
    удаляет все дерево каталогов на сервере
    """
    def __init__(self):
        self.fcount = self.dcount = 0

    def getlocaldir(self):
        return None            # не имеет смысла здесь

    def getcleanall(self):
```

```
        return True            # само собой разумеется здесь

    def cleanDir(self):
        """
        для каждого элемента в текущем каталоге на сервере
        удаляет простые файлы, выполняет рекурсивный спуск и удаляет
        подкаталоги, метод dir() объекта FTP передает каждую строку
        указанной функции или методу
        """
        lines = []                      # на каждом уровне свой список строк
        self.connection.dir(lines.append)  # список текущего каталога
                                           # на сервере
        for line in lines:
            parsed = line.split()       # разбить по пробельным символам
            permiss = parsed[0]         # предполагается формат:
            fname = parsed[-1]          # 'drw... ... filename'
            if fname in ('.', '..'):        # некоторые серверы включают cwd
                continue                    # и родительский каталог
            elif permiss[0] != 'd':         # простой файл: удалить
                print('file', fname)
                self.connection.delete(fname)
                self.fcount += 1
            else:                           # каталог: удалить рекурсивно
                print('directory', fname)
                self.connection.cwd(fname)  # переход в каталог на сервере
                self.cleanDir()             # очистить подкаталог
                self.connection.cwd('..')   # возврат на уровень выше
                self.connection.rmd(fname)  # удалить пустой каталог
                                            # на сервере
                self.dcount += 1
                print('directory exited')

if __name__ == '__main__':
    ftp = CleanAll()
    ftp.configTransfer(site='learning-python.com',
                       rdir='training', user='lutz')
    ftp.run(cleanTarget=ftp.cleanDir)
    print('Done:', ftp.fcount, 'files and', ftp.dcount,
          'directories cleaned.')
```

Помимо рекурсивного алгоритма, позволяющего обрабатывать деревья каталогов произвольной формы, главной хитростью здесь является анализ вывода содержимого каталога на сервере. Метод `nlst` класса `FTP`, использовавшийся нами ранее, возвращает простой список имен файлов. Здесь мы используем метод `dir`, чтобы получить дополнительную информацию, как показано ниже:

```
C:\...\PP4E\Internet\Ftp> ftp learning-python.com
ftp> cd training
ftp> dir
drwxr-xr-x   11 5693094  450       4096 May 4 11:06 .
drwx---r-x   19 5693094  450       8192 May 4 10:59 ..
```

```
-rw----r--    1 5693094  450      15825 May 4 11:02 2009-public-classes.htm
-rw----r--    1 5693094  450      18084 May 4 11:02 2010-public-classes.html
drwx---r-x    3 5693094  450       4096 May 4 11:02 books
-rw----r--    1 5693094  450       3783 May 4 11:02 calendar-save-aug09.html
-rw----r--    1 5693094  450       3923 May 4 11:02 calendar.html
drwx---r-x    2 5693094  450       4096 May 4 11:02 images
-rw----r--    1 5693094  450       6143 May 4 11:02 index.html
```

...часть строк опущена...

Формат вывода этого метода потенциально зависит от конкретного сервера, поэтому проверьте формат вывода на своем сервере, прежде чем приступать к использованию этого сценария. Для данного сервера провайдера, работающего под управлением Unix, если первый символ первого элемента строки является символом «d», это означает, что имя файла в конце строки является именем каталога. Анализ строки заключается в простом ее разбиении по пробельным символам и извлечении составляющих ее частей.

Обратите внимание, что этот сценарий, как и предшествующие ему, должен пропускать символические обозначения «.» и «..» текущего и родительского каталогов для корректной работы с данным сервером. Это может показаться странным, но появление этих имен в выводе также может зависеть от сервера – некоторые серверы, которые я использовал при опробовании примеров для этой книги, не включали эти специальные имена в списки. Мы можем проверить особенности сервера в этом отношении, используя `ftplib` в интерактивном сеансе, как если бы это был переносимый клиент FTP:

```
C:\...\PP4E\Internet\Ftp> python
>>> from ftplib import FTP
>>> f = FTP('ftp.rmi.net')
>>> f.login('lutz', 'xxxxxxxx')      # вывод строк опущен
>>> for x in f.nlst()[:3]: print(x)  # в списках отсутствуют имена . и ..
...
2004-longmont-classes.html
2005-longmont-classes.html
2006-longmont-classes.html

>>> L = []
>>> f.dir(L.append)              # тот же список, но с подробной информацией
>>> for x in L[:3]: print(x)
...
-rw-r--r--   1 ftp      ftp     8173 Mar 19 2006 2004-longmont-classes.html
-rw-r--r--   1 ftp      ftp     9739 Mar 19 2006 2005-longmont-classes.html
-rw-r--r--   1 ftp      ftp      805 Jul  8 2006 2006-longmont-classes.html
```

С другой стороны, сервер, который я использовал в этом разделе, включал специальные имена, состоящие из точек. Для надежности сценарий должен пропускать эти имена в списке, возвращаемом удаленным сервером, на случай, если он будет взаимодействовать с сервером, включающем их в список (здесь эта проверка обязательна, чтобы избежать

попадания в бесконечный рекурсивный цикл!). Подобную осторожность не требуется проявлять при работе со списками содержимого локальных каталогов, потому что функция os.listdir никогда не включает имена «.» и «..» в свои результаты, но положение вещей на «Диком Западе», каким сейчас является Интернет, не является таким же согласованным:

```
>>> f = FTP('learning-python.com')
>>> f.login('lutz', 'xxxxxxxx')      # вывод строк опущен
>>> for x in f.nlst()[:5]: print(x)  # включает имена . и .. здесь
...
.
..
.hcc.thumbs
2009-public-classes.htm
2010-public-classes.html

>>> L = []
>>> f.dir(L.append)              # тот же список, но с подробной информацией
>>> for x in L[:5]: print(x)
...
drwx---r-x  19 5693094  450     8192 May  4 10:59 .
drwx---r-x  19 5693094  450     8192 May  4 10:59 ..
drwx------   2 5693094  450     4096 Feb 18 05:38 .hcc.thumbs
-rw----r--   1 5693094  450    15824 May  1 14:39 2009-public-classes.htm
-rw----r--   1 5693094  450    18083 May  4 09:05 2010-public-classes.html
```

Ниже приводится вывод сценария cleanall – он появляется в окне консоли в процессе работы сценария. Добиться того же эффекта можно с помощью команды rm -rf Unix, выполнив ее в окне сеанса SSH или Telnet, но сценарий Python выполняется на стороне клиента и не требует наличия возможности удаленного доступа, кроме поддержки протокола FTP на стороне клиента:

```
C:\PP4E\Internet\Ftp\Mirror> cleanall.py
Password for lutz on learning-python.com:
connecting...
file 2009-public-classes.htm
file 2010-public-classes.html
file Learning-Python-interview.doc
file Python-registration-form-010.pdf
file PythonPoweredSmall.gif
directory _derived
file 2009-public-classes.htm_cmp_DeepBlue100_vbtn.gif
file 2009-public-classes.htm_cmp_DeepBlue100_vbtn_p.gif
file 2010-public-classes.html_cmp_DeepBlue100_vbtn_p.gif
file 2010-public-classes.html_cmp_deepblue100_vbtn.gif
directory _vti_cnf
file 2009-public-classes.htm_cmp_DeepBlue100_vbtn.gif
file 2009-public-classes.htm_cmp_DeepBlue100_vbtn_p.gif
file 2010-public-classes.html_cmp_DeepBlue100_vbtn_p.gif
```

```
file 2010-public-classes.html_cmp_deepblue100_vbtn.gif
directory exited
directory exited
```

...часть строк опущена...

```
file priorclients.html
file public_classes.htm
file python_conf_ora.gif
file topics.html
Done: 366 files and 18 directories cleaned.
```

Загрузка деревьев каталогов с сервера

Сценарий удаления деревьев каталогов на сервере можно также дополнить возможностью загрузки деревьев вместе с подкаталогами: в процессе обхода дерева каталогов на сервере достаточно вместо удаления просто создавать локальные каталоги, соответствующие удаленным, и загружать обычные файлы. Однако мы оставим этот заключительный шаг в качестве самостоятельного упражнения, отчасти потому, что зависимость от формата списка содержимого каталога, воспроизводимого сервером, усложняет возможность надежной реализации, а отчасти потому, что в моей практике эта операция практически не требуется. Для меня более типично вести разработку сайта на моем компьютере и выгружать дерево каталогов на сервер, чем загружать его оттуда.

Однако, если вы захотите поэкспериментировать с рекурсивной загрузкой, обязательно ознакомьтесь с подсказками в сценарии *Tools\Scripts\ftpmirror.py*, в каталоге установки Python или в дереве с исходными текстами. Этот сценарий реализует загрузку дерева каталогов с сервера по протоколу FTP и обрабатывает различные форматы списков содержимого каталогов, описание которых мы опустим в интересах экономии места в книге. Теперь пришло время перейти к следующему протоколу – электронной почты.

Обработка электронной почты

Имеется целый набор часто используемых высокоуровневых протоколов Интернета, которые предназначены для чтения и отправки сообщений электронной почты: POP и IMAP для получения почты с серверов, SMTP для отправки новых сообщений, а также дополнительные спецификации, такие как RFC822, которые определяют содержимое и формат сообщений электронной почты. При использовании стандартных инструментов электронной почты обычно нет необходимости знать о существовании этих акронимов, но для выполнения ваших запросов такие программы, как Microsoft Outlook и системы электронной почты с веб-интерфейсом, обычно обмениваются данными с серверами POP и SMTP.

Суть протоколов электронной почты, как и FTP, в конечном счете заключается в форматировании команд и потоков байтов, передаваемых через сокеты и порты (порт 110 для POP; 25 для SMTP). Независимо от природы содержимого и вложений сообщения электронной почты – это чуть больше, чем простые строки байтов, отправляемые и принимаемые через сокеты. Но, так же, как для FTP, в Python есть стандартные модули, которые упрощают все стороны обработки электронной почты:

- `poplib` и `imaplib` – для получения электронной почты
- `smtplib` – для отправки
- пакет `email` – для анализа и конструирования сообщений электронной почты

Эти модули связаны между собой: для обработки нетривиальных сообщений обычно используется пакет `email`, позволяющий проанализировать текст сообщения, полученного с помощью `poplib`, и сконструировать сообщение для отправки с помощью `smtplib`. Пакет `email` также позволяет решать такие задачи, как анализ адресов, форматирование даты и времени, форматирование и извлечение вложений, а также кодирование и декодирование содержимого сообщения электронной почты (например, uuencode, Base64). Для решения более специфических задач применяются другие модули (например, `mimetypes` для отображения имен файлов в их типы и обратно).

В следующих нескольких разделах мы исследуем интерфейсы POP и SMTP для получения и отправки почты на серверы и интерфейсы пакета `email` анализа и конструирования сообщений электронной почты. Другие интерфейсы электронной почты в Python аналогичны и описаны в справочном руководстве по библиотеке Python.[1]

Поддержка Юникода в Python 3.X и инструменты электронной почты

В предыдущих разделах этой главы мы рассматривали, какое влияние оказывают кодировки Юникода на использование инструментов FTP из модуля `ftplib`, потому что это иллюстрирует влияние модели строк Юникода в Python 3.X на практическое программирование. Вкратце:

---

[1] Протокол IMAP, или Internet Message Access Protocol (протокол доступа к сообщениям Интернета), был разработан как альтернатива протоколу POP, но он до сих пор не получил широкого распространения и поэтому не будет обсуждаться в этой книге. Например, основные коммерческие поставщики услуг, серверы которых используются в примерах этой книги, поддерживают только протокол POP (или веб-интерфейс) для доступа к электронной почте. Подробности, касающиеся серверного интерфейса к протоколу IMAP, вы найдете в руководстве по библиотеке Python. В Python также имеется модуль RFC822, но в версии 3.X он был интегрирован в пакет `email`.

- Для двоичных режимов передачи локальные входные и выходные файлы должны открываться в двоичном режиме (режимы `wb` и `rb`).
- При загрузке в текстовом режиме локальные выходные файлы должны открываться в текстовом режиме с явным указанием кодировки (режим `w`, с аргументом `encoding`, который по умолчанию принимает значение `latin1` в модуле `ftplib`).
- При выгрузке в текстовом режиме локальные входные файлы должны открываться в двоичном режиме (режим `rb`).

Мы уже выясняли, почему следует неукоснительно следовать этим правилам. Последние два пункта в этом списке отличаются для сценариев, изначально написанных для работы под управлением Python 2.X. Как вы уже наверняка догадались, учитывая, что данные через сокеты передаются в виде строк байтов, работа с электронной почтой также осложняется из-за особенностей поддержки Юникода в Python 3.X. В качестве предварительного знакомства:

Получение

Модуль `poplib` возвращает полученное сообщение электронной почты в виде строки `bytes`. Текст команд кодируется внутри модуля и передается серверу в кодировке `UTF8`, но ответы возвращаются как простые строки `bytes`, а не как декодированный текст `str`.

Отправка

Модуль `smtplib` принимает сообщение электронной почты для отправки в виде строки `str`. Перед передачей, внутри модуля, текст `str` кодируется в двоичное представление `bytes` с использованием схемы кодирования `ascii`. Имеется возможность передачи методу отправки уже закодированной строки `bytes`, что обеспечивает более явное управление кодированием.

Составление

При создании сообщений электронной почты, готовых к отправке с помощью модуля `smtplib`, пакет `email` воспроизводит строки `str` Юникода, содержащие простой текст, и принимает необязательные указания о кодировках для сообщений и их частей, которые будут применяться в соответствии со стандартами электронной почты. Заголовки сообщений также могут кодироваться в соответствии с соглашениями электронной почты, MIME и Юникода.

Анализ

Пакет `email` в версии 3.1 в настоящее время требует, чтобы строки байтов, возвращаемые модулем `poplib`, были декодированы в строки Юникода `str` до того, как будут переданы объекту сообщения для анализа. Такое декодирование перед анализом может выполняться по умолчанию, исходя из предпочтений пользователя, содержимого заголовков или иных обоснованных предположений. Поскольку данное требование порождает сложные проблемы для клиентов пакета, оно может быть снято в будущих версиях `email` и Python.

Навигация

Пакет `email` возвращает большинство компонентов сообщений в виде строк `str`. Однако части сообщений, декодированные с применением Base64 и других схем кодирования электронной почты, могут возвращаться в виде строк `bytes`. Части, извлекаемые без такого декодирования, могут возвращаться в виде строк `str` или `bytes`, при этом некоторые части в виде строк `str` перед обработкой могут кодироваться внутри пакета в строки `bytes` с использованием схемы `raw-unicode-escape`. Заголовки сообщения также могут декодироваться пакетом по запросу.

Если вы переносите свои сценарии электронной почты (или свой образ мышления) с версии 2.X, вам следует интерпретировать текст сообщений электронной почты, получаемых с сервера, как строки байтов и декодировать их перед анализом. Обычно это не касается сценариев, которые составляют и отправляют сообщения электронной почты (это, возможно, подавляющее большинство сценариев на языке Python, поддерживающих электронную почту), однако содержимое сообщений может потребоваться интерпретировать особым образом, если оно может возвращаться в виде строк байтов.

Таково положение дел в Python 3.1, которое, конечно же, может измениться со временем. Далее в этом разделе будет показано, как эти ограничения воплощаются в программном коде. Достаточно сказать, что текст в Интернете уже не так прост, как прежде, хотя, вероятно, он таковым и не должен был быть.

POP: чтение электронной почты

Признаюсь, что вплоть до 2000 года я использовал требующий минимального вовлечения подход к электронной почте. Я предпочитал читать свою электронную почту, подключаясь к провайдеру по Telnet и используя простой интерфейс командной строки. Конечно, для почты с вложениями, картинками это не идеальный вариант, но переносимость впечатляет – поскольку Telnet работает почти на любом компьютере, подключенном к сети, я мог быстро и просто проверить свою почту, находясь в любой точке Земного Шара. С учетом того, что моя жизнь проходит в постоянных поездках по всему миру с преподаванием языка Python, такая широкая доступность была большим преимуществом.

Как уже говорилось в разделе со сценариями для копирования веб-сайтов, времена изменились. В какой-то момент большинство провайдеров стало предоставлять веб-интерфейс для доступа к электронной почте, обеспечивающий такой же уровень переносимости, при этом одновременно закрыв доступ по Telnet. Когда мой провайдер закрыл доступ по Telnet, я лишился также доступа к электронной почте. К счастью, и здесь на помощь пришел Python – создав сценарии Python для доступа к электронной почте, я снова могу читать и отправлять сообщения

с любого компьютера, где есть Python и соединение с Интернетом. Python может быть таким же переносимым решением, как Telnet, но намного более мощным.

Кроме того, я могу использовать эти сценарии в качестве альтернативы средствам, которые предлагает мне провайдер. Помимо того, что я не большой любитель передавать средства управления коммерческим продуктам больших компаний, возможности закрытых инструментов электронной почты, предлагаемые пользователям, не всегда идеальны и иногда их бывает недостаточно. Во многом мотивация к созданию сценариев Python для работы с электронной почты остается той же, что для крупных графических интерфейсов, представленных в главе 11: *простота изменения* программ на языке Python может оказаться решающим преимуществом.

Например, Microsoft Outlook по умолчанию загружает почту на ваш ПК и удаляет ее с почтового сервера после обращения к нему. В результате ваш почтовый ящик занимает мало места (и ваш провайдер доволен), но по отношению к путешественникам, пользующимся разными компьютерами, это не очень вежливо – получив доступ к письму, вы уже не сможете сделать это повторно, кроме как с того же компьютера, куда оно первоначально было загружено. Хуже того, веб-интерфейс к почтовому ящику, предлагаемый моим провайдером, иногда оказывается недоступным, что оставляет меня отрезанным от электронной почты (и, как правило, это случается в самый неподходящий момент).

Следующие два сценария представляют возможное решение таких проблем переносимости и надежности (с другими решениями мы познакомимся далее в этой и последующих главах). Первый из них, *popmail.py*, является простым инструментом чтения почты, который загружает и выводит содержимое каждого сообщения, находящегося в почтовом ящике. Этот сценарий явно примитивен, но позволяет читать электронную почту с любого компьютера, где имеется Python и сокеты. Кроме того, он оставляет почту на сервере в целости. Второй сценарий, *smtpmail.py*, служит одновременно для создания и отправки новых сообщений электронной почты.

Далее в этой главе мы реализуем интерактивный клиент электронной почты командной строки (pymail), а впоследствии создадим полнофункциональный инструмент электронной почты с графическим интерфейсом (PyMailGUI) и собственную программу с веб-интерфейсом (PyMailCGI). А пока начнем с самых основ.

Модуль настройки электронной почты

Прежде чем перейти к сценариям, рассмотрим общий модуль, который они импортируют и используют. Модуль в примере 13.17 используется для настройки параметров электронной почты для конкретного пользователя. Это просто ряд инструкций присваивания значений переменным, используемым во всех почтовых программах, представленных

в этой книге, – каждый почтовый клиент пользуется собственной версией этого модуля, благодаря чему содержимое версий может отличаться. Выделение параметров настройки в отдельный модуль упрощает настройку представленных в книге почтовых программ для использования конкретным пользователем и ликвидирует необходимость редактировать логику самих программ.

Если вы захотите пользоваться какими-либо почтовыми программами из этой книги для работы со своей электронной почтой, исправьте инструкции присваивания в этом модуле, чтобы они отражали ваши серверы, названия учетных записей и так далее (в представленном виде они соответствуют учетным записям электронной почты, использовавшимся при работе над книгой). Не все настройки из этого модуля используются в последующих сценариях. Некоторые из них будут описываться, когда мы вернемся к этому модулю в более поздних примерах.

Обратите внимание, что некоторые провайдеры могут требовать, чтобы вы подключались непосредственно к их системам для использования их серверов SMTP при отправке электронной почты. Например, в прошлом, когда я пользовался коммутируемым подключением, я мог использовать сервер моего провайдера непосредственно, но после перехода на широкополосное подключение я должен был направлять запросы через провайдера кабельного подключения к Интернету. Вам может потребоваться изменить эти настройки в соответствии с параметрами вашего подключения – обращайтесь к своему провайдеру Интернета, чтобы получить параметры доступа к серверам POP и SMTP. Кроме того, некоторые серверы SMTP могут проверять доменные имена в адресах и могут требовать проходить процедуру аутентификации – подробности смотрите в разделе, посвященном протоколу SMTP далее в этой главе.

Пример 13.17. PP4E\Internet\Email\mailconfig.py

```
"""
пользовательские параметры настройки для различных программ электронной
почты (версия pymail/mailtools); сценарии электронной почты получают
имена серверов и другие параметры настройки из этого модуля: измените его,
чтобы он отражал имена ваших серверов и ваши предпочтения;
"""


#-----------------------------------------------------------------------
# (требуется для загрузки, удаления: для всех) имя сервера POP3,
# имя пользователя
#-----------------------------------------------------------------------

popservername = 'pop.secureserver.net'
popusername = 'PP4E@learning-python.com'

#-----------------------------------------------------------------------
# (требуется для отправки: для всех) имя сервера SMTP
# смотрите модуль Python smtpd, где приводится класс сервера SMTP,
```

```
# выполняемого локально;
#-----------------------------------------------------------------------------

smtpservername = 'smtpout.secureserver.net'

#-----------------------------------------------------------------------------
# (необязательные параметры: для всех) персональная информация,
# используемая клиентами для заполнения полей в сообщениях,
# если эти параметры определены;
# подпись - может быть блоком в тройных кавычках, игнорируется,
# если пустая строка;
# адрес   - используется в качестве начального значения поля "From",
# если непустая строка, больше не пытается определить значение
# поля From для ответов: это имело переменный успех;
#-----------------------------------------------------------------------------

myaddress = 'PP4E@learning-python.com'
mysignature = ('Thanks,\n'
    '--Mark Lutz (http://learning-python.com, http://rmi.net/~lutz)')

#-----------------------------------------------------------------------------
# (необязательные параметры: mailtools) могут потребоваться для отправки;
# имя пользователя/пароль SMTP, если требуется аутентификация;
# если аутентификация не требуется, установите переменную smtpuser
# в значение None или ''; в переменную smtppasswdfile запишите имя файла
# с паролем SMTP или пустую строку, чтобы вынудить программу
# запрашивать пароль (в консоли или в графическом интерфейсе);
#-----------------------------------------------------------------------------

smtpuser = None        # зависит от провайдера
smtppasswdfile = ''    # установите в значение '', чтобы обеспечить запрос

#-----------------------------------------------------------------------------
# (необязательный параметр: mailtools) имя локального однострочного
# текстового файла с паролем pop; если содержит пустую строку или файл
# не может быть прочитан, при первом соединении пароль запрашивается
# у пользователя; пароль не шифруется: оставьте это значение пустым
# на компьютерах общего пользования;
#-----------------------------------------------------------------------------

poppasswdfile = r'c:\temp\pymailgui.txt' # установите в значение '',
                                         # чтобы обеспечить запрос

#-----------------------------------------------------------------------------
# (обязательный параметр: mailtools) локальный файл, где некоторые клиенты
# сохраняют отправленные сообщения;
#-----------------------------------------------------------------------------

sentmailfile = r'.\sentmail.txt'        # . означает текущий рабочий каталог
```

```
#-----------------------------------------------------------------------
# (обязательный параметр: pymail, pymail2) локальный файл, где pymail
# сохраняет почту по запросу;
#-----------------------------------------------------------------------

savemailfile = r'c:\temp\savemail.txt'  # не используется в PyMailGUI: диалог

#-----------------------------------------------------------------------
# (обязательные параметры: pymail, mailtools) fetchEncoding -
# это имя кодировки, используемой для декодирования байтов сообщения,
# а также для кодирования/декодирования текста сообщения, если они
# сохраняются в текстовых файлах; подробности смотрите в главе 13:
# это временное решение, пока не появится новый пакет email,
# с более дружественным отношением к строкам bytes;
# headersEncodeTo - для отправки заголовков: смотрите главу 13;
#-----------------------------------------------------------------------

fetchEncoding = 'utf8' # 4E: как декодировать и хранить текст сообщений
                       # (или latin1?)
headersEncodeTo = None # 4E: как кодировать не-ASCII заголовки при отправке
                       # (None=utf8)

#-----------------------------------------------------------------------
# (необязательный параметр: mailtools)максимальное количество заголовков
# или сообщений, загружаемых за один запрос; если указать значение N,
# mailtools будет извлекать не более N самых последних электронных писем;
# более старые сообщения, не попавшие в это число, не будут извлекаться
# с сервера, но будут возвращаться как пустые письма;
# если этой переменной присвоить значение None (или 0), будут загружаться
# все сообщения; используйте этот параметр, если вам может поступать
# очень большое количество писем, а ваше подключение к Интернету
# или сервер слишком медленные, чтобы можно было выполнить загрузку
# сразу всех сообщений; кроме того, некоторые клиенты загружают
# только самые свежие письма, но этот параметр никак не связан
# с данной особенностью;
#-----------------------------------------------------------------------

fetchlimit = 25    # 4E: максимальное число загружаемых заголовков/сообщений
```

Сценарий чтения почты с сервера POP

Переходим к чтению электронной почты в Python: сценарий в примере 13.18 пользуется стандартным модулем Python `poplib`, реализующим интерфейс клиента POP – Post Office Protocol (почтовый протокол). POP четко определяет способ получения электронной почты с сервера через сокеты. Данный сценарий соединяется с сервером POP, реализуя простой, но переносимый инструмент загрузки и отображения электронной почты.

Пример 13.18. PP4E\Internet\Email\popmail.py

```
#!/usr/local/bin/python
"""
##############################################################################
использует модуль POP3 почтового интерфейса Python для просмотра сообщений
почтовой учетной записи pop; это простой листинг - смотрите реализацию
клиента с большим количеством функций взаимодействий с пользователем
в pymail.py и сценарий отправки почты в smtpmail.py; протокол POP
используется для получения почты и на сервере выполняется на сокете
с портом номер 110, но модуль Python poplib скрывает все детали протокола;
для отправки почты используйте модуль smtplib (или os.popen('mail...')).
Смотрите также модуль imaplib, реализующий альтернативный протокол IMAP,
и программы PyMailGUI/PyMailCGI, реализующие дополнительные особенности;
##############################################################################
"""

import poplib, getpass, sys, mailconfig

mailserver = mailconfig.popservername  # например: 'pop.rmi.net'
mailuser   = mailconfig.popusername     # например: 'lutz'
mailpasswd = getpass.getpass('Password for %s?' % mailserver)

print('Connecting...')
server = poplib.POP3(mailserver)
server.user(mailuser)                   # соединение, регистрация на сервере
server.pass_(mailpasswd)                # pass - зарезервированное слово

try:
    print(server.getwelcome())          # вывод приветствия
    msgCount, msgBytes = server.stat()
    print('There are', msgCount, 'mail messages in', msgBytes, 'bytes')
    print(server.list())
    print('-' * 80)
    input('[Press Enter key]')

    for i in range(msgCount):
        hdr, message, octets = server.retr(i+1)    # octets - счетчик байтов
        for line in message: print(line.decode())  # читает, выводит
                                                   # все письма
        print('-' * 80)                 # в 3.X сообщения - bytes
        if i < msgCount - 1:            # почтовый ящик блокируется
            input('[Press Enter key]')  # до вызова quit
finally:                                # снять блокировку с ящика
    server.quit()                       # иначе будет разблокирован
print('Bye.')                           # по тайм-ауту
```

Несмотря на свою примитивность, этот сценарий иллюстрирует основы чтения электронной почты в Python. Чтобы установить соединение с почтовым сервером, мы сначала создаем экземпляр объекта `poplib.POP3`, передавая ему имя сервера в виде строки:

```
server = poplib.POP3(mailserver)
```

Если этот вызов не возбудил исключение, значит, мы соединились (через сокет) с POP-сервером, который слушает запросы на порту POP с номером 110 на сервере, где у нас имеется учетная запись электронной почты.

Следующее, что необходимо сделать перед получением сообщений, – это сообщить серверу имя пользователя и пароль. Обратите внимание, что метод передачи пароля называется pass_ – без символа подчеркивания в конце идентификатор pass оказался бы зарезервированным словом и вызвал бы синтаксическую ошибку:

```
server.user(mailuser)                  # соединение, регистрация на сервере
server.pass_(mailpasswd)               # pass - зарезервированное слово
```

Для простоты и относительной безопасности этот сценарий всегда запрашивает пароль в интерактивном режиме. Чтобы ввести, но не отображать на экране строку пароля при вводе пользователем, используется модуль getpass, знакомый по разделу этой главы, посвященному FTP.

Сообщив серверу имя пользователя и пароль, мы можем с помощью метода stat получить информацию о почтовом ящике (количество сообщений, суммарное количество байтов в сообщениях) и извлечь полный текст конкретного сообщения с помощью метода retr (передав номер сообщения – нумерация начинается с 1). Полный текст включает все заголовки, следующую за ними пустую строку, текст письма и все вложения. Метод retr возвращает кортеж, включающий список строк с содержимым письма:

```
msgCount, msgBytes = server.stat()
hdr, message, octets = server.retr(i+1)   # octets - счетчик байтов
```

Закончив работу с почтой, закрываем соединение с почтовым сервером, вызывая метод quit объекта POP:

```
server.quit()          # иначе будет разблокирован по тайм-ауту
```

Обратите внимание, что этот вызов помещен в предложение finally инструкции try, охватывающей основную часть сценария. Для борьбы со сложностями, связанными с изменениями данных, POP-серверы блокируют почтовый ящик на время между первым соединением и закрытием соединения (или до истечения устанавливаемого системой тайм-аута). Так как метод quit тоже разблокирует почтовый ящик, важно выполнить его перед завершением работы независимо от того, было ли во время обработки электронной почты возбуждено исключение. Заключив действия в инструкцию try/finally, мы гарантируем вызов метода quit при выходе, чтобы разблокировать почтовый ящик и сделать его доступным другим процессам (например, для доставки входящей почты).

Извлечение сообщений

Ниже приводится сеанс работы сценария popmail из примера 13.18, во время которого выводятся два сообщения из моего почтового ящика на сервере *pop.secureserver.net* – доменное имя почтового сервера моего провайдера, где располагается сайт с доменным именем *learning-python.com*, указанное в модуле mailconfig. Чтобы сократить размер вывода, я опустил или обрезал несколько строк заголовков сообщений, не относящихся к делу, включая типичные заголовки Received:, которые описывают хронологию передачи электронной почты. Запустите этот сценарий у себя, чтобы увидеть все многочисленные подробности, входящие в состав сообщения электронной почты:

```
C:\...\PP4E\Internet\Email> popmail.py
Password for pop.secureserver.net?
Connecting...
b'+OK <1314.1273085900@p3pop01-02.prod.phx3.gdg>'
There are 2 mail messages in 3268 bytes
(b'+OK ', [b'1 1860', b'2 1408'], 16)
------------------------------------------------------------------------------
[Press Enter key]
Received: (qmail 7690 invoked from network); 5 May 2010 15:29:43 -0000
X-IronPort-Anti-Spam-Result: AskCAG4r4UvRVllAlGdsb2JhbACDF44FjCkVAQEBAQkLC
AkRAx+
Received: from 72.236.109.185 by webmail.earthlink.net with HTTP; Wed, 5 May
201
Message-ID: <27293081.1273073376592.JavaMail.root@mswamui-thinleaf.atl.sa.
earthl
Date: Wed, 5 May 2010 11:29:36 -0400 (EDT)
From: lutz@rmi.net
Reply-To: lutz@rmi.net
To: pp4e@learning-python.com
Subject: I'm a Lumberjack, and I'm Okay
Mime-Version: 1.0
Content-Type: text/plain; charset=UTF-8
Content-Transfer-Encoding: 7bit
X-Mailer: EarthLink Zoo Mail 1.0
X-ELNK-Trace: 309f369105a89a174e761f5d55cab8bca866e5da7af650083cf64d888edc
8b5a35
X-Originating-IP: 209.86.224.51
X-Nonspam: None

I cut down trees, I skip and jump,
I like to press wild flowers...


------------------------------------------------------------------------------
[Press Enter key]
Received: (qmail 17482 invoked from network); 5 May 2010 15:33:47 -0000
X-IronPort-Anti-Spam-Result: AlIBAIss4UthSoc7mWdsb2JhbACDF44FjD4BAQEBAQYNC
gcRIq1
```

```
Received: (qmail 4009 invoked by uid 99); 5 May 2010 15:33:47 -0000
Content-Transfer-Encoding: quoted-printable
Content-Type: text/plain; charset="utf-8"
X-Originating-IP: 72.236.109.185
User-Agent: Web-Based Email 5.2.13
Message-Id: <20100505083347.deec9532fd532622acfef00cad639f45.0371a89d29.
wbe@emai
From: lutz@learning-python.com
To: PP4E@learning-python.com
Cc: lutz@learning-python.com
Subject: testing
Date: Wed, 05 May 2010 08:33:47 -0700
Mime-Version: 1.0
X-Nonspam: None

Testing Python mail tools.


---------------------------------------------------------------------------
Bye.
```

Этот пользовательский интерфейс предельно прост – после соединения с сервером он выводит полный необработанный текст очередного сообщения, делая после каждого сообщения остановку и ожидая нажатия клавиши Enter. Для ожидания нажатия клавиши между выводом сообщений вызывается встроенная функция `input`. Благодаря этой паузе не происходит слишком быстрой прокрутки сообщений на экране. Чтобы зрительно выделить сообщения, они также отделяются строками дефисов.

Можно было бы сделать отображение более изощренным (например, мы могли бы с помощью пакета `email` анализировать заголовки, тела и вложения – смотрите примеры в этой и последующих главах), но здесь мы просто выводим целиком отправленное сообщение. Этот сценарий успешно справляется с простыми сообщениями, такими как эти два, но его неудобно использовать для чтения больших сообщений с вложениями – мы внесем дополнительные улучшения в этом направлении в последующих клиентах.

В этой книге не рассматриваются все возможные заголовки, которые только могут появляться в сообщениях электронной почты, тем не менее, некоторые из них мы будем использовать. Например, строка заголовка *X-Mailer*, если она присутствует, обычно идентифицирует программу-отправитель. Мы будем использовать этот заголовок для идентификации почтового клиента, написанного на языке Python. Типичные заголовки, такие как *From* и *Subject*, являются более важными для сообщений. На самом деле в тексте сообщения может пересылаться множество других заголовков. Например, заголовки *Received* трассируют те серверы, через которые прошло сообщение по пути к целевому почтовому ящику.

Поскольку сценарий popmail выводит весь текст сообщения в необработанном виде, здесь видны все заголовки, но в почтовых клиентах с графическим интерфейсом, ориентированных на конечного пользователя, таких как Outlook или веб-интерфейс, по умолчанию можно видеть лишь некоторые из них. Кроме того, по необработанному тексту сообщений отчетливо видно структуру электронного письма, о которой рассказывалось выше: электронное сообщение в целом состоит из множества заголовков, как в этом примере, за которыми следует пустая строка и основной текст сообщения. Однако, как мы увидим далее, они могут быть более сложными, если присутствуют альтернативные части или вложения.

Сценарий в примере 13.18 никогда не удаляет почту с сервера. Почта лишь загружается и выводится на экран и будет снова показана при следующем запуске сценария (если, конечно, не удалить ее другими средствами). Чтобы действительно удалить почту, требуется вызывать другие методы (например, server.dele(msgnum)), но такие возможности лучше отложить до момента разработки более интерактивных почтовых инструментов.

Обратите внимание, сценарий чтения декодирует каждую строку содержимого сообщения вызовом метода line.decode в строку str перед отображением. Как уже упоминалось выше, модуль poplib в версии 3.X возвращает содержимое в виде строк bytes. Фактически, если убрать вызов метода декодирования из сценария, это сразу отразится на его выводе:

```
[Press Enter key]
...часть строк опущена...
b'Date: Wed, 5 May 2010 11:29:36 -0400 (EDT)'
b'From: lutz@rmi.net'
b'Reply-To: lutz@rmi.net'
b'To: pp4e@learning-python.com'
b"Subject: I'm a Lumberjack, and I'm Okay"
b'Mime-Version: 1.0'
b'Content-Type: text/plain; charset=UTF-8'
b'Content-Transfer-Encoding: 7bit'
b'X-Mailer: EarthLink Zoo Mail 1.0'
b''
b'I cut down trees, I skip and jump,'
b'I like to press wild flowers...'
b''
```

Как будет показано далее, нам придется выполнять подобное декодирование, чтобы обеспечить возможность анализа этого текста с помощью инструментов из пакета email. Также в следующем разделе демонстрируется интерфейс, основанный на использовании строк байтов.

Чтение почты из интерактивной оболочки

Если вы не гнушаетесь вводить программный код и читать необработанные сообщения POP-сервера, тогда в качестве простого клиента электронной почты можно использовать даже интерактивную оболочку Python. В следующем сеансе используются два дополнительных интерфейса, которые мы задействуем в последующих примерах:

`conn.list()`

Возвращает список строк «номер-сообщения размер-сообщения».

`conn.top(N , 0)`

Извлекает только текстовые части заголовков для сообщения с номером *N*.

Метод `top` также возвращает кортеж, включающий список строк тела сообщения. Его второй аргумент сообщает серверу, сколько дополнительных строк после блока заголовков следует вернуть. Если вам необходимо только получить заголовки, метод `top` может оказаться более быстрым решением, чем извлекающий полный текст письма метод `retr`, – при условии, что ваш почтовый сервер реализует команду TOP (большинство серверов реализуют ее):

```
C:\...\PP4E\Internet\Email> python
>>> from poplib import POP3
>>> conn = POP3('pop.secureserver.net')    # соединиться с сервером
>>> conn.user('PP4E@learning-python.com')  # зарегистрироваться
b'+OK '
>>> conn.pass_('xxxxxxxx')
b'+OK '

>>> conn.stat()                            # кол-во сообщений, кол-во байтов
(2, 3268)
>>> conn.list()
(b'+OK ', [b'1 1860', b'2 1408'], 16)

>>> conn.top(1, 0)
(b'+OK 1860 octets ', [b'Received: (qmail 7690 invoked
                                   from network); 5 May 2010
...часть строк опущена...
b'X-Originating-IP: 209.86.224.51', b'X-Nonspam: None', b'', b''], 1827)

>>> conn.retr(1)
(b'+OK 1860 octets ', [b'Received: (qmail 7690 invoked
                                   from network); 5 May 2010
...часть строк опущена...
b'X-Originating-IP: 209.86.224.51', b'X-Nonspam: None', b'',
b'I cut down trees, I skip and jump,', b'I like to press wild flowers...',
b'', b''], 1898)

>>> conn.quit()
b'+OK '
```

Вывести полный текст сообщения в интерактивной оболочке после его получения достаточно легко: просто декодируйте каждую строку в обычное представление в процессе вывода, как это делает наш сценарий popmail, или объедините строки, возвращаемые методом retr или top, добавляя символы перевода строки между ними. Подойдет любой из следующих способов обработки открытого объекта соединения с POP-сервером:

```
>>> info, msg, oct = connection.retr(1)       # извлечь первое сообщение

>>> for x in msg: print(x.decode())           # четыре способа отображения
>>> print(b'\n'.join(msg).decode())           # строк сообщения
>>> x = [print(x.decode()) for x in msg]
>>> x = list(map(print, map(bytes.decode, msg)))
```

Анализ сообщения электронной почты с целью извлечь заголовки и компоненты реализовать значительно сложнее, особенно когда сообщения могут содержать вложения и кодированные части, такие как изображения. Как будет показано далее в этой главе, анализ полного текста сообщения или его заголовков после получения с помощью модуля poplib (или imaplib) можно реализовать с помощью стандартного пакета email.

Описание других инструментов, имеющихся в стандартном модуле поддержки протокола POP, вы найдете в руководстве по библиотеке Python. Начиная с версии 2.4, в модуле poplib имеется также класс POP3_SSL, который соединяется через сокет с SSL-шифрованием с портом 995 сервера (стандартный номер порта при использовании протокола POP через SSL). Он имеет идентичный интерфейс, но использует безопасные сокеты для обмена данными с серверами, поддерживающими такую возможность.

SMTP: отправка электронной почты

Среди хакеров бытует высказывание, что каждая полезная компьютерная программа рано или поздно дорастает до реализации возможности отправки электронной почты. Независимо от того, оправдывается ли эта старинная мудрость на практике, возможность автоматического создания в программе электронных писем является мощным инструментом.

Например, системы тестирования могут автоматически посылать по электронной почте сообщения об отказах, программы с пользовательским интерфейсом могут посылать заказы на поставки и так далее. Кроме того, переносимый почтовый сценарий на языке Python можно использовать для отправки сообщений с любого компьютера, где есть Python и соединение с Интернетом. Независимость от почтовых программ, таких как Outlook, весьма привлекательна, когда вы зарабатываете себе на жизнь, разъезжая по свету и обучая языку Python на самых разнообразных компьютерах.

К счастью, отправить почту из сценария Python столь же просто, как и прочесть ее. Фактически существует не менее четырех способов сделать это:

Вызов `os.popen` *для запуска почтовой программы командной строки*

На некоторых системах отправить электронное письмо из сценария можно с помощью вызова следующего вида:

```
os.popen('mail -s "xxx" a@b.c', 'w').write(text)
```

Как мы уже видели ранее, функция `popen` выполняет командную строку, переданную ей в первом аргументе, и возвращает объект, похожий на файл, соединенный с запущенной программой. При открытии в режиме `w` происходит соединение со стандартным потоком ввода программы – здесь мы записываем текст нового почтового сообщения в поток ввода программы `mail` – стандартной в системе Unix. Результат получается таким же, каким он был бы при запуске `mail` в интерактивном режиме, но достигается внутри работающего сценария Python.

Запуск программы `sendmail`

Программа с открытыми исходными текстами `sendmail` предоставляет еще один способ отправки электронной почты из сценария. Если она установлена в вашей системе и настроена, ее можно запускать с помощью таких средств Python, как функция `os.popen` из предыдущего абзаца.

Применение стандартного модуля Python `smtplib`

Стандартная библиотека Python поставляется с поддержкой клиентского интерфейса к SMTP – Simple Mail Transfer Protocol (простой протокол передачи почты) – стандарту Интернета высокого уровня для отправки почты через сокеты. Подобно модулю `poplib`, который мы рассматривали в предыдущем разделе, `smtplib` скрывает все детали, касающиеся сокетов и протокола, и может применяться для отправки почты с любого компьютера, где есть Python и соединение с Интернетом на базе сокетов.

Получение и использование пакетов и инструментов сторонних разработчиков

В библиотеке программного обеспечения с открытыми исходными текстами есть и другие пакеты Python для обработки почты на более высоком уровне. В большинстве своем они основаны на одном из трех вышеперечисленных приемов.

Из всех четырех вариантов наиболее переносимым и непосредственным является использование модуля `smtplib`. Запуск почтовой программы с помощью `os.popen` обычно действует только в Unix-подобных системах, но не в Windows (требуется наличие почтовой программы командной строки). Программа `sendmail`, хотя и обладает значительной мощью,

также имеет определенный уклон в сторону Unix, сложна и может быть установлена не во всех Unix-подобных системах.

Напротив, модуль smtplib работает на любом компьютере, где есть Python и соединение с Интернетом, поддерживающее доступ по SMTP, в том числе в Unix, Linux и Windows. Он отправляет электронную почту через сокеты внутри самого процесса, вместо того чтобы запускать отдельную программу. Кроме того, SMTP позволяет в значительной мере управлять форматированием и маршрутизацией электронной почты.

Сценарий отправки электронной почты по SMTP

Поскольку можно утверждать, что SMTP – это лучший вариант отправки почты из сценариев на языке Python, рассмотрим простую почтовую программу, иллюстрирующую его интерфейсы. Сценарий Python, представленный в примере 13.19, предназначен для использования из интерактивной командной строки. Он читает новое почтовое сообщение, вводимое пользователем, и отправляет почту по SMTP с помощью модуля Python smtplib.

Пример 13.19. PP4E\Internet\Email\smtpmail.py

```
#!/usr/local/bin/python
"""
##############################################################################
использует модуль Python почтового интерфейса SMTP для отправки сообщений;
это простой сценарий одноразовой отправки - смотрите pymail, PyMailGUI
и PyMailCGI, реализующие клиентов с более широкими возможностями
взаимодействия с пользователями; смотрите также popmail.py - сценарий
получения почты, и пакет mailtools, позволяющий добавлять вложения
и форматировать сообщения с помощью стандартного пакета email;
##############################################################################
"""

import smtplib, sys, email.utils, mailconfig
mailserver = mailconfig.smtpservername    # например: smtp.rmi.net

From = input('From? ').strip()      # или импортировать из mailconfig
To   = input('To? ').strip()        # например: python-list@python.org
Tos  = To.split(';')                # допускается список получателей
Subj = input('Subj? ').strip()
Date = email.utils.formatdate()     # текущие дата и время, rfc2822

# стандартные заголовки, за которыми следует пустая строка и текст
text = ('From: %s\nTo: %s\nDate: %s\nSubject: %s\n\n' % (From, To,
                                                         Date, Subj))
print('Type message text, end with line=[Ctrl+d (Unix), Ctrl+z (Windows)]')
while True:
    line = sys.stdin.readline()
    if not line:
        break                       # выход по ctrl-d/z
```

```
        #if line[:4] == 'From':
        #    line = '>' + line           # серверы могут экранировать
         text += line

    print('Connecting...')
    server = smtplib.SMTP(mailserver)   # соединиться без регистрации
    failed = server.sendmail(From, Tos, text)
    server.quit()
    if failed:                          # smtplib может возбуждать исключения
        print('Failed recipients:', failed)   # но здесь они не обрабатываются
    else:
        print('No errors.')
    print('Bye.')
```

Большая часть этого сценария реализует интерфейс пользователя – вводится адрес отправителя («From»), один или более адресов получателя («To», разделенные символом «;») и строка темы сообщения. Дата отправки формируется с помощью стандартного модуля Python `time`, строки стандартных заголовков форматируются, и цикл `while` читает строки сообщения, пока пользователь не введет символ конца файла (`Ctrl+Z` в Windows, `Ctrl+D` в Linux).

Для надежности убедитесь, что добавили пустую строку между строками заголовков и текстом тела сообщения, – этого требует протокол SMTP, и некоторые серверы SMTP считают это обязательным. Наш сценарий удовлетворяет это требование, вставляя пустую строку в виде пары символов `\n\n` в конце строки выражения форматирования – первый символ `\n` завершает текущую строку, а второй образует пустую строку. Модуль `smtplib` преобразует символы `\n` в пары символов `\r\n` в стиле Интернета перед передачей, поэтому здесь вполне можно использовать краткую форму. Далее в этой главе мы будем формировать наши сообщения с помощью пакета Python `email`, который автоматически выполняет такие тонкости.

В остальной части сценария происходят все чудеса SMTP: для отправки почты по SMTP нужно выполнить следующие два типа вызовов:

`server = smtplib.SMTP(mailserver)`

Создать экземпляр объекта SMTP, передав ему имя сервера SMTP, который первым отправит сообщение. Если при этом не возникнет исключения, значит, при возврате из конструктора вы окажетесь соединены с SMTP-сервером через сокет. Технически соединение с сервером устанавливает метод `connect`, но конструктор объекта SMTP автоматически вызывает этот метод, когда ему передается имя сервера.

`failed = server.sendmail(From, Tos, text)`

Вызвать метод `sendmail` объекта `SMTP` с передачей ему адреса отправителя, одного или более адресов получателя и собственно текста сообщения со всеми строками стандартных почтовых заголовков, какие вы зададите.

Закончив передачу, обязательно вызовите метод quit, чтобы отключиться от сервера и завершить сеанс. Обратите внимание, что при неудаче метод sendmail может возбудить исключение или возвратить список адресов получателей, которые не были приняты; сценарий обрабатывает последний случай, но позволяет исключительным ситуациям прервать работу сценария с выводом сообщения об ошибке.

Одна тонкая особенность: вызов метода quit объекта SMTP после возбуждения исключения модулем sendmail может иногда не действовать – метод quit фактически может зависнуть до истечения тайм-аута, если операция передачи потерпела неудачу и оставила интерфейс в непредусмотренном состоянии. Например, такое положение вещей может возникнуть в случае появления ошибок кодирования Юникода при трансляции исходящего сообщения в байты по схеме ASCII (запрос сброса rset в данном случае также зависает). Альтернативный метод close просто закрывает сокет на стороне клиента, не пытаясь отправить команду завершения на сервер – метод quit вызывает метод close на последнем этапе (предполагается, что команда завершения может быть отправлена!).

Дополнительно объекты SMTP предоставляют методы, не использованные в этом примере:

- server.login(user, password) предоставляет интерфейс к серверам SMTP, которые требуют и поддерживают *аутентификацию*; пример использования этого метода вы найдете в примере пакета mailtools далее в этой главе.
- server.starttls([keyfile[, certfile]]) переводит соединение SMTP в режим соблюдения безопасности на транспортном уровне (Transport Layer Security, TLS) – все команды шифруются с использованием модуля ssl, реализующего обертку SSL сокетов, при условии, что сервер поддерживает этот режим.

Обращайтесь к руководству по библиотеке Python за дополнительной информацией об этих и других методах, не упомянутых здесь.

Отправка сообщений

Попробуем послать несколько сообщений. Сценарий smtpmail является одноразовым инструментом: при каждом запуске сценария можно послать только одно сообщение. Как и большинство клиентских инструментов, представленных в этой главе, его можно запустить на любом компьютере с Python и соединением с Интернетом, поддерживающим SMTP (многие компьютеры имеют подключение к Интернету, но многие компьютеры общего доступа ограничены возможностью работы только с протоколом HTTP [Веб] или требуют специальной настройки доступа к серверу SMTP). Ниже приводится пример его выполнения в Windows:

```
C:\...\PP4E\Internet\Email> smtpmail.py
From? Eric.the.Half.a.Bee@yahoo.com
```

```
To?   PP4E@learning-python.com
Subj? A B C D E F G
Type message text, end with line=[Ctrl+d (Unix), Ctrl+z (Windows)]
Fiddle de dum, Fiddle de dee,
Eric the half a bee.
^Z
Connecting...
No errors.
Bye.
```

Это письмо посылается по электронному адресу этой книги (*PP4E@learning-python.com*), поэтому в конечном счете оно окажется в почтовом ящике на сервере моего интернет-провайдера, но лишь пройдя через неопределенное число серверов Сети и сетевые соединения неизвестной длины. На нижнем уровне все устроено сложно, но обычно Интернет «просто работает».

Обратите, однако, внимание на адрес «From» – он абсолютно фиктивный (по крайней мере, мне так кажется). Оказывается, обычно можно задать произвольный адрес «From», так как SMTP не проверяет его существование (проверяется только общий формат). Далее, в отличие от POP, в SMTP нет понятия имени пользователя и пароля, поэтому отправителя установить труднее. Необходимо лишь передать электронную почту на любой компьютер, где есть сервер, слушающий на порту SMTP, и не требуется иметь учетную запись на этом сервере. В данном случае *Eric.the.Half.a.Bee@yahoo.com* вполне годится в качестве адреса отправителя; с таким же успехом можно было бы указать адрес *Marketing.Geek.From.Hell@spam.com*.

На самом деле я специально не импортировал адрес «From» из модуля *mailconfig.py*, потому что я хотел продемонстрировать эту особенность – она лежит в основе появления в ящике всего этого раздражающего почтового мусора без настоящего адреса отправителя.[1] Торговцы, помешанные на мысли разбогатеть с помощью Интернета, рассылают рекламу по всем известным им адресам и не указывают действительный адрес «From», чтобы скрыть свои следы.

Обычно необходимо использовать один и тот же адрес «To» в сообщении и вызове SMTP и указывать свой действительный почтовый адрес в качестве значения «From» (иначе люди не смогут ответить на ваше послание). Более того, за исключением случаев, когда вы дразните вашу «вто-

[1] Все мы знаем, что такую мусорную почту обычно называют спамом (spam), но не все знают, что это название намекает на сценку из Monty Python, в которой люди пытаются заказать в ресторане завтрак, и их все время заглушает группа викингов, которые все громче и громче поют хором: «spam, spam, spam...». Так это название увязывается с мусорной электронной почтой. В примерах программ на языке Python имя spam часто используется в качестве имени некоторой обобщенной переменной, хотя ее происхождение также восходит к этой комической сценке.

рую половину», отправка фальшивого адреса явно нарушает правила добропорядочного поведения в Интернете. Запустим сценарий снова, чтобы отправить еще одно письмо с более политически корректными координатами:

```
C:\...\PP4E\Internet\Email> smtpmail.py
From? PP4E@learning-python.com
To?   PP4E@learning-python.com
Subj? testing smtpmail
Type message text, end with line=[Ctrl+d (Unix), Ctrl+z (Windows)]
Lovely Spam! Wonderful Spam!
^Z
Connecting...
No errors.
Bye.
```

Проверка получения

После этого можно запустить инструмент электронной почты, который обычно используется для доступа к почтовому ящику, и проверить результат этих двух операций отправки. В почтовом ящике должны появиться два новых письма независимо от того, какой почтовый клиент используется для их просмотра. Но поскольку мы уже написали сценарий Python для чтения почты, воспользуемся им в качестве средства проверки – при запуске сценария popmail из предыдущего раздела в конце списка писем обнаруживаются два наших новых сообщения (опять же, для экономии места и чтобы избавить вас от ненужной информации, часть вывода здесь была обрезана):

```
C:\...\PP4E\Internet\Email> popmail.py
Password for pop.secureserver.net?
Connecting...
b'+OK <29464.1273155506@pop08.mesa1.secureserver.net>'
There are 4 mail messages in 5326 bytes
(b'+OK ', [b'1 1860', b'2 1408', b'3 1049', b'4 1009'], 32)
------------------------------------------------------------------------------
[Press Enter key]

...первые два сообщения опущены...

Received: (qmail 25683 invoked from network); 6 May 2010 14:12:07 -0000
Received: from unknown (HELO p3pismtp01-018.prod.phx3.secureserver.net)
          ([10.6.1 (envelope-sender <Eric.the.Half.a.Bee@yahoo.com>)
          by p3plsmtp06-04.prod.phx3.secureserver.net (qmail-1.03) with SMTP
          for <PP4E@learning-python.com>; 6 May 2010 14:12:07 -0000
...часть строк опущена...
Received: from [66.194.109.3] by smtp.mailmt.com (ArGoSoft Mail
                                                  Server .NET v.1.
        for <PP4E@learning-python.com>; Thu, 06 May 2010 10:12:12 -0400
From: Eric.the.Half.a.Bee@yahoo.com
To: PP4E@learning-python.com
```

```
Date: Thu, 06 May 2010 14:11:07 -0000
Subject: A B C D E F G
Message-ID: <jdlohzf0j8dp8z4x06052010101212@SMTP>
X-FromIP: 66.194.109.3
X-Nonspam: None

Fiddle de dum, Fiddle de dee,
Eric the half a bee.


-----------------------------------------------------------------------------
[Press Enter key]
Received: (qmail 4634 invoked from network); 6 May 2010 14:16:57 -0000
Received: from unknown (HELO p3pismtp01-025.prod.phx3.secureserver.net)
          ([10.6.1 (envelope-sender <PP4E@learning-python.com>)
          by p3plsmtp06-05.prod.phx3.secureserver.net (qmail-1.03) with SMTP
          for <PP4E@learning-python.com>; 6 May 2010 14:16:57 -0000
...часть строк опущена...
Received: from [66.194.109.3] by smtp.mailmt.com (ArGoSoft Mail
                                                  Server .NET v.1.
        for <PP4E@learning-python.com>; Thu, 06 May 2010 10:17:03 -0400
From: PP4E@learning-python.com
To: PP4E@learning-python.com
Date: Thu, 06 May 2010 14:16:31 -0000
Subject: testing smtpmail
Message-ID: <8fad1n462667fik006052010101703@SMTP>
X-FromIP: 66.194.109.3
X-Nonspam: None

Lovely Spam! Wonderful Spam!


-----------------------------------------------------------------------------
Bye.
```

Обратите внимание, что значения полей, которые мы вводим во время работы сценария, обнаруживаются в виде заголовков и текста электронного письма, доставленного получателю. Технически некоторые интернет-провайдеры проверяют адрес получателя, чтобы убедиться, что хотя бы доменное имя в адресе отправителя (часть адреса после «@») является действительным, допустимым именем домена, и не выполняют доставку, если это не так. Как уже упоминалось выше, некоторые также требуют, чтобы отправитель имел непосредственное подключение к их сети и могут требовать пройти аутентификацию с указанием имени и пароля (как описывалось в конце предыдущего раздела). При работе над вторым изданием книги я пользовался услугами интернет-провайдера, который позволял мне творить и более бессмысленные вещи, но это во многом зависит от настроек сервера. С тех пор правила существенно ужесточились, чтобы ограничить распространение спама.

Манипулирование заголовками From и To

Первым письмом здесь было то, которое мы послали с фиктивным адресом отправителя. Вторым было более легитимное сообщение. Как и в адресах отправителя, в строках заголовков SMTP также допускает некоторый произвол. Сценарий smtpmail автоматически добавляет строки заголовков «From» и «To» в текст сообщения с теми адресами, которые были переданы интерфейсу SMTP, но только в порядке жеста вежливости. Иногда, однако, нельзя узнать даже, кому было послано письмо – чтобы скрыть круг получателей или поддержать законные списки адресов, отправители могут манипулировать этими заголовками в тексте сообщения.

Например, если изменить сценарий smtpmail так, чтобы он не создавал автоматически строку заголовка «To:» с теми же адресами, которые передаются вызову интерфейса SMTP:

```
text = ('From: %s\nDate: %s\nSubject: %s\n' % (From, Date, Subj))
```

можно будет вручную ввести заголовок «To:», отличающийся от настоящего адреса получателя – методу отправки модуля smtplib будет передаваться действительный список адресатов, а строка заголовка «To:» в тексте сообщения – это то, что будут отображать большинство почтовых клиентов (в сценарии *smtpmail-noTo.py* в дереве примеров приводится реализация поддержки такого анонимного поведения, и не забудьте ввести пустую строку после ввода строки заголовка «To:»):

```
C:\...\PP4E\Internet\Email> smtpmail-noTo.py
From? Eric.the.Half.a.Bee@aol.com
To?   PP4E@learning-python.com
Subj? a b c d e f g
Type message text, end with line=(ctrl + D or Z)
To: nobody.in.particular@marketing.com

Spam; Spam and eggs; Spam, spam, and spam.
^Z
Connecting...
No errors.
Bye.
```

В некоторых отношениях адреса «From» и «To» в вызове метода отправки и в заголовках сообщения аналогичны адресу на конверте и письму в конверте. Первое используется для пересылки, а второе представляет то, что видит читатель. Здесь оба адреса «From» являются фиктивными. Кроме того, я указал действительный адрес «To», но в строке заголовка «To:» я вручную ввел фиктивное имя – первый адрес определяет действительное место доставки сообщения, а второй отображается клиентами электронной почты. Если ваш клиент отображает строку «To:», такое письмо при просмотре будет выглядеть немного странным.

Например, если посмотреть, как отображается письмо, которое мы только что отправили, в моем почтовом ящике на сайте *learning-python.com*, то будет очень сложно сказать, кем и кому было отправлено это письмо, в том веб-интерфейсе, который предоставляет мой интернет-провайдер, как показано на рис. 13.5.

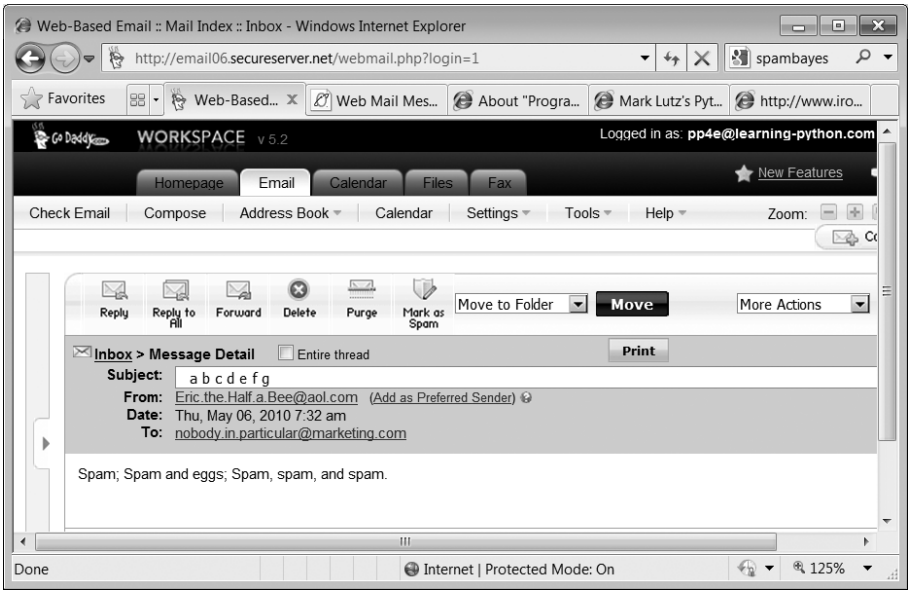


Рис. 13.5 *Анонимное письмо в клиенте с веб-интерфейсом (смотрите также пример PyMailGUI далее)*

Кроме того, и необработанный текст сообщения не поможет нам в этом, разве только внимательнее посмотреть на заголовки «Received:», добавляемые серверами, через которые пересылалось письмо:

```
C:\...\PP4E\Internet\Email> popmail.py
Password for pop.secureserver.net?
Connecting...
b'+OK <4802.1273156821@p3plpop03-03.prod.phx3.secureserver.net>'
There are 5 mail messages in 6364 bytes
(b'+OK ', [b'1 1860', b'2 1408', b'3 1049', b'4 1009', b'5 1038'], 40)
------------------------------------------------------------------------
[Press Enter key]

...первые три письма опущены...


Received: (qmail 30325 invoked from network); 6 May 2010 14:33:45 -0000
Received: from unknown (HELO p3pismtp01-004.prod.phx3.secureserver.net)
          ([10.6.1 (envelope-sender <Eric.the.Half.a.Bee@aol.com>)
```

```
          by p3plsmtp06-03.prod.phx3.secureserver.net (qmail-1.03) with SMTP
          for <PP4E@learning-python.com>; 6 May 2010 14:33:45 -0000
```

...часть строк опущена...

```
Received: from [66.194.109.3] by smtp.mailmt.com (ArGoSoft Mail
                                                  Server .NET v.1.
        for <PP4E@learning-python.com>; Thu, 06 May 2010 10:33:16 -0400
From: Eric.the.Half.a.Bee@aol.com
Date: Thu, 06 May 2010 14:32:32 -0000
Subject: a b c d e f g
To: nobody.in.particular@marketing.com
Message-ID: <66koqg66e0q1c8hl06052010103316@SMTP>
X-FromIP: 66.194.109.3
X-Nonspam: None

Spam; Spam and eggs; Spam, spam, and spam.


------------------------------------------------------------------------
Bye.
```

Еще раз повторю – не делайте так без веских причин. Я показываю это только с целью помочь вам понять, какую роль играют заголовки письма в процессе обработки электронной почты. Чтобы написать автоматический спам-фильтр, удаляющий нежелательную входящую почту, например, вам потребуется изучить контрольные признаки и научиться отыскивать их в тексте сообщения. Приемы, применяемые при рассылке спама, становятся все более изощренными, и по сложности давно превзошли простую подделку адресов отправителя и получателя (более подробную информацию по этой теме вы найдете в Интернете в целом и в почтовом фильтре *SpamBayes*, написанном на языке Python), но это одна из наиболее распространенных хитростей.

Кроме того, фокусы с адресами «To» могут оказаться полезными в контексте законных *списков рассылки* – при просмотре сообщения в заголовке «To:» выводится название списка, а не список потенциально большого количества получателей, указанных в вызове метода отправки сообщения. Как будет показано в следующем разделе, почтовый клиент легко может отправить сообщение всем получателям в списке, но вставить общее название списка в заголовок «To:».

Но в других ситуациях отправка электронной почты с фальшивыми строками «From:» и «To:» эквивалентна анонимным телефонным звонкам. Большинство почтовых программ не позволяет изменить строку «From» и не делает отличия между адресом получателя и строкой заголовка «To», хотя SMTP широко открыт в этом отношении. Ведите себя хорошо, договорились?

Знает ли кто-нибудь в действительности, который час?

В предыдущей версии сценария smtpmail, представленной в примере 13.19, в заголовок «Date» записывалось текущее время и дата в простом формате, который не совсем соответствует стандарту форматирования даты в SMTP:

```
>>> import time
>>> time.asctime()
'Wed May 05 17:52:05 2010'
```

Большинство серверов не обращают внимания на этот заголовок и позволяют вставлять в него любой текст с датой или даже могут сами добавлять его при необходимости. Клиенты тоже часто проявляют подобную беззаботность, но не всегда. Одна из программ с веб-интерфейсом, предоставляемая моим интернет-провайдером, корректно отображает даты в любом случае, но другая оставляет такие неправильно отформатированные значения пустыми при отображении. При желании как можно точнее соответствовать стандартам вы можете форматировать заголовок с датой с помощью следующего программного кода (результат которого может быть проанализирован любыми стандартными инструментами, такими как функция time.strptime):

```
import time
gmt = time.gmtime(time.time())
fmt = '%a, %d %b %Y %H:%M:%S GMT'
str = time.strftime(fmt, gmt)
hdr = 'Date: ' + str
print(hdr)
```

После выполнения этого фрагмента значение переменной hdr будет иметь следующий вид:

```
Date: Wed, 05 May 2010 21:49:32 GMT
```

Функция time.strftime позволяет формировать строку с датой и временем в любом формате; функция time.asctime реализует единственный стандартный формат. Еще лучше использовать тот же прием, который используется в сценарии smtpmail, – в современном пакете email (описывается далее в этой главе) имеется модуль email.utils, который предоставляет функции для корректного форматирования даты и времени. Сценарий smtpmail использует первый из следующих вариантов форматирования:

```
>>> import email.utils
>>> email.utils.formatdate()
'Wed, 05 May 2010 21:54:28 -0000'
>>> email.utils.formatdate(localtime=True)
'Wed, 05 May 2010 17:54:52 -0400'
>>> email.utils.formatdate(usegmt=True)
'Wed, 05 May 2010 21:55:22 GMT'
```

Смотрите описание примеров `pymail` и `mailtools` в этой главе, где вы увидите дополнительные способы использования. Последний повторно используется в крупных почтовых клиентах PyMailGUI и PyMailCGI, представленных далее в этой книге.

Отправка почты из интерактивной оболочки

Итак, где мы сейчас находимся в абстрактной модели Интернета? Со всеми этими приемами получения и отправки почты легко за деревьями не увидеть леса. Имейте в виду, что почта передается через сокеты (еще помните о них?), поэтому они являются фундаментом всего происходящего. Всякое полученное или отправленное письмо в конечном итоге состоит из форматированных последовательностей байтов, передаваемых по сети между компьютерами через сокеты. Однако, как мы уже видели, интерфейсы Python к протоколам POP и SMTP скрывают все мелкие детали. Кроме того, сценарии, которые мы начали писать, скрывают даже интерфейсы Python и предоставляют высокоуровневые интерактивные инструменты.

Оба сценария, `popmail` и `smtpmail`, предоставляют переносимые инструменты электронной почты, но удобство их использования не соответствует уровню, принятому в нынешние времена. Далее в этой главе с помощью увиденного мы реализуем более интерактивный инструмент для работы с электронной почтой. В следующей главе мы также создадим графический интерфейс к электронной почте с использованием tkinter, а в другой главе построим веб-интерфейс. Все эти инструменты отличаются, главным образом, только интерфейсом пользователя. Каждый из них, в конечном счете, применяет рассмотренные нами почтовые модули для передачи почтовых сообщений через Интернет посредством сокетов.

Прежде чем двинуться дальше, еще одно примечание, касающееся протокола SMTP: мы можем использовать интерактивную оболочку Python не только для чтения, но и для отправки почты, вызывая соответствующие функции и методы вручную. Ниже, например, демонстрируется отправка сообщения через SMTP-сервер моего интернет-провайдера по двум адресам, которые, как предполагаются, включены в список рассылки:

```
C:\...\PP4E\Internet\Email> python
>>> from smtplib import SMTP
>>> conn = SMTP('smtpout.secureserver.net')
>>> conn.sendmail(
... 'PP4E@learning-python.com',                    # истинный отправитель
... ['lutz@rmi.net', 'PP4E@learning-python.com'], # истинные получатели
... """From: PP4E@learning-python.com
... To: maillist
... Subject: test interactive smtplib
...
... testing 1 2 3...
... """)
{}
>>> conn.quit()              # вызов quit() необходим,
                             # дата добавляется автоматически
(221, b'Closing connection. Good bye.')
```

Мы проверим доставку этого сообщения с помощью программы почтового клиента, которую напишем далее. В поле «To» почтовые клиенты получателя будут отображать адрес «maillist» – результат вполне допустимого приема манипулирования заголовком. Фактически того же эффекта можно добиться при использовании сценария smtpmail-noTo, введя в ответ на запрос «To?» список адресов, разделенных точкой с запятой (например, *lutz@rmi.net; PP4E@learning-python.com*), и указав название списка рассылки в строке заголовка «To:». Почтовые клиенты, поддерживающие списки рассылки, автоматизируют этот шаг.

Однако отправка таким способом электронных писем из интерактивной оболочки может оказаться непростым делом – формат представления строк заголовков регулируются стандартами: например, пустая строка после строки заголовка с темой сообщения является обязательной и играет важную роль, а заголовок «Date» может вообще отсутствовать (он будет добавлен автоматически). Кроме того, при наличии вложений форматирование сообщений становится намного более сложным. На практике для создания сообщений электронной почты часто используется пакет email из стандартной библиотеки, после чего они передаются модулю smtplib. Пакет позволяет конструировать электронные письма, присваивая заголовки и присоединяя части, возможно кодированные, и создает корректно отформатированный текст письма. Чтобы узнать, как это делается, перейдем к следующему разделу.

Пакет email: анализ и составление электронных писем

Во втором издании этой книги использовалось множество модулей из стандартной библиотеки (rfc822, StringIO и другие) для анализа содержимого сообщений и простой обработки текста для их составления. Кроме того, это издание содержало раздел об извлечении и декодирова-

нии вложений, прикрепленных к сообщению, с использованием таких модулей, как mhlib, mimetools и base64.

При работе над третьим изданием эти инструменты все еще были доступны, но, откровенно говоря, они были несколько неудобны и провоцировали появление ошибок. Анализ вложений, например, был достаточно запутанным делом, а составление даже простых сообщений было утомительным (фактически первые печатные выпуски предыдущего издания содержали потенциальную ошибку – в операции форматирования строки отсутствовал один символ \n). Возможность добавления вложений даже не рассматривалась из-за связанных с этим сложностей форматирования. Большинство этих инструментов отсутствовало в составе Python 3.X, когда я работал над четвертым изданием, отчасти из-за их сложности, а отчасти из-за того, что они устарели.

К счастью, в настоящее время ситуация значительно улучшилась. После выхода второго издания в Python появился новый пакет email – коллекция мощных инструментов, автоматизирующих большинство задач, связанных с анализом и составлением сообщений электронной почты. Этот пакет обеспечивает объектно-ориентированный интерфейс доступа к сообщениям и обрабатывает все тонкости структуры текстовых сообщений, упрощая их анализ и составление. Он не только устраняет опасность появления целого класса ошибок, но обеспечивает более сложные способы обработки электронной почты.

С его помощью, например, работа с вложениями становится доступной для простых смертных (и для авторов с ограниченным пространством книги). Это позволило в третьем издании убрать целый раздел, посвященный анализу и декодированию вложений вручную, – эти операции были автоматизированы пакетом email. Новый пакет позволяет анализировать и конструировать заголовки и вложения, генерировать корректный текст электронных писем, применять алгоритмы декодирования и кодирования данных, такие как Base64, quoted-printable и uuencoded, и многое другое.

В этой книге мы не будем рассматривать пакет email полностью. Он достаточно хорошо описан в руководстве по библиотеке Python. Наша цель здесь – исследовать некоторые примеры его использования, которые вы сможете изучать параллельно с руководствами. Но чтобы дать вам точку опоры, начнем с краткого обзора. В двух словах, основой пакета email является объект Message, который обеспечивает:

Анализ сообщений

Полный текст электронного письма, полученного с помощью модуля poplib или imaplib, преобразуется в новый объект Message с прикладным интерфейсом для обработки различных его компонентов. В объекте заголовки письма приобретают вид ключей словаря, а компоненты превращаются в «полезный груз» (информационное наполнение) – их можно обойти, применяя интерфейс генераторов (подробнее об этих компонентах рассказывается чуть ниже).

Создание сообщения

Новые письма конструируются за счет создания новых объектов `Message`, использования его методов для создания заголовков и частей и получения печатного представления объектов – корректно отформатированного текста электронного письма, готового к отправке с помощью модуля `smtplib`. Заголовки добавляются инструкциями присваивания по ключу, а вложения – вызовами методов.

Иными словами, объект `Message` используется и для доступа к существующим сообщениям, и для создания новых. В обоих случаях пакет `email` способен автоматически обрабатывать такие детали, как кодирование содержимого (например, вложения двоичных изображений могут интерпретироваться как текст с применением схемы кодирования Base64), определение типов содержимого и многие другие.

Объекты Message

Объект `Message` является основой пакета `email`, поэтому для начала вам необходимо хотя бы в общих чертах познакомиться с ним. Вкратце, этот объект отражает структуру готового сообщения электронной почты. Каждый объект `Message` состоит из трех основных информационных частей:

Тип

Тип содержимого (простой текст, разметка HTML, изображение JPEG и так далее), кодированного в соответствии с основным типом MIME и подтипом. Например, «text/html» означает, что основным типом является текст, а подтипом – разметка HTML (веб-страница); тип «image/jpeg» соответствует изображению JPEG. Тип «multipart/mixed» соответствует вложенным частям с текстом сообщения.

Заголовки

Интерфейс отображения, напоминающий ключи словарей, когда каждому ключу (From, To и так далее) соответствует отдельный заголовок. Этот интерфейс поддерживает большинство операций, типичных для словарей, и позволяет извлекать и устанавливать заголовки с помощью обычных операций индексирования по ключу.

Содержимое

Информационное наполнение почтового сообщения. Это может быть строка (`bytes` или `str`) в простых сообщениях или список дополнительных объектов `Message`, содержащий вложения или альтернативные части. Для некоторых необычных типов информационным наполнением может быть объект Python `None`.

Тип MIME объекта сообщения `Message` является ключом к пониманию его содержимого. Например, электронные письма с вложенными изображениями могут состоять из основного объекта `Message` верхнего уровня (типа `multipart/mixed`) с тремя дополнительными объектами `Message`

в качестве информационного наполнения – один для основного текста (типа text/plain), другие два для изображений (типа image/jpeg). Изображения могут кодироваться при передаче в текст с применением алгоритма Base64 или другого; способ кодирования, а также оригинальное имя файла изображения указываются в заголовках частей.

Аналогично электронное письмо, включающее простой текст и альтернативную разметку HTML, будет представлено в виде корневого объекта Message типа multipart/alternative с двумя вложенными объектами Message – объект с простым текстом (типа text/plain) и объект с разметкой HTML (типа text/html). Программа почтового клиента сама определит, какую часть отображать, зачастую опираясь на предпочтения пользователя.

Простейшие сообщения могут состоять из единственного корневого объекта Message типа text/plain или text/html, представляющего все тело сообщения. Информационным наполнением в таких электронных письмах является обычная строка. Они могут вообще не иметь явно указанного типа; в таких случаях по умолчанию сообщения интерпретируются, как имеющие тип text/plain. Некоторые сообщения, состоящие из единственной части, имеют тип text/html без указания типа text/plain – для просмотра они требуют использования веб-броузера или другого инструмента отображения разметки HTML (или особого настроя глаз у пользователя).

Возможны и другие комбинации, включая типы, которые нечасто встретишь на практике, такие как message/delivery. Большинство сообщений имеют основную текстовую часть, хотя и не обязательно, и могут вкладываться в сообщения, состоящие из нескольких частей или имеющие другую структуру.

Во всех случаях почтовое сообщение представляет собой простую линейную последовательность, и структура сообщения определяется автоматически, когда производится анализ текста письма, или создается вызовами методов, когда формируется новое сообщение. Например, при создании сообщений метод attach объекта сообщения добавляет дополнительные части, а метод set_payload сохраняет строку в качестве информационного наполнения для простых сообщений.

Объекты Message также обладают разнообразными свойствами (например, имя файла вложения) и предоставляют удобный метод-генератор walk, который при каждом обращении в цикле for или в других итерационных контекстах возвращает следующий объект Message в цепочке вложенных объектов. Поскольку возвращаемый этим методом генератор обхода первым возвращает корневой объект Message (то есть self), сообщения, состоящие из единственной части, не приходится обрабатывать как особый случай – фактически сообщения, состоящие из единственной части, можно интерпретировать как объект Message с единственным элементом информационного наполнения – самим собой.

В конечном счете, структура объекта `Message` близко отражает текстовое представление сообщений электронной почты. Специальные строки заголовков в тексте письма определяют его тип (например, простой текст или состоящее из нескольких частей), а между вложенными частями содержимого помещается специальный разделитель. Все особенности текстового представления автоматически обрабатываются пакетом `email`, поэтому при анализе и составлении сообщений нам не требуется помнить о деталях форматирования.

Если вам интересно увидеть, как все это соотносится с настоящими электронными письмами, то лучше всего будет ознакомиться со структурой необработанного текста сообщения, отображаемого почтовым клиентом, а также с некоторыми примерами, с которыми мы встретимся в этой книге. Фактически, мы уже видели несколько писем – простыми примерами могут служить результаты, выведенные сценарием получения почты по протоколу POP выше. За более подробными сведениями об объекте `Message` и о пакете `email` в целом обращайтесь к разделу с описанием пакета `email` в руководстве по библиотеке Python. Ради экономии места мы опустим такие детали, как перечень доступных схем кодирования и классов объекта MIME.

Кроме пакета `email` в стандартной библиотеке Python имеются и другие инструменты для работы с электронной почтой. Например, модуль `mimetypes` отображает имена файлов в тип MIME и обратно[1]:

`mimetypes.guess_type(filename)`

Отображает имя файла в тип MIME. Имя *spam.txt* отображается в тип «text/plan».

`mimetypes.guess_extension(contype)`

Отображает тип MIME в расширение имени файла. Тип «text/html» отображается в расширение *.html*.

Мы уже использовали модуль `mimetypes` ранее в этой главе, когда с его помощью по имени файла определяли режим его передачи по FTP (пример 13.10), а также в главе 6, где он применялся для выбора программы-проигрывателя по имени файла (смотрите примеры в той главе, включая сценарий *playfile.py*, пример 6.23). При обработке электронной почты эти функции могут пригодиться в реализации процедуры прикрепления файлов к новым сообщениям (`guess_type`) и сохранения вложений, для которых не указано имя файла (`guess_extension`). Фактически исходный программный код этого модуля может служить исчерпывающим справочником по типам MIME. Дополнительные подробности об этих инструментах смотрите в руководстве по библиотеке.

---

[1] Точнее, отображаются расширения в именах файлов, базовая часть имени файла перед расширением не анализируется и не влияет на результат, анализируется только расширение. – *Прим. перев.*

Базовые интерфейсы пакета email в действии

У нас нет возможности привести здесь исчерпывающую справочную информацию, тем не менее, рассмотрим несколько простых интерактивных сеансов, иллюстрирующих основы обработки электронной почты. Чтобы *сконструировать* полный текст сообщения, готового к отправке, например с помощью модуля smtplib, создайте объект Message, присвойте заголовки как ключи этого объекта и определите информационное наполнение в теле сообщения. Операция преобразования получившегося объекта в строку дает в результате полный текст почтового сообщения. Этот процесс намного более прост и не так подвержен ошибкам, как выполнение текстовых операций вручную, которые мы применяли в примере 13.19 для составления сообщений в виде строк:

```
>>> from email.message import Message
>>> m = Message()
>>> m['from'] = 'Jane Doe <jane@doe.com>'
>>> m['to'] = 'PP4E@learning-python.com'
>>> m.set_payload('The owls are not what they seem...')
>>>
>>> s = str(m)
>>> print(s)
from: Jane Doe <jane@doe.com>
to: PP4E@learning-python.com

The owls are not what they seem...
```

Анализ текста сообщения – подобный тому, что возвращает модуль poplib, – выполняется так же просто и по сути представляет обратную процедуру: мы получаем из текста объект Message с ключами, представляющими заголовки, и с информационным наполнением в теле:

```
>>> s                    # та же строка, что и в предыдущем сеансе
'from: Jane Doe <jane@doe.com>\nto: PP4E@learning-python.com\n\nThe owls...'

>>> from email.parser import Parser
>>> x = Parser().parsestr(s)
>>> x
<email.message.Message object at 0x015EA9F0>
>>>
>>> x['From']
'Jane Doe <jane@doe.com>'
>>> x.get_payload()
'The owls are not what they seem...'
>>> x.items()
[('from', 'Jane Doe <jane@doe.com>'), ('to', 'PP4E@learning-python.com')]
```

Пока что продемонстрированные приемы мало отличаются от приемов использования устаревшего и ныне отсутствующего в библиотеке модуля rfc822, но, как будет показано чуть ниже, все становится намного интереснее, когда на сцене появляются сообщения, состоящие из не-

скольких частей. Для простых сообщений, как в примере выше, генератор walk объекта сообщения интерпретирует его как сообщение, состоящее из единственной части простого текстового типа:

```
>>> for part in x.walk():
...     print(x.get_content_type())
...     print(x.get_payload())
...
text/plain
The owls are not what they seem...
```

Обработка сообщений, состоящих из нескольких частей

Для создания письма с *вложениями* не требуется прикладывать особых усилий: мы просто создаем корневой объект Message и присоединяем к нему вложенные объекты Message, созданные из объектов типов MIME, соответствующих типам присоединяемых данных. Класс MIMEText, например, является подклассом Message, который предназначен для создания текстовых частей и знает, как генерировать правильные заголовки при выводе. Классы MIMEImage и MIMEAudio являются похожими расширениями класса Message для изображений и аудиофрагментов и могут применять к двоичным данным алгоритм кодирования MIME Base64 и другие. Корневым объектом сообщения является тот, который хранит главные заголовки почтового сообщения, и вместо того, чтобы конструировать все основное информационное наполнение сообщения целиком, мы прикрепляем к корневому объекту составные части – теперь информационное наполнение приобретает вид списка, а не строки. Класс MIMEMultipart – это подкласс Message, который предоставляет методы, необходимые корневому объекту:

```
>>> from email.mime.multipart import MIMEMultipart # подкласс Message
>>> from email.mime.text import MIMEText   # дополнительные заголовки+методы
>>>
>>> top = MIMEMultipart()                      # корневой объект Message
>>> top['from'] = 'Art <arthur@camelot.org>'   # по умолчанию подтип=mixed
>>> top['to'] = 'PP4E@learning-python.com'
>>>
>>> sub1 = MIMEText('nice red uniforms...\n')  # вложения
>>> sub2 = MIMEText(open('data.txt').read())
>>> sub2.add_header('Content-Disposition', 'attachment', filename='data.txt')
>>> top.attach(sub1)
>>> top.attach(sub2)
```

Если попытаться запросить текст сообщения, в ответ будет возвращен корректно отформатированный полный текст сообщения с разделителями и всем остальным, готовый к отправке с помощью smtplib, что весьма непросто, если попробовать сделать это вручную:

```
>>> text = top.as_string()        # или то же самое: str(top) или print(top)
>>> print(text)
Content-Type: multipart/mixed; boundary="===============1574823535==="
```

```
MIME-Version: 1.0
from: Art <arthur@camelot.org>
to: PP4E@learning-python.com

--================1574823535==
Content-Type: text/plain; charset="us-ascii"
MIME-Version: 1.0
Content-Transfer-Encoding: 7bit

nice red uniforms...

--================1574823535==
Content-Type: text/plain; charset="us-ascii"
MIME-Version: 1.0
Content-Transfer-Encoding: 7bit
Content-Disposition: attachment; filename="data.txt"

line1
line2
line3

--================1574823535==--
```

Если теперь отправить это сообщение и получить его с помощью `poplib`, при попытке проанализировать его полный текст мы получим объект `Message`, точно такой же, как тот, что был сконструирован для отправки. Генератор `walk` объекта сообщения позволяет обойти все его части, определять их типы и извлекать информационное наполнение:

```
>>> text           # тот же самый текст, что был получен в предыдущем сеансе
'Content-Type: multipart/mixed; boundary="================1574823535==\nMI...'

>>> from email.parser import Parser
>>> msg = Parser().parsestr(text)
>>> msg['from']
'Art <arthur@camelot.org>'

>>> for part in msg.walk():
...     print(part.get_content_type())
...     print(part.get_payload())
...     print()
...
multipart/mixed
[<email.message.Message object at 0x015EC610>,
<email.message.Message object at0x015EC630>]

text/plain
nice red uniforms...


text/plain
```

```
line1
line2
line3
```

Альтернативные части составных сообщений (с текстом и разметкой HTML, содержащими одну и ту же информацию) можно создавать и извлекать похожим способом. Поскольку простой объектно-ориентированный API позволяет клиентам электронной почты легко и просто анализировать и конструировать сообщения, они могут сосредоточиться на пользовательском интерфейсе, а не на обработке текста.

Юникод, интернационализация и пакет email в Python 3.1

Теперь, когда я показал вам, насколько «крутыми» возможностями обладает пакет email, я, к сожалению, должен отметить, что в Python 3.1 он недостаточно функционален. Пакет email действует, как было показано выше, применительно к простым сообщениям, но некоторые его аспекты испытывают существенное влияние дихотомии строковых типов str/bytes в Python 3.X.

В двух словах: реализация пакета email в Python 3.1 до сих ориентирована на работу в царстве текстовых строк str Python 2.X. Так как в версии 3.X эти строки превратились в строки Юникода, а также потому, что многие инструменты, используемые пакетом email, ныне ориентированы на работу со строками байтов, которые не могут свободно смешиваться со строками str, неожиданно возникает множество конфликтов, вызывающих проблемы в программах, использующих этот пакет.

К моменту написания этих строк велись работы по созданию новой версии пакета email, который лучше будет обрабатывать строки bytes и поддерживать кодировки Юникода, но по общему признанию включение новой версии пакета в библиотеку Python произойдет не раньше выхода версии 3.3 – намного позже выхода этого издания книги. Некоторые исправления могут быть включены уже в версию 3.2, тем не менее, необходимо понимать, что решение всех проблем, которые порождает пакет, требует полной его модернизации.

Справедливости ради следует отметить, что это фундаментальная проблема. Электронная почта исторически была ориентирована на передачу текста, состоящего из однобайтовых символов ASCII, и обобщение ее до уровня Юникода оказалось совсем непростым делом. Фактически то же самое относится и к большей части современного Интернета – как уже говорилось выше в этой главе, протоколы FTP, POP, SMTP и даже веб-страницы, получаемые с помощью протокола HTTP, подвержены тем же проблемам. Интерпретировать пересылаемые по сети байты как текст очень просто, пока они один в один отображается в символы, но включение поддержки различных кодировок текста Юникода открывает ящик Пандоры, заполненный проблемами. В настоящее время это

влечет дополнительные сложности, но, как видно на примере пакета email, это поистине грандиозная задача.

Откровенно говоря, я не хотел выпускать это издание книги, пока не будут решены проблемы в этом пакете, но решился на этот шаг, потому что на подготовку новой версии пакета email могут уйти годы (судя по всему – две версии Python). Кроме того, эти проблемы относятся к категории проблем, с которыми вам придется столкнуться при разработке полномасштабных приложений. Все течет, все изменяется, и программное обеспечение не является исключением.

Вместо этого в примерах данной книги демонстрируется новая поддержка Юникода и интернационализации, но при этом везде, где возможно, приводятся обходные решения проблем. Программы в книге в первую очередь предназначены для обучения, а не для коммерческого использования. Однако, учитывая состояние пакета email, который используется в примерах, решения, демонстрируемые здесь, могут оказаться недостаточно универсальными, и могут таить в себе дополнительные проблемы, связанные с поддержкой Юникода. На будущее – следите за информацией на веб-сайте книги (описывается в предисловии), где будут приводиться дополнительные замечания и программный код примеров с появлением новой версии пакета email. Здесь же мы будем работать с тем, что имеется.

Однако все не так плохо. В этой книге мы сможем использовать пакет email в текущем его виде для создания весьма сложных и полнофункциональных клиентов электронной почты. Как бы то ни было, он предлагает множество великолепных инструментов, включая преобразование в формат MIME и обратно, формирование и анализ сообщений, конструирование и извлечение интернационализированных заголовков, и многие другие. Есть один неприятный момент: нам придется применять некоторые обходные решения, которые, возможно, потребуется изменить в будущем, хотя некоторые программные проекты свободны от этих недостатков.

Поскольку ограничения пакета email будут оказывать влияние на примеры, следующие далее в книге, я коротко пробегусь по ним в этом разделе. Некоторые из них можно было бы спокойно оставить на будущее, но части последующих примеров будет сложно понять, если не иметь представления об этих ограничениях. В этом есть и свои плюсы – исследование ограничений также послужит более глубокому пониманию интерфейсов пакета email в целом.

Необходимость декодирования сообщений перед анализом

Первая проблема в пакете email, связанная с поддержкой Юникода в Python3.1, является практически непреодолимым препятствием в некоторых контекстах: строки bytes, подобные тем, которые возвращает модуль poplib при извлечении сообщения электронной почты, должны

декодироваться в строки str перед анализом их с помощью пакета email. К сожалению, из-за недостатка информации о том, как декодировать байты сообщения в Юникод, в некоторых клиентах этого пакета может потребоваться обеспечить определение всех кодировок, использованных в сообщении, перед тем, как анализировать его. В самых тяжелых случаях, помимо электронной почты, – в которых требуется использовать данные смешанных типов, текущая версия пакета вообще не сможет применяться. Эта проблема демонстрируется ниже:

```
>>> text                        # из предыдущего примера в этом разделе
'Content-Type: multipart/mixed; boundary="==============1574823535==\nMI...'
>>> btext = text.encode()
>>> btext
b'Content-Type: multipart/mixed; boundary="==============1574823535==\nM...'

>>> msg = Parser().parsestr(text)  # Parser ожидает получить строку Юникода,
>>> msg = Parser().parsestr(btext) # но poplib возвращает строку bytes!
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "C:\Python31\lib\email\parser.py", line 82, in parsestr
    return self.parse(StringIO(text), headersonly=headersonly)
TypeError: initial_value must be str or None, not bytes
```

(TypeError: initial_value должно иметь тип str или None, а не bytes)

```
>>> msg = Parser().parsestr(btext.decode())        # кодировка по умолчанию
>>> msg = Parser().parsestr(btext.decode('utf8'))  # текст ascii (по умолч.)
>>> msg = Parser().parsestr(btext.decode('latin1'))  # текст ascii остается
>>> msg = Parser().parsestr(btext.decode('ascii'))   # одинаковым
                                                     # во всех 3 случаях
```

Это не самое идеальное решение, так как пакет email, ориентированный на работу со строками bytes, мог бы обеспечить более непосредственное применение кодировок к сообщениям. Однако, как уже упоминалось выше, пакет email в Python 3.1 недостаточно функционален из-за того, что он ориентирован на работу с типом str и из-за существенных отличий между текстом Юникода и строками байтов в Python 3.X. В данном случае объект, выполняющий анализ, должен был бы принимать строку bytes, а не ждать, что клиенты будут знать, как выполнить ее декодирование.

По этой причине в реализациях клиентов электронной почты, представленных в этой книге, предпринят упрощенный подход к декодированию байтов сообщений, которые необходимо проанализировать с помощью пакета email. В частности, сначала выполняется попытка декодировать полный текст сообщения с применением кодировки, указанной пользователем; в случае неудачи используются эвристические алгоритмы определения кодировки и, наконец, предпринимается попытка декодировать только заголовки сообщения.

Этого будет достаточно для примеров, но для более широкого применения данный прием может потребоваться расширить. В некоторых слу-

чаях кодировку можно определить другими способами, например проанализировав заголовки сообщения (если они присутствуют), выполнив структурный анализ байтов или запросив кодировку у пользователя. Попытка добавить подобные улучшения, действующие достаточно надежно, может оказаться слишком сложной для книжных примеров, но в любом случае для этой цели лучше всего использовать инструменты стандартной библиотеки.

На самом деле в настоящее время надежное декодирование текста сообщения может оказаться вообще невозможным, если для этого требуется проверить заголовки, – мы не сможем извлечь информацию о кодировке сообщения, не выполнив его анализ, но мы не сможем проанализировать сообщение с помощью пакета `email` в Python 3.1, пока не узнаем кодировку. То есть чтобы узнать кодировку, сценарию может потребоваться проанализировать сообщение, но, чтобы выполнить анализ, необходимо знать кодировку! Строки байтов, возвращаемые `poplib`, и строки символов Юникода, с которыми работает пакет `email`, в Python 3.1 имеют фундаментальные отличия. Даже внутри стандартной библиотеки изменения, внесенные в Python 3.X, создали проблему «курицы и яйца», которая никуда не делась даже спустя почти два года после выхода версии 3.0.

Кроме разработки собственного механизма анализа сообщений электронной почты или реализации других подобных по сложности подходов лучшим решением на сегодняшний день выглядит определение кодировки из пользовательских настроек и использование значений по умолчанию. Именно этот подход будет применяться в этом издании. Клиент PyMailGUI из главы 14, например, будет позволять указывать кодировку полного текста почтового сообщения для каждого сеанса в отдельности.

Настоящая проблема, конечно же, состоит в том, что электронная почта изначально усложнена необходимостью поддержки произвольных кодировок текста. Вдобавок к проблеме выбора кодировки полного текста сообщения по мере его анализа мы должны не ошибиться при выборе кодировок его отдельных текстовых компонентов – текстовых частей и заголовков. Двинемся дальше, чтобы увидеть, почему.

Подобная проблема в CGI-сценариях: следует также отметить, что проблема декодирования полного текста сообщения может оказаться не настолько существенной, как для некоторых клиентов пакета `email`. Оригинальные стандарты электронной почты предусматривают обмен текстом ASCII и требуют выполнять преобразование двоичных данных в формат MIME, поэтому большинство электронных писем наверняка будут корректно декодироваться с применением 7- или 8-битовой кодировки, такой как Latin-1.

Однако, как мы увидим в главе 15, перед серверными веб-сценариями, поддерживающими *выгрузку файлов по CGI*, встает похожая и еще более непреодолимая проблема. Для анализа составных форм

данных модуль CGI в языке Python использует пакет email. Пакет email требует декодировать данные в строки str для анализа. А сами данные, в свою очередь, могут представлять собой смесь текстовых и двоичных данных (включая простые двоичные данные, к которым *не* применялось преобразование в формат MIME, текст в любой кодировке и даже произвольные их комбинации). Поэтому в Python 3.1 операция отправки таких форм будет терпеть неудачу, если они будут включать двоичные или несовместимые текстовые файлы. Еще до того, как у сценария появится шанс вмешаться в происходящее, внутри модуля cgi будет возбуждено исключение ошибки декодирования Юникода или неверного типа.

Выгрузка по CGI работает в Python 2.X лишь по той простой причине, что тип str в этой версии может представлять и закодированный текст, и двоичные данные. Сохранения такого типа содержимого в двоичном файле в виде строки байтов в версии 2.X было достаточно для обработки произвольного текста и двоичных данных, таких как изображения. По той же причине в 2.X не возникает проблем с анализом сообщений электронной почты. Хорошо это или плохо, но разделение типов str/bytes в 3.X делает это обобщение невозможным.

Иными словами, в целом мы можем обойти требование пакета email и представлять для анализа почтовые сообщения в виде строк str, декодируя их с применением 8-битовой кодировки, однако эта проблема является весьма болезненной для всех сфер современного веб-программирования. Смотрите дополнительные подробности по этой теме в главе 15 и обязательно следите за изменениями, которые, возможно, уже будут реализованы к тому моменту, когда вы читаете эти строки.

Кодировка текстового содержимого: обработка результатов смешанного типа

Следующая проблема в пакете email, связанная с поддержкой Юникода, в некоторой степени даже противоречит модели программирования на языке Python в целом: типы данных содержимого объектов сообщений могут отличаться в зависимости от того, как они извлекаются. Это особенно усложняет реализацию программ, которые выполняют обход и обработку частей сообщения с содержимым.

В частности, метод get_payload объекта Message, который мы использовали выше, принимает необязательный аргумент decode, управляющий автоматическим декодированием данных в формате MIME (например, Base64, uuencode, quoted-printable). Если в этом аргументе передать число 1 (или True), содержимое будет декодироваться при извлечении, если это необходимо. Это настолько удобно при обработке сложных сообщений с произвольными частями, что в этом аргументе обычно всегда передается значение 1. Двоичные части сообщения, как правило, всегда кодируются в формат MIME, но даже текстовые части могут быть пред-

ставлены в одном из форматов MIME, таком как Base64, если значения их байтов выходят за рамки стандартов электронной почты. Некоторые типы текста Юникода, например, требуют кодирования в формат MIME.

В результате метод get_payload, обычно возвращающий строки str для текстовых частей str, будет возвращать строки bytes, если его аргумент decode имеет значение True – даже когда заранее известно, что обрабатываемая часть сообщения по своей природе является текстовой. Если этот аргумент не используется, тип содержимого будет зависеть от его исходного типа: str или bytes. Поскольку в Python 3.X не допускается смешивать типы str и bytes в операциях, клиенты, которые предусматривают сохранение результатов в файлах или их обработку как текста, должны учитывать эти различия. Выполним некоторый программный код для иллюстрации:

```
>>> from email.message import Message
>>> m = Message()
>>> m['From'] = 'Lancelot'
>>> m.set_payload('Line?...')

>>> m['From']
'Lancelot'
>>> m.get_payload()         # str, если содержимое имеет тип str
'Line?...'
>>> m.get_payload(decode=1) # bytes, (то же, что и decode=True) выполняется
b'Line?...'                 # декодирование MIME, если необходимо
```

Сочетание этих различий в типах возвращаемых данных со строгим разделением типов str/bytes в Python 3.X может вызывать проблемы обработки результатов при невнимательном отношении к декодированию:

```
>>> m.get_payload(decode=True) + 'spam'           # нельзя смешивать в 3.X!
TypeError: can't concat bytes to str
(TypeError: невозможно объединить bytes и str)
>>> m.get_payload(decode=True).decode() + 'spam'  # преобразовать,
'Line?...spam'                                    # если необходимо
```

Чтобы вы могли понять смысл этих примеров, запомните, что к тексту сообщения электронной почты применяются два разных понятия термина «кодировка»:

- *Кодировки MIME электронной почты*, такие как Base64, uuencode и quoted-printable, применяются к двоичным данным и другому необычному содержимому, чтобы обеспечить возможность их передачи в тексте сообщения электронной почты.
- *Кодировки Юникода* текстовых строк в целом применяются к тексту сообщения, а также к его частям, и могут потребоваться для текстовых частей после декодирования из формата MIME.

Пакет email обрабатывает кодировки MIME автоматически, когда при анализе и извлечении содержимого указывается аргумент decode=1 или когда создается сообщение, содержащее непечатаемые части, однако сценарии по-прежнему должны принимать во внимание кодировки Юникода из-за важных отличий между строковыми типами в Python 3.X. Например, следующая инструкция выполняет первое декодирование из формата MIME, а второе – в строку символов Юникода:

```
m.get_payload(decode=True).decode() # в bytes через MIME, затем в str
```

Даже когда аргумент decode не используется, тип содержимого все равно может отличаться при сохранении в разных форматах:

```
>>> m = Message(); m.set_payload('spam'); m.get_payload()  # извлекаемый тип
'spam'                                                     # соответствует
>>> m = Message(); m.set_payload(b'spam'); m.get_payload() # сохраненному
b'spam'                                                    # типу
```

То же самое справедливо для текстовых подклассов MIME (хотя, как мы увидим далее в этом разделе, мы не можем передавать строки bytes их конструкторам, чтобы принудительно создать двоичное содержимое):

```
>>> from email.mime.text import MIMEText
>>> m = MIMEText('Line...?')
>>> m['From'] = 'Lancelot'
>>> m['From']
'Lancelot'
>>> m.get_payload()
'Line...?'
>>> m.get_payload(decode=1)
b'Line...?'
```

К сожалению, тот факт, что в настоящее время содержимое может быть типа str или bytes, не только противоречит нейтральной к типам модели программирования на языке Python, но и усложняет программный код – в сценариях может оказаться необходимым выполнять преобразования, когда контекст требует использования того или иного типа. Например, библиотеки графических интерфейсов могут позволять использовать оба типа, но операции сохранения файлов и создания содержимого веб-страниц могут оказаться менее гибкими. В наших примерах программ мы будем обрабатывать содержимое как строки типа bytes везде, где это возможно, и декодировать в строки типа str в случаях, где это необходимо, используя информацию о кодировке, доступную для прикладного интерфейса заголовков, описываемого в следующем разделе.

Кодировка текстового содержимого: использование информации в заголовках для декодирования

Если копнуть глубже, текст в почтовых сообщениях может быть еще разнообразнее, чем предполагалось до сих пор. В принципе, текстовые

части внутри одного и того же почтового сообщения могут быть представлены в различных кодировках (например, три HTML-файла, вложенные в письмо, в разных кодировках, возможно, отличающихся от кодировки полного текста сообщения). Обращение с таким текстом как со строками двоичных байтов иногда позволяет ловко обойти проблемы кодирования, однако при сохранении таких частей в текстовых файлах необходимо учитывать оригинальные кодировки. Кроме того, любые операции по обработке текста, применяемые к этим частям, точно так же зависят от типа.

К счастью, пакет email добавляет заголовки с названием кодировок при создании текста сообщения и позволяет получить информацию о кодировках для частей, где она указана, при анализе текста сообщения. Например, при добавлении вложений с текстом, который не является текстом ASCII, нужно просто указать имя кодировки – соответствующие заголовки будут добавлены автоматически при создании полного текста сообщения, а кодировки можно будет получить непосредственно с помощью метода get_content_charset:

```
>>> s = b'A\xe4B'
>>> s.decode('latin1')
'AäB'

>>> from email.message import Message
>>> m = Message()
>>> m.set_payload(b'A\xe4B', charset='latin1') # или 'latin-1': см. далее
>>> t = m.as_string()
>>> print(t)
MIME-Version: 1.0
Content-Type: text/plain; charset="latin1"
Content-Transfer-Encoding: base64

QeRC

>>> m.get_content_charset()
'latin1'
```

Обратите внимание, что пакет email автоматически применяет MIME-кодирование Base64 к частям с текстом, который не является текстом ASCII, чтобы обеспечить соответствие стандартам электронной почты. То же самое справедливо для более специализированных подклассов с поддержкой преобразования в формат MIME класса Message:

```
>>> from email.mime.text import MIMEText
>>> m = MIMEText(b'A\xe4B', _charset='latin1')
>>> t = m.as_string()
>>> print(t)
Content-Type: text/plain; charset="latin1"
MIME-Version: 1.0
Content-Transfer-Encoding: base64
```

```
QeRC

>>> m.get_content_charset()
'latin1'
```

Если теперь проанализировать текст этого сообщения с помощью пакета email, мы получим новый объект Message, содержимое которого в формате MIME Base64 представляет строку Юникода с символами не из диапазона ASCII. Если затребовать содержимое с декодированием из формата MIME, указав аргумент decode=1, будет возвращена строка байтов, которую мы ранее вложили в сообщение:

```
>>> from email.parser import Parser
>>> q = Parser().parsestr(t)
>>> q
<email.message.Message object at 0x019ECA50>
>>> q.get_content_type()
'text/plain'
>>> q._payload
'QeRC\n'
>>> q.get_payload()
'QeRC\n'
>>> q.get_payload(decode=1)
b'A\xe4B'
```

Однако попытка получить текст декодированием этой строки байтов с использованием кодировки по умолчанию в Windows (UTF8) потерпит неудачу. Для большей точности и поддержки большего разнообразия типов текста необходимо использовать информацию о кодировках, сохраненную при анализе и прикрепленную к объекту Message. Это особенно важно, когда может возникнуть необходимость сохранить данные в файл – мы должны будем либо сохранять данные в двоичном режиме, либо указывать корректную (или, по крайней мере, совместимую) кодировку, чтобы сохранять такие строки в текстовых файлах. Декодирование вручную выполняется точно так же:

```
>>> q.get_payload(decode=1).decode()
UnicodeDecodeError: 'utf8' codec can't decode bytes in position 1-2:
unexpected...
```

(UnicodeDecodeError: кодек 'utf8' не может декодировать байты в позиции 1–2: неожиданный...)

```
>>> q.get_content_charset()
'latin1'
>>> q.get_payload(decode=1).decode('latin1')                # известный тип
'AäB'
>>> q.get_payload(decode=1).decode(q.get_content_charset())  # для любого
                                                            # типа
'AäB'
```

На самом деле, в объектах Message имеется информация обо всех заголовках, нужно только знать, как ее найти. Информация о кодировке может вообще отсутствовать, и в этом случае возвращается значение None. Клиенты должны предусматривать политику обращения с таким неоднозначным текстом (они могут попытаться применить наиболее распространенные кодировки, определить кодировку по содержимому или обращаться с данными, как с простой строкой байтов):

```
>>> q['content-type']                    # интерфейс отображения
'text/plain; charset="latin1"'
>>> q.items()
[('Content-Type', 'text/plain; charset="latin1"'), ('MIME-Version', '1.0'),
('Content-Transfer-Encoding', 'base64')]

>> q.get_params(header='Content-Type')  # интерфейс параметров
[('text/plain', ''), ('charset', 'latin1')]
>>> q.get_param('charset', header='Content-Type')
'latin1'

>>> charset = q.get_content_charset()   # информация может отсутствовать
>>> if charset:
...     print(q.get_payload(decode=1).decode(charset))
...
AäB
```

Так обрабатываются кодировки текстовых частей, полученных в результате анализа электронных писем. При составлении новых почтовых сообщений нам по-прежнему необходимо применять пользовательские настройки сеанса или дать пользователю возможность указывать кодировку для каждой части в интерактивном режиме. В некоторых клиентах электронной почты, представленных в этой книге, преобразование содержимого выполняется по мере необходимости. При этом используется информация о кодировках в заголовках сообщений, полученных в результате анализа, и предоставленная пользователями в ходе составления новых электронных писем.

Кодировка заголовков сообщения: поддержка в пакете email

Кроме того, пакет email предоставляет также поддержку кодирования и декодирования самих заголовков сообщений (таких как «From», «Subject») в соответствии со стандартами электронной почты, когда они не являются простым текстом. Такие заголовки иногда называют *интернационализированными* (или *i18n*) заголовками, потому что они поддерживают включение в почтовые сообщения символов, не входящих в набор ASCII. Этот термин иногда используется также для обозначения кодированного текста содержимого в сообщениях. Однако, в отличие от заголовков сообщений, при создании содержимого сообщений кодиро-

ванию подвергается не только текст с интернациональными символами Юникода, но и действительно двоичные данные, такие как изображения (как будет показано в следующем разделе).

Как и информационное наполнение, интернационализированные заголовки в почтовых сообщениях кодируются особым образом и могут также кодироваться с применением кодировок Юникода. Например, следующий фрагмент демонстрирует, как можно декодировать заголовок темы сообщения, возможно, являющегося спамом, которое только что появилось в моем почтовом ящике. Преамбула =?UTF-8?Q? в заголовке свидетельствует, что данные, следующие за ней, являются текстом Юникода, закодированным с помощью кодировки UTF-8, который дополнительно был преобразован в MIME-формат quoted-printable для передачи по электронной почте (в отличие от примеров содержимого, приводившихся в предыдущем разделе, кодировка которых объявляется в отдельных заголовках, сами заголовки могут встраивать свои схемы кодирования Юникода и MIME прямо в свое содержимое, как в данном случае):

```
>>> rawheader = '=?UTF-8?Q?Introducing=20Top=20Values=3A=20A=20Special=20Selec
tion=20of=20Great=20Money=20Savers?='

>>> from email.header import decode_header # выполнить декодирование
>>> decode_header(rawheader)               # email+MIME
[(b'Introducing Top Values: A Special Selection of Great Money Savers',
'utf-8')]

>>> bin, enc = decode_header(rawheader)[0] # декодировать в Юникод
>>> bin, enc
(b'Introducing Top Values: A Special Selection of Great Money Savers',
'utf-8')
>>> bin.decode(enc)
'Introducing Top Values: A Special Selection of Great Money Savers'
```

Важно отметить, что пакет email может возвращать несколько частей, если в заголовке имеется несколько подстрок, каждую из которых требуется декодировать отдельно, и затем объединять их, чтобы получить полный текст заголовка. Обратите также внимание, что в версии 3.1 этот пакет возвращает строку bytes, если в заголовке имеется хотя бы одна закодированная подстрока (или весь заголовок), но возвращает строку str, если заголовок не был закодирован, а некодированные подстроки возвращаются в виде строк bytes при применении кодировки «raw-unicode-escape», которую удобно использовать для преобразования строки str в строку bytes, когда никакое кодирование не применяется:

```
>>> from email.header import decode_header
>>> S1 = 'Man where did you get that assistant?'
>>> S2 = '=?utf-8?q?Man_where_did_you_get_that_assistant=3F?='
```

```
>>> S3 = 'Man where did you get that =?UTF-8?Q?assistant=3F?='

# str: не требуется вызывать decode()
>>> decode_header(S1)
[('Man where did you get that assistant?', None)]

# bytes: требуется вызвать decode()
>>> decode_header(S2)
[(b'Man where did you get that assistant?', 'utf-8')]

# bytes: требуется вызвать decode(), пакет применит
# кодировку raw-unicode-escape
>>> decode_header(S3)
[(b'Man where did you get that', None), (b'assistant?', 'utf-8')]

# объединить декодированные части, если их несколько
>>> parts = decode_header(S3)
>>> ' '.join(abytes.decode('raw-unicode-escape' if enc == None else enc)
...          for (abytes, enc) in parts)
'Man where did you get that assistant?'
```

К логике, подобной использованной здесь на последнем шаге, мы будем обращаться в пакете `mailtools` и далее, но также будем оставлять подстроки `str` нетронутыми, не пытаясь их декодировать.

Самые последние новости: В середине 2010 года, когда я пишу эти строки, кажется вполне возможным, что такое неполиморфическое и, честно признаться, непитоническое поведение прикладного интерфейса, смешивающего разные типы данных, изменится в будущих версиях Python. В ответ на тираду, посланную в список рассылки разработчиков Python автором книги, с работами которого вы, возможно, знакомы, разгорелось жаркое обсуждение этой темы. Среди прочих родилась идея создать тип, подобный `bytes`, который явно будет нести в себе название кодировки – в некоторых ситуациях такой тип позволил бы организовать обработку текстовых данных более универсальным образом. Заранее невозможно предсказать, во что выльются такие идеи, но приятно видеть, когда они активно обсуждаются. Следите за информацией на веб-сайте книги, где будет сообщаться об изменениях в API библиотеки Python 3.X и обо всем, что касается поддержки Юникода.

Кодировка заголовков с адресами: при анализе и создании сообщения

Примечание к предыдущему разделу: в заголовках сообщений, содержащих *адреса электронной почты* (такие как «From»), компонент «имя» в паре имя/адрес также может кодироваться подобным образом. Поскольку пакет `email` при анализе ожидает, что кодированные подстроки будут завершаться пробельным символом или простираться до конца

строки, мы не можем выполнить декодирование всего заголовка с адресом – кавычки вокруг компонента с именем будут вызывать неудачу.

Для поддержки таких интернационализированных заголовков с адресами необходимо извлечь первую часть адреса электронной почты и затем декодировать ее. Прежде всего, необходимо с помощью инструментов в пакете email извлечь имя и адрес:

```
>>> from email.utils import parseaddr, formataddr
>>> p = parseaddr('"Smith, Bob" <bob@bob.com>')  # разбить пару имя/адрес
>>> p                                            # адрес - некодированный
('Smith, Bob', 'bob@bob.com')
>>> formataddr(p)
'"Smith, Bob" <bob@bob.com>'

>>> parseaddr('Bob Smith <bob@bob.com>')         # имя без кавычек
('Bob Smith', 'bob@bob.com')
>>> formataddr(parseaddr('Bob Smith <bob@bob.com>'))
'Bob Smith <bob@bob.com>'

>>> parseaddr('bob@bob.com')                     # простой адрес, без имени
('', 'bob@bob.com')
>>> formataddr(parseaddr('bob@bob.com'))
'bob@bob.com'
```

В заголовках с несколькими адресами (например, «To») адреса отделяются друг от друга запятыми. Так как имена в адресах также могут содержать запятые, простое разбиение по запятым не всегда дает желаемый результат. Однако для анализа отдельных адресов можно использовать еще одну утилиту: функция getaddresses игнорирует запятые в именах и разбивает содержимое заголовка на отдельные адреса. Функция parseaddr делает то же самое, потому что она просто возвращает первую пару из результата, полученного вызовом функции getaddresses (в следующем примере кое-где были добавлены пустые строки для большей удобочитаемости):

```
>>> from email.utils import getaddresses
>>> multi = '"Smith, Bob" <bob@bob.com>, Bob Smith <bob@bob.com>, bob@bob.com,
"Bob" <bob@bob.com>'

>>> getaddresses([multi])
[('Smith, Bob','bob@bob.com'),('Bob Smith','bob@bob.com'),('','bob@bob.com'),
('Bob', 'bob@bob.com')]

>>> [formataddr(pair) for pair in getaddresses([multi])]
['"Smith, Bob" <bob@bob.com>', 'Bob Smith <bob@bob.com>', 'bob@bob.com',
'Bob <bob@bob.com>']

>>> ', '.join([formataddr(pair) for pair in getaddresses([multi])])
'"Smith, Bob" <bob@bob.com>, Bob Smith <bob@bob.com>, bob@bob.com,
Bob <bob@bob.com>'
```

```
>>> getaddresses(['bob@bob.com'])  # также корректно обрабатывает простые
('', 'bob@bob.com')]               # адреса без имени
```

Итак, декодирование адресов электронной почты включает всего лишь один шаг перед применением обычной логики декодирования заголовков, которую мы видели выше, и один шаг после:

```
>>> rawfromheader = '"=?UTF-8?Q?Walmart?=" <newsletters@walmart.com>'

>>> from email.utils import parseaddr, formataddr
>>> from email.header import decode_header

>>> name, addr = parseaddr(rawfromheader)   # разбить на части имя/адрес
>>> name, addr
('=?UTF-8?Q?Walmart?=', 'newsletters@walmart.com')

>>> abytes, aenc = decode_header(name)[0]   # выполнить
                                            # декодирование email+MIME
>>> abytes, aenc
(b'Walmart', 'utf-8')

>>> name = abytes.decode(aenc)              # декодировать в Юникод
>>> name
'Walmart'

>>> formataddr((name, addr))                # объединить компоненты адреса
'Walmart <newsletters@walmart.com>'
```

Заголовки «From» обычно содержат только один адрес, однако для большей надежности необходимо применять этот прием ко всем заголовкам с адресами, таким как «To», «Cc» и «Bcc». Напомню, что утилита `getaddresses` корректно распознает запятые внутри имен и не путает их с запятыми, отделяющими разные адреса, а так как она корректно обрабатывает случай единственного адреса, ее можно применять для обработки заголовков «From»:

```
>>> rawfromheader = '"=?UTF-8?Q?Walmart?=" <newsletters@walmart.com>'
>>> rawtoheader = rawfromheader + ', ' + rawfromheader
>>> rawtoheader
'"=?UTF-8?Q?Walmart?=" <newsletters@walmart.com>, "=?UTF-8?Q?Walmart?="
<newsletters@walmart.com>'

>>> pairs = getaddresses([rawtoheader])
>>> pairs
[('=?UTF-8?Q?Walmart?=', 'newsletters@walmart.com'), ('=?UTF-8?Q?Walmart?=',
'newsletters@walmart.com')]

>>> addrs = []
>>> for name, addr in pairs:
...     abytes, aenc = decode_header(name)[0]  # декодирование email+MIME
...     name = abytes.decode(aenc)             # декодирование в Юникод
...     addrs.append(formataddr((name, addr))) # один или более адресов
```

```
...
>>> ', '.join(addrs)
'Walmart <newsletters@walmart.com>, Walmart <newsletters@walmart.com>'
```

Эти инструменты в целом могут принимать некодированное содержимое и возвращают его нетронутым. Однако для большей надежности в последней части примера выше необходимо также учесть, что decode_header может возвращать несколько частей (для кодированных подстрок), значение None в качестве имен кодировок (для некодированных подстрок) и подстроки типа str вместо bytes (для полностью некодированных имен).

В данном алгоритме предусматриваются обе разновидности декодирования полученных почтовых сообщений – из формата MIME и в текст Юникода. Создание корректно закодированных заголовков для включения в новые почтовые сообщения выполняется так же просто:

```
>>> from email.header import make_header
>>> hdr = make_header([(b'A\xc4B\xe4C', 'latin-1')])
>>> print(hdr)
AÄBäC
>>> print(hdr.encode())
=?iso-8859-1?q?A=C4B=E4C?=
>>> decode_header(hdr.encode())
[(b'A\xc4B\xe4C', 'iso-8859-1')]
```

Этот прием может применяться как к целым заголовкам, таким как «Subject», так и к отдельным их компонентам, например к именам в заголовках с адресами электронной почты, таких как «From» и «To» (при необходимости используйте getaddresses, чтобы разбить содержимое заголовка на отдельные адреса). Объект заголовка предоставляет альтернативный интерфейс – оба приема учитывают дополнительные детали, такие как длина строк, за которыми я отсылаю вас к руководствам по языку Python:

```
>>> from email.header import Header
>>> h = Header(b'A\xe4B\xc4X', charset='latin-1')
>>> h.encode()
'=?iso-8859-1?q?A=E4B=C4X?='
>>>
>>> h = Header('spam', charset='ascii') # то же, что и Header('spam')
>>> h.encode()
'spam'
```

Пакет mailtools, описываемый далее, и его клиент PyMailGUI, который будет представлен в главе 14, используют эти интерфейсы для автоматического декодирования заголовков полученных почтовых сообщений в отображаемое содержимое и для кодирования заголовков отправляемых сообщений, которые содержат символы не из диапазона ASCII. Кроме того, в последнем случае кодирование применяется также к компоненту имени в адресах электронной почты, и предполагается, что

серверы SMTP правильно воспринимают такие адреса. На некоторых серверах SMTP это может приводить к проблемам, решение которых мы не имеем возможности описать из-за нехватки места в книге. Дополнительные сведения по обработке заголовков на серверах SMTP ищите в Интернете. Кроме того, дополнительную информацию о декодировании заголовков можно найти в файле *_test-i18n-headers.py* в пакете с примерами. В нем с помощью методов `mailtools` реализовано декодирование дополнительных заголовков, имеющих отношение к теме сообщения и к адресам, и их отображение в виджете `Text` из библиотеки tkinter – демонстрируя, как они будут отображаться в PyMailGUI.

Обходное решение: ошибка создания текста сообщения при наличии двоичных вложений

Последние две проблемы с поддержкой Юникода в пакете `email` являются самыми настоящими ошибками, из-за которых сейчас приходится искать обходные решения, но которые наверняка будут исправлены в одной из будущих версий Python. Первая из них препятствует созданию текста практически всех, кроме самых тривиальных, сообщений – нынешняя версия пакета `email` больше не поддерживает возможность создания полного текста сообщений, содержащих любые двоичные части, такие как изображения или аудиофайлы. Без обходного решения с помощью пакета `email` в Python 3.1 можно создавать только самые простые почтовые сообщения, содержащие исключительно текстовые части, – любое двоичное содержимое в формате MIME вызывает ошибку при создании полного текста сообщения.

Это сложно понять без детального изучения исходных текстов пакета `email` (что, к счастью, нам доступно в мире открытых исходных текстов). Но чтобы продемонстрировать проблему, сначала посмотрите, как отображается простое текстовое содержимое в полный текст сообщения при печати:

```
C:\...\PP4E\Internet\Email> python
>>> from email.message import Message       # обобщенный объект сообщения
>>> m = Message()
>>> m['From'] = 'bob@bob.com'
>>> m.set_payload(open('text.txt').read())  # содержимое - текст str
>>> print(m)                                # print использует as_string()
From: bob@bob.com

spam
Spam
SPAM!
```

Мы также уже видели, что для удобства пакет `email` предоставляет специализированные подклассы класса `Message`, предназначенные для добавления заголовков с дополнительной описательной информацией, которая используется почтовыми клиентами, чтобы определить, как обрабатывать данные:

```
>>> from email.mime.text import MIMEText  # подкласс Message с заголовками
>>> text = open('text.txt').read()
>>> m = MIMEText(text)                    # содержимое - текст str
>>> m['From'] = 'bob@bob.com'
>>> print(m)
Content-Type: text/plain; charset="us-ascii"
MIME-Version: 1.0
Content-Transfer-Encoding: 7bit
From: bob@bob.com

spam
Spam
SPAM!
```

С текстом никаких проблем не возникает, но посмотрите, что произойдет, если попытаться отобразить часть сообщения с действительно двоичными данными, такими как изображение, которое не может быть декодировано в текст Юникода:

```
>>> from email.message import Message  # обобщенный объект сообщения
>>> m = Message()
>>> m['From'] = 'bob@bob.com'
>>> bytes = open('monkeys.jpg', 'rb').read()  # прочитать байты (не Юникод)
>>> m.set_payload(bytes)               # установить, что содержимое двоичное
>>> print(m)
Traceback (most recent call last):
  ...часть строк опущена...
  File "C:\Python31\lib\email\generator.py", line 155, in _handle_text
    raise TypeError('string payload expected: %s' % type(payload))
TypeError: string payload expected: <class 'bytes'>
(TypeError: ожидается строковое содержимое: <class 'bytes'>)

>>> m.get_payload()[:20]
b'\xff\xd8\xff\xe0\x00\x10JFIF\x00\x01\x01\x01\x00x\x00x\x00\x00'
```

Проблема состоит в том, что механизм создания полного текста сообщения в пакете email предполагает получить данные для содержимого в виде строки str в MIME-кодировке Base64 (или подобной ей), но не bytes. В действительности, в данном случае, вероятно, мы сами виноваты в появлении ошибки, потому что мы устанавливаем двоичное содержимое вручную. Мы должны использовать подкласс MIMEImage с поддержкой кодирования в формат MIME, предназначенный для изображений, – в этом случае пакет email автоматически выполнит MIME-кодирование Base64 для данных в процессе создания объекта сообщения. К сожалению, само содержимое останется при этом строкой bytes, а не str, несмотря на то, что главная цель кодирования Base64 как раз и состоит в том, чтобы преобразовать двоичные данные в текст (хотя точная разновидность Юникода, в которую должен трансформироваться этот текст, может оставаться неясной). Это ведет к появлению других ошибок в Python 3.1:

```
>>> from email.mime.image import MIMEImage    # подкласс Message
                                              # с заголовками+base64
>>> bytes = open('monkeys.jpg', 'rb').read()  # снова прочитать
                                              # двоичные данные
>>> m = MIMEImage(bytes)                 # MIME-класс выполнит кодирование
>>> print(m)                             # данных в формат Base64
Traceback (most recent call last):
  ...часть строк опущена...
  File "C:\Python31\lib\email\generator.py", line 155, in _handle_text
    raise TypeError('string payload expected: %s' % type(payload))
TypeError: string payload expected: <class 'bytes'>
(TypeError: ожидается строковое содержимое: <class 'bytes'>)

>>> m.get_payload()[:40]                    # это уже текст в формате Base64
b'/9j/4AAQSkZJRgABAQEAeAB4AAD/2wBDAAIBAQIB'

>>> m.get_payload()[:40].decode('ascii')    # но в действительности
'/9j/4AAQSkZJRgABAQEAeAB4AAD/2wBDAAIBAQIB'  # это строка bytes!
```

Иными словами, разделение типов str/bytes в Python 3.X не только не поддерживается полностью пакетом email в Python 3.1, но и фактически нарушает его работоспособность. К счастью, данная проблема поддается решению.

Чтобы решить эту конкретную проблему, я создал собственную функцию MIME-кодирования двоичных вложений и передал ее всем подклассам с поддержкой кодирования в формат MIME, предназначенным для создания двоичных вложений. Эта функция реализована в пакете mailtools, который будет представлен далее в этой главе (пример 13.23). Поскольку пакет email использует ее для преобразования данных типа bytes в текст на этапе инициализации, появляется возможность декодировать текст ASCII в Юникод после преобразования двоичных данных в формат Base64 и установить заголовки с информацией о кодировке содержимого. То обстоятельство, что пакет email не выполняет это дополнительное декодирование в Юникод, можно считать ошибкой в пакете (хотя она и вызвана изменениями где-то в другом месте в стандартной библиотеке Python), но обходное решение справляется с заданием:

```
# в модуле mailtools.mailSender, далее в этой главе...
def fix_encode_base64(msgobj):
    from email.encoders import encode_base64
    encode_base64(msgobj)         # пакет email оставляет данные в виде bytes
    bytes = msgobj.get_payload()  # операция создания текста терпит неудачу
                                  # при наличии двоичных данных
    text  = bytes.decode('ascii') # декодировать в str, чтобы обеспечить
                                  # создание текста
    ...логика разбиения строк опущена...
    msgobj.set_payload('\n'.join(lines))
```

```
>>> from email.mime.image import MIMEImage
>>> from mailtools.mailSender import fix_encode_base64  # использовать
                                                        # решение
>>> bytes = open('monkeys.jpg', 'rb').read()
>>> m = MIMEImage(bytes, _encoder=fix_encode_base64)    # преобразовать
                                                        # в ascii str
>>> print(m.as_string()[:500])
Content-Type: image/jpeg
MIME-Version: 1.0
Content-Transfer-Encoding: base64

/9j/4AAQSkZJRgABAQEAeAB4AAD/2wBDAAIBAQIBAQICAgICAgICAwUDAwMDAwYEBAMFBwYHBwcG
BwcICQsJCAgKCAcHCgOKCgsMDAwMBwkODwOMDgsMDAz/2wBDAQICAgMDAwYDAwYMCAcIDAwMDAwM
DAwMDAwMDAwMDAwMDAwMDAwMDAwMDAwMDAwMDAwMDAwMDAwMDAwMDAwMDAwMDAz/wAARCAHoAvQDASIA
AhEBAxEB/8QAHwAAAQUBAQEBAQEAAAAAAAAAAECAwQFBgcICQoL/8QAtRAAAgEDAwIEAwUFBAQA
AAF9AQIDAAQRBRIhMUEGE1FhByJxFDKBkaEIIOKxwRVSOfAkM2JyggkKFhcYGRolJicoKSoONTY3
ODk6QORFRkdISUpTVFVWV1hZWmNkZWZnaGlqc

>>> print(m)   # вывести все сообщение: очень длинное
```

Другое возможное обходное решение вовлекает определение собственного класса `MIMEImage`, похожего на оригинальный, но не выполняющего кодирование Base64 на этапе создания. При таком подходе мы должны сами выполнять кодирование и преобразование данных в тип `str` перед созданием объекта сообщения, но по-прежнему можем использовать логику создания заголовков из оригинального класса. Однако если вы пойдете таким путем, то обнаружите, что он требует повторения (то есть простого копирования) слишком большого объема использованной в оригинале логики, чтобы считаться достаточно разумным, – в этом повторяющемся программном коде придется в будущем отражать все изменения в пакете `email`:

```
>>> from email.mime.nonmultipart import MIMENonMultipart
>>> class MyImage(MIMENonMultipart):
...     def __init__(self, imagedata, subtype):
...         MIMENonMultipart.__init__(self, 'image', subtype)
...         self.set_payload(_imagedata)

...повторить всю логику кодирования base64 с дополнительным
декодированием в Юникод...
>>> m = MyImage(text_from_bytes)
```

Интересно отметить, что эта регрессия в пакете `email` фактически отражает изменения в модуле `base64`, выполненные в 2007 году и никак не связанные с поддержкой Юникода, которые были вполне разумными, пока не было сделано строгое разделение типов `bytes/str` в Python 3.X. До этого механизм кодирования электронной почты работал в Python 2.X по той простой причине, что тип `bytes` в действительности был типом `str`. Однако в версии 3.X, так как `base64` возвращает строки `bytes`, механизм кодирования электронной почты в пакете `email` также оставляет

содержимое в виде строк bytes, даже при том, что в результате кодирования получается текст в формате Base64. Это в свою очередь нарушает работу механизма создания полного текста сообщения в пакете email, потому что в данном случае он ожидает получить содержимое в виде текста и требует, чтобы оно было строкой str. Как и в любых крупномасштабных программных системах, эффект воздействия некоторых изменений в 3.X, возможно, было трудно предугадать или полностью согласовать.

Анализ двоичных вложений (в противоположность созданию текста из них), напротив, прекрасно работает в Python 3.X, потому что полученное содержимое сохраняется в объектах сообщений в виде строк str в формате Base64, а не в виде строк bytes, и преобразуется в тип bytes только при извлечении. Эта ошибка, скорее всего, также будет исправлена в будущих версиях Python и пакета email (вероятно, даже в виде простой «заплаты» в Python 3.2), но она является более серьезной по сравнению с другими проблемами декодирования в Юникод, описанными здесь, потому что препятствует составлению почтовых сообщений, кроме самых простых.

Гибкость, предоставляемая пакетом и языком Python, позволяет разрабатывать подобные обходные решения, внешние по отношению к пакету, вместо того, чтобы вторгаться непосредственно в программный код пакета. С открытым программным обеспечением и прикладным интерфейсом, терпимым к ошибкам, вы редко будете попадать в безвыходные ситуации.

Самые последние новости: Ошибка, описанная в этом разделе, намечена к исправлению в версии Python 3.2[1], представленные здесь обходные решения могут оказаться ненужными в этой и в следующих версиях Python. Это выяснилось в ходе диалога с членами специальной заинтересованной группы по проблемам пакета email (в списке рассылки «email-sig»).

К сожалению, это исправление не появилось до момента окончания работы над этой главой и над примерами в ней. Я с удовольствием удалил бы обходное решение и его описание полностью, но эта книга основана на версии Python 3.1, которая существовала до и останется существовать после внесения этих исправлений.

[1] Действительно, описываемая проблема была исправлена в версии Python 3.2. И, как пишет автор на сайте поддержки книги: «Исправления в библиотеке версии 3.2 включают важные улучшения в пакете email по сравнению с версией 3.1: решение проблемы декодирования в тип str, а также другие обходные решения для пакета email, которые приводятся в главе 13, стали излишними в Python 3.2. В то же время, эти обходные решения, необходимые при работе с пакетом email в версии 3.1, можно безопасно использовать в версии 3.2 и их следует рассматривать как примеры решения проблем, с которыми приходится сталкиваться в практической деятельности, что является основной темой книги». – *Прим. перев.*

Однако, чтобы это обходное решение работало также под управлением альфа-версии Python 3.2, его реализация была скорректирована непосредственно перед публикацией книги, куда была добавлена проверка на принадлежность типу bytes перед декодированием. Кроме того, в обходном решении по-прежнему необходимо вручную разбивать строки в данных Base64, потому что это еще не реализовано в версии 3.2.

Обходное решение: ошибка при создании текстовых частей с символами не из диапазона ASCII

Последняя проблема, связанная с поддержкой Юникода в пакете email, является столь же серьезной, как и предыдущая: изменения, подобные тем, о которых рассказывалось в предыдущем разделе, вызвали еще одну регрессию в работе механизма создания новых почтовых сообщений. Она выражается в невозможности создавать текстовые части сообщений без адаптации под различные кодировки Юникода.

Тексты некоторых типов автоматически подвергаются преобразованию в формат MIME, пригодный для отправки. К сожалению, из-за строгого разделения типов str/bytes класс MIMEText в пакете email теперь требует указывать разные типы строковых объектов для различных кодировок Юникода. В результате вам теперь необходимо точно знать, как пакет email обрабатывает текстовые данные при создании объектов текстовых сообщений, или повторять значительную часть логики этого класса в своих подклассах.

Например, ниже показано, как в настоящее время необходимо обрабатывать текст наиболее распространенных типов, чтобы корректно создать заголовки с информацией о кодировках и применить необходимое преобразование MIME:

```
>>> m = MIMEText('abc', _charset='ascii')    # передать текст
                                             # для кодировки ascii
>>> print(m)
MIME-Version: 1.0
Content-Type: text/plain; charset="us-ascii"
Content-Transfer-Encoding: 7bit

abc
>>> m = MIMEText('abc', _charset='latin-1')  # текст для кодировки latin-1
>>> print(m)                                 # но не для 'latin1': см. далее
MIME-Version: 1.0
Content-Type: text/plain; charset="iso-8859-1"
Content-Transfer-Encoding: quoted-printable

abc
>>> m = MIMEText(b'abc', _charset='utf-8')  # передать строку bytes для utf8
>>> print(m)
```

```
Content-Type: text/plain; charset="utf-8"
MIME-Version: 1.0
Content-Transfer-Encoding: base64

YWJj
```

Все работает, но если посмотреть внимательно, можно заметить, что в первых двух случаях мы должны передавать строку `str`, а в третьем – строку `bytes`. Требование такого особого подхода к типам представления Юникода исходит из особенностей внутренней реализации пакета. Передача типов, отличных от тех, что ожидаются для представления Юникода, вызывает ошибку, что обусловлено появлением внутри пакета `email` в версии 3.1 недопустимых комбинаций типов `str`/`bytes`:

```
>>> m = MIMEText('abc', _charset='ascii')
>>> m = MIMEText(b'abc', _charset='ascii')  # ошибка: предполагается str 2.X
Traceback (most recent call last):
  ...часть строк опущена...
  File "C:\Python31\lib\email\encoders.py", line 60, in encode_7or8bit
    orig.encode('ascii')
AttributeError: 'bytes' object has no attribute 'encode'
```
(AttributeError: объект 'bytes' не имеет атрибута 'encode')

```
>>> m = MIMEText('abc', _charset='latin-1')
>>> m = MIMEText(b'abc', _charset='latin-1')  # ошибка: qp использует str
Traceback (most recent call last):
  ...часть строк опущена...
  File "C:\Python31\lib\email\quoprimime.py", line 176, in body_encode
    if line.endswith(CRLF):
TypeError: expected an object with the buffer interface
```
(TypeError: ожидается объект с интерфейсом буфера)

```
>>> m = MIMEText(b'abc', _charset='utf-8')
>>> m = MIMEText('abc', _charset='utf-8')  # ошибка: base64 использует bytes
Traceback (most recent call last):
  ...часть строк опущена...
  File "C:\Python31\lib\email\base64mime.py", line 94, in body_encode
    enc = b2a_base64(s[i:i + max_unencoded]).decode("ascii")
TypeError: must be bytes or buffer, not str
```
(TypeError: должен быть объект типа bytes или buffer, а не str)

Кроме того, пакет `email` более придирчив к синонимам имен кодировок, чем Python и большинство других инструментов: кодировка «latin-1» определяется как MIME-тип quoted-printable, но кодировка «latin1» считается неизвестной и поэтому она определяется как MIME-тип по умолчанию Base64. Именно по этой причине для кодировки «latin1» выбирался MIME-тип Base64 в приведенных ранее примерах этого раздела – такой выбор формата MIME считается некорректным для любых получателей, которые распознают синоним «latin1», включая сам Python. К сожалению, это также означает, что при использовании синони-

ма, который не распознается пакетом, мы должны передавать строку другого типа:

```
>>> m = MIMEText('abc', _charset='latin-1') # str для 'latin-1'
>>> print(m)
MIME-Version: 1.0
Content-Type: text/plain; charset="iso-8859-1"
Content-Transfer-Encoding: quoted-printable

abc
>>> m = MIMEText('abc', _charset='latin1')
Traceback (most recent call last):
  ...часть строк опущена...
  File "C:\Python31\lib\email\base64mime.py", line 94, in body_encode
    enc = b2a_base64(s[i:i + max_unencoded]).decode("ascii")
TypeError: must be bytes or buffer, not str
```

(TypeError: должен быть объект типа bytes или buffer, а не str)

```
>>> m = MIMEText(b'abc', _charset='latin1') # bytes для 'latin1'!
>>> print(m)
Content-Type: text/plain; charset="latin1"
MIME-Version: 1.0
Content-Transfer-Encoding: base64

YWJj
```

Существуют разные способы добавления новых типов кодировок и псевдонимов в пакет email, но они не являются стандартными. Программы, для которых надежность имеет большое значение, должны перепроверять имена кодировок, введенные пользователем, которые могут быть допустимыми для Python, но незнакомыми пакету email. То же относится и к данным, которые вообще не являются символами ASCII, – вам сначала придется декодировать их в текст, чтобы использовать ожидаемое имя «latin-1», потому что для преобразования в формат MIME quoted-printable требуется строка str, хотя для преобразования в формат MIME Base64, когда указывается кодировка «latin1», требуется строка bytes:

```
>>> m = MIMEText(b'A\xe4B', _charset='latin1')
>>> print(m)
Content-Type: text/plain; charset="latin1"
MIME-Version: 1.0
Content-Transfer-Encoding: base64

QeRC
>>> m = MIMEText(b'A\xe4B', _charset='latin-1')
Traceback (most recent call last):
  ...часть строк опущена...
  File "C:\Python31\lib\email\quoprimime.py", line 176, in body_encode
    if line.endswith(CRLF):
```

```
TypeError: expected an object with the buffer interface
```
(TypeError: ожидается объект с интерфейсом буфера)

```
>>> m = MIMEText(b'A\xe4B'.decode('latin1'), _charset='latin-1')
>>> print(m)
MIME-Version: 1.0
Content-Type: text/plain; charset="iso-8859-1"
Content-Transfer-Encoding: quoted-printable

A=E4B
```

Фактически объект текстового сообщения не проверяет, являются ли данные, предназначенные вами для преобразования в формат MIME, допустимыми символами Юникода – мы сможем отправить недопустимый текст в кодировке UTF, но у получателя могут возникнуть проблемы при попытке декодировать его:

```
>>> m = MIMEText(b'A\xe4B', _charset='utf-8')
>>> print(m)
Content-Type: text/plain; charset="utf-8"
MIME-Version: 1.0
Content-Transfer-Encoding: base64

QeRC
>>> b'A\xe4B'.decode('utf8')
UnicodeDecodeError: 'utf8' codec can't decode bytes in position 1-2:
unexpected...
```
(UnicodeDecodeError: кодек 'utf8' не может декодировать байты в позиции 1-2: неожиданный...)

```
>>> import base64
>>> base64.b64decode(b'QeRC')
b'A\xe4B'
>>> base64.b64decode(b'QeRC').decode('utf')
UnicodeDecodeError: 'utf8' codec can't decode bytes in position 1-2:
unexpected...
```
(UnicodeDecodeError: кодек 'utf8' не может декодировать байты а позиции 1-2: неожиданный...)

Так как же быть, если потребуется прикрепить текст к составляемому почтовому сообщению, когда выбор типа данных косвенно зависит от имени его кодировки? Обобщенный суперкласс `Message` не сможет оказать прямую помощь, если мы укажем кодировку, так как он проявляет точно такое же поведение, зависимое от кодировки:

```
>>> m = Message()
>>> m.set_payload('spam', charset='us-ascii')
>>> print(m)
MIME-Version: 1.0
Content-Type: text/plain; charset="us-ascii"
Content-Transfer-Encoding: 7bit
```

```
spam
>>> m = Message()
>>> m.set_payload(b'spam', charset='us-ascii')
AttributeError: 'bytes' object has no attribute 'encode'
```
(AttributeError: объект 'bytes' не имеет атрибута 'encode')

```
>>> m.set_payload('spam', charset='utf-8')
TypeError: must be bytes or buffer, not str
```
(TypeError: должен быть объект типа bytes или buffer, а не str)

Мы могли бы попробовать обойти эти проблемы, повторив большую часть программного кода, который выполняет пакет email, но эта избыточность накрепко привязала бы нас к текущей реализации и ввела бы в зависимость от изменений в будущем. Следующий пример как попугай повторяет шаги, выполняемые пакетом email при создании объекта текстового сообщения с текстом в кодировке ASCII. В отличие от приема с использованием класса MIMEText, данный подход позволяет читать любые данные из файлов в виде строк двоичных байтов, даже если эти данные – простой текст ASCII:

```
>>> m = Message()
>>> m.add_header('Content-Type', 'text/plain')
>>> m['MIME-Version'] = '1.0'
>>> m.set_param('charset', 'us-ascii')
>>> m.add_header('Content-Transfer-Encoding', '7bit')
>>> data = b'spam'
>>> m.set_payload(data.decode('ascii'))  # данные читаются в двоичном виде
>>> print(m)
MIME-Version: 1.0
Content-Type: text/plain; charset="us-ascii"
Content-Transfer-Encoding: 7bit

spam
>>> print(MIMEText('spam', _charset='ascii')) # то же самое,
MIME-Version: 1.0                             # но уже зависит от типа
Content-Type: text/plain; charset="us-ascii"
Content-Transfer-Encoding: 7bit

spam
```

Чтобы сделать то же самое с другими видами текста, требующими преобразования в формат MIME, просто добавьте дополнительный шаг преобразования. Несмотря на то, что здесь мы говорим о текстовых частях, аналогичный подражательный подход можно было бы предпринять для решения проблемы создания текста сообщения с двоичными частями, описанной выше:

```
>>> m = Message()
>>> m.add_header('Content-Type', 'text/plain')
>>> m['MIME-Version'] = '1.0'
```

```
>>> m.set_param('charset', 'utf-8')
>>> m.add_header('Content-Transfer-Encoding', 'base64')
>>> data = b'spam'
>>> from binascii import b2a_base64   # преобразование MIME, если необходимо
>>> data = b2a_base64(data)           # здесь также читаются двоичные данные
>>> m.set_payload(data.decode('ascii'))
>>> print(m)
MIME-Version: 1.0
Content-Type: text/plain; charset="utf-8"
Content-Transfer-Encoding: base64

c3BhbQ==
>>> print(MIMEText(b'spam', _charset='utf-8'))  # то же самое
Content-Type: text/plain; charset="utf-8"       # но уже зависит от типа
MIME-Version: 1.0
Content-Transfer-Encoding: base64

c3BhbQ==
```

Этот прием действует, но помимо избыточности и зависимости, создаваемыми им, для широкого использования такого подхода его необходимо обобщить, – чтобы его можно было применять со всеми возможными кодировками Юникода и форматами MIME, как это уже делает пакет email. Нам может также потребоваться реализовать поддержку синонимов имен кодировок для большей гибкости и тем самым еще больше увеличить избыточность. Иными словами, потребуется выполнить массу дополнительной работы, и в конце нам все равно необходимо будет специализировать свою реализацию для различных типов Юникода.

Какой бы путь мы ни избрали, похоже, что в настоящее время нам не удастся избежать некоторой зависимости от текущей реализации. Пожалуй, самое лучшее, что можно сделать в такой ситуации, кроме как сидеть и надеяться на появление улучшений в пакете email через несколько лет, – это указывать кодировки Юникода в вызовах конструкторов текстовых сообщений и надеяться, что имена кодировок будут известны пакету, а данные в сообщениях будут совместимы с выбранной кодировкой. Ниже приводится, возможно, немного непонятный программный код, который используется в следующем ниже пакете mailtools (пример 13.23) для выбора текстовых типов:

```
>>> from email.charset import Charset, BASE64, QP
>>> for e in ('us-ascii', 'latin-1', 'utf8', 'latin1', 'ascii'):
...     cset = Charset(e)
...     benc = cset.body_encoding
...     if benc in (None, QP):
...         print(e, benc, 'text')    # прочитать/получить данные как str
...     else:
...         print(e, benc, 'binary')  # прочитать/получить данные как bytes
...
```

```
us-ascii None text
latin-1 1 text
utf8 2 binary
latin1 2 binary
ascii None text
```

Мы продолжим действовать в этом направлении далее в книге, помня, что это практически наверняка потребует внесения изменений в программный код в будущем, потому что при таком подходе образуется тесная связь с текущей реализацией email.

Самые последние новости: Есть сведения, что эта ошибка, как и в предыдущем разделе, также будет исправлена в Python 3.2, что сделает данное обходное решение ненужным в этой и в последующих версиях Python. Однако пока неизвестно, каким будет это исправление, и нам по-прежнему требуется решение для версии Python, текущей на момент, когда писалась эта глава. Перед самой публикацией книги уже вышла альфа-версия Python 3.2, в которой по-прежнему присутствует некоторая зависимость от типа, но теперь принимаются текстовые данные в виде str или bytes и при необходимости вызывается преобразование в формат Base64, вместо простого сохранения данных в виде строки bytes.

Итоги: решения и обходные приемы

Пакет email в Python 3.1 содержит мощные инструменты для анализа и составления почтовых сообщений и может использоваться с применением небольшого количества обходных решений как основа полнофункциональных клиентов электронной почты, подобных тем, что приводятся в этой книге. Однако, как вы могли убедиться, на сегодняшний день он недостаточно функционален. Вследствие этого необходима дальнейшая специализация его текущего прикладного интерфейса, что, впрочем, является временным решением. Кроме создания собственных механизмов анализа и составления почтовых сообщений (которые практически невозможно описать в книге конечного размера!) здесь вполне уместно пойти на некоторые компромиссы. Более того, изначальная сложность поддержки Юникода в email устанавливает определенные рамки на дальнейшее развитие этой темы в книге.

В этом издании мы обеспечим поддержку кодировок Юникода для текстовых частей и заголовков при составлении сообщений и будем учитывать их в текстовых частях и заголовках принимаемых сообщений. Однако, чтобы обеспечить такие возможности с «хромающим» пакетом email в Python 3.1, мы будем применять в клиентах электронной почты, представленных в этой книге, следующие приемы работы с Юникодом:

- Для предварительного декодирования полного текста полученного почтового сообщения и для кодирования текстового содержимого отправляемого сообщения будут использоваться пользовательские настройки и умолчания.

- Для декодирования двоичного содержимого, возвращаемого методом `get_payload`, когда потребуется обращаться с текстовыми частями как со строками `str`, будет использоваться информация в заголовках, но в других контекстах мы будем использовать файлы двоичного режима, чтобы избежать ненужных проблем.
- Для кодирования и декодирования заголовков сообщений, таких как «From» и «Subject», если они не являются простым текстом, будут использоваться форматы, предусматриваемые стандартами электронной почты.
- Для решения проблемы создания сообщений с двоичными вложениями будет применяться описанное выше обходное решение.
- В зависимости от типов Юникода и с учетом поведения пакета `email` объекты текстовых сообщений будут конструироваться особым образом.

Эти решения нельзя назвать полными. Например, некоторые клиенты электронной почты в этом издании учитывают кодировки Юникода для текстовых вложений и заголовков, но они не предпринимают никаких дополнительных шагов по кодированию полного текста отправляемых сообщений, кроме тех, что диктуются модулем `smtplib`, и реализуют приемы, которые могут оказаться неудобными в некоторых ситуациях. Но, как мы увидим далее, несмотря на ограничения, наши клиенты по-прежнему будут способны решать сложные задачи и обрабатывать очень широкий круг почтовых сообщений.

Так как развитие Python продолжается непрерывно, следите за информацией на веб-сайте книги, где будет сообщаться об изменениях, которые могут вызвать необходимость корректировки программного кода, использующего пакет `email`. Новые версии пакета `email`, возможно, обеспечат более полную поддержку Юникода. Однако при этом, как это случилось с Python 3.X, в жертву может быть принесена обратная совместимость, что повлечет за собой необходимость изменения программного кода из этой книги. Более подробную информацию по этой проблеме ищите в Интернете и в примечаниях к новым версиям Python.

Этот краткий обзор позволяет получить некоторое представление об основных особенностях интерфейса, тем не менее, чтобы получить более полное представление о возможностях пакета `email`, необходимо перейти к исследованию более крупных примеров. Первый такой пример приводится в следующем разделе.

Почтовый клиент командной строки

Теперь объединим вместе все, что мы узнали о получении, отправке, анализе и составлении сообщений электронной почты, в простом, но функциональном инструменте командной строки для электронной почты. Сценарий в примере 13.20 реализует интерактивный сеанс электронной почты – пользователи могут вводить команды для чтения, от-

правки и удаления электронных писем. Для получения и отправки писем он использует модули poplib и smtplib, а для анализа и составления новых сообщений – пакет email.

Пример 13.20. PP4E\Internet\Email\pymail.py

```
#!/usr/local/bin/python
"""
##############################################################################
pymail - простой консольный клиент электронной почты на языке Python;
использует модуль Python poplib для получения электронных писем,
smtplib для отправки новых писем и пакет email для извлечения
заголовков с содержимым и составления новых сообщений;
##############################################################################
"""

import poplib, smtplib, email.utils, mailconfig
from email.parser import Parser
from email.message import Message
fetchEncoding = mailconfig.fetchEncoding

def decodeToUnicode(messageBytes, fetchEncoding=fetchEncoding):
    """
    4E, Py3.1: декодирует извлекаемые строки bytes в строки str Юникода
    для отображения или анализа; использует глобальные настройки
    (или значения по умолчанию для платформы, исследует заголовки, делает
    обоснованные предположения); в Python 3.2/3.3 этот шаг может оказаться
    необязательным: в этом случае достаточно будет просто вернуть
    сообщение нетронутым;
    """
    return [line.decode(fetchEncoding) for line in messageBytes]

def splitaddrs(field):
    """
    4E: разбивает список адресов по запятым, учитывает возможность
    появления запятых в именах
    """
    pairs = email.utils.getaddresses([field])               # [(name,addr)]
    return [email.utils.formataddr(pair) for pair in pairs] # [name <addr>]

def inputmessage():
    import sys
    From = input('From? ').strip()
    To   = input('To? ').strip()     # заголовок Date
                                     # устанавливается автоматически
    To   = splitaddrs(To)            # допускается множество, name+<addr>
    Subj = input('Subj? ').strip()   # не разбивать вслепую по ',' или ';'
    print('Type message text, end with line="."')
    text = ''
    while True:
        line = sys.stdin.readline()
```

```
        if line == '.\n': break
        text += line
    return From, To, Subj, text

def sendmessage():
    From, To, Subj, text = inputmessage()
    msg = Message()
    msg['From']    = From
    msg['To']      = ', '.join(To)   # для заголовка, не для отправки
    msg['Subject'] = Subj
    msg['Date']    = email.utils.formatdate() # текущие дата
                                              # и время, rfc2822
    msg.set_payload(text)
    server = smtplib.SMTP(mailconfig.smtpservername)
    try:
        failed = server.sendmail(From, To, str(msg))  # может также
    except:                                           # возбудить исключение
        print('Error - send failed')
    else:
        if failed: print('Failed:', failed)

def connect(servername, user, passwd):
    print('Connecting...')
    server = poplib.POP3(servername)
    server.user(user)           # соединиться, зарегистрироваться на сервере
    server.pass_(passwd)        # pass - зарезервированное слово
    print(server.getwelcome())  # print выведет возвращаемое приветствие
    return server

def loadmessages(servername, user, passwd, loadfrom=1):
    server = connect(servername, user, passwd)
    try:
        print(server.list())
        (msgCount, msgBytes) = server.stat()
        print('There are', msgCount, 'mail messages in', msgBytes, 'bytes')
        print('Retrieving...')
        msgList = []                          # получить почту
        for i in range(loadfrom, msgCount+1):  # пусто, если low >= high
            (hdr, message, octets) = server.retr(i)  # сохранить текст
                                                     # в списке
            message = decodeToUnicode(message) # 4E, Py3.1: bytes в str
            msgList.append('\n'.join(message)) # оставить письмо на сервере
    finally:
        server.quit()                         # разблокировать почтовый ящик
    assert len(msgList) == (msgCount - loadfrom) + 1  # нумерация с 1
    return msgList

def deletemessages(servername, user, passwd, toDelete, verify=True):
    print('To be deleted:', toDelete)
    if verify and input('Delete?')[:1] not in ['y', 'Y']:
```

```
        print('Delete cancelled.')
    else:
        server = connect(servername, user, passwd)
        try:
            print('Deleting messages from server...')
            for msgnum in toDelete:     # повторно соединиться
                                        # для удаления писем
                server.dele(msgnum)     # ящик будет заблокирован
                                        # до вызова quit()
        finally:
            server.quit()

def showindex(msgList):
    count = 0                           # вывести некоторые заголовки
    for msgtext in msgList:
        msghdrs = Parser().parsestr(msgtext, headersonly=True) # ожидается
                                                               # тип
        count += 1                      # str в 3.1
        print('%d:\t%d bytes' % (count, len(msgtext)))
        for hdr in ('From', 'To', 'Date', 'Subject'):
            try:
                print('\t%-8s=>%s' % (hdr, msghdrs[hdr]))
            except KeyError:
                print('\t%-8s=>(unknown)' % hdr)
        if count % 5 == 0:
            input('[Press Enter key]') # приостановка через каждые 5 писем

def showmessage(i, msgList):
    if 1 <= i <= len(msgList):
       #print(msgList[i-1])     # устар.: вывести целиком - заголовки+текст
        print('-' * 79)
        msg = Parser().parsestr(msgList[i-1])  # ожидается тип str в 3.1
        content = msg.get_payload()            # содержимое: строка
                                               # или [Messages]
        if isinstance(content, str):           # сохранить только самый
            content = content.rstrip() + '\n'  # последний символ
                                               # конца строки
        print(content)
        print('-' * 79)          # получить только текст, см. email.parsers
    else:
        print('Bad message number')

def savemessage(i, mailfile, msgList):
    if 1 <= i <= len(msgList):
        savefile = open(mailfile, 'a', encoding=mailconfig.fetchEncoding) # 4E
        savefile.write('\n' + msgList[i-1] + '-'*80 + '\n')
    else:
        print('Bad message number')

def msgnum(command):
    try:
```

```
        return int(command.split()[1])
    except:
        return -1           # предполагается, что это ошибка

helptext = """
Available commands:
i    - index display
l n? - list all messages (or just message n)
d n? - mark all messages for deletion (or just message n)
s n? - save all messages to a file (or just message n)
m    - compose and send a new mail message
q    - quit pymail
?    - display this help text
"""

def interact(msgList, mailfile):
    showindex(msgList)
    toDelete = []
    while True:
        try:
            command = input('[Pymail] Action? (i, l, d, s, m, q, ?) ')
        except EOFError:
            command = 'q'
        if not command: command = '*'

        #завершение
        if command == 'q':
            break

        # оглавление
        elif command[0] == 'i':
            showindex(msgList)

        # содержимое письма
        elif command[0] == 'l':
            if len(command) == 1:
                for i in range(1, len(msgList)+1):
                    showmessage(i, msgList)
            else:
                showmessage(msgnum(command), msgList)

        # сохранение
        elif command[0] == 's':
            if len(command) == 1:
                for i in range(1, len(msgList)+1):
                    savemessage(i, mailfile, msgList)
            else:
                savemessage(msgnum(command), mailfile, msgList)

        # удаление
        elif command[0] == 'd':
```

```
            if len(command) == 1:                       # удалить все позднее
                toDelete = list(range(1, len(msgList)+1))  # в 3.x требуется
            else:                                          # вызвать list()
                delnum = msgnum(command)
                if (1 <= delnum <= len(msgList))
                                and (delnum not in toDelete):
                    toDelete.append(delnum)
                else:
                    print('Bad message number')

        # составление нового письма
        elif command[0] == 'm':          # отправить новое сообщение
                                         # через SMTP
            sendmessage()
            #execfile('smtpmail.py', {}) # альтернатива: запустить
                                         # в собственном пространстве имен
        elif command[0] == '?':
            print(helptext)
        else:
            print('What? -- type "?" for commands help')
    return toDelete

if __name__ == '__main__':
    import getpass, mailconfig
    mailserver = mailconfig.popservername   # например: 'pop.rmi.net'
    mailuser = mailconfig.popusername       # например: 'lutz'
    mailfile = mailconfig.savemailfile      # например: r'c:\stuff\savemail'
    mailpswd = getpass.getpass('Password for %s?' % mailserver)
    print('[Pymail email client]')
    msgList = loadmessages(mailserver, mailuser, mailpswd)  # загрузить все
    toDelete = interact(msgList, mailfile)
    if toDelete: deletemessages(mailserver, mailuser, mailpswd, toDelete)
    print('Bye.')
```

Нового здесь немного – просто сочетание логики интерфейса пользователя, уже знакомых нам инструментов и некоторых новых приемов:

Загрузка

Этот клиент загружает с сервера всю электронную почту в находящийся в оперативной памяти список Python только один раз, при начальном запуске. Чтобы получить вновь поступившую почту, необходимо завершить программу и запустить ее снова.

Сохранение

По требованию сценарий `pymail` сохраняет необработанный текст выбранного сообщения в локальном файле, имя которого указано в модуле `mailconfig` из примера 13.17.

Удаление

Теперь, наконец, поддерживается удаление почты с сервера по соответствующему запросу: сценарий `pymail` позволяет выбирать письма

для удаления по номерам, но все же физически они удаляются с сервера только при выходе и только при подтверждении операции. Благодаря удалению только при выходе из программы удается избежать изменения номеров почтовых сообщений во время сеанса – в POP удаление почты из середины списка ведет к уменьшению номеров всех сообщений, следующих за тем, которое удаляется. Так как pymail кэширует все сообщения в памяти, последующие операции с пронумерованными сообщениями в памяти могут быть неправильно применены, если осуществлять удаления незамедлительно.[1]

Анализ и составление сообщений

По командам вывода сообщений сценарий pymail выводит только содержательный текст сообщения, а не весь исходный текст, а при выводе оглавления почтового ящика отображаются только выбранные заголовки, выделенные из каждого сообщения. Для извлечения заголовков и содержимого писем используется пакет email, как показано в предыдущем разделе. Кроме того, сценарий использует пакет email также для составления сообщений, запрашивая строку, которая будет отправлена в виде письма.

Я думаю, что сейчас вы уже достаточно хорошо знаете язык Python, чтобы прочесть этот сценарий и разобраться, как он работает, поэтому вместо лишних слов о его устройстве перейдем к интерактивному сеансу pymail и посмотрим его в действии.

Работа с клиентом командной строки pymail

Запустим сценарий pymail, чтобы с его помощью прочесть и удалить письма на нашем почтовом сервере и отправить новые сообщения. Сценарий pymail может выполняться на любом компьютере с Python и сокетами, загружать почту с любого почтового сервера с интерфейсом POP, на котором у вас имеется учетная запись, и отправлять почту через сервер SMTP, указанный в модуле mailconfig, который мы написали ранее (пример 13.17).

Ниже приводится сеанс работы со сценарием на моем ноутбуке в Windows. На других компьютерах он работает идентичным образом, благодаря переносимости Python и его стандартной библиотеки.

---

[1] Подробнее о номерах сообщений в POP будет сказано при рассмотрении модуля mailtools, далее в этой главе. Интересно, что список номеров удаляемых сообщений не должен сортироваться – номера остаются действительными на протяжении всего сеанса соединения с сервером POP. Мы также познакомимся с некоторыми тонкими проблемами, которые могут возникать, если письма будут удалены из почтового ящика на сервере без ведома pymail (например, вашим интернет-провайдером или с помощью другого клиента электронной почты). А кроме того, хотя это случается крайне редко, иногда все же операция удаления, реализованная в сценарии, выполняется неточно.

Во-первых, мы запускаем сценарий, вводим пароль POP (напомню, что для серверов SMTP пароль обычно не требуется) и ждем, когда pymail выведет индекс списка сообщений почтового ящика – данная версия сценария загружает полный текст всех сообщений при запуске:

```
C:\...\PP4E\Internet\Email> pymail.py
Password for pop.secureserver.net?
[Pymail email client]
Connecting...
b'+OK <8927.1273263898@p3pop01-10.prod.phx3.gdg>'
(b'+OK ', [b'1 1860', b'2 1408', b'3 1049', b'4 1009', b'5 1038',
b'6 957'], 47)
There are 6 mail messages in 7321 bytes
Retrieving...
1:      1861 bytes
        From    =>lutz@rmi.net
        To      =>pp4e@learning-python.com
        Date    =>Wed, 5 May 2010 11:29:36 -0400 (EDT)
        Subject =>I'm a Lumberjack, and I'm Okay
2:      1409 bytes
        From    =>lutz@learning-python.com
        To      =>PP4E@learning-python.com
        Date    =>Wed, 05 May 2010 08:33:47 -0700
        Subject =>testing
3:      1050 bytes
        From    =>Eric.the.Half.a.Bee@yahoo.com
        To      =>PP4E@learning-python.com
        Date    =>Thu, 06 May 2010 14:11:07 -0000
        Subject =>A B C D E F G
4:      1010 bytes
        From    =>PP4E@learning-python.com
        To      =>PP4E@learning-python.com
        Date    =>Thu, 06 May 2010 14:16:31 -0000
        Subject =>testing smtpmail
5:      1039 bytes
        From    =>Eric.the.Half.a.Bee@aol.com
        To      =>nobody.in.particular@marketing.com
        Date    =>Thu, 06 May 2010 14:32:32 -0000
        Subject =>a b c d e f g
[Press Enter key]
6:      958 bytes
        From    =>PP4E@learning-python.com
        To      =>maillist
        Date    =>Thu, 06 May 2010 10:58:40 -0400
        Subject =>test interactive smtplib
[Pymail] Action? (i, l, d, s, m, q, ?) l 6
------------------------------------------------------------------------------
testing 1 2 3...

------------------------------------------------------------------------------
[Pymail] Action? (i, l, d, s, m, q, ?) l 3
```

```
------------------------------------------------------------------------
Fiddle de dum, Fiddle de dee,
Eric the half a bee.

------------------------------------------------------------------------
[Pymail] Action? (i, l, d, s, m, q, ?)
```

Как только pymail загрузит электронную почту в список Python на локальном компьютере, можно вводить буквы команд для ее обработки. Команда l выводит содержимое сообщения с указанным номером. В примере выше с ее помощью были выведены два сообщения, которые мы отправили в предыдущем разделе с помощью сценария smtpmail.

Сценарий pymail позволяет также получать подсказку по командам, удалять сообщения (фактическое удаление происходит на сервере при выходе из программы) и сохранять сообщения в локальном текстовом файле, имя которого указано в модуле mailconfig, который мы видели выше:

```
[Pymail] Action? (i, l, d, s, m, q, ?) ?

Available commands:
i    - index display
l n? - list all messages (or just message n)
d n? - mark all messages for deletion (or just message n)
s n? - save all messages to a file (or just message n)
m    - compose and send a new mail message
q    - quit pymail
?    - display this help text

[Pymail] Action? (i, l, d, s, m, q, ?) s 4
[Pymail] Action? (i, l, d, s, m, q, ?) d 4
```

Теперь выберем команду m, чтобы составить новое почтовое сообщение – pymail предложит ввести части письма, сконструирует полный текст сообщения с помощью модуля email и отправит его с помощью smtplib. Допускается указывать несколько адресов получателей, разделяя их запятыми и использовать краткую форму записи адреса «адрес» или полную «имя <адрес>». Так как отправка писем осуществляется по протоколу SMTP, в заголовке «From» можно использовать произвольные адреса; но обычно не следует этого делать (если, конечно, вы не пытаетесь получить интересные примеры для книги).

```
[Pymail] Action? (i, l, d, s, m, q, ?) m
From? Cardinal@hotmail.com
To?   PP4E@learning-python.com
Subj? Among our weapons are these
Type message text, end with line="."
Nobody Expects the Spanish Inquisition!
.
[Pymail] Action? (i, l, d, s, m, q, ?) q
To be deleted: [4]
```

```
Delete?y
Connecting...
b'+OK <16872.1273264370@p3pop01-17.prod.phx3.secureserver.net>'
Deleting messages from server...
Bye.
```

Как уже говорилось, удаление в действительности происходит только при выходе. При завершении pymail командой q сценарий сообщает, какие сообщения помещены в очередь на удаление, и просит подтвердить запрос. При подтверждении pymail снова соединяется с почтовым сервером и производит вызовы POP для удаления отобранных почтовых сообщений. Из-за того, что операция удаления изменяет порядковые номера входящих сообщений в почтовом ящике на сервере, откладывание фактического удаления до момента выхода из сценария упрощает обработку уже загруженных сообщений (мы улучшим реализацию этой операции в клиенте PyMailGUI в следующей главе).

Так как сценарий pymail загружает почту с сервера в локальный список Python только один раз при начальном запуске, необходимо заново запустить его, чтобы снова получить почту с сервера, если нужно посмотреть результат отправки почты и произведенных удалений. В следующем примере новое письмо показывается под номером 6, а первоначальное письмо, имевшее номер 4, отсутствует:

```
C:\...\PP4E\Internet\Email> pymail.py
Password for pop.secureserver.net?
[Pymail email client]
Connecting...
b'+OK <11563.1273264637@p3pop01-26.prod.phx3.secureserver.net>'
(b'+OK ', [b'1 1860', b'2 1408', b'3 1049', b'4 1038', b'5 957',
           b'6 1037'], 47)
There are 6 mail messages in 7349 bytes
Retrieving...
1:      1861 bytes
        From    =>lutz@rmi.net
        To      =>pp4e@learning-python.com
        Date    =>Wed, 5 May 2010 11:29:36 -0400 (EDT)
        Subject =>I'm a Lumberjack, and I'm Okay
2:      1409 bytes
        From    =>lutz@learning-python.com
        To      =>PP4E@learning-python.com
        Date    =>Wed, 05 May 2010 08:33:47 -0700
        Subject =>testing
3:      1050 bytes
        From    =>Eric.the.Half.a.Bee@yahoo.com
        To      =>PP4E@learning-python.com
        Date    =>Thu, 06 May 2010 14:11:07 -0000
        Subject =>A B C D E F G
4:      1039 bytes
        From    =>Eric.the.Half.a.Bee@aol.com
        To      =>nobody.in.particular@marketing.com
```

```
        Date    =>Thu, 06 May 2010 14:32:32 -0000
        Subject =>a b c d e f g
5:      958 bytes
        From    =>PP4E@learning-python.com
        To      =>maillist
        Date    =>Thu, 06 May 2010 10:58:40 -0400
        Subject =>test interactive smtplib
[Press Enter key]
6:      1038 bytes
        From    =>Cardinal@hotmail.com
        To      =>PP4E@learning-python.com
        Date    =>Fri, 07 May 2010 20:32:38 -0000
        Subject =>Among our weapons are these
[Pymail] Action? (i, l, d, s, m, q, ?) l 6
------------------------------------------------------------------------
Nobody Expects the Spanish Inquisition!

------------------------------------------------------------------------
[Pymail] Action? (i, l, d, s, m, q, ?) q
Bye.
```

Хотя это и не показано в данном примере сеанса, тем не менее, существует возможность отправлять письма сразу нескольким адресатам и включать в заголовок с адресом пары имя/адрес. Это возможно благодаря тому, что для разбиения списков адресов и их анализа сценарий использует утилиты из пакета `email`, описанные выше, которые корректно распознают запятые и как разделители, и как символы внутри имени. Следующие последовательности операций, например, использующие главным образом полные адреса, отправят сообщения двум и трем адресатам соответственно:

```
[Pymail] Action? (i, l, d, s, m, q, ?) m
From? "moi 1" <pp4e@learning-python.com>
To?   "pp 4e" <pp4e@learning-python.com>, "lu,tz" <lutz@learning-python.com>

[Pymail] Action? (i, l, d, s, m, q, ?) m
From? The Book <pp4e@learning-python.com>
To?   "pp 4e" <pp4e@learning-python.com>, "lu,tz" <lutz@learning-python.com>,
lutz@rmi.net
```

Наконец, если вы запускали этот сценарий, вы также обнаружите, что он сохранил на вашем компьютере один файл, содержащий письмо, которое мы попросили сохранить в предыдущем сеансе; файл просто содержит необработанный текст сохраненных сообщений со строками-разделителями. Этот файл могут читать как человек, так и машина – в принципе, можно загрузить сохраненную в этом файле почту в список Python в другом сценарии, применив метод `string.split` к тексту файла и указав строку-разделитель. Как показано в этой книге, файл сохраняется с именем *C:\temp\savemail.txt*, но вы можете выбрать любое другое имя, указав его в модуле `mailconfig`.

Вспомогательный пакет mailtools

Пакет email, используемый сценарием pymail из предыдущего раздела, является обширной коллекцией инструментов – фактически даже слишком обширной, чтобы запомнить все, что в ней имеется. Для начала, наличие некоторого шаблонного программного кода, реализующего типичные случаи использования пакета, может помочь освободить вас от вникания во все его особенности, – изолируя операции использования модуля, такой программный код может также упростить переход к изменениям в пакете email, возможным в будущем. Чтобы упростить взаимодействие с пакетом для нужд более сложных клиентов электронной почты и для дальнейшей демонстрации использования инструментов электронной почты, имеющихся в стандартной библиотеке, я создал собственные вспомогательные модули, которые будут представлены в этом разделе, – пакет mailtools.

mailtools – это пакет модулей Python: каталог, содержащий модули, по одному на класс инструментов, и модуль инициализации, который автоматически выполняется при первой попытке импортировать пакет. Данный пакет модулей по сути является всего лишь оберткой вокруг пакета email, а также вокруг модулей poplib и smtplib из стандартной библиотеки. Они строят некоторые достаточно обоснованные предположения о том, как должен использоваться пакет email, и позволяют нам забыть о некоторых сложностях использования инструментов стандартной библиотеки.

Проще говоря, пакет mailtools предоставляет три класса – для получения, отправки и анализа почтовых сообщений. Эти классы могут использоваться как *суперклассы* с целью подмешивания их методов в более специализированные прикладные классы или как *самостоятельные* или *встроенные* объекты, экспортирующие свои методы для непосредственного использования. Далее мы увидим, как используются эти классы в обоих качествах.

Примеры использования инструментов этого пакета приводятся в модуле *selftest.py*, который играет роль сценария самотестирования пакета. При запуске он отправляет письмо с вашего адреса вам же, в которое вложен сам файл *selftest.py*. Кроме того, он извлекает и выводит некоторые заголовки сообщений, а также проанализированное и необработанное содержимое. Эти интерфейсы, с добавлением некоторой магии для создания пользовательского интерфейса, приведут нас к полнофункциональным почтовым клиентам и веб-сайтам далее в этой главе.

Предварительно следует отметить две конструктивные особенности. Во-первых, реализация этого пакета ничего не знает о пользовательском интерфейсе, с которым она будет использоваться (консоль, графический интерфейс, веб-интерфейс или какой-то другой), и не делает никаких предположений об использовании в многопоточной среде выполнения – это просто комплект инструментов. Как мы увидим далее, все

решения по использованию пакета должны приниматься его клиентами. Сосредоточившись исключительно на обработке электронной почты, мы также способствуем упрощению программного кода программ, использующих его.

Во-вторых, каждый из основных модулей в пакете иллюстрирует приемы решения проблем поддержки Юникода, с которыми приходится сталкиваться программному коду для Python 3.X, в частности при использовании пакета `email` в Python 3.1:

- *Отправитель* должен заботиться о кодировках текста основного сообщения, о вложении входных файлов, о сохранении почты в выходных файлах и о заголовках сообщения.
- *Получатель* должен определять кодировку полного текста сообщения при извлечении его с сервера.
- *Механизм анализа* должен заботиться о кодировках текстового содержимого анализируемых сообщений, а также о кодировках заголовков.

Кроме того, отправитель должен предоставить реализацию обходных решений проблем создания двоичных и текстовых частей в почтовых сообщениях, описанных выше в этой главе. Поскольку в целом это очень важные проблемы использования Юникода и они не имеют такого универсального решения, как хотелось бы, из-за ограничений текущей версии пакета `email`, я буду тщательно разъяснять каждый сделанный выбор.

В следующих нескольких разделах приводится исходный программный код модулей из пакета `mailtools`. Все вместе его файлы содержат примерно 1050 строк кода, включая пустые строки и комментарии. Мы не будем рассматривать все детали реализации пакета, поэтому ищите дополнительные подробности в листингах и загляните в модуль самотестирования, где приводятся примеры использования пакета. Кроме того, дополнительную информацию и примеры вы найдете в трех клиентах, использующих этот пакет, – в измененной версии *pymail2.py*, которая следует сразу же за листингами, в клиенте PyMailGUI из главы 14 и в реализации сервера PyMailCGI из главы 16. Совместно используя этот пакет, все три системы получают в наследство все его инструменты, а также любые расширения в будущем.

Файл инициализации

Модуль в примере 13.21 реализует логику инициализации пакета `mailtools`. Как обычно, этот модуль выполняется автоматически при первой попытке импортировать пакет. Обратите внимание, что при использовании инструкций `from *` этот файл помещает содержимое всех вложенных модулей в пространство имен каталога – поскольку первые версии пакета `mailtools` состояли из единственного файла *.py*, такой подход обеспечивает обратную совместимость с уже существующими клиента-

ми. Мы также должны использовать здесь синтаксис импортирования относительно пакета (`from .module`), потому что Python 3.X больше не включает собственный каталог пакета в путь поиска модулей (только родительский каталог пакета находится в пути поиска). Так как это корневой модуль, в него помещены также глобальные комментарии.

Пример 13.21. PP4E\Internet\Email\mailtools__init__.py

```
""""
##############################################################################
пакет mailtools: интерфейс к почтовому серверу, используется клиентами
pymail2, PyMailGUI и PyMailCGI; реализует загрузку, отправку, анализ,
составление, удаление, возможность добавления вложений, кодирование
(оба вида - MIME и Юникода) и так далее; классы, реализующие анализ,
получение и отправку, могут подмешиваться в подклассы,
использующие их методы, или использоваться как встраиваемые
или самостоятельные объекты;

этот пакет также включает удобные подклассы для работы в немом режиме
и многое другое; загружает все почтовые сообщения, если сервер POP
не устанавливает верхнюю границу; не содержит специальной поддержки
многопоточной модели выполнения или графического интерфейса
и позволяет подклассам предоставлять свою реализацию запроса пароля;
функция обратного вызова progress получает признак состояния; в случае
ошибки все методы возбуждают исключения - они должны обрабатываться
клиентом с графическим/другим интерфейсом; этот набор инструментов был
преобразован из простого модуля в пакет: вложенные модули импортируются
здесь для обратной совместимости;

4E: необходимо использовать синтаксис импортирования относительно
пакета, потому что в Py 3.X каталог пакета больше не включается
в путь поиска модулей при импортировании пакета, если пакет
импортируется из произвольного каталога (из другого каталога,
который использует этот пакет); кроме того, выполняется декодирование
Юникода в тексте письма при его получении (смотрите mailFetcher),
а также в некоторых частях с текстовым содержимым, которые, возможно,
были закодированы в формат MIME (смотрите mailParser);

TBD: в saveparts, возможно, следовало бы открывать файл в текстовом режиме,
когда основной тип определяется, как text?
TBD: в walkNamedParts, возможно, следовало бы перешагивать через нетипичные
типы, такие как message/delivery-status?
TBD: поддержка Юникода не подвергалась всеобъемлющему тестированию:
обращайтесь к главе 13 за дополнительными сведениями о пакете email в Py3.1,
его ограничениях и о приемах, используемых здесь;
##############################################################################
"""

# собрать содержимое всех модулей здесь, если импортируется каталог пакета
from .mailFetcher import *
```

```
from .mailSender import *      # 4E: импортирование относительно пакета
from .mailParser import *

# экспортировать вложенные модули для инструкций from mailtools import *
__all__ = 'mailFetcher', 'mailSender', 'mailParser'

# программный код самотестирования находится в файле selftest.py, чтобы
# позволить установить путь к модулю mailconfig перед тем, как будут
# выполнены инструкции импортирования вложенных модулей выше
```

Класс MailTool

В примере 13.22 представлена реализация суперкласса, общего для всех остальных классов в пакете. Отчасти такая организация предусмотрена с целью поддержки возможности расширения. В настоящее время эта возможность используется, только чтобы включать и выключать вывод трассировочных сообщений (в некоторых клиентах, таких как веб-приложения, может быть нежелательным, чтобы в стандартный поток вывода выводился бы посторонний текст). Подклассы, подмешивающие «немую» версию, лишаются возможности вывода.

Пример 13.22. PP4E\Internet\Email\mailtools\mailTool.py

```
"""
##############################################################################
общие суперклассы: используются для включения и отключения вывода
трассировочных сообщений
##############################################################################
"""


class MailTool:                # суперкласс всех инструментов
                               # электронной почты
    def trace(self, message):  # переопределите, чтобы запретить
        print(message)         # или выводить в файл журнала

class SilentMailTool:          # для подмешивания, а не наследования
    def trace(self, message):
        pass
```

Класс MailSender

В примере 13.23 представлен класс, используемый для составления и отправки сообщений. Этот модуль предоставляет удобный интерфейс, объединяющий в себе инструменты из стандартной библиотеки, с которыми мы уже встречались в этой главе, – пакет `email` для составления сообщений с вложениями и их кодирования и модуль `smtplib` – для отправки получившегося текста сообщений. Вложения передаются в виде списка имен файлов – типы MIME и любые необходимые кодировки определяются автоматически, с помощью модуля `mimetypes`. Кроме того,

с помощью функций из модуля email.utils автоматизировано создание строк с датой и временем, а заголовки, содержащие символы не из диапазона ASCII, кодируются в соответствии со стандартами электронной почты, MIME и Юникода. Более подробные сведения об особенностях работы класса вы найдете в программном коде и в комментариях внутри этого файла.

Проблемы поддержки Юникода при работе с вложениями заголовками и при сохранении файлов

Этот класс также открывает и вкладывает файлы, генерирует полный текст сообщений и сохраняет отправленные сообщения в локальном файле. Большинство файлов вложений открывается в двоичном режиме, но, как мы уже видели, некоторые текстовые вложения необходимо открывать в текстовом режиме, потому что текущая версия пакета email требует передавать их конструкторам объектов сообщений в виде строк str. Кроме того, выше мы также видели, что пакет email требует, чтобы вложения были представлены в виде строк str, когда позднее будет генерироваться полный текст сообщения, возможно, как результат преобразования в формат MIME.

Чтобы удовлетворить эти требования пакета email в Python 3.1, необходимо применить два обходных решения, описанных выше, – исходя из того, как пакет email обрабатывает данные, передавать функции open флаг текстового или двоичного режима (и тем самым обеспечить чтение данных в виде строки str или bytes) и выполнять преобразование двоичных данных в формат MIME, чтобы потом декодировать результат в текст ASCII. Последняя операция также разбивает на строки текст с двоичными частями в формате Base64 (в отличие от email), потому что в противном случае будет отправлена одна длинная строка, что может быть допустимо в некоторых контекстах, но может вызывать проблемы в некоторых текстовых редакторах при просмотре необработанного текста.

Вдобавок к этим обходным решениям клиенты могут передавать имена кодировок Юникода для основной текстовой части и для каждого текстового вложения в отдельности. В приложении PyMailGUI, которое будет представлено в главе 14, выбор кодировки осуществляется с помощью модуля с пользовательскими настройками mailconfig, и всякий раз, когда с помощью пользовательских настроек оказывается невозможным закодировать текстовую часть, применяется кодировка UTF-8. В принципе, можно было бы также перехватывать ошибки декодирования файла и возвращать строку с сообщением об ошибке (как это делается в классе получения почты, обсуждаемом далее), но отправка недопустимого вложения может иметь более печальные последствия, чем его отображение. Поэтому в случае появления ошибок вся операция отправки терпит неудачу.

Наконец, реализованы также новая поддержка кодирования заголовков с символами не из диапазона ASCII (полных заголовков и компо-

нентов имен в заголовках с адресами) с применением кодировки, выбираемой клиентом, или UTF-8 по умолчанию, и сохранение отправленного сообщения в файл, открытый с указанием той же кодировки, определяемой с помощью модуля `mailconfig`, которая использовалась для декодирования поступающих сообщений.

Последнее правило, применяемое при сохранении отправленных сообщений, используется потому, что позднее файл с отправленными сообщениями может быть открыт клиентами, применяющими ту же схему кодирования, для извлечения полного текста почтового сообщения в этой же кодировке. Это отражает способ, каким клиенты, такие как PyMailGUI, сохраняют полный текст сообщения в локальном файле, чтобы позднее его можно было открыть и вывести содержимое. Этот прием может приводить к неудаче, если механизм извлечения почты попытается применить другую, несовместимую кодировку. Он предполагает, что в файле не будет сохранено ни одно сообщение в несовместимой кодировке, даже при многократных сохранениях. Мы могли бы попробовать создать отдельный файл для каждой кодировки, при этом предполагая, что одна кодировка относится ко всему полному тексту сообщения. Стандартом предусмотрено, что полный текст сообщения должен быть в кодировке ASCII, поэтому скорее всего это будет 7- или 8-битовый текст.

Пример 13.23. PP4E\Internet\Email\mailtools\mailSender.py

```
"""
##############################################################################
отправляет сообщения, добавляет вложения (описание и тест приводятся
в модуле __init__)
##############################################################################
"""

import mailconfig                              # клиентские настройки
import smtplib, os, mimetypes                  # mime: имя в тип
import email.utils, email.encoders             # строка с датой, base64
from .mailTool import MailTool, SilentMailTool # 4E: относительно пакета

from email.message          import Message     # объект сообщения, obj->text
from email.mime.multipart   import MIMEMultipart  # специализированные
                                                  # объекты
from email.mime.audio       import MIMEAudio   # вложений с поддержкой
from email.mime.image       import MIMEImage   # форматирования/кодирования
from email.mime.text        import MIMEText
from email.mime.base        import MIMEBase
from email.mime.application import MIMEApplication  # 4E: использовать новый
                                                    # класс приложения
def fix_encode_base64(msgobj):
    """
    4E: реализация обходного решения для ошибки в пакете email в Python 3.1,
    препятствующей созданию полного текста сообщения с двоичными частями,
```

```
    преобразованными в формат base64 или другой формат электронной почты;
    функция email.encoder, вызываемая конструктором, оставляет содержимое
    в виде строки bytes, даже при том, что оно находится в текстовом формате
    base64; это препятствует работе механизма создания полного текста
    сообщения, который предполагает получить текстовые данные и поэтому
    требует, чтобы они имели тип str; в результате этого с помощью пакета
    email в Py 3.1 можно создавать только простейшие текстовые части
    сообщений - любая двоичная часть в формате MIME будет вызывать
    ошибку на этапе создания полного текста сообщения; есть сведения,
    что эта ошибка будет устранена в будущих версиях Python и пакета email,
    в этом случае данная функция не должна выполнять
    никаких действий; подробности смотрите в главе 13;
    """

    linelen = 76 # согласно стандартам MIME
    from email.encoders import encode_base64

    encode_base64(msgobj)         # что обычно делает email: оставляет bytes
    text = msgobj.get_payload()   # bytes выз. ош. в email
                                  # при создании текста
    if isinstance(text, bytes):     # содержимое - bytes в 3.1, str в 3.2
        text = text.decode('ascii') # декодировать в str,
                                    # чтобы сгенерировать текст

    lines = []                    # разбить на строки, иначе 1 большая строка
    text = text.replace('\n', '')   # в 3.1 нет \n, но что будет потом!
    while text:
        line, text = text[:linelen], text[linelen:]
        lines.append(line)
    msgobj.set_payload('\n'.join(lines))

def fix_text_required(encodingname):
    """
    4E: обходное решение для ошибки, вызываемой смешиванием str/bytes
    в пакете email; в Python 3.1 класс MIMEText требует передавать
    ему строки разных типов для текста в разных кодировках,
    что обусловлено преобразованием некоторых типов текста
    в разные форматы MIME; смотрите главу 13;
    единственная альтернатива - использовать обобщенный класс Message
    и повторить большую часть программного кода;
    """

    from email.charset import Charset, BASE64, QP

    charset = Charset(encodingname) # так email опр., что делать
                                    # для кодировки
    bodyenc = charset.body_encoding # utf8 и др. требует данные типа bytes
    return bodyenc in (None, QP)    # ascii, latin1 и др. требует
                                    # данные типа str
```

```
class MailSender(MailTool):
    """
    отправляет сообщение: формирует сообщение, соединяется с SMTP-сервером;
    работает на любых компьютерах с Python+Интернет, не использует клиента
    командной строки; не выполняет аутентификацию: смотрите MailSenderAuth,
    если требуется аутентификация;
    4E: tracesize - количество символов в трассировочном сообщении: 0=нет,
        большое значение=все;
    4E: поддерживает кодирование Юникода для основного текста и текстовых
        частей;
    4E: поддерживает кодирование заголовков – и полных, и компонента имени
        в адресах;
    """
    def __init__(self, smtpserver=None, tracesize=256):
        self.smtpServerName = smtpserver or mailconfig.smtpservername
        self.tracesize = tracesize

    def sendMessage(self, From, To, Subj, extrahdrs, bodytext, attaches,
                                    saveMailSeparator=(('=' * 80) + 'PY\n'),
                                    bodytextEncoding='us-ascii',
                                    attachesEncodings=None):
        """"
        формирует и отправляет сообщение: блокирует вызывающую программу,
        в графических интерфейсах следует вызывать в отдельном потоке
        выполнения;
        bodytext - основной текст, attaches - список имен файлов,
        extrahdrs - список кортежей (имя, значение) добавляемых заголовков;
        возбуждает исключение, если отправка не удалась по каким-либо
        причинам; в случае успеха сохраняет отправленное сообщение
        в локальный файл; предполагается, что значения
        для заголовков To, Cc, Bcc являются списками
        из 1 или более уже декодированных адресов (возможно, в полном
        формате имя+<адрес>); клиент должен сам выполнять анализ,
        чтобы разбить их по разделителям или использовать
        многострочный ввод;
        обратите внимание, что SMTP допускает использование полного формата
        имя+<адрес> в адресе получателя;
        4E: адреса Bcc теперь используются для отправки, а заголовок
            отбрасывается;
        4E: повторяющиеся адреса получателей отбрасываются, иначе они будут
            получать несколько копий письма;
        предупреждение: не поддерживаются сообщения multipart/alternative,
                        только /mixed;
        """"

        # 4E: предполагается, что основной текст уже в требуемой кодировке;
        # клиенты могут декодировать, используя кодировку по выбору
        # пользователя, по умолчанию или utf8;
        # так или иначе, email требует передать либо str, либо bytes;
```

```
if fix_text_required(bodytextEncoding):
    if not isinstance(bodytext, str):
        bodytext = bodytext.decode(bodytextEncoding)
else:
    if not isinstance(bodytext, bytes):
        bodytext = bodytext.encode(bodytextEncoding)

# создать корень сообщения
if not attaches:
    msg = Message()
    msg.set_payload(bodytext, charset=bodytextEncoding)
else:
    msg = MIMEMultipart()
    self.addAttachments(msg, bodytext, attaches,
                            bodytextEncoding, attachesEncodings)

# 4E: не-ASCII заголовки кодируются; кодировать только имена
# в адресах, иначе smtp может отвергнуть сообщение;
# кодирует все имена в аргументе To (но не адреса),
# предполагается, что это допустимо для сервера;
# msg.as_string сохраняет все разрывы строк,
# добавленные при кодировании заголовков;

hdrenc = mailconfig.headersEncodeTo or 'utf-8' # по умолчанию=utf8
Subj   = self.encodeHeader(Subj, hdrenc)       # полный заголовок
From   = self.encodeAddrHeader(From, hdrenc)   # имена в адресах
To     = [self.encodeAddrHeader(T, hdrenc) for T in To] # каждый
                                                        # адрес
Tos    = ', '.join(To)                         # заголовок+аргумент

# добавить заголовки в корень сообщения
msg['From']    = From
msg['To']      = Tos     # возможно несколько: список адресов
msg['Subject'] = Subj    # серверы отвергают разделитель ';'
msg['Date']    = email.utils.formatdate() # дата+время, rfc2822 utc
recip          = To
for name, value in extrahdrs:  # Cc, Bcc, X-Mailer и др.
    if value:
        if name.lower() not in ['cc', 'bcc']:
            value = self.encodeHeader(value, hdrenc)
            msg[name] = value
        else:
            value = [self.encodeAddrHeader(V, hdrenc) for V in value]
            recip += value     # некоторые серверы отвергают [''']
            if name.lower() != 'bcc': # 4E: bcc получает почту,
                                      # без заголовка
                msg[name] = ', '.join(value) # доб. зап. между cc

recip = list(set(recip))    # 4E: удалить дубликаты
fullText = msg.as_string()  # сформировать сообщение
```

```
        # вызов sendmail возбудит исключение, если все адреса Tos ошибочны,
        # или вернет словарь с ошибочными адресами Tos

        self.trace('Sending to...' + str(recip))
        self.trace(fullText[:self.tracesize])      # вызов SMTP
                                                   # для соединения
        server = smtplib.SMTP(self.smtpServerName,
                              timeout=15)          # также может дать ошибку
        self.getPassword()                         # если сервер требует
        self.authenticateServer(server)            # регистрация в подклассе
        try:
            failed = server.sendmail(From, recip, fullText) # искл.
                                                            # или словарь
        except:
            server.close()                # 4E: заверш. может подвесить!
            raise                         # повторно возбудить исключение
        else:
            server.quit()                 # соединение + отправка, успех
        self.saveSentMessage(fullText, saveMailSeparator) # 4E: в первую
                                                          # очередь
        if failed:
            class SomeAddrsFailed(Exception): pass
            raise SomeAddrsFailed('Failed addrs:%s\n' % failed)
        self.trace('Send exit')

    def addAttachments(self, mainmsg, bodytext, attaches,
                                      bodytextEncoding, attachesEncodings):
        """
        формирует сообщение, состоящее из нескольких частей, добавляя
        вложения attachments; использует для текста указанную кодировку
        Юникода, если была передана;
        """
        # добавить главную часть text/plain
        msg = MIMEText(bodytext, _charset=bodytextEncoding)
        mainmsg.attach(msg)

        # добавить части с вложениями
        encodings = attachesEncodings or (['us-ascii'] * len(attaches))
        for (filename, fileencode) in zip(attaches, encodings):
            # имя файла может содержать абсолюный или относительный путь
            if not os.path.isfile(filename):   # пропустить каталоги и пр.
                continue

            # определить тип содержимого по расширению имени файла,
            # игнорировать кодировку
            contype, encoding = mimetypes.guess_type(filename)
            if contype is None or encoding is not None: # не определено, сжат?
                contype = 'application/octet-stream'    # универсальный тип
            self.trace('Adding ' + contype)
```

```
            # сконструировать вложенный объект Message соответствующего типа
            maintype, subtype = contype.split('/', 1)
            if maintype == 'text':       # 4E: текст требует кодирования
                if fix_text_required(fileencode): # требуется str или bytes
                    data = open(filename, 'r', encoding=fileencode)
                else:
                    data = open(filename, 'rb')
                msg = MIMEText(data.read(), _subtype=subtype,
                               _charset=fileencode)
                data.close()

            elif maintype == 'image':
                data = open(filename, 'rb') # 4E: обходной прием для двоичных
                msg  = MIMEImage(data.read(), _subtype=subtype,
                                 _encoder=fix_encode_base64)
                data.close()

            elif maintype == 'audio':
                data = open(filename, 'rb')
                msg  = MIMEAudio(data.read(), _subtype=subtype,
                                 _encoder=fix_encode_base64)
                data.close()

            elif maintype == 'application':    # новый тип в 4E
                data = open(filename, 'rb')
                msg  = MIMEApplication(data.read(), _subtype=subtype,
                                       _encoder=fix_encode_base64)
                data.close()

            else:
                data = open(filename, 'rb')       # тип application/* мог бы
                msg  = MIMEBase(maintype, subtype) # обрабатываться здесь
                msg.set_payload(data.read())
                data.close()                       # создание универс. типа
                fix_encode_base64(msg)             # также было нарушено!
               #email.encoders.encode_base64(msg)  # преобразовать в base64

            # установить имя файла и присоединить к контейнеру
            basename = os.path.basename(filename)
            msg.add_header('Content-Disposition',
                           'attachment', filename=basename)
            mainmsg.attach(msg)

        # текст за пределами структуры mime, виден клиентам,
        # которые не могут декодировать формат MIME
        mainmsg.preamble = 'A multi-part MIME format message.\n'
        mainmsg.epilogue = ''     # гарантировать завершение сообщения
                                  # переводом строки

    def saveSentMessage(self, fullText, saveMailSeparator):
        """
```

```
    добавляет отправленное сообщение в конец локального файла,
    если письмо было отправлено хотя бы одному адресату;
    клиент: определяет строку-разделитель, используемую приложением;
    предупреждение: пользователь может изменить файл во время работы
                    сценария (маловероятно);
    """
    try:
        sentfile = open(mailconfig.sentmailfile, 'a',
                        encoding=mailconfig.fetchEncoding)     # 4E
        if fullText[-1] != '\n': fullText += '\n'
        sentfile.write(saveMailSeparator)
        sentfile.write(fullText)
        sentfile.close()
    except:
        self.trace('Could not save sent message') # не прекращает работу
                                                  # сценария
def encodeHeader(self, headertext, unicodeencoding='utf-8'):
    """
    4E: кодирует содержимое заголовков с символами не из диапазона ASCII
    в соответствии со стандартами электронной почты и Юникода, применяя
    кодировку пользователя или UTF-8; метод header.encode автоматически
    добавляет разрывы строк, если необходимо;
    """
    try:
        headertext.encode('ascii')
    except:
        try:
            hdrobj = email.header.make_header([(headertext,
                                                unicodeencoding)])
            headertext = hdrobj.encode()
        except:
            pass      # автоматически разбивает на несколько строк
    return headertext # smtplib может потерпеть неудачу, если не будет
                      # закодировано в ascii

def encodeAddrHeader(self, headertext, unicodeencoding='utf-8'):
    """
    4E: пытается закодировать имена в адресах электронной почты
    с символами не из диапазона ASCII в соответствии со стандартами
    электронной почты, MIME и Юникода; если терпит неудачу, компонент
    имени отбрасывается и используется только часть
    с фактическим адресом;
    если не может получить даже адрес, пытается декодировать целиком,
    иначе smtplib может столкнуться с ошибками, когда попытается
    закодировать все почтовое сообщение как ASCII; в большинстве случаев
    кодировки utf-8 вполне достаточно, так как она предусматривает
    довольно широкое разнообразие кодовых пунктов;

    вставляет символы перевода строки, если строка заголовка слишком
    длинная, иначе метод hdr.encode разобьет имена на несколько строк,
    но он может не замечать некоторые строки, длиннее максимального
```

```
            значения (улучшите меня); в данном случае метод Message.as_string
            форматирования не будет пытаться разбивать строки;
            смотрите также метод decodeAddrHeader в модуле mailParser,
            реализующий обратную операцию;
            """
            try:
                pairs = email.utils.getaddresses([headertext]) # разбить
                                                               # на части
                encoded = []
                for name, addr in pairs:
                    try:
                        name.encode('ascii') # использовать, как есть,
                                             # если ascii
                    except UnicodeError:     # иначе закодировать
                                             # компонент имени
                        try:
                            uni  = name.encode(unicodeencoding)
                            hdr  = email.header.make_header([(uni,
                                                              unicodeencoding)])
                            name = hdr.encode()
                        except:
                            name = None # отбросить имя, использовать только адрес
                    joined = email.utils.formataddr((name, addr)) # заключить
                    encoded.append(joined)                    # имя в кавычки,
                                                              # если необходимо
                fullhdr = ', '.join(encoded)
                if len(fullhdr) > 72 or '\n' in fullhdr:      # не одна короткая
                                                              # строка?
                    fullhdr = ',\n '.join(encoded)       # попробовать несколько
                                                         # строк
                return fullhdr
            except:
                return self.encodeHeader(headertext)

        def authenticateServer(self, server):
            pass     # этот класс/сервер не предусматривает аутентификацию

        def getPassword(self):
            pass     # этот класс/сервер не предусматривает аутентификацию

##############################################################################
# специализированные подклассы
##############################################################################

class MailSenderAuth(MailSender):
    """
    используется для работы с серверами, требующими аутентификацию;
    клиент: выбирает суперкласс MailSender или MailSenderAuth, опираясь
    на параметр mailconfig.smtpuser (None?)
    """
```

```
    smtpPassword = None  # 4E: в классе, не в self, совместно используется
                         #     всеми экземплярами
    def __init__(self, smtpserver=None, smtpuser=None):
        MailSender.__init__(self, smtpserver)
        self.smtpUser = smtpuser or mailconfig.smtpuser
        #self.smtpPassword = None   # 4E: заставит PyMailGUI запрашивать
                                    # пароль при каждой операции отправки!
    def authenticateServer(self, server):
        server.login(self.smtpUser, self.smtpPassword)

    def getPassword(self):
        """
        get получает пароль для аутентификации на сервере SMTP, если он еще
        не известен; может вызываться суперклассом автоматически или
        клиентом вручную: не требуется до момента отправки, но не следует
        вызывать из потока выполнения графического интерфейса; пароль
        извлекается из файла на стороне клиента или методом подкласса
        """
        if not self.smtpPassword:
            try:
                localfile = open(mailconfig.smtppasswdfile)
                MailSenderAuth.smtpPassword = localfile.readline()[:-1] # 4E
                self.trace('local file password' + repr(self.smtpPassword))
            except:
                MailSenderAuth.smtpPassword = self.askSmtpPassword()   # 4E

    def askSmtpPassword(self):
        assert False, 'Subclass must define method'

class MailSenderAuthConsole(MailSenderAuth):
    def askSmtpPassword(self):
        import getpass
        prompt = 'Password for %s on %s?' % (self.smtpUser,
                                             self.smtpServerName)
        return getpass.getpass(prompt)

class SilentMailSender(SilentMailTool, MailSender):
    pass # отключает трассировку
```

Класс MailFetcher

Класс, представленный в примере 13.24, взаимодействует с почтовым сервером POP – выполняет операции загрузки, удаления и синхронизации. Этот класс заслуживает дополнительных пояснений.

Основы использования

Этот модуль имеет дело только с текстом почтовых сообщений – анализ писем после их получения делегируется другому модулю в пакете. Кроме того, этот модуль не кэширует уже загруженные данные – клиенты должны добавлять собственные инструменты сохранения почты, если

это необходимо. Клиенты должны также обеспечивать ввод пароля или его передачу, если они не могут использовать подкласс, реализующий ввод пароля в консоли (например, графические или веб-интерфейсы).

В операциях загрузки и удаления используется модуль `poplib` из стандартной библиотеки, как мы уже видели выше в этой главе. Обратите внимание, что имеются также интерфейсы, извлекающие только заголовки сообщений, использующие операцию TOP протокола POP, если почтовый сервер поддерживает ее. Это дает существенную экономию времени, когда клиенту требуется получить только самые основные сведения об имеющейся почте. Кроме того, средства получения заголовков и полных почтовых сообщений способны также загружать сообщения более свежие, чем сообщение с определенным номером (удобно использовать при выполнении начальной загрузки), и ограничивать количество загружаемых сообщений, главным образом, наиболее свежих (удобно при большом количестве входящих сообщений и медленном соединении с Интернетом или малопроизводительном сервере).

Этот модуль поддерживает также понятие индикаторов хода выполнения операции – методам, выполняющим загрузку и удаление множества сообщений, вызывающая программа может передать функцию, которая будет вызываться по мере обработки каждого сообщения. Этой функции будут передаваться текущее и общее количество шагов. Это позволяет вызывающей программе показать ход выполнения операции в консоли, в графическом или в любом другом пользовательском интерфейсе.

Декодирование Юникода полного текста сообщений при получении

Кроме того, в этом модуле применяется действующий в рамках сеанса прием декодирования байтов сообщений в строку Юникода, что необходимо для дальнейшего их анализа, о чем уже говорилось выше в этой главе. Для декодирования используется кодировка, определенная пользователем в модуле `mailconfig`, и эвристические алгоритмы. Так как декодирование выполняется немедленно, сразу после получения сообщения, все клиенты этого пакета могут быть уверены, что текст сообщения будет возвращен в виде строки Юникода `str`, и могут опираться на этот факт при выполнении дальнейших операций, включая анализ, отображение или сохранение. В дополнение к настройкам в модуле `mailconfig` выполняется также попытка угадать кодировку поочередным применением некоторых распространенных кодировок, хотя при этом вероятен вариант, когда при использовании параметров из `mailconfig` окажется невозможным сохранить в файл сообщение, декодированное с помощью угаданной кодировки.

Как уже говорилось, такой подход, когда настройки кодировки действуют в рамках всего сеанса, не является идеальным, но он может быть скорректирован в сеансе работы клиента и отражает текущие ограни-

чения пакета `email` в Python 3.1 – механизм анализа в пакете принимает уже декодированные строки Юникода, а операции получения почты возвращают двоичные строки. Если декодировать сообщение не удалось, предпринимается попытка декодировать только заголовки в текст ASCII (или в другую распространенную кодировку) или в кодировку по умолчанию для данной платформы, а в тело письма вставляется сообщение об ошибке. Этот эвристический алгоритм делает все возможное, чтобы избежать аварийного завершения клиента из-за исключения (тест этой логики вы найдете в файле *_test-decoding.py* в пакете с примерами). На практике обычно бывает вполне достаточно 8-битовой кодировки, такой как Latin-1, потому что изначально стандарты электронной почты ограничивались только кодировкой ASCII.

В принципе, можно попробовать отыскать информацию о кодировке в заголовках сообщения, выполняя анализ письма по частям. В этом случае можно ограничить действие настроек кодировки при декодировании полного текста сообщения только текущим письмом, а не сеансом, и сопровождать каждое письмо своим именем кодировки для выполнения дальнейших операций, таких как сохранение. Однако это еще больше осложнит реализацию, так как для сообщений, сохраняемых в один файл, можно определить только одну (совместимую) кодировку, общую для всего файла, а не для каждого сообщения в отдельности. Кроме того, указанные в заголовках кодировки могут относиться лишь к отдельным компонентам, а не ко всему тексту сообщения. Поскольку большинство почтовых отправлений будут соответствовать 7- или 8-битовому стандартам и в будущих версиях пакета `email` эта проблема, скорее всего, будет решена, дополнительное усложнение может оказаться излишним.

Имейте также в виду, что здесь выполняется декодирование полного текста сообщения, получаемого с сервера. В действительности это лишь один из этапов на пути декодирования сообщений в современном мире Юникода. Кроме того:

- Содержимое анализируемых частей сообщений по-прежнему может возвращаться в виде строк байтов и требовать специальной обработки или дальнейшего декодирования в Юникод (смотрите модуль *mailParser.py* далее).
- Текстовые части и вложения также налагают определенные требования к кодированию при составлении сообщений (смотрите модуль *mailSender.py* выше).
- Для заголовков сообщений имеются собственные соглашения по кодированию, и интернационализированные заголовки могут быть закодированы с применением кодировки Юникода и преобразованы в формат MIME (смотрите модули *mailParser.py* и *mailSender.py*).

Инструменты синхронизации почтового ящика

Приступив к изучению примера 13.24, вы также заметите, что значительная часть программного кода посвящена обнаружению ошибок синхронизации списка входящих почтовых отправлений на стороне клиента и текущего состояния почтового ящика на стороне POP-сервера. Обычно входящим электронным письмам присваиваются относительные порядковые номера, и в список входящих сообщений добавляются только вновь поступившие письма. Благодаря этому имеющиеся порядковые номера сообщений обычно можно использовать для удаления ненужных писем и извлечения писем впоследствии.

Однако, хотя это случается достаточно редко, номера входящих писем в почтовом ящике на стороне сервера могут измениться, что сделает недействительными номера сообщений, уже полученных клиентом. Например, сообщения могут быть удалены другим клиентом или сервер сам может перевести входящие сообщения в состояние недоставленных, если при загрузке произошла ошибка (это поведение может отличаться для разных интернет-провайдеров). В обоих случаях сообщения могут быть удалены из середины почтового ящика, в результате чего наступает рассинхронизация ранее присвоенных номеров сообщений между сервером и клиентом.

Эта ситуация может привести к ошибочному извлечению сообщений почтовым клиентом – пользователи будут принимать не те сообщения, которые они выбрали. Хуже того, эта проблема может привести к удалению не тех сообщений – если почтовый клиент использует в запросе на удаление относительные номера, изменение почтового ящика на сервере с момента последнего получения оглавления может привести к удалению не тех сообщений, которые были выбраны на стороне клиента.

Чтобы как-то помочь клиентам, модуль в примере 13.24 включает инструменты, которые сопоставляют заголовки удаляемых сообщений для обеспечения правильности и выполняют общую синхронизацию входящей почты по требованию. Эти инструменты полезны только для почтовых клиентов, которые сохраняют в качестве информации о состоянии список полученных сообщений. Мы будем использовать их в реализации клиента PyMailGUI в главе 14. Таким образом, операции удаления используют безопасный интерфейс, а операция загрузки при необходимости выполняет синхронизацию – при обнаружении рассинхронизации оглавление почтового ящика будет загружено автоматически. А теперь рассмотрите исходный программный код в примере 13.24 и комментарии к нему.

Обратите внимание, что при проверке синхронизации применяются различные приемы сопоставления, но все они требуют наличия полного текста заголовков и в самом тяжелом случае вынуждены выполнять анализ заголовков и сопоставлять содержимое множества полей. Во многих случаях было бы достаточно сопоставить ранее извлеченное поле заголовка `message-id` с входящими сообщениями на сервере. Но по-

скольку это поле является необязательным и может быть подделано, оно не обеспечивает достаточно надежную идентификацию сообщений. Иными словами, совпадение значений message-id не гарантирует соответствие сообщений, однако это поле можно использовать для выяснения несовпадений – в примере 13.24 поле message-id используется для исключения сообщений, в которых оно присутствует и его значение не совпадает с искомым. Эта проверка выполняется перед тем, как перейти к более медленной операции анализа и сопоставления множества заголовков.

Пример 13.24. PP4E\Internet\Email\mailtools\mailFetcher.py

```
"""
###############################################################################
получает, удаляет, сопоставляет почту с POP-сервера (описание и тест
приводятся в модуле __init__)
###############################################################################
"""

import poplib, mailconfig, sys         # клиентский mailconfig в sys.path
print('user:', mailconfig.popusername) # в каталоге сценария, в PYTHONPATH

from .mailParser import MailParser     # сопоставление заголовков (4E: .)
from .mailTool import MailTool, SilentMailTool  # суперкл., упр.
                                                # трассир. (4E: .)

# рассинхронизация номеров сообщений
class DeleteSynchError(Exception): pass  # обнаружена рассинхр-я при удалении
class TopNotSupported(Exception): pass   # невозможно выполнить
                                         # проверку синхронизации
class MessageSynchError(Exception): pass # обнаружена рассинхр-я оглавления

class MailFetcher(MailTool):
    """
    получение почты: соединяется, извлекает заголовки+содержимое, удаляет
    работает на любых компьютерах с Python+Интернет; создайте подкласс,
    чтобы реализовать кэширование средствами протокола POP;
    для поддержки протокола IMAP требуется создать новый класс;
    4E: предусматривает декодирование полного текста сообщений
    для последующей передачи его механизму анализа;
    """
    def __init__(self, popserver=None, popuser=None, poppswd=None,
                 hastop=True):
        self.popServer   = popserver or mailconfig.popservername
        self.popUser     = popuser or mailconfig.popusername
        self.srvrHasTop  = hastop
        self.popPassword = poppswd      # если имеет значение None,
                                        # пароль будет запрошен позднее

    def connect(self):
        self.trace('Connecting...')
```

```
        self.getPassword()              # файл, GUI или консоль
        server = poplib.POP3(self.popServer, timeout=15)
        server.user(self.popUser)       # соединиться, зарегистрироваться
        server.pass_(self.popPassword)  # pass - зарезервированное слово
        self.trace(server.getwelcome()) # print выведет приветствие
        return server

    # использовать настройки из клиентского mailconfig, находящегося в пути
    # поиска; при необходимости можно изменить в классе или в экземплярах;
    fetchEncoding = mailconfig.fetchEncoding

    def decodeFullText(self, messageBytes):
        """
        4E, Py3.1: декодирует полный текст сообщения, представленный в виде
        строки bytes, в строку Юникода str; выполняется на этапе получения
        для последующего отображения или анализа (после этого полный текст
        почтового сообщения всегда будет обрабатываться как строка Юникода);
        декодирование выполняется в соответствии с настройками в классе или
        в экземпляре или применяются наиболее распространенные кодировки;
        можно было бы также попробовать определить кодировку из заголовков
        или угадать ее, проанализировав структуру байтов; в Python 3.2/3.3
        этот этап может оказаться излишним: в этом случае измените метод
        так, чтобы он возвращал исходный список строк сообщения нетронутым;
        дополнительные подробности смотрите в главе 13;

        для большинства сообщений достаточно будет простой 8-битовой
        кодировки, такой как latin-1, потому что стандартной считается
        кодировка ASCII; этот метод применяется ко всему тексту сообщения -
        это лишь один из этапов на пути декодирования сообщений: содержимое
        и заголовки сообщений могут также находиться в формате MIME и быть
        закодированы в соответствии со стандартами электронной почты
        и Юникода; смотрите подробности в главе 13, а также реализацию
        модулей mailParser и mailSender;
        """
        text  = None
        kinds = [self.fetchEncoding]        # сначала настройки пользователя
        kinds += ['ascii', 'latin1', 'utf8']  # затем наиб. распр. кодировки
        kinds += [sys.getdefaultencoding()] # и по умолч. (может отличаться)
        for kind in kinds:            # может вызывать ошибку при сохранении
            try:
                text = [line.decode(kind) for line in messageBytes]
                break
            except (UnicodeError, LookupError): # LookupError: неверное имя
                pass

        if text == None:
            # пытается вернуть заголовки + сообщение об ошибке, иначе
            # исключение может вызвать аварийное завершение клиента;
            # пытается декодировать заголовки как ascii,
            # с применением других кодировок или с помощью
            # кодировки по умолчанию для платформы;
```

```
        blankline = messageBytes.index(b'')
        hdrsonly  = messageBytes[:blankline]
        commons   = ['ascii', 'latin1', 'utf8']
        for common in commons:
            try:
                text = [line.decode(common) for line in hdrsonly]
                break
            except UnicodeError:
                pass
        else:              # не подошла ни одна кодировка
            try:
                text = [line.decode() for line in hdrsonly] # по умолч.?
            except UnicodeError:
               text= ['From: (sender of unknown Unicode format headers)']
        text += ['',
                 '--Sorry: mailtools cannot decode this mail content!--']
    return text

def downloadMessage(self, msgnum):
    """
    загружает полный текст одного сообщения по указанному относительному
    номеру POP msgnum; анализ содержимого выполняет вызывающая программа
    """
    self.trace('load ' + str(msgnum))
    server = self.connect()
    try:
        resp, msglines, respsz = server.retr(msgnum)
    finally:
        server.quit()
    msglines = self.decodeFullText(msglines) # декодировать bytes в str
    return '\n'.join(msglines)               # объединить строки

def downloadAllHeaders(self, progress=None, loadfrom=1):
    """
    получает только размеры и заголовки для всех или только
    для сообщений с номерами от loadfrom и выше;
    используйте loadfrom для загрузки
    только новых сообщений; для последующей загрузки полного текста
    сообщений используйте downloadMessage; progress - это функция,
    которая вызывается с параметрами (счетчик, всего);
    возвращает: [текст заголовков], [размеры сообщений],
    флаг "сообщения загружены полностью"

    4E: добавлена проверка параметра mailconfig.fetchlimit для поддержки
    почтовых ящиков с большим количеством входящих сообщений: если он
    не равен None, извлекается только указанное число заголовков, вместо
    остальных возвращаются пустые заголовки; иначе пользователи,
    получающие большое количество сообщений, как я (4K сообщений),
    будут испытывать неудобства;
    4E: передает loadfrom методу downloadAllMessages (чтобы хоть
        немного облегчить положение);
```

```
        """
        if not self.srvrHasTop: # не все серверы поддерживают команду TOP
            # загрузить полные сообщения
            return self.downloadAllMsgs(progress, loadfrom)
        else:
            self.trace('loading headers')
            fetchlimit = mailconfig.fetchlimit
            server = self.connect()     # ящик теперь заблокирован до вызова
                                        # метода quit
            try:
                resp, msginfos, respsz = server.list() # список строк
                                                       # 'номер размер'
                msgCount = len(msginfos) # альтернатива методу srvr.stat[0]
                msginfos = msginfos[loadfrom-1:]      # пропустить уже загр.
                allsizes = [int(x.split()[1]) for x in msginfos]
                allhdrs = []
                for msgnum in range(loadfrom, msgCount+1):  # возможно пустой
                    if progress: progress(msgnum, msgCount) # вызвать progress
                    if fetchlimit and (msgnum <= msgCount - fetchlimit):
                        # пропустить, добавить пустой заголовок
                        hdrtext = 'Subject: --mail skipped--\n\n'
                        allhdrs.append(hdrtext)
                    else:
                        # получить, только заголовки
                        resp, hdrlines, respsz = server.top(msgnum, 0)
                        hdrlines = self.decodeFullText(hdrlines)
                        allhdrs.append('\n'.join(hdrlines))
            finally:
                server.quit()            # разблокировать почтовый ящик
            assert len(allhdrs) == len(allsizes)
            self.trace('load headers exit')
            return allhdrs, allsizes, False

    def downloadAllMessages(self, progress=None, loadfrom=1):
        """
        загрузить все сообщения целиком с номерами loadfrom..N,
        независимо от кэширования, которое может выполняться вызывающей
        программой; намного медленнее, чем downloadAllHeaders,
        если требуется загрузить только заголовки;

        4E: поддержка mailconfig.fetchlimit: смотрите downloadAllHeaders;
        можно было бы использовать server.list() для получения размеров
        пропущенных сообщений, но клиентам скорее всего этого не требуется;
        """
        self.trace('loading full messages')
        fetchlimit = mailconfig.fetchlimit
        server = self.connect()
        try:
            (msgCount, msgBytes) = server.stat()   # ящик на сервере
            allmsgs  = []
            allsizes = []
```

```
        for i in range(loadfrom, msgCount+1):  # пусто, если low >= high
            if progress: progress(i, msgCount)
            if fetchlimit and (i <= msgCount - fetchlimit):
                # пропустить, добавить пустое сообщение
                mailtext = 'Subject: --mail skipped--\n\nMail skipped.\n'
                allmsgs.append(mailtext)
                allsizes.append(len(mailtext))
            else:
                # получить полные сообщения
                (resp, message, respsz) = server.retr(i) # сохр. в списке
                message = self.decodeFullText(message)
                allmsgs.append('\n'.join(message)) # оставить на сервере
                allsizes.append(respsz)            # отлич. от len(msg)
    finally:
        server.quit()                          # разблокировать ящик
    assert len(allmsgs) == (msgCount - loadfrom) + 1    # нумерация с 1
   #assert sum(allsizes) == msgBytes           # если не loadfrom > 1
    return allmsgs, allsizes, True             # и если нет fetchlimit

def deleteMessages(self, msgnums, progress=None):
    """
    удаляет несколько сообщений на сервере; предполагается, что номера
    сообщений в ящике не изменялись с момента последней
    синхронизации/загрузки; используется, если заголовки сообщения
    недоступны; выполняется быстро, но может быть опасен: смотрите
    deleteMessagesSafely
    """
    self.trace('deleting mails')
    server = self.connect()
    try:                                        # не устанавливать
        for (ix, msgnum) in enumerate(msgnums): # соединение для каждого
            if progress: progress(ix+1, len(msgnums))
            server.dele(msgnum)
    finally:                            # номера изменились: перезагрузить
        server.quit()

def deleteMessagesSafely(self, msgnums, synchHeaders, progress=None):
    """
    удаляет несколько сообщений на сервере, но перед удалением выполняет
    проверку заголовка с помощью команды TOP; предполагает, что почтовый
    сервер поддерживает команду TOP протокола POP, иначе возбуждает
    исключение TopNotSupported - клиент может вызвать deleteMessages;

    используется, если почтовый ящик на сервере мог измениться с момента
    последней операции получения оглавления и соответственно могли
    измениться номера POP-сообщений; это может произойти при удалении
    почты с помощью другого клиента; кроме того, некоторые провайдеры
    могут перемещать почту из ящика входящих сообщений в ящик
    недоставленных сообщений в случае ошибки во время загрузки;

    аргумент synchHeaders должен быть списком уже загруженных
```

```
        заголовков, соответствующих выбранным сообщениям
        (обязательная информация);
        возбуждает исключение, если обнаруживается рассинхронизация
        с почтовым сервером; доступ к входящей почте
        блокируется до вызова метода quit, поэтому номера не могут
        измениться между командой TOP и фактическим
        удалением: проверка синхронизации должна выполняться здесь,
        а не в вызывающей программе; может оказаться недостаточным
        вызвать checkSynchError+deleteMessages, но здесь проверяется
        каждое сообщение, на случай удаления или вставки
        сообщений в середину почтового ящика;
        """
        if not self.srvrHasTop:
            raise TopNotSupported('Safe delete cancelled')

        self.trace('deleting mails safely')
        errmsg  = 'Message %s out of synch with server.\n'
        errmsg += 'Delete terminated at this message.\n'
        errmsg += 'Mail client may require restart or reload.'

        server = self.connect()                # блокирует ящик до quit
        try:                                   # не устан. соед. для каждого
            (msgCount, msgBytes) = server.stat() # объем входящей почты
            for (ix, msgnum) in enumerate(msgnums):
                if progress: progress(ix+1, len(msgnums))
                if msgnum > msgCount:            # сообщения были удалены
                    raise DeleteSynchError(errmsg % msgnum)
                resp, hdrlines, respsz = server.top(msgnum, 0) # только загол.
                hdrlines = self.decodeFullText(hdrlines)
                msghdrs = '\n'.join(hdrlines)
                if not self.headersMatch(msghdrs, synchHeaders[msgnum-1]):
                    raise DeleteSynchError(errmsg % msgnum)
                else:
                    server.dele(msgnum)   # безопасно удалить это сообщение
        finally:                          # номера изменились: перезагрузить
            server.quit()                 # разблокировать при выходе

    def checkSynchError(self, synchHeaders):
        """
        сопоставляет уже загруженные заголовки в списке synchHeaders с теми,
        что находятся на сервере, с использованием команды TOP
        протокола POP, извлекающей текст заголовков;
        используется, если содержимое почтового ящика могло измениться,
        например в результате удаления сообщений с помощью другого клиента
        или в результате автоматических действий, выполняемых
        почтовым сервером; возбуждает исключение в случае обнаружения
        рассинхронизации или ошибки во время взаимодействия с сервером;

        для повышения скорости проверяется только последний в последнем:
        это позволяет обнаружить факт удаления из ящика, но предполагает,
        что сервер не мог вставить новые сообщения перед последним (верно
        для входящих сообщений); сначала проверяется объем входящей почты:
```

```
        если меньше - были только удаления; иначе, если сообщения удалялись
        и в конец добавлялись новые, результат top будет отличаться;
        результат этого метода можно считать действительным только на момент
        его работы: содержимое ящика входящих сообщений может
        измениться после возврата;
        """
        self.trace('synch check')
        errormsg  = 'Message index out of synch with mail server.\n'
        errormsg += 'Mail client may require restart or reload.'
        server = self.connect()
        try:
            lastmsgnum = len(synchHeaders)        # 1..N
            (msgCount, msgBytes) = server.stat()  # объем входящей почты
            if lastmsgnum > msgCount:             # теперь меньше?
                raise MessageSynchError(errormsg) # нечего сравнивать
            if self.srvrHasTop:
                resp, hdrlines, respsz = server.top(lastmsgnum, 0) # только
                hdrlines = self.decodeFullText(hdrlines)         # заголовки
                lastmsghdrs = '\n'.join(hdrlines)
                if not self.headersMatch(lastmsghdrs, synchHeaders[-1]):
                    raise MessageSynchError(errormsg)
        finally:
            server.quit()

    def headersMatch(self, hdrtext1, hdrtext2):
        """
        для сопоставления недостаточно простого сравнения строк: некоторые
        серверы добавляют заголовок "Status:", который изменяется с течением
        времени; у одного провайдера он устанавливался изначально
        как "Status: U" (unread - непрочитанное) и заменялся на "Status: RO"
        (read, old - прочитано, старое) после загрузки сообщения -
        это сбивает с толку механизм проверки синхронизации,
        если после загрузки нового оглавления, но непосредственно
        перед удалением или проверкой последнего сообщения клиентом
        было загружено новое сообщение;
        теоретически значение заголовка "Message-id:" является уникальным
        для сообщения, но сам заголовок является необязательным и может
        быть подделан; сначала делается попытка выполнить более типичное
        сопоставление; анализ - дорогостоящая операция, поэтому
        выполняется последним
        """
        # попробовать просто сравнить строки
        if hdrtext1 == hdrtext2:
            self.trace('Same headers text')
            return True

        # попробовать сопоставить без заголовков Status
        split1 = hdrtext1.splitlines() # s.split('\n'), но без последнего
        split2 = hdrtext2.splitlines() # элемента пустой строки (''')
        strip1 = [line for line in split1 if not line.startswith('Status:')]
        strip2 = [line for line in split2 if not line.startswith('Status:')]
```

```
        if strip1 == strip2:
            self.trace('Same without Status')
            return True

        # попробовать найти несовпадения заголовков message-id,
        # если они имеются
        msgid1 = [line for line in split1
                       if line[:11].lower() == 'message-id:']
        msgid2 = [line for line in split2
                       if line[:11].lower() == 'message-id:']
        if (msgid1 or msgid2) and (msgid1 != msgid2):
            self.trace('Different Message-Id')
            return False

        # выполнить полный анализ заголовков и сравнить наиболее типичные
        # из них, если заголовки message-id отсутствуют или в них
        # были найдены различия
        tryheaders = ('From', 'To', 'Subject', 'Date')
        tryheaders += ('Cc', 'Return-Path', 'Received')
        msg1 = MailParser().parseHeaders(hdrtext1)
        msg2 = MailParser().parseHeaders(hdrtext2)
        for hdr in tryheaders:     # возможно несколько адресов в Received
            if msg1.get_all(hdr) != msg2.get_all(hdr): # без учета регистра,
                self.trace('Diff common headers')      # по умолчанию None
                return False

        # все обычные заголовки совпадают
        # и нет отличающихся заголовков message-id
        self.trace('Same common headers')
        return True

    def getPassword(self):
        """
        получает пароль POP, если он еще не известен
        не требуется до обращения к серверу из файла
        на стороне клиента или вызовом метода подкласса
        """
        if not self.popPassword:
            try:
                localfile = open(mailconfig.poppasswdfile)
                self.popPassword = localfile.readline()[:-1]
                self.trace('local file password' + repr(self.popPassword))
            except:
                self.popPassword = self.askPopPassword()

    def askPopPassword(self):
        assert False, 'Subclass must define method'

##############################################################################
# специализированные подклассы
##############################################################################
```

```
class MailFetcherConsole(MailFetcher):
    def askPopPassword(self):
        import getpass
        prompt = 'Password for %s on %s?' % (self.popUser, self.popServer)
        return getpass.getpass(prompt)

class SilentMailFetcher(SilentMailTool, MailFetcher):
    pass # отключает трассировку
```

Класс MailParser

В примере 13.25 представлена реализация последнего основного класса в пакете mailtools – получая (уже декодированный) текст электронного письма, этот инструмент выполняет его анализ и преобразует в объект Message с заголовками и декодированными частями. В значительной степени этот модуль является всего лишь оберткой вокруг пакета email из стандартной библиотеки, но он добавляет ряд удобных инструментов, позволяющих отыскивать основную текстовую часть сообщения, генерировать имена файлов для вложенных частей сообщения, сохранять вложенные части в файлы, декодировать заголовки, разбивать списки адресов и так далее. За дополнительной информацией обращайтесь к исходному программному коду. Обратите также внимание, как здесь выполняется обход частей: определив логику поиска в одном месте в виде функции-генератора, мы гарантировали, что все три клиента этой логики в этом модуле, а также любые другие клиенты будут выполнять обход совершенно одинаково.

Декодирование Юникода для текстового содержимого частей и заголовков

Этот модуль обеспечивает поддержку декодирования заголовков сообщений в соответствии со стандартами электронной почты (полных заголовков и компонентов имен в заголовках с адресами) и выполняет декодирование текстовых частей с применением их собственных кодировок. Заголовки декодируются с применением инструментов из пакета email в соответствии с их содержимым – каждый заголовок сам определяет свой формат MIME и кодировку Юникода, поэтому для их декодирования не требуется вмешательства пользователя. Для удобства реализации клиентов класс также реализует декодирование основных текстовых частей с целью преобразовать их из типа bytes в тип str, если это необходимо.

Последняя операция, декодирование основного текста, заслуживает отдельного описания. Как уже обсуждалось выше в этой главе, объекты Message (основные и вложенные) могут возвращать свое содержимое в виде строк bytes, если извлекать его с аргументом decode=1 или если это содержимое изначально имело тип bytes. В других случаях содержимое может возвращаться в виде строки str. В общем случае нам тре-

буется декодировать строку bytes, чтобы иметь возможность обращаться с содержимым, как с текстом.

В самом пакете mailtools содержимое str текстовых частей автоматически кодируется в байты при использовании аргумента decode=1 и затем сохраняется в файлы в двоичном режиме, чтобы избежать проблем с кодировками, но текст основного содержимого декодируется в строку str, если он был представлен строкой bytes. Эта операция декодирования основного текста выполняется с использованием имени кодировки, извлекаемой из заголовка сообщения этой части (если указана и корректна), или кодировки по умолчанию для текущей платформы, или делается попытка угадать кодировку. Как мы узнали в главе 9, графические интерфейсы способны отображать строки bytes, тем не менее, строки str в целом обеспечивают более широкую поддержку Юникода. Кроме того, строки str иногда необходимо получить, чтобы обеспечить дальнейшую обработку, например, добавить символы перевода строки или сгенерировать веб-страницу.

Поскольку пакет не может делать какие-либо предположения относительно других частей, кроме главной текстовой части, их кодирование и декодирование полностью возлагается на клиентов. Например, другие текстовые части, сохраняемые здесь в двоичном режиме, могут требовать, чтобы для их отображения кодировки Юникода брались из заголовков сообщения. Этот прием будет использовать приложение PyMailGUI в главе 14 для открытия текстовых частей по требованию, передавая информацию о кодировке из заголовка сообщения редактору PyEdit для декодирования содержимого в текст.

Некоторые из текстовых преобразований, выполняемых здесь, являются лишь частичными решениями (в некоторых частях сообщения могут отсутствовать требуемые заголовки, а применение кодировки по умолчанию для текущей платформы может приводить к неудаче) и, возможно, должны быть усовершенствованы. Поскольку есть основания надеяться, что эти проблемы будут решены в будущей версии пакета email, мы пока примем наши допущения.

Пример 13.25. PP4E\Internet\Email\mailtools\mailParser.py

```
"""
##############################################################################
разбор и извлечение, анализ, сохранение вложения (описание и тест
приводятся в модуле __init__)
##############################################################################
"""


import os, mimetypes, sys          # mime: отображение типа в имя
import email.parser                # анализ текста в объекте Message
import email.header                # 4E: кодирование/декодирование заголовков
import email.utils                 # 4E: кодирование/декодирование
                                   #     заголовков с адресами
```

```
from email.message import Message # обход объектов Message
from .mailTool import MailTool    # 4E: относительно пакета

class MailParser(MailTool):
    """
    методы анализа текста сообщения, вложений

    важное замечание: содержимое объекта Message может быть простой строкой
    в простых несоставных сообщениях или списком объектов Message
    в сообщениях, состоящих из нескольких частей (возможно, вложенных);
    мы не будем различать эти два случая, потому что генератор walk
    объекта Message всегда первым возвращает сам объект и прекрасно
    обрабатывает простые, несоставные объекты
    (выполняется обход единственного объекта);

    в случае простых сообщений тело сообщения всегда рассматривается здесь
    как единственная часть сообщения; в случае составных сообщений список
    частей включает основной текст сообщения, а также все вложения;
    это позволяет обрабатывать в графических интерфейсах простые нетекстовые
    сообщения как вложения (например, сохранять, открывать);
    иногда, в редких случаях, содержимым частей объекта Message
    может быть None;

    4E примечание: в Py 3.1 содержимое текстовых частей возвращается в виде
    строки bytes, когда передается аргумент decode=1, в других случаях может
    возвращаться строка str; в модуле mailtools текст хранится в виде строки
    bytes, чтобы упростить сохранение в файлах, но основное текстовое
    содержимое декодируется в строку str в соответствии с информацией
    в заголовках или с применением кодировки по умолчанию+предполагаемой;
    при необходимости клиенты должны сами декодировать остальные части:
    для декодирования частей, сохраненных в двоичных файлах, PyMailGUI
    использует информацию в заголовках;

    4E: добавлена поддержка автоматического декодирования заголовков
    сообщения в соответствии с их содержимым - как полных заголовков,
    таких как Subject, так и компонентов имен в заголовках
    с адресами, таких как From и To;
    клиент должен запрашивать эту операцию после анализа полного
    текста сообщения, перед отображением: механизм анализа
    не выполняет декодирование;
    """

    def walkNamedParts(self, message):
        """
        функция-генератор, позволяющая избежать повторения логики выбора
        именованных частей; пропускает заголовки multipart, извлекает
        имена файлов частей; message - это уже созданный из сообщения
        объект email.message.Message; не пропускает части необычного типа:
        содержимым может быть None, при сохранении следует обрабатывать
        такую возможность; некоторые части некоторых других типов также
        может потребоваться пропустить;
```

```
        """
        for (ix, part) in enumerate(message.walk()): # walk включает сообщение
            fulltype = part.get_content_type() # ix включает пропущенные части
            maintype = part.get_content_maintype()
            if maintype == 'multipart':            # multipart/*: контейнер
                continue
            elif fulltype == 'message/rfc822': # 4E: пропустить message/rfc822
                continue                       # пропустить все message/* ?
            else:
                filename, contype = self.partName(part, ix)
                yield (filename, contype, part)

    def partName(self, part, ix):
        """
        извлекает имя файла и тип содержимого из части сообщения;
        имя файла: сначала пытается определить из параметра
        filename заголовка Content-Disposition, затем из параметра name
        заголовка Content-Type и под конец генерирует имя файла из типа,
        определяемого с помощью модуля mimetypes;
        """
        filename = part.get_filename()       # имя файла в заголовке?
        contype = part.get_content_type()    # тип/подтип, в нижнем регистре
        if not filename:
            filename = part.get_param('name') # проверить параметр name
        if not filename:                      # заголовка content-type
            if contype == 'text/plain':       # расширение текстового файла
                ext = '.txt'                  # иначе будет предложено .ksh!
            else:
                ext = mimetypes.guess_extension(contype)
                if not ext: ext = '.bin'      # универсальное по умолчанию
            filename = 'part-%03d%s' % (ix, ext)
        return (filename, contype)

    def saveParts(self, savedir, message):
        """
        сохраняет все части сообщения в файлахв локальном каталоге;
        возвращает список [('тип/подтип, 'имя файла')] для использования
        в вызывающей программе, но не открывает какие-либо части или
        вложения; метод get_payload декодирует содержимое с применением
        кодировок base64, quoted-printable, uuencoded; механизм анализа
        почтовых сообщений может вернуть содержимое None для некоторых
        необычных типов частей, которые, вероятно, следует пропустить:
        здесь преобразовать в str для безопасности;
        """
        if not os.path.exists(savedir):
            os.mkdir(savedir)
        partfiles = []
        for (filename, contype, part) in self.walkNamedParts(message):
            fullname = os.path.join(savedir, filename)
            fileobj = open(fullname, 'wb')        # двоичный режим
            content = part.get_payload(decode=1)  # декодирует base64,qp,uu
```

```
        if not isinstance(content, bytes):    # 4E: bytes для rb
            content = b'(no content)'    # decode=1 возвращает bytes,
        fileobj.write(content)           # но для некоторых типов - None
        fileobj.close()                  # 4E: не str(content)
        partfiles.append((contype, fullname)) # для открытия
    return partfiles                          # в вызывающей программе

def saveOnePart(self, savedir, partname, message):
    """
    то же самое, но отыскивает по имени только одну часть
    и сохраняет ее
    """
    if not os.path.exists(savedir):
        os.mkdir(savedir)
    fullname = os.path.join(savedir, partname)
    (contype, content) = self.findOnePart(partname, message)
    if not isinstance(content, bytes):  # 4E: bytes для rb
        content = b'(no content)'       # decode=1 возвращает bytes,
    open(fullname, 'wb').write(content) # но для некоторых типов - None
    return (contype, fullname)          # 4E: не str(content)

def partsList(self, message):
    """
    возвращает список имен файлов для всех частей уже
    проанализированного сообщения, используется та же логика определения
    имени файла, что и в saveParts, но не сохраняет части в файлы
    """
    validParts = self.walkNamedParts(message)
    return [filename for (filename, contype, part) in validParts]

def findOnePart(self, partname, message):
    """
    отыскивает и возвращает содержимое части по ее имени;
    предназначен для совместного использования с методом partsList;
    можно было бы также использовать mimetypes.guess_type(partname);
    необходимости поиска можно было бы избежать, сохраняя данные
    в словаре;
    4E: содержимое может иметь тип str или bytes - преобразовать
    при необходимости;
    """
    for (filename, contype, part) in self.walkNamedParts(message):
        if filename == partname:
            content = part.get_payload(decode=1) # декодирует base64,qp,uu
            return (contype, content)            # может быть текст
                                                 # в двоичном виде
def decodedPayload(self, part, asStr=True):
    """
    4E: декодирует текстовую часть, представленную в виде
    строки bytes, в строку str Юникода для отображения,
    разбиения на строки и так далее;
    аргумент part - это объект Message; (decode=1) декодирует
    из формата MIME (base64, uuencode, qp), bytes.decode() выполняет
```

```
        дополнительное декодирование в текстовые строки Юникода;
        прежде чем вернуть строку с ошибкой, сначала пытается применить
        кодировку, указанную в заголовках сообщения (если имеется
        и соответствует), затем пытается применить кодировку по умолчанию
        для текущей платформы и несколько предполагаемых кодировок;
        """
        payload = part.get_payload(decode=1)     # может быть строка bytes
        if asStr and isinstance(payload, bytes): # decode=1 возвращает bytes
            tries = []
            enchdr = part.get_content_charset()  # сначала проверить
            if enchdr:                           # заголовки сообщения
                tries += [enchdr]
            tries += [sys.getdefaultencoding()]  # то же, что и bytes.decode()
            tries += ['latin1', 'utf8']        # попр. 8-битовые, вкл. ascii
            for trie in tries:                 # попр. utf8 (умолч. Windows)
                try:
                    payload = payload.decode(trie)  # подошла?
                    break
                except (UnicodeError, LookupError): # lookuperr:
                    pass                            # недопустимое имя
            else:
                payload = '--Sorry: cannot decode Unicode text--'
        return payload

    def findMainText(self, message, asStr=True):
        """
        для текстовых клиентов возвращает первую текстовую часть в виде str;
        в содержимом простого сообщения или во всех частях составного
        сообщения отыскивает часть типа text/plain, затем text/html, затем
        text/*, после чего принимается решение об отсутствии текстовой
        части, пригодной для отображения; это эвристическое решение,
        но оно охватывает простые, а также multipart/alternative
        и multipart/mixed сообщения;
        если это не простое сообщение, текстовая часть по умолчанию имеет
        заголовок content-type со значением text/plain;

        обрабатывает вложенные сообщения, выполняя обход начиная с верхнего
        уровня, вместо сканирования списка; если это не составное сообщение,
        но имеет тип text/html, возвращает разметку HTML
        как текст типа HTML: ызывающая программа может
        в открыть его в веб-броузере, извлечь простой
        текст и так далее; если это простое сообщение и текстовая часть
        не найдена, следовательно, нет текста для отображения: предусмотрите
        сохранение/открытие содержимого в графическом интерфейсе;
        предупреждение: не пытайтесь объединить несколько встроенных
        частей типа text/plain, если они имеются;
        4E: текстовое содержимое может иметь тип bytes -
        декодирует в str здесь;
        4E: передайте asStr=False, чтобы получить разметку HTML в двоичном
        представлении для сохранения в файл;
        """
```

```
        # отыскать простой текст
        for part in message.walk():        # walk выполнит обход всех частей
            type = part.get_content_type() # если не составное
            if type == 'text/plain':       # может иметь формат base64,qp,uu
                return type, self.decodedPayload(part, asStr) # bytes в str?

        # отыскать часть с разметкой HTML
        for part in message.walk():
            type = part.get_content_type() # html отображается вызывающей ф.
            if type == 'text/html':
                return type, self.decodedPayload(part, asStr)

        # отыскать части любого другого текстового типа, включая XML
        for part in message.walk():
            if part.get_content_maintype() == 'text':
                return part.get_content_type(),self.decodedPayload(part,asStr)

        # не найдено: можно было бы использовать первую часть,
        # но она не помечена как текстовая
        failtext ='[No text to display]' if asStr else b'[No text to display]'
        return 'text/plain', failtext

    def decodeHeader(self, rawheader):
        """
        4E: декодирует текст заголовка i18n в соответствии со стандартами
        электронной почты и Юникода и их содержимым; в случае ошибки
        при декодировании возвращает в первоначальном виде; клиент должен
        вызывать этот метод для подготовки заголовка к отображению: объект
        Message не декодируется;
        пример: '=?UTF-8?Q?Introducing=20Top=20Values=20..Savers?=';
        пример: 'Man where did you get that =?UTF-8?Q?assistant=3F?=';

        метод decode_header автоматически обрабатывает любые разрывы строк
        в заголовке, может возвращать несколько частей, если в заголовке
        имеется несколько подстрок, закодированных по-разному, и возвращает
        все части в виде списка строк bytes, если кодировки были найдены
        (некодированные части возвращаются как закодированные
        в raw-unicode-escape, со значением enc=None), но возвращает
        единственную часть с enc=None, которая является строкой str,
        а не bytes в Py3.1, если весь заголовок
        оказался незакодированным (должен обрабатывать смешанные типы);
        дополнительные подробности/примеры смотрите в главе 13;

        следующей реализации было бы достаточно, если бы не возможность
        появления подстрок, кодированных по-разному,
        или если бы в переменной enc не возвращалось
        значение None (возбуждает исключение, в результате
        которого аргумент rawheader возвращается в исходном виде):

        hdr, enc = email.header.decode_header(rawheader)[0]
```

```
        return hdr.decode(enc) # ошибка, если enc=None: нет имени кодировки
                               # или кодированных подстрок
        """
        try:
            parts = email.header.decode_header(rawheader)
            decoded = []
            for (part, enc) in parts:               # для всех подстрок
                if enc == None:                     # некодированная часть?
                    if not isinstance(part, bytes): # str: некодир. заголовок
                        decoded += [part]           # иначе декодир. в Юникод
                    else:
                        decoded += [part.decode('raw-unicode-escape')]
                else:
                    decoded += [part.decode(enc)]
            return ' '.join(decoded)
        except:
            return rawheader                        # вернуть как есть!

    def decodeAddrHeader(self, rawheader):
        """
        4E: декодирует заголовок i18n с адресами в соответствии
        со стандартами электронной почты и Юникода и их содержимым;
        должен анализировать первую часть адреса, чтобы получить
        интернационализированную часть:
        '"=?UTF-8?Q?Walmart?=" <newsletters@walmart.com>';
        заголовок From скорее всего будет содержать единственный адрес,
        но заголовки To, Cc, Bcc могут содержать несколько адресов;

        метод decodeHeader обрабатывает вложенные подстроки в разных
        кодировках внутри заголовка, но мы не можем напрямую вызвать
        его здесь для обработки всего заголовка, потому что он будет
        завершаться с ошибкой, если закодированная строка
        с именем будет заканчиваться кавычкой ", а не пробелом
        или концом строки; смотрите также метод encodeAddrHeader
        в модуле mailSender, реализующий обратную операцию;

        ниже приводится первая реализация, которая терпела неудачу
        при обработке некодированных подстрок в имени и возбуждала
        исключение при встрече некодированных частей типа bytes,
        если в адресе имеется хоть одна закодированная подстрока;

        namebytes, nameenc = email.header.decode_header(name)[0] (email+MIME)
        if nameenc: name = namebytes.decode(nameenc)             (Юникод?)
        """
        try:
            pairs = email.utils.getaddresses([rawheader]) # разбить на части
            decoded = []                                  # учитывает запятые
            for (name, addr) in pairs:                    # в именах
                try:
                    name = self.decodeHeader(name)        # email+MIME+Юникод
```

```
            except:
                name = None     # исп. кодиров. имя при возб. искл.
                                # в decodeHeader
            joined = email.utils.formataddr((name, addr))  # объединить
            decoded.append(joined)
        return ', '.join(decoded)                   # более 1 адреса
    except:
        return self.decodeHeader(rawheader)  # попробовать декодировать
                                             # всю строку
def splitAddresses(self, field):
    """
    4E: используйте в графическом интерфейсе запятую как
    символ-разделитель адресов и функцию getaddresses
    для корректного разбиения, которая позволяет использовать
    запятые в компонентах имен адресов;
    используется программой PyMailGUI для разбиения содержимого
    заголовков To, Cc, Bcc, обработки ввода пользователя и копий
    заголовков; возвращает пустой список, если аргумент field пуст
    или возникло какое-либо исключение;
    """
    try:
        pairs = email.utils.getaddresses([field]) # [(имя,адр)]
        return [email.utils.formataddr(pair)      # [имя <адр>]
                for pair in pairs]
    except:
        return '' # синтаксическая ошибка в поле, введенном
                  # пользователем?, и так далее

# возвращаются, когда анализ завершается неудачей
errorMessage = Message()
errorMessage.set_payload('[Unable to parse message - format error]')

def parseHeaders(self, mailtext):
    """
    анализирует только заголовки, возвращает корневой объект
    email.message.Message; останавливается сразу после анализа
    заголовков, даже если за ними ничего не следует (команда top);
    объект email.message.Message является отображением заголовков
    сообщения; в качестве содержимого объекта сообщения устанавливается
    значение None, а не необработанный текст тела
    """
    try:
        return email.parser.Parser().parsestr(mailtext, headersonly=True)
    except:
        return self.errorMessage

def parseMessage(self, fulltext):
    """
    анализирует все сообщение, возвращает корневой объект
    email.message.Message; содержимым объекта сообщения является строка,
    если is_multipart() возвращает False; при наличии нескольких частей
```

```
        содержимым объекта сообщения является множество объектов Message;
        метод, используемый здесь, действует так же, как функция
        email.message_from_string()
        """
        try:                                                # может потерпеть
            return email.parser.Parser().parsestr(fulltext) # неудачу!
        except:                       # или дать возможность обработать
            return self.errorMessage # в вызывающей программе? можно
                                      # проверить возвращаемое значение

    def parseMessageRaw(self, fulltext):
        """
        анализирует только заголовки, возвращает корневой объект
        email.message.Message; останавливается сразу
        после анализа заголовков для эффективности
        (здесь не используется); содержимым объекта
        сообщения является необработанный текст письма,
        следующий за заголовками
        """
        try:
            return email.parser.HeaderParser().parsestr(fulltext)
        except:
            return self.errorMessage
```

Сценарий самотестирования

Последний файл в пакете `mailtools` содержит программный код самотестирования, представленный в примере 13.26. Он оформлен в виде отдельного файла, чтобы обеспечить возможность манипулирования путем поиска модулей – он имитирует работу клиента, который, как предполагается, имеет собственный модуль `mailconfig.py` в своем каталоге (каждый клиент может иметь собственную версию этого модуля).

Пример 13.26. PP4E\Internet\Email\mailtools\selftest.py

```
"""
##############################################################################
когда этот файл запускается как самостоятельный сценарий, выполняет
тестирование пакета
##############################################################################
"""


#
# обычно используется модуль mailconfig, находящийся в каталоге клиента
# или в пути sys.path; для нужд тестирования берется модуль
# из каталога Email уровнем выше
#
import sys
sys.path.append('..')
import mailconfig
print('config:', mailconfig.__file__)
```

```
# получить из __init__
from mailtools import (MailFetcherConsole,
                       MailSender, MailSenderAuthConsole,
                       MailParser)

if not mailconfig.smtpuser:
    sender = MailSender(tracesize=5000)
else:
    sender = MailSenderAuthConsole(tracesize=5000)

sender.sendMessage(From      = mailconfig.myaddress,
                   To        = [mailconfig.myaddress],
                   Subj      = 'testing mailtools package',
                   extrahdrs = [('X-Mailer', 'mailtools')],
                   bodytext  = 'Here is my source code\n',
                   attaches  = ['selftest.py'],
                  )

               # bodytextEncoding='utf-8',      # дополнительные тесты
               # attachesEncodings=['latin-1'], # проверка текста заголовков
               # attaches=['monkeys.jpg'])      # проверка Base64
               # to='i18n adddr list...',     # тест заголовков mime/unicode

# измените параметр fetchlimit в модуле mailconfig,
# чтобы проверить ограничение на количество получаемых сообщений
fetcher = MailFetcherConsole()
def status(*args): print(args)

hdrs, sizes, loadedall = fetcher.downloadAllHeaders(status)
for num, hdr in enumerate(hdrs[:5]):
    print(hdr)
    if input('load mail?') in ['y', 'Y']:
        print(fetcher.downloadMessage(num+1).rstrip(), '\n', '-'*70)

last5 = len(hdrs)-4
msgs, sizes, loadedall = fetcher.downloadAllMessages(status, loadfrom=last5)
for msg in msgs:
    print(msg[:200], '\n', '-'*70)

parser = MailParser()
for i in [0]:                 # попробуйте [0 , len(msgs)]
    fulltext = msgs[i]
    message  = parser.parseMessage(fulltext)
    ctype, maintext = parser.findMainText(message)
    print('Parsed:', message['Subject'])
    print(maintext)
input('Press Enter to exit')  # пауза на случай запуска
                              # щелчком мыши в Windows
```

Запуск сценария самотестирования

Ниже приводятся результаты запуска сценария самотестирования. Он выводит большое количество строк, значительная часть из которых была удалена при представлении в книге, – как обычно, запустите его на своем компьютере, чтобы получить более полное представление:

```
C:\...\PP4E\Internet\Email\mailtools> selftest.py
config: ..\mailconfig.py
user: PP4E@learning-python.com
Adding text/x-python
Sending to...['PP4E@learning-python.com']
Content-Type: multipart/mixed; boundary="===============0085314748==" 
MIME-Version: 1.0
From: PP4E@learning-python.com
To: PP4E@learning-python.com
Subject: testing mailtools package
Date: Sat, 08 May 2010 19:26:22 -0000
X-Mailer: mailtools

A multi-part MIME format message.

----------------0085314748==
Content-Type: text/plain; charset="us-ascii"
MIME-Version: 1.0
Content-Transfer-Encoding: 7bit

Here is my source code

----------------0085314748==
Content-Type: text/x-python; charset="us-ascii"
MIME-Version: 1.0
Content-Transfer-Encoding: 7bit
Content-Disposition: attachment; filename="selftest.py"

……
##############################################################################
когда этот файл запускается как самостоятельный сценарий, выполняет
тестирование пакета
##############################################################################
……
```

...часть строк опущена...

```
print(maintext)
input('Press Enter to exit')   # пауза, на случай запуска
                               # щелчком мыши в Windows

----------------0085314748==--

Send exit
loading headers
```

```
Connecting...
Password for PP4E@learning-python.com on pop.secureserver.net?
b'+OK <28121.1273346862@p3pop01-07.prod.phx3.gdg>'
(1, 7)
(2, 7)
(3, 7)
(4, 7)
(5, 7)
(6, 7)
(7, 7)
load headers exit
Received: (qmail 7690 invoked from network); 5 May 2010 15:29:43 -0000
Received: from unknown (HELO p3pismtp01-026.prod.phx3.secureserver.net)
([10.6.1
```

...часть строк опущена...

```
load mail?y
load 1
Connecting...
b'+OK <29205.1273346957@p3pop01-10.prod.phx3.gdg>'
Received: (qmail 7690 invoked from network); 5 May 2010 15:29:43 -0000
Received: from unknown (HELO p3pismtp01-026.prod.phx3.secureserver.net)
([10.6.1
```

...часть строк опущена...

```
load mail?
loading full messages
Connecting...
b'+OK <31655.1273347055@p3pop01-25.prod.phx3.secureserver.net>'
(3, 7)
(4, 7)
(5, 7)
(6, 7)
(7, 7)
Received: (qmail 25683 invoked from network); 6 May 2010 14:12:07 -0000
Received: from unknown (HELO p3pismtp01-018.prod.phx3.secureserver.net)
([10.6.1
```

...часть строк опущена...

```
Parsed: A B C D E F G
Fiddle de dum, Fiddle de dee,
Eric the half a bee.

Press Enter to exit
```

Обновление клиента командной строки pymail

В качестве последнего примера работы с электронной почтой и для демонстрации более полного сценария использования пакета `mailtools`, представленного в предыдущих разделах, в примере 13.27 приводится обновленная версия программы `pymail`, с которой мы встречались выше

(пример 13.20). Для доступа к электронной почте эта версия программы применяет пакет mailtools вместо непосредственного использования пакета email из стандартной библиотеки Python. Сравните реализацию оригинальной версии pymail с этой версией, чтобы увидеть, как можно использовать mailtools на практике. Здесь вы увидите, насколько упростилась логика загрузки и отправки сообщений.

Пример 13.27. PP4E\Internet\Email\pymail2.py

```
#!/usr/local/bin/python
"""
###############################################################################
pymail2 - простой консольный клиент электронной почты на языке Python; эта
версия использует пакет mailtools, который в свою очередь использует модули
poplib, smtplib и пакет email для анализа и составления электронных писем;
отображает только первую текстовую часть электронных писем, а не весь полный
текст; изначально загружает только заголовки сообщений, используя команду
TOP; полный текст загружается только для писем, выбранных для отображения;
кэширует уже загруженные письма; предупреждение: не предусматривает
возможность обновления оглавления; напрямую использует объекты из пакета
mailtools, однако они точно так же могут использоваться как суперклассы;
###############################################################################
"""

import mailconfig, mailtools
from pymail import inputmessage
mailcache = {}

def fetchmessage(i):
    try:
        fulltext = mailcache[i]
    except KeyError:
        fulltext = fetcher.downloadMessage(i)
        mailcache[i] = fulltext
    return fulltext

def sendmessage():
    From, To, Subj, text = inputmessage()
    sender.sendMessage(From, To, Subj, [], text, attaches=None)

def deletemessages(toDelete, verify=True):
    print('To be deleted:', toDelete)
    if verify and input('Delete?')[:1] not in ['y', 'Y']:
        print('Delete cancelled.')
    else:
        print('Deleting messages from server...')
        fetcher.deleteMessages(toDelete)

def showindex(msgList, msgSizes, chunk=5):
    count = 0
    for (msg, size) in zip(msgList, msgSizes): # email.message.Message, int
        count += 1                              # в 3.x - итератор
```

```
        print('%d:\t%d bytes' % (count, size))
        for hdr in ('From', 'To', 'Date', 'Subject'):
            print('\t%-8s=>%s' % (hdr, msg.get(hdr, '(unknown)')))
        if count % chunk == 0:
            input('[Press Enter key]') # пауза после каждой группы сообщений

def showmessage(i, msgList):
    if 1 <= i <= len(msgList):
        fulltext = fetchmessage(i)
        message  = parser.parseMessage(fulltext)
        ctype, maintext = parser.findMainText(message)
        print('-' * 79)
        print(maintext.rstrip() + '\n') # главная текстовая часть,
                                        # не все письмо
        print('-' * 79)                 # и никаких вложений после
    else:
        print('Bad message number')

def savemessage(i, mailfile, msgList):
    if 1 <= i <= len(msgList):
        fulltext = fetchmessage(i)
        savefile = open(mailfile, 'a', encoding=mailconfig.fetchEncoding) # 4E
        savefile.write('\n' + fulltext + '-'*80 + '\n')
    else:
        print('Bad message number')

def msgnum(command):
    try:
        return int(command.split()[1])
    except:
        return -1                # предполагается, что это ошибка

helptext = """
Available commands:
i    - index display
l n? - list all messages (or just message n)
d n? - mark all messages for deletion (or just message n)
s n? - save all messages to a file (or just message n)
m    - compose and send a new mail message
q    - quit pymail
?    - display this help text
"""

def interact(msgList, msgSizes, mailfile):
    showindex(msgList, msgSizes)
    toDelete = []
    while True:
        try:
            command = input('[Pymail] Action? (i, l, d, s, m, q, ?) ')
        except EOFError:
            command = 'q'
```

```
        if not command: command = '*'

        if command == 'q':         # завершение
            break

        elif command[0] == 'i':    # оглавление
            showindex(msgList, msgSizes)

        elif command[0] == 'l':    # содержимое письма
            if len(command) == 1:
                for i in range(1, len(msgList)+1):
                    showmessage(i, msgList)
            else:
                showmessage(msgnum(command), msgList)

        elif command[0] == 's':    # сохранение
            if len(command) == 1:
                for i in range(1, len(msgList)+1):
                    savemessage(i, mailfile, msgList)
            else:
                savemessage(msgnum(command), mailfile, msgList)

        elif command[0] == 'd':    # пометить для удаления позднее
            if len(command) == 1:  # в 3.x требуется вызвать list(): итератор
                toDelete = list(range(1, len(msgList)+1))
            else:
                delnum = msgnum(command)
                if (1 <= delnum <= len(msgList)) and (delnum not in toDelete):
                    toDelete.append(delnum)
                else:
                    print('Bad message number')

        elif command[0] == 'm':    # отправить новое сообщение через SMTP
            try:
                sendmessage()
            except:
                print('Error - mail not sent')

        elif command[0] == '?':
            print(helptext)
        else:
            print('What? -- type "?" for commands help')
    return toDelete

def main():
    global parser, sender, fetcher
    mailserver = mailconfig.popservername
    mailuser   = mailconfig.popusername
    mailfile   = mailconfig.savemailfile

    parser     = mailtools.MailParser()
```

```
    sender     = mailtools.MailSender()
    fetcher    = mailtools.MailFetcherConsole(mailserver, mailuser)

    def progress(i, max):
        print(i, 'of', max)

    hdrsList, msgSizes, ignore = fetcher.downloadAllHeaders(progress)
    msgList = [parser.parseHeaders(hdrtext) for hdrtext in hdrsList]

    print('[Pymail email client]')
    toDelete = interact(msgList, msgSizes, mailfile)

    if toDelete: deletemessages(toDelete)

if __name__ == '__main__': main()
```

Работа с клиентом командной строки pymail2

Как и оригинал, эта программа используется в интерактивном режиме. Вывод этой версии почти идентичен выводу оригинальной версии, поэтому не будем углубляться в его описание. Ниже приводится листинг сеанса работы со сценарием. Запустите его на своем компьютере, чтобы получить информацию из первых рук:

```
C:\...\PP4E\Internet\Email> pymail2.py
user: PP4E@learning-python.com
loading headers
Connecting...
Password for PP4E@learning-python.com on pop.secureserver.net?
b'+OK <24460.1273347818@pop15.prod.mesa1.secureserver.net>'
1 of 7
2 of 7
3 of 7
4 of 7
5 of 7
6 of 7
7 of 7
load headers exit
[Pymail email client]
1:      1860 bytes
        From    =>lutz@rmi.net
        To      =>pp4e@learning-python.com
        Date    =>Wed, 5 May 2010 11:29:36 -0400 (EDT)
        Subject =>I'm a Lumberjack, and I'm Okay
2:      1408 bytes
        From    =>lutz@learning-python.com
        To      =>PP4E@learning-python.com
        Date    =>Wed, 05 May 2010 08:33:47 -0700
        Subject =>testing
3:      1049 bytes
        From    =>Eric.the.Half.a.Bee@yahoo.com
        To      =>PP4E@learning-python.com
```

```
        Date    =>Thu, 06 May 2010 14:11:07 -0000
        Subject =>A B C D E F G
4:      1038 bytes
        From    =>Eric.the.Half.a.Bee@aol.com
        To      =>nobody.in.particular@marketing.com
        Date    =>Thu, 06 May 2010 14:32:32 -0000
        Subject =>a b c d e f g
5:     957 bytes
        From    =>PP4E@learning-python.com
        To      =>maillist
        Date    =>Thu, 06 May 2010 10:58:40 -0400
        Subject =>test interactive smtplib
[Press Enter key]
6:      1037 bytes
        From    =>Cardinal@hotmail.com
        To      =>PP4E@learning-python.com
        Date    =>Fri, 07 May 2010 20:32:38 -0000
        Subject =>Among our weapons are these
7:      3248 bytes
        From    =>PP4E@learning-python.com
        To      =>PP4E@learning-python.com
        Date    =>Sat, 08 May 2010 19:26:22 -0000
        Subject =>testing mailtools package
[Pymail] Action? (i, l, d, s, m, q, ?) l 7
load 7
Connecting...
b'+OK <20110.1273347827@pop07.prod.mesa1.secureserver.net>'
------------------------------------------------------------------------------
Here is my source code

------------------------------------------------------------------------------
[Pymail] Action? (i, l, d, s, m, q, ?) d 7
[Pymail] Action? (i, l, d, s, m, q, ?) m
From? lutz@rmi.net
To?   PP4E@learning-python.com
Subj? test pymail2 send
Type message text, end with line="."
Run away! Run away!
.
Sending to...['PP4E@learning-python.com']
From: lutz@rmi.net
To: PP4E@learning-python.com
Subject: test pymail2 send
Date: Sat, 08 May 2010 19:44:25 -0000

Run away! Run away!

Send exit
[Pymail] Action? (i, l, d, s, m, q, ?) q
To be deleted: [7]
Delete?y
```

```
Deleting messages from server...
deleting mails
Connecting...
b'+OK <11553.1273347873@pop17.prod.mesa1.secureserver.net>'
```

В почтовом ящике теперь имеются сообщения, отправленные с помощью самых разных клиентов, таких как веб-клиенты интернет-провайдеров, простой сценарий SMTP, интерактивный сеанс Python, сценарий самотестирования из пакета `mailtools` и два клиента командной строки. В последующих главах мы еще больше расширим этот список. Все электронные письма, отправленные с их помощью, выглядят для нашего сценария совершенно одинаковыми. Ниже выполняется получение письма, которое только что было отправлено (вторая попытка получить письмо обнаружит, что оно уже в кэше):

```
C:\...\PP4E\Internet\Email> pymail2.py
user: PP4E@learning-python.com
loading headers
Connecting...
...часть строк опущена...

[Press Enter key]
6:      1037 bytes
        From    =>Cardinal@hotmail.com
        To      =>PP4E@learning-python.com
        Date    =>Fri, 07 May 2010 20:32:38 -0000
        Subject =>Among our weapons are these
7:      984 bytes
        From    =>lutz@rmi.net
        To      =>PP4E@learning-python.com
        Date    =>Sat, 08 May 2010 19:44:25 -0000
        Subject =>test pymail2 send
[Pymail] Action? (i, l, d, s, m, q, ?) l 7
load 7
Connecting...
b'+OK <31456.1273348189@p3pop01-03.prod.phx3.gdg>'
------------------------------------------------------------------------
Run away! Run away!

------------------------------------------------------------------------
[Pymail] Action? (i, l, d, s, m, q, ?) l 7
------------------------------------------------------------------------
Run away! Run away!

------------------------------------------------------------------------
[Pymail] Action? (i, l, d, s, m, q, ?) q
```

Изучите программный код сценария `pymail2`, чтобы глубже проникнуть в его суть. Вы увидите, что в этой версии исчезли некоторые сложности, такие как ручное форматирование текста электронного письма. Кроме того, она иначе выводит текст письма – вместо того чтобы вслепую вы-

водить полный текст сообщения (вместе со всеми вложениями), она с помощью mailtools извлекает и выводит первую текстовую часть сообщения. Использованные здесь сообщения слишком просты, чтобы увидеть разницу, но при чтении писем с вложениями новая версия более избирательна к тому, что следует отобразить.

Кроме того, так как интерфейс к электронной почте инкапсулирован в модули пакета mailtools, при необходимости что-то изменить в нем достаточно будет внести изменения только в конкретный модуль независимо от того, как много клиентов электронной почты используют эти инструменты. А поскольку программный код пакета mailtools используется совместно, то если известно, что он работает в одном клиенте, можно быть уверенными, что он будет работать и в другом – отпадает необходимость отлаживать новый программный код.

С другой стороны, сценарий pymail2 на самом деле не использует всю мощь ни пакета mailtools, ни лежащего в его основе пакета email. Например, сценарий вообще никак не обрабатывает вложения и интернационализированные заголовки, не поддерживает синхронизацию почтового ящика входящих сообщений, и иногда декодированный текст основной части, подготовленный к выводу, может содержать символы, несовместимые с терминалом. Чтобы увидеть возможности пакета email в полном объеме, необходимо исследовать более крупную систему электронной почты, такую как PyMailGUI или PyMailCGI. Первая из них является темой следующей главы, а со второй мы познакомимся в главе 16. Однако перед этим коротко рассмотрим еще несколько дополнительных инструментов для работы с клиентскими протоколами.

NNTP: доступ к телеконференциям

До настоящего момента в этой главе мы рассматривали инструменты Python для работы с FTP и электронной почтой и попутно познакомились с рядом модулей, используемых на стороне клиента: ftplib, poplib, smtplib, mimetypes, urllib и так далее. Этот набор хорошо представляет библиотечные инструменты Python для передачи и обработки информации в Интернете, но он далеко не полон.

Более или менее полный список модулей Python, связанных с Интернетом, приводится в начале предыдущей главы. Среди прочего Python содержит также вспомогательные библиотеки для поддержки на стороне клиента телеконференций Интернета, Telnet, HTTP, XML-RPC и других стандартных протоколов. Большинство из них аналогично модулям, с которыми мы уже встречались, – они предоставляют объектно-ориентированный интерфейс, автоматизирующий операции с сокетами и структурами сообщений.

Например, модуль Python nntplib поддерживает интерфейс клиента к протоколу NNTP (Network News Transfer Protocol – протокол передачи новостей по сети), используемый для чтения и передачи статей в те-

леконференции Usenet в Интернете. Как и другие протоколы, NNTP выполняется поверх сокетов и просто определяет стандартный протокол обмена сообщениями. Как и другие модули, `nntplib` скрывает большую часть деталей протокола и предоставляет сценариям на языке Python объектно-ориентированный интерфейс.

Мы не станем здесь вдаваться в тонкости протокола, но кратко отметим, что серверы NNTP хранят ряд статей, обычно в плоском файле базы данных. Если знать доменное имя или IP-адрес компьютера, на котором выполняется программа NNTP-сервера, прослушивающая порт NNTP, то можно написать сценарий, получающий или передающий статьи с любого компьютера, на котором установлен Python и имеется соединение с Интернетом. Например, сценарий в примере 13.28 по умолчанию получает и отображает последние 10 статей из телеконференции Python *comp.lang.python* с сервера NNTP *news.rmi.net* моего интернет-провайдера.

Пример 13.28. PP4E\Internet\Other\readnews.py

```
"""
получает и выводит сообщения из телеконференции comp.lang.python с помощью
модуля nntplib, который в действительности действует поверх сокетов;
nntplib поддерживает также отправку новых сообщений и так далее;
примечание: после прочтения сообщения не удаляются;
"""


listonly = False
showhdrs = ['From', 'Subject', 'Date', 'Newsgroups', 'Lines']
try:
    import sys
    servername, groupname, showcount = sys.argv[1:]
    showcount  = int(showcount)
except:
    servername = nntpconfig.servername # присвойте этому
                                       # параметру имя сервера
    groupname  = 'comp.lang.python'    # арг. ком. строки или знач.
                                       # по умолчанию
    showcount  = 10                    # показать последние showcount сообщ.

# соединиться с сервером nntp
print('Connecting to', servername, 'for', groupname)
from nntplib import NNTP
connection = NNTP(servername)
(reply, count, first, last, name) = connection.group(groupname)
print('%s has %s articles: %s-%s' % (name, count, first, last))

# запросить только заголовки
fetchfrom = str(int(last) - (showcount-1))
(reply, subjects) = connection.xhdr('subject', (fetchfrom + '-' + last))

# вывести заголовки, получить заголовки+тело
for (id, subj) in subjects:    # [-showcount:] для загрузки всех заголовков
```

```
        print('Article %s [%s]' % (id, subj))
        if not listonly and input('=> Display?') in ['y', 'Y']:
            reply, num, tid, list = connection.head(id)
            for line in list:
                for prefix in showhdrs:
                    if line[:len(prefix)] == prefix:
                        print(line[:80])
                        break
            if input('=> Show body?') in ['y', 'Y']:
                reply, num, tid, list = connection.body(id)
                for line in list:
                    print(line[:80])
        print()
    print(connection.quit())
```

Как и при использовании инструментов FTP и электронной почты, этот сценарий создает объект NNTP и вызывает его методы для получения информации телеконференции, а также заголовков и текста статей. Например, метод xhdr загружает выбранные заголовки из указанного диапазона сообщений.

При использовании NNTP-серверов, требующих аутентификацию, вам может также потребоваться передать конструктору NNTP имя пользователя, пароль и, возможно, флаг режима чтения. Более подробную информацию о других параметрах конструктора NNTP и методах объекта вы найдете в руководстве по библиотеке Python.

В интересах экономии места и времени я не буду здесь приводить вывод этого сценария. При запуске он соединяется с сервером и выводит строку с темой каждой статьи, останавливаясь для запроса, следует ли получить и показать информационные строки заголовка статьи (только заголовки, перечисленные в переменной showhdrs) и текст тела. Этот сценарий можно использовать и другим способом, явно передавая ему в командной строке имя сервера, название телеконференции и количество отображаемых сообщений. Потрудившись еще немного, можно было бы превратить этот сценарий в полноценный интерфейс новостей. Например, можно было бы посылать из сценария Python новые статьи с помощью следующего программного кода (предполагается присутствие в локальном файле надлежащих строк заголовков NNTP):

```
# для отправки выполните следующее
# (но только если действительно хотите отправить сообщение!)
connection = NNTP(servername)
localfile  = open('filename')   # в файле содержатся правильные заголовки
connection.post(localfile)      # послать текст в телеконференцию
connection.quit()
```

Можно также добавить к этому сценарию графический интерфейс на основе tkinter, чтобы облегчить работу с ним, но это расширение мы добавим к списку упражнений для самостоятельного решения (смотрите также в конце следующей главы предлагаемые расширения к интер-

фейсу PyMailGui – электронная почта и телеконференции имеют похожую структуру).

HTTP: доступ к веб-сайтам

Стандартная библиотека Python (то есть модули, устанавливаемые вместе с интерпретатором) содержит также поддержку протокола HTTP (Hypertext Transfer Protocol – гипертекстовый транспортный протокол) на стороне клиента – стандарта структуры сообщений и портов, используемых для передачи информации в World Wide Web. Вкратце, это тот протокол, который использует ваш веб-броузер (например, Internet Explorer, Firefox, Chrome или Safari) для получения веб-страниц и запуска приложений на удаленных серверах при веб-серфинге. По сути, он просто определяет порядок обмена байтами через порт 80.

Чтобы действительно понять, как передаются данные по протоколу HTTP, необходимо знать некоторые темы, относящиеся к выполнению сценариев на стороне сервера, рассматриваемые в главе 15 (например, как вызываются сценарии и какие схемы адресации используются в Интернете), поэтому данный раздел может оказаться менее полезным для читателей, не имеющих соответствующей подготовки. К счастью, основные интерфейсы HTTP в Python достаточно просты для начального их понимания даже на данном этапе, поэтому мы сейчас кратко их рассмотрим.

Стандартный модуль Python `http.client` в значительной мере автоматизирует использование протокола HTTP и позволяет сценариям получать веб-страницы, почти как в веб-броузерах. Как мы увидим в главе 15, модуль `http.server` позволяет также создавать веб-серверы для работы с другой стороны соединения. В частности, сценарий в примере 13.29 может получить любой файл с любого компьютера, на котором выполняется программа веб-сервера HTTP. Как обычно, файл (и строки заголовков описания) в конечном счете передается через стандартный порт сокета, но большая часть сложных деталей скрыта в модуле `http.client` (сравните с нашей реализацией диалога с HTTP-сервером через порт 80 с применением простых сокетов в главе 12).

Пример 13.29. PP4E\Internet\Other\http-getfile.py

```
"""
получает файл с сервера HTTP (web) через сокеты с помощью модуля http.
client; параметр с именем файла может содержать полный путь к каталогу
и быть именем любого сценария CGI с параметрами запроса в конце,
отделяемыми символом ?, для вызова удаленной программы; содержимое
полученного файла или вывод удаленной программы можно сохранить
в локальном файле, имитируя поведение FTP, или анализировать
с помощью модуля str.find или html.parser; смотрите также описание
метода http.client request(method, url, body=None, hdrs={});
"""
```

```
import sys, http.client
showlines = 6
try:
    servername, filename = sys.argv[1:] # аргументы командной строки?
except:
    servername, filename = 'learning-python.com', '/index.html'

print(servername, filename)
server = http.client.HTTPConnection(servername) # соединиться с http-сервером
server.putrequest('GET', filename)      # отправить запрос и заголовки
server.putheader('Accept', 'text/html') # можно также отправить запрос POST
server.endheaders()                     # как и имена файлов сценариев CGI

reply = server.getresponse()            # прочитать заголовки+данные ответа
if reply.status != 200:                 # код 200 означает успех
    print('Error sending request', reply.status, reply.reason)
else:
    data = reply.readlines()            # объект файла для получаемых данных
    reply.close()                       # вывести строки с eoln в конце
    for line in data[:showlines]:       # чтобы сохранить, запишите в файл
        print(line)                     # строки уже содержат \n,
                                        # но являются строками bytes
```

Требуемые имена серверов и файлов можно передать в командной строке, переопределив тем самым значения по умолчанию, определенные в программном коде. Чтобы до конца понять этот сценарий, необходимо иметь представление протокола HTTP, но в общем он довольно прост для расшифровки. При выполнении на стороне клиента этот сценарий создает объект `HTTP`, который соединяется с сервером, посылает запрос GET и допустимые типы ответов, а затем читает ответ сервера. Подобно исходному тесту сообщения электронной почты, ответ сервера http обычно начинается с нескольких строк заголовков описания, за которыми следует содержимое запрошенного файла. Метод `getfile` объекта `HTTP` возвращает объект файла, из которого можно читать загруженные данные.

Давайте получим с помощью этого сценария несколько файлов. Как и другие клиентские сценарии на языке Python, данный сценарий может выполняться на любом компьютере, где установлен Python и имеется соединение с Интернетом (в данном случае он выполняется на клиенте Windows). Если все пройдет хорошо, будут выведены несколько первых строк загруженного файла. В более реалистичном приложении получаемый текст можно было бы сохранить в локальном файле, проанализировать с помощью модуля Python `html.parser` (будет представлен в главе 19) и так далее. При запуске без аргументов сценарий просто загрузит начальную страницу HTML с сайта *http://learning-python.com*, который находится на сервере моего коммерческого провайдера:

```
C:\...\PP4E\Internet\Other> http-getfile.py
learning-python.com /index.html
b'<HTML>\n'
```

```
b' \n'
b'<HEAD>\n'
b"<TITLE>Mark Lutz's Python Training Services</TITLE>\n"
b'<!--mstheme--><link rel="stylesheet" type="text/css" href="_themes/blends/
blen...'
b'</HEAD>\n'
```

Обратите внимание, что в Python 3.X данные поступают клиенту в виде строк типа bytes, а не str. Так как модуль html.parser, выполняющий анализ разметки HTML, с которым мы познакомимся в главе 19, требует текстовые строки str, а не строки bytes, вам наверняка придется решить, какую кодировку Юникода применить, чтобы обеспечить возможность анализа, – практически так же, как мы делали это при обработке почтовых сообщений выше. Как и прежде, для декодирования строки bytes в str можно было бы использовать кодировку по умолчанию, извлекать ее из настроек, предлагать сделать выбор пользователю, определять из заголовков или на основе анализа структуры байтов. Поскольку данные через сокеты всегда передаются в виде простых байтов, мы постоянно будем оказываться перед выбором при передаче через них текстовых данных. Поддержка Юникода подразумевает выполнение дополнительных действий, за исключением случаев, когда тип текста известен заранее или всегда имеет простую форму.

Имя сервера и файла, который должен быть получен, также можно указать в командной строке. В следующем примере показано применение этого сценария для получения файлов с двух разных веб-сайтов, имена которых указываются в командных строках (я обрезал некоторые из строк, чтобы уместить их по ширине страницы). Обратите внимание, что аргумент имени файла может содержать произвольный путь к удаленному каталогу с нужным файлом, как в последней из приведенных попыток:

```
C:\...\PP4E\Internet\Other> http-getfile.py www.python.org /index.html
www.python.org /index.html
b'<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://....'
b'\n'
b'\n'
b'<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">\n'
b'\n'
b'<head>\n'

C:\...\PP4E\Internet\Other> http-getfile.py www.python.org index.html
www.python.org index.html
Error sending request 400 Bad Request

C:\...\PP4E\Internet\Other> http-getfile.py www.rmi.net /~lutz
www.rmi.net /~lutz
Error sending request 301 Moved Permanently
```

```
C:\...\PP4E\Internet\Other> http-getfile.py www.rmi.net /~lutz/index.html
www.rmi.net /~lutz/index.html
b'<HTML>\n'
b'\n'
b'<HEAD>\n'
b"<TITLE>Mark Lutz's Book Support Site</TITLE>\n"
b'</HEAD>\n'
b'<BODY BGCOLOR="#f1f1ff">\n'
```

Обратите внимание на вторую и третью попытки в этом примере: в случае неудачи сценарий получает и выводит код ошибки HTTP, возвращаемый сервером (во второй попытке мы забыли добавить ведущий символ слэша, а в третьей – имя файла «index.html», совершенно необходимые для данного сервера и интерфейса). При использовании низкоуровневых интерфейсов HTTP необходимо точно указывать, что требуется получить.

Технически переменная `filename` в сценарии может ссылаться на простой статический файл веб-страницы или на программу сервера, генерирующую разметку HTML. Такие серверные программы обычно называются сценариями CGI – они являются темой глав 15 и 16. Пока лишь запомните, что если `filename` указывает на сценарий, то данная программа может запустить другую программу, находящуюся на удаленном сервере. В таком случае после символа `?`, вслед за именем программы, можно также указать параметры (называемые строкой запроса), которые должны быть переданы удаленной программе.

В следующем примере мы передаем параметр `language=Python` сценарию CGI, с которым мы познакомимся в главе 15 (для этого примера нам также необходимо сначала запустить локальный веб-сервер на языке Python, используя сценарий, с которым мы впервые встретились в главе 1 и к которому еще вернемся в главе 15):

```
В другом окне
C:\...\PP4E\Internet\Web> webserver.py
webdir ".", port 80

C:\...\PP4E\Internet\Other> http-getfile.py localhost
                                /cgi-bin/languages.py?language=Python
localhost /cgi-bin/languages.py?language=Python
b'<TITLE>Languages</TITLE>\n'
b'<H1>Syntax</H1><HR>\n'
b'<H3>Python</H3><P><PRE>\n'
b" print('Hello World')                \n"
b'</PRE></P><BR>\n'
b'<HR>\n'
```

В этой книге еще много будет говориться о разметке HTML, сценариях CGI и смысле запроса HTTP GET, использовавшегося в примере 13.29 (наряду с методом POST являющимся одним из способов форматирова-

ния информации, посылаемой серверу HTTP), поэтому сейчас мы опустим дополнительные детали.

Достаточно сказать, однако, что с помощью интерфейсов HTTP можно написать собственный веб-броузер и создать сценарии, использующие веб-сайты так, как если бы они были подпрограммами. Путем отправки параметров удаленным программам и анализа возвращаемых результатов можно заставить веб-сайты играть роль простых функций, выполняющихся в том же процессе (хотя и значительно более медленно и косвенно).

Еще раз о пакете urllib

Модуль `http.client`, с которым мы только что познакомились, предоставляет клиентам HTTP механизмы низкого уровня. Однако при работе с объектами, находящимися в Сети, часто оказывается проще организовать загрузку файлов с помощью стандартного модуля Python `urllib.request`, который был представлен в разделе этой главы, посвященном FTP. Так как этот модуль обеспечивает еще одну возможность обмена данными по протоколу HTTP, остановимся здесь на его интерфейсах.

Вспомните, что при наличии адреса URL модуль `urllib.request` либо загружает запрашиваемый объект из Сети в локальный файл, либо создает объект файла, который позволяет осуществлять чтение его содержимого. Благодаря этому сценарий в примере 13.30 выполняет ту же работу, что и сценарий с использованием `http.client`, который мы только что написали, но он при этом значительно короче.

Пример 13.30. PP4E\Internet\Other\http-getfile-urllib1.py

```
"""
получает файл с сервера HTTP (web) через сокеты с помощью модуля urllib;
urllib поддерживает протоколы HTTP, FTP, HTTPS и обычные файлы в строках
адресов URL; для HTTP в строке URL можно указать имя файла или удаленного
сценария CGI; смотрите также пример использования urllib в разделе FTP
и вызов сценария CGI в последующей главе; Python позволяет получать файлы
из сети самыми разными способами, различающимися сложностью и требованиями
к серверам: через сокеты, FTP, HTTP, urllib и вывод CGI;
предостережение: имена файлов следует обрабатывать функцией
urllib.parse.quote, чтобы экранировать специальные символы,
если это не делается в программном коде, - смотрите следующие главы;
"""


import sys
from urllib.request import urlopen
showlines = 6
try:
    servername, filename = sys.argv[1:]      # аргументы командной строки?
except:
    servername, filename = 'learning-python.com', '/index.html'
```

```
remoteaddr = 'http://%s%s' % (servername, filename) # может быть именем
                                                    # CGI-сценария
print(remoteaddr)
remotefile = urlopen(remoteaddr)                # объект файла для ввода
remotedata = remotefile.readlines()             # чтение данных напрямую
remotefile.close()
for line in remotedata[:showlines]: print(line) # строка bytes
                                               # со встроенными символами \n
```

Почти все тонкости пересылки по HTTP скрыты здесь за интерфейсом urllib.request. Данная версия работает примерно так же, как версия с http.client, которую мы написали выше, но должна сконструировать и передать адрес URL (созданный URL представлен в первой строке вывода сценария). Как отмечалось в разделе этой главы об FTP, функция urlopen из модуля urllib.request возвращает объект, похожий на файл, из которого можно читать данные, находящиеся на сервере. Но поскольку созданный здесь адрес URL начинается с «http://», модуль urllib.request автоматически применяет для загрузки запрашиваемого файла низкоуровневые интерфейсы HTTP, а не FTP:

```
C:\...\PP4E\Internet\Other> http-getfile-urllib1.py
http://learning-python.com/index.html
b'<HTML>\n'
b' \n'
b'<HEAD>\n'
b"<TITLE>Mark Lutz's Python Training Services</TITLE>\n"
b'<!--mstheme--><link rel="stylesheet" type="text/css" href="_themes/blends/
blen...'
b'</HEAD>\n'

C:\...\PP4E\Internet\Other> http-getfile-urllib1.py www.python.org /index
http://www.python.org/index
b'<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://
www.w3....'
b'\n'
b'\n'
b'<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">\n'
b'\n'
b'<head>\n'

C:\...\PP4E\Internet\Other> http-getfile-urllib1.py www.rmi.net /~lutz
http://www.rmi.net/~lutz
b'<HTML>\n'
b'\n'
b'<HEAD>\n'
b"<TITLE>Mark Lutz's Book Support Site</TITLE>\n"
b'</HEAD>\n'
b'<BODY BGCOLOR="#f1f1ff">\n'
```

```
C:\...\PP4E\Internet\Other> http-getfile-urllib1.py
                              localhost /cgi-bin/languages.py?language=Java
http://localhost/cgi-bin/languages.py?language=Java
b'<TITLE>Languages</TITLE>\n'
b'<H1>Syntax</H1><HR>\n'
b'<H3>Java</H3><P><PRE>\n'
b' System.out.println("Hello World"); \n'
b'</PRE></P><BR>\n'
b'<HR>\n'
```

Как и раньше, аргумент с именем файла может содержать не только имя простого файла, но и имя вызываемой программы с дополнительными параметрами. Если внимательно изучить этот вывод, можно заметить, что сценарий работает, даже если опустить «index.html» в конце пути к каталогу на сайте (в третьей команде). В отличие от версии, реализованной на основе использования протокола чистого HTTP, интерфейс на основе URL достаточно сообразителен, чтобы сделать именно то, что требуется.

Другие интерфейсы urllib

Еще одна версия: в следующем сценарии загрузки на базе модуля `urllib.request` используется более высокоуровневый интерфейс `urlretrieve` из этого модуля, автоматически сохраняющий загруженный файл или вывод сценария в локальном файле на компьютере клиента. Этот интерфейс удобно использовать, когда действительно требуется сохранить полученные данные (например, чтобы сымитировать протокол FTP). Однако, если планируется немедленная обработка загружаемых данных, такая форма загрузки может оказаться менее удобной, чем только что рассмотренная версия: потребуется открыть и прочитать содержимое сохраненного файла. Кроме того, потребуется реализовать дополнительный протокол определения или получения имен локальных файлов, как в примере 13.31.

Пример 13.31. PP4E\Internet\Other\http-getfile-urllib2.py

```
"""
получает файл с сервера HTTP (web) через сокеты с помощью urlllib;
в этой версии используется интерфейс, сохраняющий полученные данные
в локальном файле в двоичном режиме; имя локального файла передается
в аргументе командной строки или выделяется из URL посредством
модуля urllib.parse: аргумент с именем файла может содержать путь
к каталогу в начале и параметры запроса в конце, поэтому функции os.path.
split будет недостаточно (отделяет только путь к каталогу);
предостережение: имя файла следует обрабатывать функцией urllib.parse.quote,
если заранее не известно, что оно не содержит недопустимых символов -
смотрите следующие главы;
"""
```

```
import sys, os, urllib.request, urllib.parse
showlines = 6
try:
    servername, filename = sys.argv[1:3]  # первые 2 арг. командной строки?
except:
    servername, filename = 'learning-python.com', '/index.html'

remoteaddr = 'http://%s%s' % (servername, filename) # любой адрес в Сети
if len(sys.argv) == 4:                              # получить имя файла
    localname = sys.argv[3]
else:
    (scheme,server,path,parms,query,frag) = urllib.parse.
urlparse(remoteaddr)
    localname = os.path.split(path)[1]

print(remoteaddr, localname)
urllib.request.urlretrieve(remoteaddr, localname) # файл или сценарий
remotedata = open(localname, 'rb').readlines()    # сохранит в локальном файле
for line in remotedata[:showlines]: print(line)  # файл - двоичный
```

Запустим этот последний вариант из командной строки. Он действует точно так же, как последние две версии: подобно предыдущей версии он конструирует URL, и, как в обеих последних версиях, можно явно указать целевой сервер и путь к файлу в командной строке:

```
C:\...\PP4E\Internet\Other> http-getfile-urllib2.py
http://learning-python.com/index.html index.html
b'<HTML>\n'
b' \n'
b'<HEAD>\n'
b"<TITLE>Mark Lutz's Python Training Services</TITLE>\n"
b'<!--mstheme--><link rel="stylesheet" type="text/css" href="_themes/blends/
blen...'
b'</HEAD>\n'

C:\...\PP4E\Internet\Other> http-getfile-urllib2.py www.python.org /index.html
http://www.python.org/index.html index.html
b'<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://
www.w3....'
b'\n'
b'\n'
b'<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">\n'
b'\n'
b'<head>\n'
```

Из-за использования в этой версии интерфейса `urllib.request`, автоматически сохраняющего загруженные данные в локальном файле, она по духу ближе к загрузке файлов по протоколу FTP. Но этот сценарий должен также каким-то образом получить имя локального файла, в который будут записаны данные. Можно позволить сценарию вырезать

из созданного URL базовое имя файла и использовать его, либо передавать имя локального файла в последнем аргументе командной строки. В предшествующем примере запуска загруженная веб-страница сохраняется в локальный файл *index.html* – базовое имя файла, выделенное из URL (сценарий выводит URL и имя локального файла в первой строке вывода). В следующем примере запуска имя локального файла передается явным образом как *py-index.html*:

```
C:\...\PP4E\Internet\Other> http-getfile-urllib2.py
                                www.python.org /index.html py-index.html
http://www.python.org/index.html py-index.html
b'<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
http://www.w3....'
b'\n'
b'\n'
b'<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">\n'
b'\n'
b'<head>\n'

C:\...\PP4E\Internet\Other> http-getfile-urllib2.py
                                www.rmi.net /~lutz books.html
http://www.rmi.net/~lutz books.html
b'<HTML>\n'
b'\n'
b'<HEAD>\n'
b"<TITLE>Mark Lutz's Book Support Site</TITLE>\n"
b'</HEAD>\n'
b'<BODY BGCOLOR="#f1f1ff">\n'

C:\...\PP4E\Internet\Other> http-getfile-urllib2.py
                                www.rmi.net /~lutz/about-pp.html
http://www.rmi.net/~lutz/about-pp.html about-pp.html
b'<HTML>\n'
b'\n'
b'<HEAD>\n'
b'<TITLE>About "Programming Python"</TITLE>\n'
b'</HEAD>\n'
b'\n'
```

Вызов программ и экранирование текста

Следующий листинг демонстрирует, как с помощью этого сценария запускается удаленная программа. Как и прежде, если не указать имя локального файла явным образом, сценарий выделит базовое имя файла из аргумента с именем удаленного файла. Это не всегда просто или уместно при запуске программ – имя файла может содержать путь к удаленному каталогу в начале и параметры, необходимые для запуска удаленной программы, в конце.

Если при запуске сценарию был передан адрес URL и не было явно указано имя выходного файла, этот сценарий извлечет содержащееся в се-

редине базовое имя файла, применяя сначала стандартный модуль urllib.parse для получения пути к файлу, а затем функцию os.path.split для отделения пути к каталогу. Однако в результате получается имя удаленного сценария, которое может оказаться непригодным для локального сохранения данных. Так, в первом примере запуска сценария, приведенном ниже, его вывод попадает в локальный файл с именем *languages.py*, полученным как имя сценария в середине URL; во втором примере запуска имя файла указано явно, как *CxxSyntax.html*, что подавляет извлечение имени файла из URL:

```
C:\...\PP4E\Internet\Other> python http-getfile-urllib2.py localhost
                            /cgi-bin/languages.py?language=Scheme
http://localhost/cgi-bin/languages.py?language=Scheme languages.py
b'<TITLE>Languages</TITLE>\n'
b'<H1>Syntax</H1><HR>\n'
b'<H3>Scheme</H3><P><PRE>\n'
b' (display "Hello World") (newline)  \n'
b'</PRE></P><BR>\n'
b'<HR>\n'

C:\...\PP4E\Internet\Other> python http-getfile-urllib2.py localhost
                           /cgi-bin/languages.py?language=C++ CxxSyntax.html
http://localhost/cgi-bin/languages.py?language=C++ CxxSyntax.html
b'<TITLE>Languages</TITLE>\n'
b'<H1>Syntax</H1><HR>\n'
b'<H3>C </H3><P><PRE>\n'
b"Sorry--I don't know that language\n"
b'</PRE></P><BR>\n'
b'<HR>\n'
```

Здесь удаленный сценарий возвращает сообщение о неудаче поиска, когда в последней команде ему передается строка «C++». Дело в том, что символ «+»в строках URL имеет специальное значение (обозначает пробел), и для надежности оба написанные нами сценария urllib должны были бы пропустить строку filename через такую штуку, как urllib.parse.quote – средство экранирования специальных символов для передачи. Подробно мы будем говорить об этом в главе 15, поэтому считайте это лишь предварительным рассмотрением. Но чтобы заставить работать эту программу, в создаваемой строке URL необходимо использовать специальные последовательности. Ниже показано, как это сделать вручную:

```
C:\...\PP4E\Internet\Other> python http-getfile-urllib2.py localhost
                        /cgi-bin/languages.py?language=C%2b%2b CxxSyntax.html
http://localhost/cgi-bin/languages.py?language=C%2b%2b CxxSyntax.html
b'<TITLE>Languages</TITLE>\n'
b'<H1>Syntax</H1><HR>\n'
b'<H3>C++</H3><P><PRE>\n'
b' cout &lt;&lt; "Hello World" &lt;&lt; endl;   \n'
b'</PRE></P><BR>\n'
b'<HR>\n'
```

Кажущиеся странными строки %2b в этой команде вполне объяснимы: как выглядит результат экранирования, требуемого для URL, можно увидеть, запустив вручную стандартные инструменты Python; это то, что данные сценарии должны делать автоматически, чтобы корректно обрабатывать все возможные случаи. Экранирование можно при необходимости отменить с помощью функции urllib.parse.unquote:

```
C:\...\PP4E\Internet\Other> python
>>> import urllib.parse
>>> urllib.parse.quote('C++')
'c%2B%2B'
```

Опять же не усердствуйте, пытаясь понять несколько последних команд, – мы вернемся к адресам URL и экранированию специальных символов в них в главе 15, когда будем изучать сценарии Python, выполняемые на сервере. Там я также объясню, почему результат для C++ был возвращен со странными символами << – экранированными последовательностями HTML для <<, сгенерированными вызовом функции cgi.escape в серверном сценарии, который произвел ответ. Обратное преобразование обычно выполняется с помощью инструментов анализа разметки HTML, включая модуль html.parser в библиотеке Python, с которым мы встретимся в главе 19:

```
>>> import cgi
>>> cgi.escape('<<')
'&lt;&lt;'
```

Кроме того, в главе 15 мы встретимся с поддержкой *прокси-серверов* и *cookie* в пакете urllib на стороне клиента. В главе 16 мы обсудим родственные концепции HTTPS – передачу данных по протоколу HTTP через защищенные сокеты, поддерживаемую модулем urllib.request на стороне клиента, если Python был скомпилирован с поддержкой SSL. А теперь пришло время завершить наш обзор Всемирной паутины и Интернета в целом со стороны клиента.

Прочие возможности создания клиентских сценариев

В данной главе мы сосредоточились на интерфейсах клиентской стороны к стандартным протоколам, действующим через сокеты, но, как уже отмечалось в одной из сносок выше, программирование для стороны клиента может также принимать другие формы. Многие из них мы отмечали в начале главы 12 – протоколы веб-служб (включая SOAP и XML-RPC); инструменты полнофункциональных интернет-приложений (Rich Internet Application) (включая Flex, Silverlight и pyjamas); фреймворки интеграции различных языков (включая Java и .NET) и многие другие.

Как уже упоминалось, большинство этих систем служат для расширения возможностей веб-броузеров и потому в конечном итоге действуют поверх протокола HTTP, который мы исследовали в этой главе. Например:

- Система *Jython* – компилятор, который поддерживает написанные на Python апплеты Java, представляющие собой программы общего назначения, загружаемые с сервера и выполняемые локально на стороне клиента при обращении к ним по URL, которые расширяют возможности веб-броузеров и взаимодействий.
- Аналогично, *полнофункциональные Интернет-приложения* предоставляют поддержку технологии взаимодействий AJAX и комплекты виджетов, позволяющие реализовать на языке JavaScript взаимодействие с пользователем внутри веб-броузеров, обеспечивая более высокую динамичность и богатство возможностей, чем способна обеспечить разметка HTML.
- В главе 19 мы также познакомимся с инструментами языка Python для обработки XML – структурированного текста, который используется в качестве носителя данных в диалогах клиент/сервер в протоколах *веб-служб*, таких как XML-RPC, позволяющего передавать объекты в формате XML по протоколу HTTP и поддерживаемого пакетом `xmlrpc` в стандартной библиотеке Python. Такие протоколы позволяют упростить интерфейс к веб-серверам на стороне клиента.

Однако, принимая во внимание отсутствие времени и места, мы не станем углубляться в детали этих и других инструментов, используемых на стороне клиента. Если вас интересует применение Python в клиентских сценариях, следует потратить немного времени и ознакомиться с перечнем инструментов Интернета, имеющимся в справочном руководстве по библиотеке Python. Все они действуют на сходных принципах, но имеют несколько различающиеся интерфейсы.

В главе 15 мы перемахнем по ту сторону барьера в мире Интернета и рассмотрим сценарии, выполняющиеся на серверах. Эти программы позволят начать лучше понимать приложения, целиком живущие в Веб и запускаемые веб-броузерами. Осуществляя этот скачок в структуре, нужно помнить, что инструментов, с которыми мы ознакомились в этой и предыдущей главах, часто достаточно для полной реализации распределенной обработки, требуемой во многих приложениях, и работать они могут в согласии со сценариями, выполняемыми на сервере. Однако для полного понимания картины мира Веб необходимо также исследовать и царство серверов.

Но до того как мы туда попадем, в следующей главе будут объединены понятия, рассматривавшиеся ранее, и представлена законченная клиентская программа – полнофункциональный клиент электронной почты с графическим интерфейсом, использующий многие инструменты, которые мы изучили и реализовали. Фактически большая часть из то-

го, что было создано в этой главе, задумывалось как фундамент, который должен лечь в основу более реалистичного и масштабного примера PyMailGUI, демонстрируемого в следующей главе. На самом деле, многое из того, что до сих пор приводилось в этой книге, было предназначено, чтобы выработать навыки, необходимые для решения этой задачи: как мы увидим далее, PyMailGUI объединяет в себе системные инструменты, графические интерфейсы и приемы работы с протоколами Интернета на стороне клиента, образуя полезную систему, которая делает настоящую работу. В качестве дополнительного вознаграждения этот пример поможет нам понять взаимоотношения между клиентскими решениями, с которыми мы встретились здесь, и серверными решениями, которые мы будем рассматривать в следующей части книги.

14

Почтовый клиент PyMailGUI

«Пользуйся исходными текстами, Люк!»

В предыдущей главе был представлен комплект инструментов на языке Python для работы с протоколами Интернета на стороне клиента – модули в стандартной библиотеке для работы с электронной почтой, FTP, телеконференциями, HTTP и многими другими в сценариях Python. Эта глава продолжает повествование с того места, где оно было закончено в предыдущей главе, и представляет законченный пример клиентского приложения PyMailGUI – программы на языке Python, которая отправляет, принимает, составляет и анализирует электронные письма.

Хотя главная задача этой главы – создание действующей программы для работы с электронной почтой, тем не менее, дополнительно она затрагивает несколько концептуальных областей, о которых следует упомянуть, прежде чем двинуться дальше:

Разработка клиентских сценариев

Программа PyMailGUI является полнофункциональным настольным приложением с графическим интерфейсом, которое выполняется на компьютере и взаимодействует с почтовым сервером. Кроме того, это сетевая клиентская программа, иллюстрирующая некоторые темы предыдущих глав, которая может использоваться для сопоставления с серверными решениями, представленными в следующей главе.

Повторное использование программного кода

Кроме того, программа PyMailGUI объединяет ряд вспомогательных модулей, которые мы написали к настоящему моменту, и демонст-

рирует мощь концепции повторного использования программного кода – она использует модуль `thread`, чтобы обеспечить одновременное выполнение нескольких операций передачи почты; комплект модулей для работы с электронной почтой для обработки содержимого сообщений и передачи их по сети; модуль протокола окон для работы с ярлыками; компонент текстового редактора и так далее. Кроме того, она наследует мощь инструментов стандартной библиотеки Python, таких как пакет `email`; реализация конструирования и анализа сообщений, например, здесь становится практически тривиальной.

Программирование в целом

И, наконец, эта глава служит иллюстрацией разработки действующего, крупномасштабного программного обеспечения. Поскольку PyMailGUI является относительно крупной и законченной программой, она может служить примером использования некоторых приемов организации программного кода, эффективность которых становится более очевидной, как только мы покидаем сферу маленьких и искусственных сценариев. Например, объектно-ориентированный стиль программирования и модульный подход проявляются здесь с самой лучшей стороны, позволяя разделить систему на небольшие автономные модули.

Однако в конечном счете программа PyMailGUI служит иллюстрацией того, чего можно достичь, объединив графические интерфейсы, сети и Python. Подобно всем программам на языке Python эта система доступна для дальнейшего усовершенствования – как только вы познакомитесь с ее общей структурой, вы легко сможете заставить ее действовать, как вам угодно, изменив ее исходный программный код. И подобно всем программам на языке Python она является переносимой – вы сможете пользоваться ею в любой системе, где установлен Python и имеется соединение с сетью, при этом вам не придется изменять ее реализацию. Все эти преимущества вы получаете автоматически, когда ваше программное обеспечение разрабатывается с открытыми исходными текстами на переносимом и удобочитаемом языке, таком как Python.

Модули с исходными текстами и их объем

Эта глава представляет собой своего рода упражнение для самостоятельного изучения. Поскольку программа PyMailGUI является достаточно крупной и в значительной мере воплощает понятия, которые мы уже изучили, мы не будем углубляться в детали ее реализации. Программный код, который приводится здесь, предназначен для самостоятельного чтения. Я предлагаю вам ознакомиться с исходными текстами и с комментариями в них и запустить эту программу у себя, чтобы

получить более полное представление о том, как она действует. В состав примеров были включены файлы с сохраненными почтовыми сообщениями, поэтому вы сможете поэкспериментировать с программой, даже не имея подключения к сети.

В процессе изучения и опробования программы для полного понимания функционирования системы вам также может потребоваться вернуться к модулям, представленным выше в книге и повторно используемым здесь. Для справки ниже перечислены основные примеры, которые вновь задействованы в этой главе:

Пример 13.21: PP4E.Internet.Email.mailtools (пакет)

Прием с сервера, передача на сервер, анализ и конструирование почтовых сообщений (глава о разработке сценариев на стороне клиента)

Пример 10.20: PP4E.Gui.Tools.threadtools.py

Управление очередями обработчиков обратного вызова в потоках выполнения для графических интерфейсов (глава об инструментах графических интерфейсов)

Пример 10.16: PP4E.Gui.Tools.windows.py

Настройка рамки окна верхнего уровня (глава об инструментах графических интерфейсов)

Пример 11.4: PP4E.Gui.TextEditor.textEditor.py

Виджет текстового редактора, используемый здесь для просмотра почты, и некоторые диалоги (глава с примерами реализации графических интерфейсов)

Некоторые из этих модулей в свою очередь используют другие примеры, представленные ранее, которые не импортируются программой PyMailGUI непосредственно (например, для создания окон и панели инструментов модуль `textEditor` использует модуль `guimaker`). Естественно, здесь также будут создаваться новые модули. Перечисленные ниже новые модули могут оказаться полезными в других программах:

`popuputil.py`

Различные всплывающие окна общего назначения

`messagecache.py`

Механизм кэширования, который следит за уже загруженными письмами

`wraplines.py`

Утилита для переноса слишком длинных строк в сообщениях

`mailconfig.py`

Пользовательские параметры настройки – имена серверов, шрифты и так далее (расширенная версия)

`html2text.py`

Простейшее средство преобразования содержимого электронных писем в формате HTML в простой текст

Наконец, ниже перечислены новые основные модули, которые будут созданы в этой главе и которые предназначены специально для программы PyMailGUI. Всего программа PyMailGUI состоит из десяти модулей, которые перечислены в этом и в предыдущих списках, наряду с несколькими менее значимыми файлами с исходными текстами, которые мы увидим в этой главе:

`SharedNames.py`

Глобальные переменные и функции, используемые разными модулями

`ViewWindows.py`

Реализация окон просмотра сообщений для операций: View (Просмотреть), Write (Написать), Reply (Ответить) и Forward (Переслать)

`ListWindows.py`

Реализация окон просмотра списка писем на сервере и в локальном файле

`PyMailGUIHelp.py`

Текст справки для пользователя, который открывается щелчком на кнопке в главном окне

`PyMailGUI.py`

Главный файл программы, который открывает главное окно

Объем программного кода

Так как PyMailGUI – пример реальной системы, то информативной станет также оценка ее размера. PyMailGUI состоит из 18 новых файлов: 10 новых модулей Python, перечисленных в двух списках выше, файла справки в формате HTML, маленького файла с настройками для PyEdit, файла инициализации, неиспользуемого в настоящее время инициализационного файла пакета и 5 коротких файлов в Python в подкаталоге, используемых для настройки альтернативных учетных записей.

Всего в 16 файлах Python содержится примерно **2400** новых строк программного кода (включая комментарии и пустые строки) плюс примерно 1700 строк справочного текста в одном файле Python и в одном файле HTML (в двух разновидностях). В число этих 4100 новых строк не входят четыре примера из книги, перечисленные в предыдущем разделе, которые повторно используются в PyMailGUI. Сами повторно используемые примеры содержат 2600 дополнительных строк программного кода на языке Python, из них примерно по 1000 строк приходится на PyEdit и `mailtools`. Таким образом, общий объем составляет **6700** строк: 4100 новых + 2600 повторно используемых. **5000** строк из общего коли-

чества приходятся на файлы с программным кодом (2400 из которых являются новыми) и 1700 – на текст справки.[1]

Количество строк я определил с помощью информационного диалога в PyEdit, а файлы открывал кнопкой code в окне PyDemos, расположенной рядом с кнопкой запуска этой программы (аналогичное действие выполняет кнопка Source в окне текстовой справки в программе PyMailGUI). Количество строк в каждом файле можно увидеть в файле электронной таблицы Excel *linecounts.xls* в подкаталоге *media*, в каталоге с программой PyMailGUI. Кроме того, этот файл использовался как отправляемое и принимаемое вложение, и потому его можно увидеть где-то в конце файла с электронными письмами *SavedEmail\version30-4E*, если открыть его в графическом интерфейсе (как открывать файлы с сохраненной в них почтой, мы узнаем чуть ниже).

Сравнение объема программного кода с предыдущими версиями также приведено в разделе, в котором описываются внесенные в программу изменения. Для подсчета количества строк в файлах можно также использовать сценарий-счетчик SLOC из главы 6, который избавляет от некоторой доли ручного труда, но не включает все взаимосвязанные файлы в один прогон и не различает программный код и текст справки.

Организация программного кода

Все эти подсчеты свидетельствуют о том, что это самый крупный пример в книге, но вас не должен пугать его размер. Благодаря использованию модульного подхода и приемов ООП программный код гораздо проще, чем можно было бы подумать:

- Модули Python позволяют разделить систему на файлы по целевому назначению с минимальными взаимозависимостями между ними – программный код проще отыскать и понять, если модули имеют логическую и автономную организацию.
- Поддержка ООП в языке Python позволяет выделить программный код для повторного использования и избежать избыточности – как можно будет убедиться, программный код специализируется, не повторяется; и классы, которые мы запрограммируем, отражают фак-

[1] И помните: эквивалентная программа на другом языке программирования, таком как C или C++, будет иметь объем в четыре или более раз больше. Если вы уже достаточно давно занимаетесь программированием, то признаете, что сам факт реализации достаточно полнофункциональной почтовой программы в 5000 строках программного кода говорит о возможностях языка Python и его библиотек красноречивее всяких слов. Для сравнения, оригинальная версия 1.0 этой программы во втором издании книги содержала всего 745 строк программного кода в 3 новых модулях, однако она была весьма ограничена в возможностях – она не поддерживала вложения версии PyMailGUI 2.X, многопоточную модель выполнения, локальные файлы почты и так далее и не имела поддержки интернационализации и всех остальных особенностей версии PyMailGUI 3.X из этого издания.

тические компоненты графического интерфейса, что делает их более простыми для понимания.

Например, реализация окон со списком сообщений легко читается и в нее легко вносить изменения, потому что она была оформлена в виде общего суперкласса, который специализируется подклассами, осуществляющими операции с почтовым сервером и с файлом, куда была сохранена почта. Поскольку они в значительной степени являются просто вариациями на одну тему, значительная часть программного кода оказывается сосредоточенной в одном месте. Аналогично программный код, реализующий окно просмотра сообщения, оформлен в виде суперкласса, разделяемого подклассами окон создания нового сообщения, ответа и пересылки, – подклассы просто приспосабливают его для создания сообщения вместо простого просмотра.

Хотя здесь эти приемы рассматриваются в контексте программы обработки электронной почты, тем не менее, они могут найти применение в любых нетривиальных программах на языке Python.

Чтобы помочь вам подступиться к этой программе, в конце этой главы приводится модуль `PyMailGUIHelp.py`, включающий текст справки, описывающей правила пользования этой программой, а также ее основные возможности. Вы можете просмотреть эту справку в текстовом виде и в формате HTML, запустив программу. Экспериментируйте с системой в процессе знакомства с ее программным кодом – это, пожалуй, самый быстрый и эффективный способ раскрыть все ее секреты.

Зачем нужен PyMailGUI?

Прежде чем мы начнем углубляться в программный код этой относительно крупной системы, несколько предварительных замечаний. PyMailGUI – это программа на языке Python с графическим интерфейсом на базе стандартной библиотеки tkinter, реализующая обработку электронной почты на стороне клиента. Она одновременно является образцом сценария на языке Python для Интернета и масштабным действующим примером, соединяющим в себе другие, уже знакомые нам, инструменты, такие как потоки выполнения и графические интерфейсы tkinter.

Подобно программе `pymail` командной строки, которую мы написали в главе 13, PyMailGUI целиком действует на локальном компьютере. Электронная почта извлекается с удаленного почтового сервера и отправляется на сервер через сокеты, но сама программа и ее пользовательский интерфейс выполняются локально. По этой причине PyMailGUI называется почтовым клиентом: подобно `pymail` она выполняется на локальном компьютере и взаимодействует с удаленными почтовыми серверами, используя клиентские инструменты Python. Однако, в отличие от `pymail`, программа PyMailGUI имеет полноценный пользовательский интерфейс: операции с электронными письмами выполняются с помощью мыши, предусматриваются дополнительные действия,

такие как вложение и сохранение файлов, и обеспечивается поддержка интернационализации.

Как и многие примеры, представленные в этой книге, программа PyMailGUI является также действующей, полезной программой. Я запускаю ее на самых разных компьютерах, чтобы проверить свою электронную почту, путешествуя по свету в процессе преподавания Python. Вряд ли PyMailGUI затмит в ближайшее время Microsoft Outlook, но она имеет две ключевые особенности, упоминавшиеся выше, не имеющие никакого отношения к электронной почте – переносимость и доступность для усовершенствования, которые сами по себе являются весьма привлекательными и заслуживают нескольких дополнительных слов:

Переносимость

Программа PyMailGUI может выполняться на любом компьютере, где имеются сокеты и установлен Python с библиотекой tkinter. Поскольку электронная почта пересылается с помощью библиотек Python, подойдет любое соединение с Интернетом, поддерживающее почтовый протокол (Post Office Protocol, POP) и простой протокол передачи почты (Simple Mail Transfer Protocol, SMTP). Более того, поскольку интерфейс пользователя реализован на основе библиотеки tkinter, программа PyMailGUI в неизменном виде должна действовать в Windows, X Window System (Unix, Linux) и Macintosh (классическая версия и OS X) при условии, что в этих системах также способен работать Python 3.X.

Пакет Microsoft Outlook более богат функциями, но он выполняется только в Windows, а точнее, на одной и том же компьютере с Windows. Поскольку он обычно удаляет загруженную электронную почту с сервера и сохраняет ее на компьютере клиента, вы не сможете использовать Outlook на нескольких компьютерах, не пересылая почту на все эти компьютеры. Напротив, PyMailGUI сохраняет и удаляет почту только по требованию, а потому несколько более дружелюбен по отношению к тем, кто проверяет свою почту по случаю, на произвольных компьютерах (подобно мне).

Доступность для усовершенствования

Программу PyMailGUI можно превратить во что угодно, поскольку она полностью доступна для изменений. Собственно, это по-настоящему непревзойденная характеристика PyMailGUI и любого открытого программного обеспечения типа Python в целом – так как исходные тексты PyMailGUI полностью доступны, вы способны полностью управлять дальнейшим ее развитием. С закрытыми коммерческими продуктами, такими как Outlook, такой степени управления у вас и близко быть не может – обычно вы получаете то, что солидная компания-разработчик отнесла к вашим потребностям, вместе с теми ошибками, которые могли оказаться в программе по вине разработчика.

Будучи сценарием Python, PyMailGUI является значительно более гибким инструментом. Например, я могу быстро изменить структуру интерфейса, отключить функции или добавить новые, внеся изменения в исходный программный код Python. Не нравится, как выводится список почтовых сообщений? Поменяйте несколько строк и настройте его, как вам хочется. Хотите автоматически сохранять и удалять почту при загрузке? Добавьте еще несколько строк и кнопки. Надоело видеть рекламную почту? Добавьте несколько строк обработки текста, чтобы загрузить функцию, отфильтровывающую спам. Это лишь некоторые примеры. Смысл в том, что, поскольку PyMailGUI написан на языке сценариев высокого уровня, который легко сопровождать, такие настройки осуществляются относительно просто и могут даже доставить массу удовольствия.

В конечном счете, именно благодаря этим особенностям я действительно *использую* эту программу на языке Python – и как основной инструмент электронной почты, и как резервный вариант, когда почтовая система с веб-интерфейсом моего интернет-провайдера оказывается недоступной (что, как я уже говорил в предыдущей главе, обычно случается в самый неподходящий момент).[1] Умение программировать на языке Python – навык, достойный, чтобы приобрести его.

Запуск PyMailGUI

Конечно, чтобы управлять работой программы PyMailGUI, необходимо иметь возможность запускать ее. Для PyMailGUI требуется лишь компьютер с каким-либо видом подключения к Интернету (достаточно ПК с широкополосным или коммутируемым доступом к Интернету) и установленным на нем Python с расширением tkinter. Версия Python для Windows содержит все необходимое, поэтому пользователи Windows имеют возможность сразу запустить эту программу, щелкнув на ее значке.

Два примечания, касающиеся запуска системы: во-первых, необходимо изменить файл *mailconfig.py* в каталоге с исходными текстами программы, чтобы отразить в нем параметры вашей учетной записи, если вы собираетесь отправлять или получать почту с действующего сервера; подробнее об этом – по ходу диалога с системой.

Во-вторых, вы все еще можете экспериментировать с системой без подключения к Интернету – для быстрого просмотра списка сообщений щелкните на кнопке Open в главном окне и откройте файлы с сохранен-

[1] Как назло, почтовая система с веб-интерфейсом моего интернет-провайдера оказалась недоступной именно в тот день, когда я должен был отослать третье издание этой книги моему издателю! Но у меня не возникло никаких проблем – я запустил программу PyMailGUI и с ее помощью отправил книгу в виде вложения через другой сервер. В некотором смысле эта книга отправила себя сама.

ными почтовыми сообщениями, которые включены в комплект программы и находятся в подкаталоге *SavedMail*. Сценарий запуска примеров PyDemos, находящийся в корневом каталоге с примерами, например, запускает программу PyMailGUI и заставляет ее открыть файлы, передавая имена файлов в командной строке. Хотя у вас со временем появится желание подключиться к своему почтовому серверу, тем не менее, для исследования особенностей системы возможности просмотра сохраненных почтовых сообщений без подключения к Интернету будет вполне достаточно, и это не потребует внесения изменений в файл с настройками.

Стратегия представления

Программа PyMailGUI является самой большой программой в этой книге, но в ней используется не так много новых библиотечных интерфейсов сверх тех, что уже были показаны. Например:

- Интерфейс PyMailGUI конструируется на основе расширения tkinter с использованием уже знакомых окон списков, кнопок и текстовых виджетов.
- Для извлечения заголовков сообщений, текста, вложений и для конструирования новых сообщений применяется пакет `email`.
- Для получения, отправки и удаления почтовых сообщений через сокеты используются библиотечные модули Python поддержки протоколов POP и SMTP.
- Если в вашем интерпретаторе Python имеется поддержка многопоточной модели выполнения, программа использует ее с целью избежать блокировки при длительных операциях с почтой.

Для просмотра и составления сообщений, а также для вывода необработанного текста вложений и исходного программного кода будет повторно использоваться объект `TextEditor`, который мы написали в главе 11. Для загрузки и удаления почты с сервера будут использоваться инструменты из пакета `mailtools`, написанного в главе 13. А для поддержки стратегии использования пользовательских параметров электронной почты, представленной в главе 13, будет применяться модуль `mailconfig`. Программа PyMailGUI в значительной мере представляет собой образец комбинирования уже имеющихся инструментов.

С другой стороны, ввиду большой длины программы мы не станем исчерпывающим образом документировать ее реализацию. Вместо этого мы начнем с описания работы PyMailGUI с точки зрения конечного пользователя. После этого мы перечислим новые модули системы без каких-либо комментариев, чтобы в дальнейшем заняться их изучением.

Как и для большинства длинных примеров этой книги, в данном разделе предполагается, что читатель достаточно хорошо знает Python, чтобы самостоятельно разобраться в программном коде. Если вы читали книгу последовательно, то должны также быть достаточно знакомы с tkinter,

потоками выполнения и почтовыми интерфейсами, чтобы понять библиотечные инструменты, которые здесь применяются. Если возникнут сложности, может потребоваться освежить в памяти эти темы, излагаемые в более ранних главах.

Основные изменения в PyMailGUI

Подобно текстовому редактору PyEdit из главы 11 программа PyMailGUI служит отличным примером развития программного продукта. Поскольку новые версии помогают документировать функциональные возможности этой системы, а главная цель этого примера состоит в том, чтобы показать принципы разработки программного обеспечения на языке Python, давайте коротко познакомимся с последними изменениями.

Новое в версиях 2.1 и 2.0 (третье издание)

Версия 2.1 программы PyMailGUI, представленная в третьем издании книги, вышедшем в начале 2006 года, все еще в значительной мере присутствует и актуальна для данного четвертого издания, вышедшего в 2010 году. Версия 2.1 лишь несколько расширила функциональные возможности версии 2.0, а версия 2.0 была полностью переписана после выхода версии 1.0 во втором издании, с существенным расширением набора возможностей.

Версия 1.0 этой программы из второго издания, написанная в начале 2000 года, содержала всего 685 строк программного кода (515 строк в основном сценарии реализации графического интерфейса и 170 строк в модуле с утилитами для работы с электронной почтой), без учета примеров, повторно использованных в ней, и всего 60 строк справочного текста. В действительности версия 1.0 была чем-то вроде прототипа (если не игрушкой), написанного, по большому счету, чтобы служить коротким примером в книге.

Хотя и без поддержки возможности обработки интернационализированного содержимого электронных писем и других расширений, появившихся в версии 3.0, тем не менее, версия PyMailGUI 2.1 в третьем издании превратилась в более практичную и более богатую возможностями программу, которую можно использовать для повседневной работы с электронной почты. Ее размер вырос почти в три раза и достиг 1800 строк нового программного кода (плюс 1700 строк во взаимосвязанных модулях и 500 дополнительных строк справочного текста). В сравнении, размер версии 3.0 вырос всего на 30% и достиг 2400 строк нового программного кода, как уже отмечалось выше (плюс 2500 во взаимосвязанных модулях и 1700 строк справочного текста). Кому интересна статистика, могут заглянуть в файл *linecounts-prior-version.xls* в подкаталоге *media*, где приводятся точные цифры количества строк по файлам в версии 2.1.

В версии 2.1 в числе новых PyMailGUI были добавлены (и посейчас входят в арсенал) следующие особенности:

- Поддержка просмотра и составления почтовых сообщений, состоящих из нескольких частей с вложениями.
- Операции отправки почты больше не блокируют работу программы и могут перекрываться во времени.
- Почтовые сообщения могут сохраняться в локальных файлах и обрабатываться без подключения к Интернету.
- Дополнительные части сообщений могут теперь автоматически открываться в графическом интерфейсе.
- Поддерживается возможность выбора нескольких сообщений одновременно для обработки в нескольких окнах.
- Первоначально загружаются только заголовки сообщений – полный текст сообщения извлекается по требованию.
- Заголовки окон просмотра и колонки в окне с оглавлением могут настраиваться.
- Удаление выполняется немедленно, а не откладывается до завершения программы.
- Для большинства операций передачи почты в графическом интерфейсе отображается индикатор хода их выполнения.
- При просмотре выполняется автоматический перенос длинных строк.
- Шрифты и цвета, используемые в окнах просмотра сообщений и оглавления, могут настраиваться пользователем.
- Поддерживается аутентификация на серверах SMTP, требующих указать имя и пароль учетной записи при отправке.
- Отправленные сообщения сохраняются в локальном файле, который можно открыть в графическом интерфейсе.
- При просмотре содержимого сообщения интеллектуально выбирается главная текстовая часть.
- Уже извлеченные заголовки и полные тексты сообщений кэшируются для большей эффективности.
- Строки с датами и адресами в составляемых почтовых сообщениях форматируются должным образом.
- В окнах просмотра сообщений теперь выводятся кнопки для быстрого доступа к вложениям/частям (2.1).
- Ошибки синхронизации ящика входящих сообщений обнаруживаются при выполнении операций удаления, получения оглавления и загрузки сообщений (2.1).
- Загрузка из файла и сохранение электронных писем в файле выполняется в отдельном потоке выполнения, чтобы избежать паузы при обработке больших файлов (2.1).

Последние три пункта в этом списке были добавлены в версии 2.1; остальные относятся к полностью переписанной версии 2.0. Некоторые из этих изменений просто наращивают возможности инструментов в стандартной библиотеке (например, поддержка вложений является прямым следствием появления нового пакета `email`), но наиболее существенные изменения были внесены в саму программу PyMailGUI. Было также внесено несколько очевидных исправлений: реализован более точный анализ адресов, а формат представления даты в отправляемых сообщениях приведен в соответствие со стандартами – благодаря тому, что для решения этих задач стали использоваться новые инструменты из пакета `email`.

Новое в версии 3.0 (четвертое издание)

Версия 3.0 программы PyMailGUI, представленная в четвертом издании книги, наследует все улучшения, внесенные в версию 2.1 и описанные в предыдущем разделе, и добавляет множество новых усовершенствований. Изменения в версии 3.0, пожалуй, нельзя назвать внушительными, однако в ней были решены некоторые важные проблемы, связанные с удобством в использовании, что в сумме вполне тянет на то, чтобы присвоить ей новый старший номер версии. Ниже приводится перечень появившихся нововведений:

Перенос на Python 3.X

Реализация была изменена так, что теперь программа выполняется только под управлением Python 3.X – для поддержки версии Python 2.X необходимо вносить изменения в программный код. Для решения одних задач по переносу в Python 3.X в программный код достаточно было внести лишь незначительные поправки, тогда как для других потребовалось вносить более серьезные идиоматические изменения. Новый подход к поддержке Юникода в Python 3.X, например, в значительной степени повлиял на реализацию поддержки интернационализации в этой версии PyMailGUI (обсуждается ниже).

Улучшения в компоновке графического интерфейса

Формы в окнах просмотра теперь создаются с применением компоновки по сетке, вместо фреймов колонок, благодаря чему улучшился внешний вид и было обеспечено единообразное отображение почтовых заголовков в разных платформах (подробнее о компоновке форм рассказывается в главе 9). Дополнительно для наглядности панели инструментов в окнах с оглавлением теперь компонуются с расширяющимися разделителями – это позволяет выделить группы кнопок по их назначению. Кроме того, теперь увеличены размеры окон с оглавлением, чтобы отобразить больше информации.

Исправления в текстовом редакторе, учитывающие изменения в библиотеке Tk

Встроенный текстовый редактор и экземпляры текстовых редакторов во всплывающих окнах, открываемых по требованию, теперь

принудительно обновляются перед тем, как в них будет вставлен текст, чтобы обеспечить точное позиционирование текстового курсора в строке 1. Подробнее об этом рассказывается в описании PyEdit, в главе 11, – необходимость этой операции вызвана недавним изменением (ошибкой?) в Tk или tkinter.

Автоматически наследуются изменения в текстовом редакторе

Поскольку для решения разных задач здесь повторно используется программа PyEdit, данная версия PyMailGUI автоматически получает все последние исправления в PyEdit. Наиболее заметными из них являются новый диалог Grep поиска во внешних файлах и поддержка отображения, открытия и сохранения текста с символами Юникода. Подробности смотрите в главе 11.

Обходное решение проблемы вывода трассировочной информации в Python 3.1

Из разряда «непонятно, но факт»: для корректной работы в Python 3.1 потребовалось изменить общую функцию вывода трассировочной информации в модуле `SharedNames.py`. Функция `print_tb` из модуля `traceback` более не способна выводить содержимое стека в поток вывода `sys.stdout`, если вызывающая ее программа была запущена другой программой в Windows, но работает, как и прежде, когда вызывается из программы, запущенной из командной строки обычным способом. Так как эта функция вызывается из основного потока выполнения в случае появления исключения в рабочем потоке, любая попытка вывести трассировочную информацию (если это разрешено) приводит к полному уничтожению графического интерфейса, если он был запущен из программ запуска демонстрационных примеров.

Для решения этой проблемы функция теперь перехватывает исключения, возникающие при вызове `print_tb`, и повторно вызывает ее, передавая действительный файл вместо `sys.stdout`. Эта проблема, скорее всего, является недостатком Python 3.X, так как тот же самый программный код выполнялся без ошибок в Python 2.5 и 2.6. В отличие от других похожих проблем, данная никак не связана с поддержкой Юникода, так как трассировочная информация содержит только символы ASCII. Еще больше расстраивает тот факт, что прямой вывод в поток `stdout` в той же функции работает прекрасно. Однако, если бы все было просто, это не называлось бы это «работой».

Адреса «Bcc» теперь добавляются в аргумент функции отправки, а сам заголовок удаляется

Небольшое изменение: адреса, вводимые пользователем в поле ввода строки заголовка «Bcc», включаются в список получателей (на «конверте»), а сам заголовок «Bcc» не включается в текст отправляемого сообщения. В противном случае адреса в заголовке «Bcc» могут оказаться видимыми в некоторых клиентах электронной почты (включая PyMailGUI), что лишает смысла этот заголовок.

Исключение возможности параллельной загрузки одного и того же сообщения

Программа PyMailGUI загружает только заголовки сообщений, а полный текст загружается позднее, когда его необходимо будет просмотреть или выполнить другую операцию; при этом допускается одновременная загрузка нескольких сообщений (загрузка выполняется в параллельных потоках). Маловероятно, но вполне возможно запустить операцию загрузки письма, когда оно уже загружается, выбрав письмо еще раз (для этого достаточно два раза подряд быстро щелкнуть на элементе списка в оглавлении). Операция обновления кэша сообщений, производимая выполняющимися потоками параллельно, является безопасной, тем не менее, такое поведение не является нормальным и приводит к пустым тратам времени.

Чтобы избавиться от этого недостатка, данная версия теперь запоминает в главном потоке выполнения, какие письма загружаются в текущий момент, чтобы избежать перекрытия во времени загрузки одного и того же письма – пока сообщение загружается, запрещаются все новые запросы на получение этого письма, пока оно не будет получено полностью. При этом сохраняется возможность одновременного выполнения нескольких операций получения, при условии, что их цели не пересекаются. Для определения пересекающихся запросов используется множество. Письма, уже загруженные и помещенные в кэш, не являются предметом этой проверки и всегда доступны для выбора независимо от того, какие письма загружаются в текущий момент.

Адреса в полях ввода графического интерфейса отделяются запятыми, а не точками с запятой

В предыдущем издании в качестве символа-разделителя адресов использовался символ «;», и при передаче адреса просто разбивались по этому символу. Это было сделано с целью избежать конфликтов в адресах с символом «,», часто используемым в компоненте имени. При создании ответов перед записью в адрес «To» компонент имени отбрасывался, если он содержал «;», но сохранялась вероятность конфликтов, если символ «;» появлялся одновременно и как разделитель адресов, и как символ в компоненте имени адреса электронной почты при ручном вводе.

В версии для этого издания в качестве разделителя адресов используется символ «,», а вместо простого разбиения строки выполняется анализ списков адресов с помощью инструментов `getaddresses` и `parseaddr` из пакета `email`. Поскольку эти инструменты специально предназначены для этого, символы «,», присутствующие в компонентах имен, не воспринимаются по ошибке за разделители адресов и конфликты не возникают. Серверы и клиенты обычно ожидают символ «,», поэтому все работает естественным образом.

С этими исправлениями запятые можно использовать и как разделители адресов, и как символы внутри компонентов имен. При создании ответов они обрабатываются автоматически: поле «To» заполняется содержимым поля «From» оригинального сообщения. При отправке новых писем разбиение адресов в заголовках «To», «Cc» и «Bcc» выполняется инструментами электронной почты автоматически (последние два игнорируются, если содержат первый символ «?»).

Отображение справки в формате HTML

Справка теперь может отображаться как простой текст в окне графического интерфейса, в формате HTML в веб-броузере или и там и там одновременно. Выбор формы отображения справки регулируется пользовательскими настройками в модуле `mailconfig`. Представление в формате HTML является новым. Для его создания выполняется простое преобразование справочного текста с добавлением ссылок на разделы и внешние сайты, а для открытия справки в веб-броузере применяется модуль `webbrowser`, обсуждавшийся ранее в этой книге. Функция отображения справки в текстовом виде теперь является избыточной, но она была оставлена, потому что при отображении HTML-версии отсутствует возможность открыть окна с исходными текстами программы.

Увеличена частота опроса очереди обработчиков в потоках выполнения

Глобальная очередь обработчиков, используемая потоками выполнения для обновления графического интерфейса, теперь позволяет вызывать большее количество обработчиков – 100 вызовов в секунду против 10 в предыдущей версии. Это было достигнуто увеличением частоты проверки очереди (20 раз в секунду против 10) и увеличением числа обработчиков, вызываемых при каждой проверке (5 против 1). В зависимости от чередования операций помещения в очередь и извлечения из нее это повышает скорость загрузки объемных почтовых ящиков примерно во столько же раз (10), причем ценой незначительного увеличения нагрузки на процессор. При этом на моем ноутбуке с Windows 7 потребление ресурса процессора программой PyMailGUI по-прежнему оставалось на уровне 0% в Диспетчере Задач (Task Manager) при простое.

Я увеличил количество вызовов обработчиков, чтобы обеспечить поддержку почтового ящика, содержащего 4800 входящих сообщений (на самом деле их оказалось еще больше к тому моменту, когда я нашел время, чтобы сделать снимок с экрана для этой главы). Без этого на начальную загрузку заголовков из данного почтового ящика, то есть на вызов 4800 обработчиков, уходило 8 минут (4800 ÷ 10 ÷ 60), даже при том, что большинство отображаемых сообщений пропускалось немедленно из-за ограничений на количество загружаемых новых писем (смотрите следующий пункт). После повышения частоты опроса очереди начальная загрузка стала занимать всего 48 се-

кунд – возможно, и не идеально, но не забывайте, что начальная загрузка заголовков выполняется всего один раз за сеанс, и данный прием позволяет сохранить баланс между нагрузкой на процессор и отзывчивостью интерфейса. Конечно, этот почтовый ящик не является типичным примером, но в любом случае множественная начальная загрузка выигрывает от повышения скорости.

В главе 10 в описании модуля `threadtools` приведены большая часть кода, связанного с этим изменением, и другие важные подробности. Можно было бы организовать вызов всех обработчиков, оказавшихся в очереди в момент возникновения события от таймера, но это может привести к блокированию графического интерфейса на неопределенное время при большой скорости добавления обработчиков в очередь.

Ограничение на количество загружаемых писем

Начиная с версии 2.1, PyMailGUI при первом запуске загружает не полный текст сообщений, а только заголовки, а затем загружает только заголовки новых сообщений. Однако в зависимости от пропускной способности вашего подключения к Интернету и быстродействия сервера может оказаться затруднительным выполнить загрузку заголовков сразу всех входящих сообщений (как уже говорилось, в одном таком моем почтовом ящике в настоящее время имеется примерно 4800 электронных писем). Для поддержки подобных случаев в модуле `mailconfig` имеется новый параметр, позволяющий ограничить количество заголовков (или полных сообщений, если команда TOP не поддерживается сервером), извлекаемых при загрузке.

Если этот параметр будет установлен в значение N, программа PyMailGUI будет загружать не более N самых последних почтовых сообщений. Более старые сообщения, не попавшие в это число, не будут загружаться с сервера, но будут отображаться в оглавлении как пустые/фиктивные сообщения, которые не могут участвовать ни в каких операциях (хотя они все могут быть загружены по требованию).

Данная особенность наследуется из реализации пакета `mailtools`, представленного в главе 13, – параметр в модуле `mailconfig`, связанный с ней, описывается далее. Обратите внимание, что – даже с этим исправлением – так как механизм обслуживания очереди, реализованный в модуле `threadtools` и используемый здесь, способен обслуживать события графического интерфейса, такие как обновление индикаторов хода выполнения операции, не более 100 раз в секунду, полная загрузка 4800 заголовков по-прежнему занимает 48 секунд. Тут либо очередь должна действовать быстрее, либо я должен удалять почту время от времени!

Извлечение основного текста в формате HTML (прототип)

Несмотря на все более широкое проникновение HTML в электронную почту в последние годы, PyMailGUI все еще остается клиентом, ориентированным на простой текст. Когда содержимое основной

(или единственной) текстовой части письма представлено разметкой HTML, оно отображается в окне веб-броузера. Однако в предыдущей версии этот текст в формате HTML отображался в компоненте текстового редактора PyEdit и считался главной текстовой частью при создании ответов или при пересылке.

Так как большинству пользователей неудобно разбирать разметку HTML, в версии для этого издания предпринята попытка реализовать извлечение простого текста из этой части в формате HTML за счет простого анализа HTML. После этого извлеченный текст отображается в окне просмотра содержимого письма и используется для цитирования при создании ответов и пересылке.

Этот механизм анализа HTML в лучшем случае можно считать только прототипом, и он включен в пример, в основном, чтобы помочь вам сделать первый шаг и приспособить его под свои потребности, но в любом случае результат, который он воспроизводит, лучше, чем отображение впрямую разметки HTML. Если этот механизм окажется не в состоянии воспроизвести простой текст, пользователи всегда смогут вернуться к просмотру в веб-броузере и копировать текст оттуда для цитирования в ответах и при пересылке. Смотрите также примечание, касающееся альтернативных решений с открытыми исходными текстами далее в этой главе, – эта проблема лежит в уже исследованной области.

Ответ рассылается всем получателям оригинального сообщения

В этой версии по умолчанию ответы действительно отправляются всем – заголовок «Cc» в окне составления ответа автоматически заполняется адресами получателей оригинального сообщения. Для этого извлекаются все адреса из заголовков «To» и «Cc» оригинального сообщения и с помощью операций над множествами удаляются дубликаты, а также адрес нового отправителя. В результате ответ рассылается всем получателям. Дополнительно сохраняется возможность ответа только отправителю посредством инициализации заголовка «To» адресом отправителя оригинального сообщения.

Данная возможность отражает типичный случай, когда электронная почта циркулирует внутри определенной группы. Но поскольку это бывает не всегда желательно, данную возможность можно отключить в модуле `mailconfig`, чтобы при создании ответа заголовок «To» инициализировался только адресом отправителя оригинального сообщения. При включенной возможности пользователям может потребоваться очищать предварительно заполненный заголовок «Cc», а при отключенной – вручную вставлять адреса из заголовка «Cc» оригинала. Оба случая выглядят примерно равновероятными. Кроме того, может так случиться, что в список получателей оригинального сообщения будут включены имена, псевдонимы или фиктивные адреса, которые будут рассматриваться, как неправильные или ошибочные при отправке ответа. Подобно возможности предварительно-

го заполнения заголовка «Bcc», как описано в следующем пункте, заголовок «Cc», предварительно заполненный при создании ответа, в случае необходимости можно отредактировать перед отправкой или вообще запретить автоматическое его заполнение. Смотрите также описание предлагаемых усовершенствований в конце этой главы – более удачным решением может оказаться реализация в графическом интерфейсе возможности включать или отключать эту функцию для каждого сообщения в отдельности.

Другие улучшения: предварительное заполнение «Bcc», регистр символов «Re» и «Fwd», размер списка, дубликаты адресов получателей

Кроме всего прочего было внесено множество небольших улучшений. Среди них: заголовок «Bcc» в окнах редактирования для удобства теперь предварительно заполняется адресом отправителя (обычная роль этого заголовка); операции Reply (Ответить) и Forward (Переслать) теперь не учитывают регистр символов, когда определяют необходимость добавления префикса «Re:» или «Fwd:» в поле темы сообщения; ширина окна с оглавлением почтового ящика теперь может настраиваться в модуле `mailconfig`; дубликаты в списке адресов получателей теперь удаляются пакетом `mailtools` при передаче, чтобы избежать отправки нескольких копий одному и тому же адресату (например, если один и тот же адрес присутствует в обоих заголовках, «To» и «Cc»); и другие незначительные улучшения, которые я не буду упоминать здесь. Обращайте внимание на метки «3.0» и «4E» в комментариях в программном коде ниже и в реализации модулей из пакета `mailtools` в главе 13, где отмечаются другие изменения в программном коде.

Поддержка Юникода (интернационализации)

Самое существенно обновление в PyMailGUI 3.0 я оставил напоследок: теперь эта программа поддерживает кодировки Юникода для получаемых, сохраняемых и отправляемых писем, насколько это позволяет пакет `email` в Python 3.1. И текстовые части сообщений, и их заголовки теперь декодируются при отображении и кодируются при отправке. Однако это слишком большое изменение, чтобы описывать его в данном формате, поэтому более подробное описание приводится в следующем разделе.

Политика поддержки Юникода в версии 3.0

Последний пункт в предыдущем списке является, пожалуй, самым важным. Как уже рассказывалось в главе 13, пользовательские настройки в модуле `mailconfig`, определяющие порядок декодирования байтов полного текста почтового сообщения в строку Юникода при получении и порядок кодирования/декодирования при сохранении сообщений в файлы, действуют на протяжении всего сеанса работы с программой.

Более определенно, при составлении: имеется возможность явно указать в модуле `mailconfig` или запросить у пользователя кодировку, кото-

рая будет применяться к основному тексту и к текстовым вложениям составляемого сообщения; при просмотре: для определения кодировки основного текста, а также текстовых частей, открываемых по требованию, используется информация из заголовков. Кроме того, декодирование интернационализированных заголовков (например, «Subject», «To» и «From») при их отображении выполняется в соответствии со стандартами электронной почты, MIME и Юникода и в зависимости от их содержимого, и они автоматически кодируются при отправке, если содержат текст с символами, не входящими в диапазон ASCII.

Здесь также наследуются и другие особенности поддержки Юникода (и исправления) из пакета `mailtools`, представленного в главе 13, – подробнее об этом рассказывается в предыдущей главе. Ниже суммируется, как все эти особенности отражаются на пользовательском интерфейсе:

Полученные письма

При *получении* электронных писем для декодирования байтов полного сообщения в строку Юникода, как того требует механизм анализа, реализованный в пакете `email`, используются пользовательские настройки, действующие на протяжении всего сеанса работы с программой. Если эта операция не увенчалась успехом, выполняется попытка угадать, применив по очереди некоторые из наиболее распространенных кодировок. Большинство почтовых сообщений наверняка будут иметь 7- или 8-битовую природу, поскольку оригинальные стандарты электронной почты требуют, чтобы текст сообщений состоял только из символов ASCII.

Текстовые части при составлении письма

При *отправке* новых писем для определения кодировки основной текстовой части и всех текстовых вложений используются настройки пользователя. Если они отсутствуют в модуле `mailconfig`, графический интерфейс предложит пользователю указать кодировку для каждой текстовой части. Они также будут использоваться для добавления заголовков с информацией о кодировках и определять формат MIME. Во всех случаях программа будет использовать кодировку UTF-8, если кодировка, указанная в `mailconfig` или введенная пользователем, окажется неприменимой к отправляемому тексту – например, если пользователь укажет кодировку ASCII для основного текста ответа или пересылаемого письма, содержащего символы за пределами набора ASCII или текстовые вложения с такими символами.

Заголовки составляемого письма

При *отправке* новых писем, если заголовки или компоненты имен в заголовках с адресами содержат символы за пределами набора ASCII, программа пытается закодировать заголовки в соответствии со стандартами интернационализации. Для этого по умолчанию используется кодировка UTF-8, но имеется возможность указать другую кодировку в модуле `mailconfig`. В адресах, если компонент имени не может быть закодирован, он отбрасывается, и используется

только компонент адреса. При этом предполагается, что серверы будут корректно обрабатывать закодированные имена в адресах.

Отображение текстовых частей

При *просмотре* полученных сообщений везде, где возможно, для декодирования текста используются имена кодировок, указанные в заголовках. Основная текстовая часть декодируется в строку `str` Юникода согласно информации в заголовке перед вставкой ее в компонент PyEdit. Содержимое всех остальных текстовых частей, а также всех двоичных частей сохраняется в виде строк `bytes` в двоичные файлы, откуда они могут быть извлечены по требованию позднее в графическом интерфейсе. При открытии по требованию таких текстовых частей они отображаются в отдельных окнах PyEdit, при этом компоненту PyEdit передается имя двоичного файла, где хранится содержимое извлекаемой части, а также имя кодировки, полученное из заголовка этой части.

Если имя кодировки отсутствует в заголовке текстовой части или декодирование с применением этой кодировки не увенчалось успехом, предпринимается попытка применить кодировку основной текстовой части. К частям сообщения, открываемым по требованию, применяется также политика поддержки Юникода, реализованная в редакторе PyEdit (как описывается в главе 11 – он может запросить кодировку у пользователя, если она неизвестна). Кроме того, текстовые части в формате HTML сохраняются в двоичном режиме и открываются в веб-броузере в надежде, что проблема декодирования будет решена механизмом поддержки кодировок в броузере, который может использовать пользовательские настройки или информацию в тегах HTML для определения кодировки.

Отображение заголовков

При *просмотре* сообщений заголовки автоматически декодируются в соответствии со стандартами электронной почты. Под этим подразумевается декодирование полных заголовков, таких как «Subject», и компонентов имен в заголовках с адресами, таких как «From», «To» и «Cc», при этом допускается, что компоненты имен будут полностью закодированными или содержать подстроки в разных кодировках. Поскольку содержимое заголовков определяет формат MIME и кодировку Юникода, для декодирования заголовков не требуется получать какую-либо информацию от пользователя.

Иными словами, программа PyMailGUI теперь поддерживает отображение и составление *интернационализированных* сообщений, включая интернационализированное содержимое или заголовки. Для большей пользы эта поддержка распределена по множеству пакетов и примеров. Например, декодирование полного текста сообщения в Юникод при получении фактически производится глубоко в классах, импортируемых из пакета `mailtools`. Благодаря этому полный (не разобранный) текст сообщения всегда будет представлен в программе в виде строки `str` Юни-

кода. Точно так же декодирование заголовков для отображения выполняется с помощью инструментов, реализованных в пакете mailtools, при этом операция кодирования заголовков инициируется и выполняется непосредственно внутри пакета mailtools во время отправки.

Реализация декодирования полного текста сообщения иллюстрирует способы выбора кодировки. Основная кодировка определяется переменной fetchEncoding в модуле mailconfig. Этот пользовательский параметр используется программой PyMailGUI для декодирования байтов получаемых сообщений в строку str, необходимую механизму анализа, и для сохранения и загрузки полного текста сообщения в файл на протяжении всего сеанса работы. Пользователи могут присвоить этой переменной строку с именем своей кодировки Юникода. Вполне разумным выбором для большинства электронных писем могут оказаться кодировки «latin-1», «utf-8» и «ascii», так как стандарты электронной почты изначально требуют, чтобы почтовые сообщения состояли из символов ASCII (хотя для декодирования некоторых старых файлов, сгенерированных предыдущей версией, требуется использовать кодировку «latin-1»). Если операция декодирования с этим именем кодировки терпит неудачу, выполняется попытка применить другие распространенные кодировки, и в самом крайнем случае отображается содержимое заголовков, если оно может быть декодировано, а тело сообщения заменяется сообщением об ошибке. Чтобы просмотреть такие письма, попробуйте запустить PyMailGUI снова, с другими настройками кодировки.

Из недостатков: в программе не делается ничего, что касается поддержки Юникода для полного текста отправляемого сообщения, сверх того, что предлагается библиотеками Python (как мы узнали в главе 13, модуль smtplib пытается использовать кодировку ASCII при отправке сообщения, что является одной из причин необходимости кодирования заголовков). И, несмотря на то, что обеспечивается полная поддержка наборов символов для содержимого почтовых сообщений, в самом графическом интерфейсе по-прежнему используются надписи на кнопках и в метках на английском языке.

Как описывалось в главе 13, данные решения, касающиеся поддержки Юникода, несмотря на всю их широту, являются неполными, так как пакет email в Python 3.1, на корректную работу которого полагается программа PyMailGUI, еще находится на пути к реализации окончательного правильного решения для некоторых случаев. Обновленная версия пакета, которая будет более точно и полно соответствовать разделению типов str/bytes в Python 3.X, наверняка появится в будущем – следите за страницей с обновлениями для книги (описывается в предисловии), где будет приводиться информация об улучшениях и изменениях, касающаяся поддержки Юникода в программе. Хотелось бы надеяться, что новая версия пакета email, лежащего в основе PyMailGUI 3.0, будет доступна в ближайшее время.

В программном обеспечении всегда есть место для улучшений (смотрите список в конце этой главы), тем не менее, программа PyMailGUI

обеспечивает вполне полнофункциональный интерфейс к электронной почте, представляет собой самый существенный пример в этой книге и является демонстрацией практического применения языка Python и принципов разработки программного обеспечения в целом. Как часто отмечают пользователи, может доставлять удовольствие сама работа с языком Python, но кроме того, он также полезен при создании практических и нетривиальных программ. Данный пример более, чем какой-то другой из приводимых в книге, свидетельствует об этом. А как именно, демонстрируется в следующем разделе.

Демонстрация PyMailGUI

Программа PyMailGUI имеет многооконный интерфейс. Он включает следующее:

- Главное окно со списком сообщений, открываемое при запуске программы, – для обработки почты на сервере.
- Одно или более окон с оглавлением содержимого в файлах с сохраненными сообщениями – для обработки почты без подключения к Интернету.
- Одно или более окон для просмотра и редактирования содержимого почтовых сообщений.
- Окна PyEdit для отображения необработанного текста почтовых сообщений, текстовых вложений и исходного программного кода системы.
- Неблокирующие диалоги с информацией о ходе выполнения операции.
- Различные диалоги, предназначенные для открытия частей сообщения, вывода справки и другие.

С технической точки зрения PyMailGUI выполняется как множество параллельных потоков, которые могут действовать одновременно: по одному для каждой операции передачи почты между программой и сервером и по одному для каждой операции загрузки, сохранения или удаления сообщений в файлах. Программа PyMailGUI поддерживает сохранение почты в файлах, автоматически сохраняет отправленные сообщения, обеспечивает возможность настройки шрифтов и цветов, позволяет просматривать и добавлять вложения, извлекает основной текст сообщения, преобразует содержимое в формате HTML в простой текст и обеспечивает множество других возможностей.

Чтобы легче было понять этот практический пример, рассмотрим вначале операции, которые выполняет программа PyMailGUI, – взаимодействие с пользователем и функции обработки почты – и только потом обратимся к программному коду Python, реализующему их. При чтении этого материала не стесняйтесь заглядывать в листинги, которые следуют за снимками экранов, но обязательно прочтите также этот раз-

дел, где я объясняю некоторые тонкости конструкции PyMailGUI. После этого раздела вы можете самостоятельно изучать листинги с программным кодом; это дает лучшее и более полное объяснение, чем написанное на естественном языке.

Запуск

Итак, пришло время испытать систему в действии. Я буду демонстрировать работу программы на моем ноутбуке с Windows 7. На других платформах (включая другие версии Windows) интерфейс программы будет выглядеть несколько иначе благодаря поддержке инструментами графического интерфейса внешнего вида, свойственного конкретной платформе, но основная функциональность при этом останется прежней.

PyMailGUI – это программа Python/tkinter, запускаемая выполнением файла сценария верхнего уровня *PyMailGUI.py*. Как и другие программы на языке Python, PyMailGUI можно запускать из командной строки, щелчком на значке с именем его файла в окне менеджера файлов или нажатием кнопки в панелях запуска PyDemos или PyGadgets. Независимо от способа запуска сначала PyMailGUI выводит окно, изображенное на рис. 14.1 и полученное после щелчка на кнопке Load (Загрузить), чтобы загрузить заголовки с почтового сервера моего интернет-провайдера. Обратите внимание на ярлык окна с изображением символов «PY»: это результат работы инструментов поддержки протоколов оконного интерфейса, которые мы написали ранее в этой книге. Обратите также внимание на строки с заголовками, содержащие символы, не входящие в диапазон ASCII, – я буду рассказывать об особенностях поддержки реализации позднее.

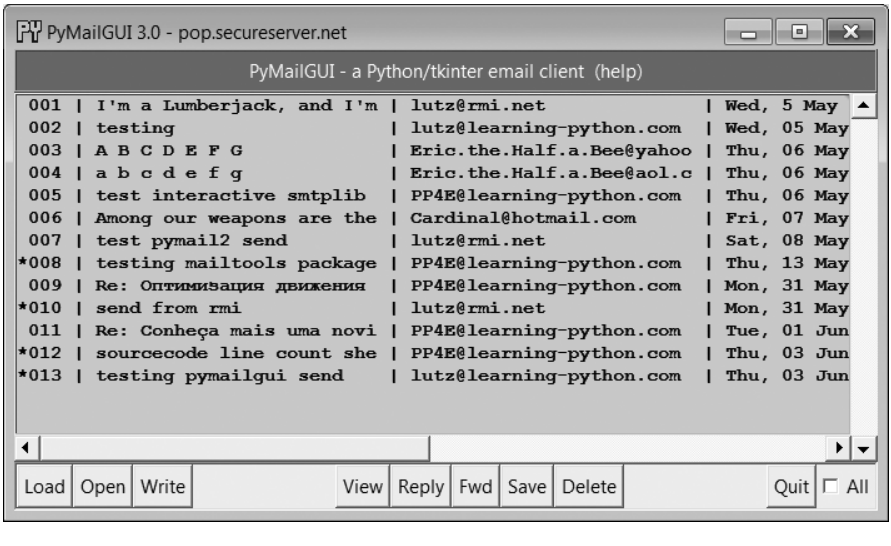


Рис. 14.1. *Главное окно PyMailGUI со списком сообщений на сервере*

Это главное окно PyMailGUI – отсюда начинаются все операции. Оно нам предоставляет:

- Кнопку справки (полоса вверху).
- Область для списка полученных почтовых сообщений (средняя часть), на которой можно щелкнуть мышью.
- Нижнюю панель с кнопками для обработки сообщений, выделенных в области списка.

Обычно пользователь загружает свою почту, выбирает щелчком сообщение в области списка и нажимает одну из нижних кнопок для его обработки. При запуске программы не выводится никаких сообщений – их необходимо загрузить, щелкнув на кнопке Load (Загрузить), после чего отображается простой диалог ввода пароля, затем диалог с информацией о ходе выполнения операции, где ведется обратный отсчет заголовков сообщений, которые осталось загрузить, и затем список заполняется сообщениями, готовыми для выбора.

Окно программы PyMailGUI с оглавлением, изображенное на рис. 14.1, отображает информацию из заголовков в колонках фиксированной ширины и может распахиваться до максимального размера. Почтовые сообщения с вложениями отмечаются символом «*» в колонке с порядковыми номерами, а шрифты и цвета в этом и других окнах PyMailGUI можно настроить в модуле `mailconfig`. На черно-белых изображениях в этой книге не видно, но большинство окон со списками сообщений, которые мы увидим, имеют красный фон, окна просмотра содержимого – светло-голубой, окна PyEdit – бежевый, вместо привычного бирюзового, а окно справки – синевато-стальной цвет. Большинство из них вы можете изменить по своему вкусу, а внешний вид окон PyEdit можно изменить в самом графическом интерфейсе (образец определения цвета смотрите в примере 8.11 и альтернативные примеры настроек – далее).

Окна со списками сообщений позволяют выбирать одновременно несколько сообщений – операция, выбранная в панели инструментов внизу окна, применяется ко всем выбранным письмам. Например, чтобы просмотреть содержимое нескольких писем, выберите их в списке и щелкните на кнопке View (Просмотреть) – программа загрузит (при необходимости) их с сервера и отобразит каждое из них в отдельном окне просмотра. Используйте флажок All (Все) в правом нижнем углу, чтобы выбрать все сообщения в списке. Чтобы выбрать не все, а только несколько сообщений, щелкайте мышью, удерживая клавишу Ctrl или Shift (стандартная операция выбора нескольких элементов из списка, используемая в Windows, – попробуйте сами).

Однако прежде чем двинуться дальше, нажмем на изображенную на рис. 14.1 полосу над списком вверху окна и посмотрим имеющуюся справку. На рис. 14.2 изображено окно текстовой справки, которое при этом появляется, – одной из двух имеющихся разновидностей справки.

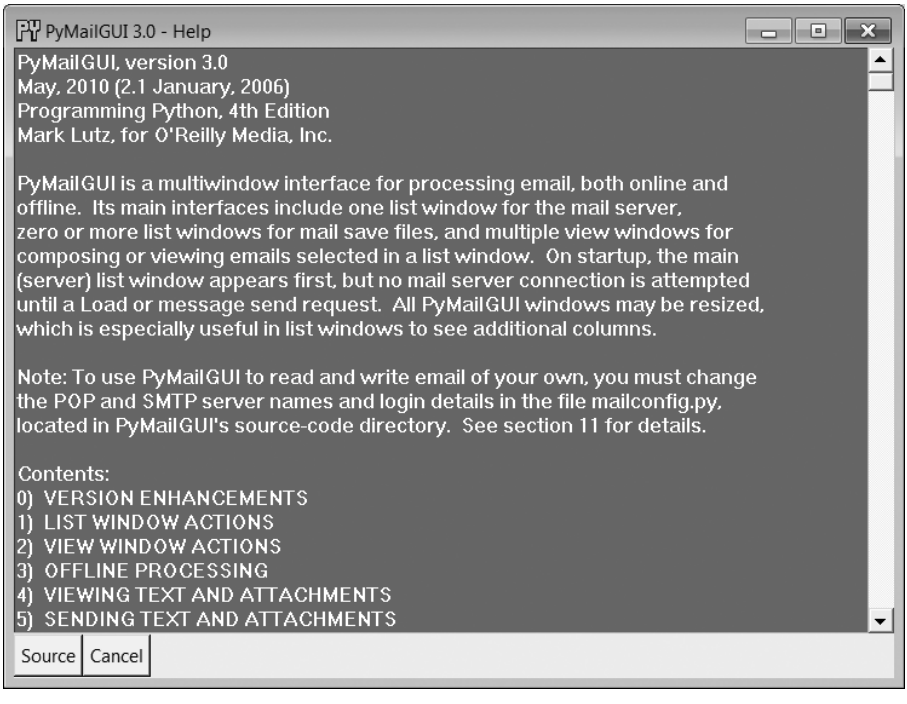


Рис. 14.2. *Всплывающее окно справки PyMailGUI*

Основную часть этого окна составляют блок текста в текстовом виджете с прокруткой, а также две кнопки внизу. Весь текст справки представлен в программе, как одна строка в тройных кавычках. Как мы увидим чуть ниже, имеется также возможность открыть этот же текст в формате HTML в веб-броузере, но для многих вполне достаточно и простого текста.[1] Кнопка Cancel (Отмена) закрывает это немодальное (то есть неблокирующее) окно. Следует отметить, что кнопка Source (Исходные тексты) открывает окна с текстовым редактором PyEdit для просмотра всех файлов с исходным программным кодом PyMailGUI. На рис. 14.3 показан один из таких файлов (их довольно много, но этот выбран наугад,

[1] В действительности вывод справки был вначале еще менее красивым: первоначально текст выводился в стандартном информационном окне, генерируемом функцией `showinfo` из библиотеки tkinter, использовавшимся ранее в этой книге. В Windows все было хорошо (по крайней мере, при небольшом объеме справочного текста), но в Linux возникала проблема из-за ограничений по умолчанию на длину строки в информационном окне – строки переносились так плохо, что становились нечитаемыми. За эти годы диалог был заменен текстовым виджетом с прокруткой, который теперь в свою очередь замещается справкой в формате HTML. Я думаю, что в следующем издании придется реализовать голографический интерфейс отображения справки...

а не для демонстрации разработки). Не всякая программа покажет вам свои исходные тексты, но PyMailGUI следует лейтмотиву открытого программного обеспечения Python.

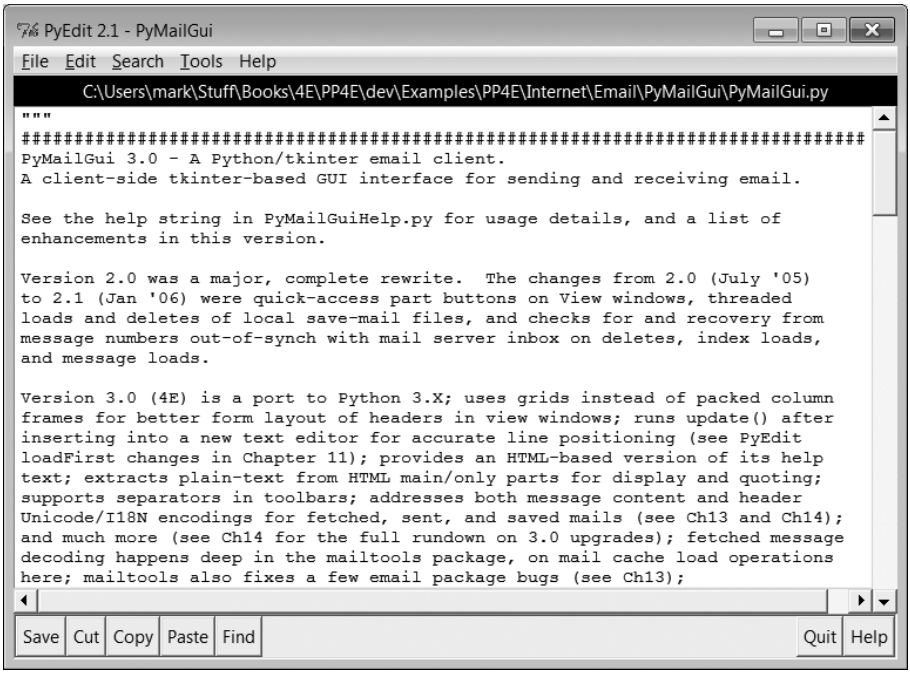


Рис. 14.3. *Окно просмотра исходного программного кода PyMailGUI*

Впервые в этом издании справка может также отображаться в формате HTML в окне веб-броузера, вместо или в дополнение к текстовой справке, которую мы только что видели. Выбор отображения справки в текстовом формате, в формате HTML или в обоих форматах одновременно определяется настройками в модуле `mailconfig`. Справка в формате HTML отображается с помощью модуля Python `webbrowser`, который открывает файл HTML в броузере на локальном компьютере, и в настоящее время не имеет кнопки открытия файлов с исходными текстами, присутствующей в окне с текстовой версией (еще одна из причин, почему может потребоваться инструмент просмотра текста). Справка в формате HTML изображена на рис. 14.4.

Когда сообщение для просмотра выбирается в главном окне и затем выполняется щелчок на кнопке View (Просмотреть), PyMailGUI загружает полный текст сообщения (если оно еще не было загружено в этом сеансе) и выводит окно просмотра сообщения, как показано на рис. 14.5, где изображено содержимое одного из сообщений в моем почтовом ящике.

Рис. 14.4. *Отображение справки PyMailGUI в формате HTML (новое в версии 3.0)*

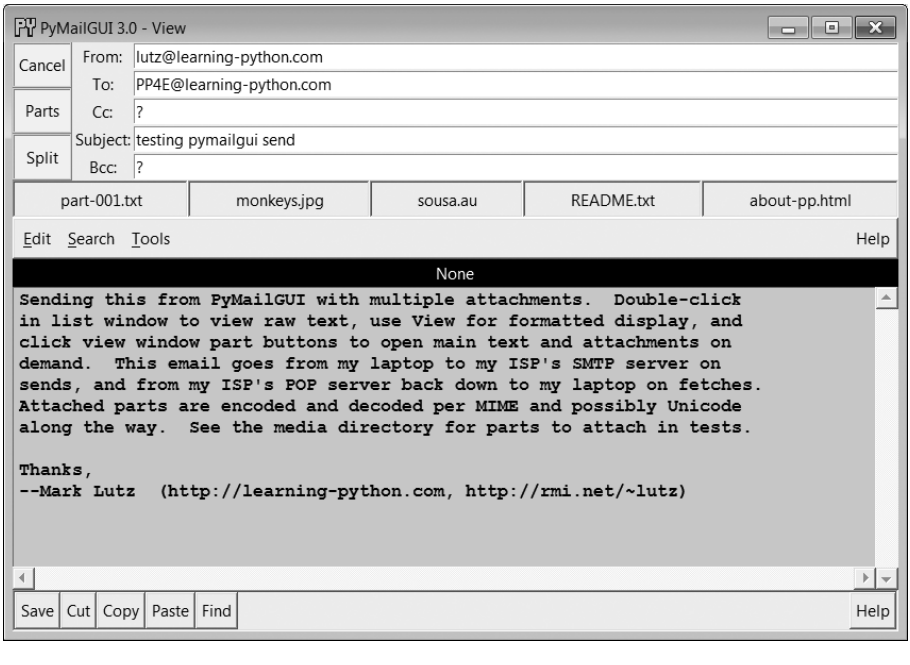


Рис. 14.5. *Окно PyMailGUI просмотра сообщения*

Окна просмотра создаются в ответ на операции в главном окне со списком и имеют следующую организацию:

- В верхней части находятся кнопки операций (Parts (Части) выводит список частей сообщения, Split (Разделить) сохраняет части сообщения в указанный каталог и открывает их и Cancel (Отмена) закрывает это немодальное окно), а также раздел для отображения заголовков сообщения («From», «To» и так далее).
- В середине находится панель с кнопками быстрого доступа для открытия частей сообщения, включая вложения. В случае щелчка на этих кнопках PyMailGUI открывает части сообщения в соответствии с их типами. Изображения могут открываться в веб-броузере или в программе просмотра изображений, текстовые части – в PyEdit, HTML – в веб-броузере, документы Windows – согласно записям в реестре Windows и так далее.
- Основу этого окна (всю нижнюю его часть) занимает повторно используемый объект класса TextEditor, который был написан для программы PyEdit в главе 11, – PyMailGUI просто вставляет экземпляр TextEditor в каждое окно для просмотра и создания нового сообщения, бесплатно получая компонент полнофункционального текстового редактора. В действительности многое в окне, изображенном на рис. 14.5, реализовано с помощью TextEditor, а не PyMailGUI.

Такое повторное использование класса редактора PyEdit означает, что мы можем использовать все его инструменты для работы с текстом почтового сообщения – копирование и вставку, поиск и переход по номеру строки, сохранение копии текста в файл и другие. Например, кнопка Save (Сохранить) редактора PyEdit в левом нижнем углу на рис. 14.5 может использоваться для сохранения основного текста сообщения (как мы увидим далее, подобную же возможность предоставляет самая левая кнопка в середине окна, соответствующая первой части сообщения, и также имеется возможность сохранить письмо целиком в окне со списком). Чтобы сделать факт повторного использования еще более очевидным, выберите в текстовой части окна пункта Info (Информация) в меню Tools (Инструменты), и вы получите стандартное окно объекта TextEditor редактора PyEdit со статистикой файла, как показано на рис. 14.6, – в точности то же всплывающее окно, что и в автономном текстовом редакторе PyEdit и в программе просмотра изображений PyView из главы 11.

Фактически, это уже третий пример повторного использования класса TextEditor в данной книге: PyEdit, PyView и теперь PyMaiGUI предоставляют пользователям один и тот же интерфейс редактирования текста, поскольку все они используют один и тот же объект TextEditor. Программа PyMailGUI использует его для решения разных задач – она прикрепляет экземпляры этого класса к окнам просмотра и создания сообщений, запускает экземпляры в независимых окнах для просмотра текстовых частей и необработанного текста сообщений, а также для отображения исходных текстов на языке Python (последний случай мы видели на

рис. 14.3, выше). Для просмотра компонентов сообщений PyMailGUI настраивает шрифт и цвет PyEdit в соответствии с его собственным модулем настроек – при открытии редактора в отдельных окнах применяются настройки из локального модуля `textConfig`.

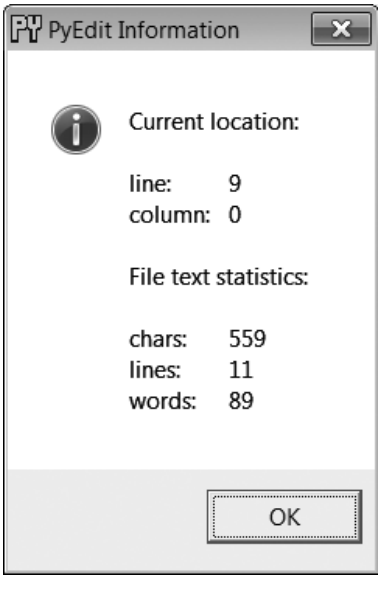


Рис. 14.6. *Диалог со статистической информацией компонента PyEdit, используемого в PyMailGUI*

Для отображения электронного письма PyMailGUI вставляет его текст в прикрепленный объект `TextEditor`; при составлении нового письма PyMailGUI предоставляет объект `TextEditor` и потом извлекает из него текст для передачи через Сеть. Помимо очевидной простоты при таком повторном использовании программного кода упрощается распространение усовершенствований и исправлений – все изменения в объекте `TextEditor` автоматически наследуются программами PyMailGUI, PyView и PyEdit.

Версия PyMailGUI в третьем издании, например, начала поддерживать операции отмены (undo) и возврата (redo) ввода просто потому, что эта особенность была реализована в PyEdit. А в этом четвертом издании все программы, импортирующие редактор PyEdit, наследуют его новый диалог `Grep` поиска в файлах, а также новую поддержку просмотра и редактирования текста Юникода в произвольной кодировке, что особенно полезно при работе с текстовыми частями электронных писем произвольного происхождения (подробнее история развития PyEdit описывается в главе 11).

Загрузка почты

Теперь вернемся к главному окну PyMailGUI со списком и щелкнем по кнопке Load (Загрузить), чтобы получить входящую почту по протоколу POP. Функция загрузки в программе PyMailGUI получает параметры учетной записи из модуля `mailconfig`, содержимое которого будет представлено ниже в этой главе, поэтому проверьте, чтобы в этом файле были отражены параметры вашей электронной почты (то есть имена серверов и имена пользователей), если, конечно, вы хотите читать свою почту с помощью PyMailGUI. Если только вам не удастся угадать пароль к учетной записи для этой книги, предустановленные настройки у вас работать не будут.

О пароле учетной записи следует сказать дополнительно. Программа PyMailGUI может получать его из двух мест:

Локальный файл

Если поместить имя локального файла, содержащего пароль, в модуль `mailconfig`, PyMailGUI при необходимости загрузит пароль из этого файла.

Диалог

Если в модуле `mailconfig` не указывать имя файла с паролем (или PyMailGUI не сможет по каким-то причинам получить пароль из этого файла), PyMailGUI запросит пароль, когда он потребуется.

На рис. 14.7 изображен диалог для ввода пароля, который выводится, если пароль не был сохранен в локальном файле. Обратите внимание: при вводе символы пароля не отображаются – параметр `show='*'` поля `Entry` в этом окне указывает библиотеке tkinter, что вводимые символы должны отображаться как звездочки (этот параметр аналогичен по духу модулю ввода с консоли `getpass`, с которым мы встречались в предыдущей главе, и параметру HTML `type=password`, с которым мы встретимся позже). Введенный пароль находится только в памяти вашего компьютера – PyMailGUI никуда не записывает его для постоянного хранения.

Обратите также внимание, что при использовании параметра с именем локального файла, где хранится пароль, требуется, чтобы пароль хранился в файле на локальном компьютере клиента в незашифрованном виде. Это удобно (не нужно каждый раз заново вводить пароль для получения почты), но не очень хорошо, если на компьютере могут работать несколько пользователей. Оставьте этот параметр в `mailconfig` пустым, если предпочитаете всегда вводить пароль в диалоговом окне.

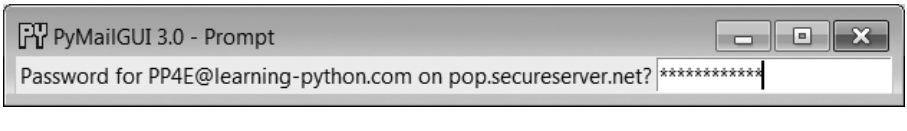


Рис. 14.7. *Окно для ввода пароля PyMailGUI*

Получив тем или иным способом настройки почты и пароль, PyMailGUI постарается загрузить заголовки всех входящих сообщений с сервера POP. Последующие операции загрузки будут загружать только вновь прибывшую почту. Для поддержки до неприличия больших почтовых ящиков (таких как у меня), программа теперь пропускает загрузку заголовков сообщений, кроме самых свежих, количество которых ограничивается настройками в модуле `mailconfig` – пропущенные заголовки в списке отображаются с текстом «--mail skipped--» в поле темы – подробности смотрите выше, в обзоре изменений в версии 3.0.

Для экономии времени PyMailGUI извлекает с сервера только заголовки сообщений, заполняя ими список в главном окне. Полный текст сообщений извлекается, только когда они выбираются для просмотра или для обработки и только если полный текст не был загружен ранее в течение текущего сеанса. PyMailGUI повторно использует инструменты загрузки почты из модуля `mailtools`, представленного в главе 13, который в свою очередь использует для получения почты стандартный модуль Python `poplib`.

Многопоточная модель выполнения

Теперь, когда почта загружена, я должен пояснить, как действует программа PyMailGUI, чтобы избежать блокирования графического интерфейса и обеспечить выполнение операций, перекрывающихся во времени. В конечном итоге передача почты выполняется через сокеты в относительно медленных сетях. В процессе загрузки графический интерфейс остается активным – вы, к примеру, в то же самое время можете составлять и отправлять другие письма. Чтобы показать ход выполнения операции, при извлечении оглавления почтового ящика программа выводит немодальный диалог, как показано на рис. 14.8.

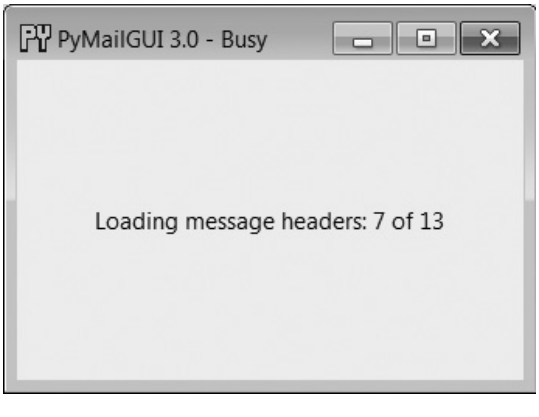


Рис. 14.8. *Неблокирующий диалог с информацией о ходе выполнения операции: загрузка*

Вообще, все операции обмена с сервером отображают такие диалоги. На рис. 14.9 приводится диалог, который отображался программой в процессе загрузки полного текста пяти выбранных сообщений, отсутствующих в локальном кэше (то есть еще не загруженных), в ответ на выполнение операции View (Просмотреть). После окончания загрузки все пять сообщений были отображены в отдельных всплывающих окнах просмотра.

Рис. 14.9. *Неблокирующий диалог с информацией о ходе выполнения операции: просмотр*

Такие операции с сервером, как и другие продолжительные операции, запускаются в *потоках выполнения*, чтобы избежать блокирования графического интерфейса. Это не мешает параллельно запускать другие операции, при условии, что вновь запускаемые операции не вступают в противоречие с уже выполняющимися. Например, одновременно может выполняться отправка нескольких писем и загрузка новой почты, параллельно с графическим интерфейсом – окна графического интерфейса могут перемещаться, перерисовываться и изменяться в размерах в процессе передачи почты. Другие операции обмена, такие как удаление писем, не должны позволять инициировать другие операции до своего завершения – операция удаления обновляет содержимое почтового ящика и внутренние кэши слишком радикально, чтобы поддерживать возможность параллельного выполнения других операций.

В системах, где нет поддержки многопоточной модели выполнения, PyMailGUI во время таких длительных операций переходит в заблокированное состояние (вместо операции порождения потока ставится заглушка, выполняющая простой вызов функции). Поскольку графический интерфейс без потоков выполнения оказывается фактически мертв, на таких платформах при перекрытии и открытии графического интерфейса другими окнами во время загрузки почты его содержимое стирается или каким-то образом искажается. По умолчанию потоки

выполнения поддерживаются на большинстве платформ, где выполняется Python (включая Windows), поэтому вряд ли вы увидите такие странности на своем компьютере.

Реализация многопоточной модели выполнения

Практически на любой платформе продолжительные операции, такие как получение или отправка почты, выполняются в параллельных потоках, благодаря чему графический интерфейс остается активным – он продолжает обновляться и откликаться на действия пользователя, пока производится передача в фоновом режиме. Это относится к большинству графических интерфейсов с поддержкой многопоточной модели выполнения; тем не менее, ниже приводятся два замечания, касающиеся особенностей ее реализации в PyMailGUI:

Обновление графического интерфейса: очередь обработчиков обратного вызова

Как мы узнали ранее в этой книге, обычно только главный поток выполнения, создавший графический интерфейс, должен обновлять его. Подробнее об этом рассказывается в главе 9 – библиотека tkinter не поддерживает возможность параллельного изменения графического интерфейса. Как результат, программа PyMailGUI не выполняет никаких действий, имеющих отношение к пользовательскому интерфейсу, внутри потоков, производящих загрузку, отправку или удаление электронных писем. Вместо этого главный поток графического интерфейса продолжает откликаться на события, порождаемые пользователем, и обновляться и использует таймер для просмотра очереди обработчиков обратного вызова, добавляемых рабочими потоками выполнения, применяя для этого инструменты, реализованные нами в главе 10 (пример 10.20). При обнаружении обработчика в очереди поток выполнения графического интерфейса извлекает и вызывает его, чтобы обеспечить возможность изменения графического интерфейса в контексте главного потока.

Такие обработчики обратного вызова, помещаемые в очередь, могут отображать полученное сообщение, обновлять оглавление, изменять индикатор хода выполнения операции, сообщать об ошибках или закрывать окно составления письма – все они планируются для выполнения рабочими потоками, но выполняются в контексте главного потока графического интерфейса. Этот прием автоматически делает безопасными операции изменения интерфейса в обработчиках: так как они выполняются только в одном потоке, подобные операции с графическим интерфейсом не перекрываются во времени.

Для простоты PyMailGUI помещает в очередь *связанные методы* объектов, которые соединяют в себе вызываемую функцию и ссылку на объект графического интерфейса. Поскольку все потоки выполняются в одном и том же процессе и области памяти, ссылка на объект графического интерфейса, помещаемая в очередь, обеспечивает

доступ к любой информации в графическом интерфейсе, которую требуется обновить, включая объекты виджетов. Кроме того, в качестве функций потоков PyMailGUI также использует связанные методы, что позволяет потокам обновлять информацию в целом, о чем рассказывается в следующем абзаце.

Изменение другой информации: блокирование операций, перекрывающихся во времени

Несмотря на то, что только что описанная схема обновления графического интерфейса с применением очереди обработчиков обратного вызова эффективно ограничивает область изменения графического интерфейса единственным главным потоком, тем не менее, она не гарантирует полную безопасность в целом. Поскольку некоторые дочерние потоки обновляют объекты, используемые совместно с другими потоками (например, кэш электронной почты), для предотвращения перекрытия операций во времени PyMailGUI также использует блокировки, если эти операции могут конфликтовать друг с другом, обращаясь к одним и тем же данным. Сюда входят операции, изменяющие совместно используемые объекты в памяти (например, загрузка заголовков и содержимого в кэш), а также операции, которые могут изменять нумерацию загруженных сообщений (например, удаление).

Везде, где одновременное выполнение потоков может приводить к проблемам, графический интерфейс проверяет состояние блокировок и выводит сообщение, когда операция не может быть выполнена немедленно. Конкретные случаи, когда применяется это правило, описываются в исходном программном коде и в тексте справки.

Такие операции, как отправка или получение для просмотра отдельных писем, в значительной степени являются независимыми и могут перекрываться во времени, но операции удаления и получения заголовков – нет.

Кроме того, некоторые операции по сохранению почты, которые порой могут выполняться продолжительное время, также выполняются в параллельных потоках, чтобы избежать блокирования графического интерфейса. А чтобы предотвратить возможность одновременного получения одних и тех же писем в параллельных потоках и выполнения лишней работы, в этой редакции используется объект множества (смотрите обзор изменений в версии 3.0 выше).

Подробнее о том, почему все это имеет большое значение, рассказывается в главах 5, 9 и 10, где обсуждаются потоки выполнения. На самом деле, программа PyMailGUI просто является конкретной реализацией концепций, исследованных нами ранее.

Интерфейс загрузки с сервера

Вернемся к операции загрузки электронной почты: поскольку она в действительности является операцией с сокетами, программа PyMailGUI автоматически соединится с почтовым сервером, используя те возможности соединения, которые имеются на компьютере, где она выполняется. Например, если соединение с Сетью осуществляется через модем и в данный момент не установлено, Windows автоматически выведет стандартный диалог соединения. При широкополосном подключении, получившем наибольшее распространение в настоящее время, соединение с почтовым сервером обычно устанавливается автоматически.

По завершении загрузки электронной почты программа PyMailGUI заполняет прокручиваемый список в главном окне сообщениями с почтового сервера и автоматически прокручивает его до появления последнего полученного письма. На рис. 14.10 показано, как выглядит главное окно после выбора одного из сообщений и изменения размеров окна – текстовая область в середине растягивается и сжимается вместе с окном, открывая по мере увеличения все большее количество колонок.

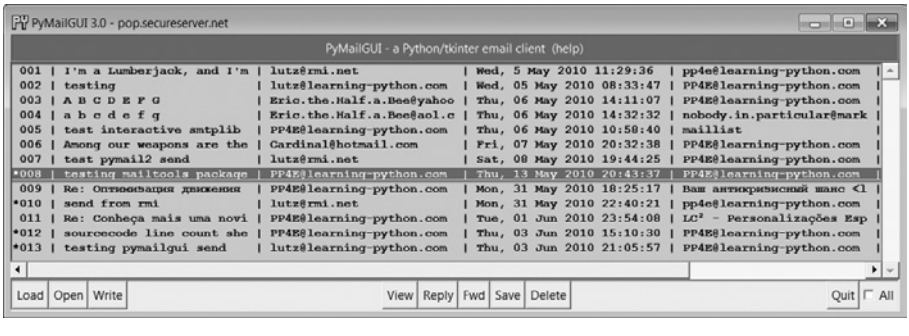


Рис. 14.10. *Главное окно PyMailGUI после изменения размеров*

Технически кнопка Load (Загрузить) при первом своем нажатии загружает всю почту, а при последующих нажатиях – только вновь поступившие сообщения. PyMailGUI следит за тем, какое сообщение было загружено последним, и при последующих загрузках запрашивает только сообщения с номерами, большими последнего. Загруженная почта хранится в памяти в списке Python, чтобы не загружать ее каждый раз снова. Программа PyMailGUI не удаляет почту с сервера при загрузке. Если вы не хотите видеть какое-то письмо при загрузке в будущем, то должны явным образом удалить его.

В записях в главном списке показано лишь то, что дает пользователю представление о содержании письма, – в каждой записи выводятся части заголовков «Subject», «From», «Date», «To» и других, разделяемые

символом |, и с номером POP-сообщения впереди (например, в данном списке имеется 13 сообщений). Размеры колонок определяются максимальным размером содержимого, необходимым для отображения любой записи, а набор отображаемых заголовков настраивается в модуле `mailconfig`. Чтобы увидеть содержимое дополнительных заголовков, таких как размер сообщения или название программы-клиента электронной почты, используйте горизонтальную прокрутку или распахните окно программы.

Как мы уже видели, основное волшебство происходит при загрузке почты с сервера – клиент (компьютер, на котором выполняется PyMailGUI) должен соединиться с сервером (где находится ваша учетная запись электронной почты) через сокет и передать байты через соединения Интернета. Если что-то пойдет не так, PyMailGUI выведет стандартный диалог с описанием ошибки и сообщит, что произошло. Например, если неправильно указать имя учетной записи или пароль (в модуле `mailconfig` или в диалоге ввода пароля), вы увидите сообщение, аналогичное приведенному на рис. 14.11. Здесь выводятся лишь тип и данные исключения Python. Дополнительная информация об ошибках, включая трассировку стека, выводится в стандартный поток вывода (окно консоли).

Рис. 14.11. *Диалог PyMailGUI с сообщением о неверном пароле*

Обработка без подключения к Интернету, сохранение и открытие

Мы видели, как получать электронные письма с сервера и просматривать их, но программу PyMailGUI можно также использовать вообще без подключения к Интернету. Чтобы сохранить почту в локальном файле для последующей работы с ней, выберите нужные сообщения в любом окне со списком и щелкните на кнопке Save (Сохранить) – выбрать для сохранения в виде единого множества можно любое количество сообщений. После щелчка появляется стандартный диалог выбора файла, как

показано на рис. 14.12, после закрытия которого почта сохраняется в выбранном текстовом файле.

Чтобы просмотреть сохраненную почту, щелкните на кнопке Open (Открыть) внизу любого окна со списком и выберите файл в диалоге открытия файла. На экране появится новое окно со списком, который будет заполнен информацией из файла с сохраненными сообщениями, – для больших файлов операция открытия может вызывать небольшую задержку из-за большого объема работы, которую необходимо выполнить. Программа PyMailGUI выполняет загрузку данных из файлов и удаление сообщений в дочерних потоках, чтобы избежать блокирования графического интерфейса, – эти потоки могут перекрываться во времени с другими операциями: открытия файлов с сохраненной почтой, передачи почты между клиентом и сервером и действиями самого графического интерфейса.

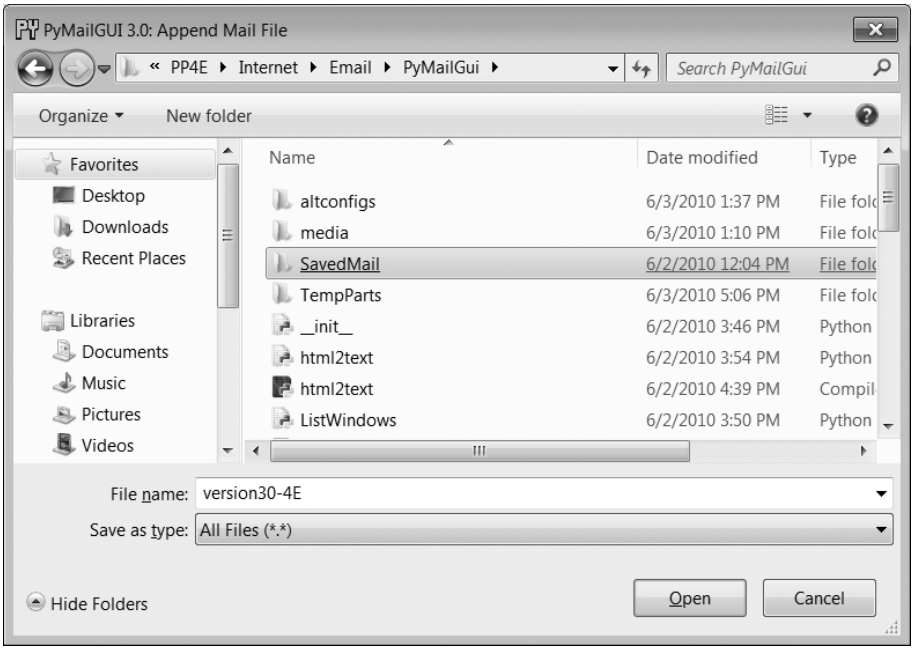


Рис. 14.12. *Диалог сохранения выбранной почты*

Пока в параллельном потоке выполняется загрузка почты из файла, в заголовке соответствующего окна отображается текст «Loading...», как индикатор состояния. Остальной графический интерфейс во время загрузки остается активным (вы можете загружать почту с сервера, удалять ее, просматривать сообщения в других файлах, писать новые сообщения и так далее). По окончании загрузки заголовок окна будет

отображать имя открытого файла. После заполнения в окне появится оглавление содержимого файла, как показано на рис. 14.13 (в этом окне были выбраны три сообщения для обработки).

В целом на экране одновременно могут находиться одно окно со списком сообщений на сервере и любое количество окон с содержимым файлов, в которые сохранялась почта. Окна с оглавлением содержимого файлов, подобные тому, что изображено на рис. 14.13, могут быть открыты в любой момент времени, даже перед получением почты с сервера. Они совершенно идентичны окну с оглавлением входящей почты на сервере, только в них отсутствует полоса вызова справки и кнопка Load (Загрузить), поскольку эти окна не представляют почту на сервере, а все остальные кнопки выполняют операции с файлом, а не с сервером.

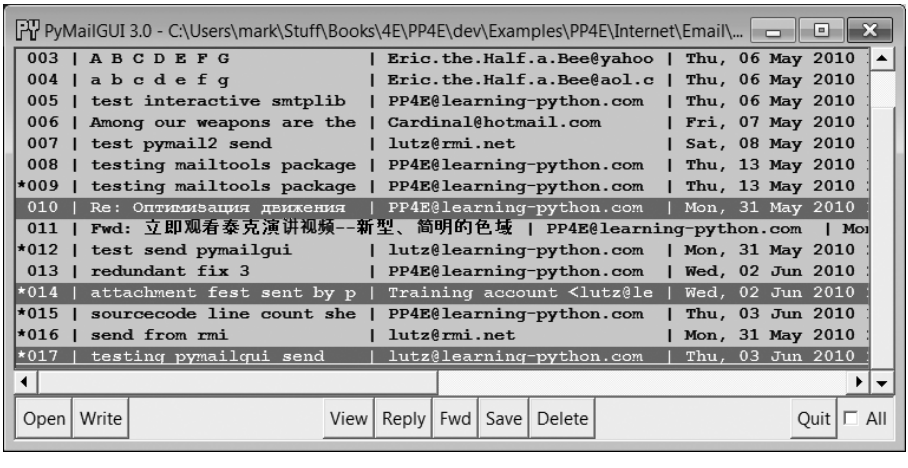


Рис. 14.13. *Окно со списком сообщений, сохраненных в файле, где выбрано несколько сообщений*

Например, кнопка View (Просмотреть) откроет выбранное сообщение в обычном окне просмотра, идентичном тому, что изображено на рис. 14.5, но само сообщение будет извлечено из локального файла. Аналогично кнопка Delete (Удалить) удалит сообщение из файла, а не с сервера. Удаление сообщений из файлов также выполняется в потоке, чтобы избежать блокирования графического интерфейса, – во время выполнения операции в заголовке окна отображается текст «Deleting...», как индикатор выполнения операции. Для индикации состояния операции загрузки почты с сервера и удаления ее на сервере используется диалоговое окно, потому что ожидание может оказаться более длительным и имеется возможность отобразить информацию о ходе выполнения операции (см. рис. 14.8).

Технически операция сохранения всегда дописывает необработанный полный текст сообщения в конец выбранного файла; файл открывается

в режиме 'a' – добавления текста в конец, то есть если файл еще не существует, создается новый файл или выполняется запись в конец существующего файла. Кроме того, операции Save (Сохранить) и Open (Открыть) запоминают последний выбранный каталог – при последующем вызове навигация будет начинаться с выбранного ранее каталога.

Почту можно также сохранить из окна со списком содержимого файла, сохраненного ранее, – для перемещения почты из файла в файл используйте кнопки Save (Сохранить) и Delete (Удалить). Кроме того, операция сохранения в файл, окно с содержимым которого уже было открыто, вызывает автоматическое обновление содержимого этого окна в графическом интерфейсе. То же относится и к файлу, куда автоматически сохраняются отправленные сообщения, который описывается в следующем разделе.

Отправка почты и вложений

После загрузки электронной почты с сервера или из сохраненного ранее файла сообщения можно обрабатывать с помощью кнопок, расположенных внизу окна со списком. Также в любой момент можно отправить новое сообщение, даже до загрузки почты или открытия файла. При нажатии кнопки Write (Написать) в любом окне со списком создается окно, в котором можно составить новое письмо, как показано на рис. 14.14.

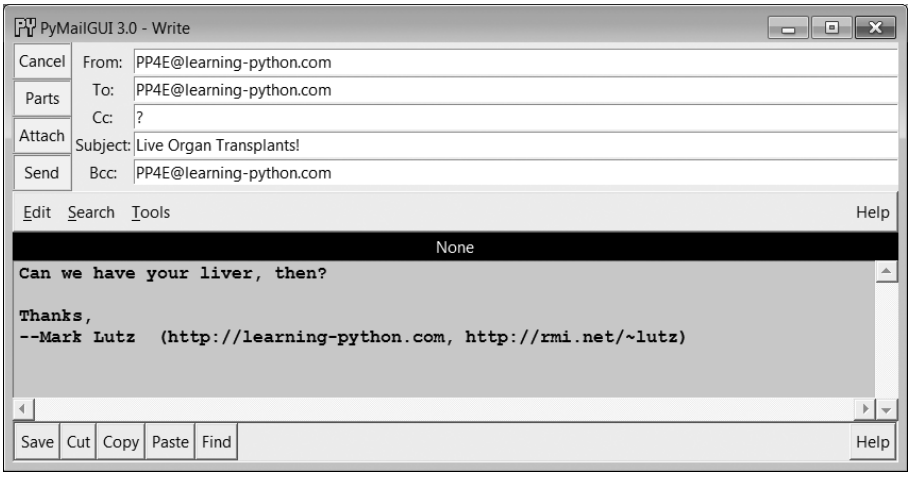


Рис. 14.14. *Окно PyMailGUI для составления сообщения*

Это такое же окно, как для просмотра сообщения, которое мы видели на рис. 14.5, за исключением того, что в нем нет кнопок быстрого доступа к вложениям в середине окна (в данном окне создается новое письмо). В нем есть поля для ввода заголовков, кнопки выполнения операций отправки и управления файлами вложений, добавляемыми в мо-

мент отправки, и встроенный объект TextEditor, в котором можно написать основной текст нового сообщения.

При использовании в таком качестве компонент текстового редактора PyEdit не имеет меню File (Файл), но имеет кнопку Save (Сохранить), что удобно для сохранения черновиков текста сообщения в файл. Позднее вы сможете выделить и вставить эту временную копию в текст нового письма, если потребуется написать текст с самого начала. К тексту сообщения, сохраняемому таким способом, применяются правила поддержки Юникода, реализованные в редакторе PyEdit (он может запросить у вас имя кодировки, как описывается в главе 11).

В операциях создания сообщений PyMailGUI автоматически заполняет строку «From» и вставляет строку подписи (последние две строки на рисунке) согласно настройкам в модуле mailconfig. Вы можете заменить их любым текстом в графическом интерфейсе, но значения по умолчанию автоматически берутся из mailconfig. При отправке сообщения форматирование строки с датой и временем выполняется функцией из модуля email.utils, вызываемой из пакета mailtools, представленного в главе 13.

Здесь также появляется новый набор кнопок в левом верхнем углу: щелчок на кнопке Cancel (Отмена) закроет окно, если пользователь подтвердит эту операцию, а кнопка Send (Отправить) выполнит отправку почты – при ее нажатии текст, введенный в тело этого окна, пересылается после удаления дубликатов по адресам, указанным в заголовках «To», «Cc» и «Bcc», с помощью модуля Python smtplib. PyMailGUI добавляет содержимое полей заголовков в качестве строк почтовых заголовков в отправляемом письме (исключение: письмо будет отправлено по адресам в заголовке «Bcc», но сам заголовок создаваться не будет).

Для отправки письма по нескольким адресам следует перечислить их, разделяя запятой в полях заголовков, при этом можно использовать полную форму адреса в виде «имя» <адрес>. В данном письме я указал в заголовке «To» свой собственный почтовый адрес, чтобы в иллюстративных целях отправить письмо самому себе. Новая версия PyMailGUI также предварительно заполняет заголовок «Bcc» собственным адресом отправителя, если это разрешено настройками в mailconfig – в результате этого копия письма будет отправлена самому отправителю (помимо того, что это письмо будет сохранено в файл отправленных сообщений), но его можно удалить при необходимости.

Новая кнопка Attach (Вложить) открывает диалог выбора файла вложения, как показано на рис. 14.15. Кнопка Parts (Части) открывает диалог, в котором отображается список файлов, уже вложенных в сообщение, как показано на рис. 14.16. При отправке сообщения текст в окне редактирования отправляется как основной текст сообщения, а все присоединенные части отправляются как вложения, корректно закодированные в соответствии с их типами.

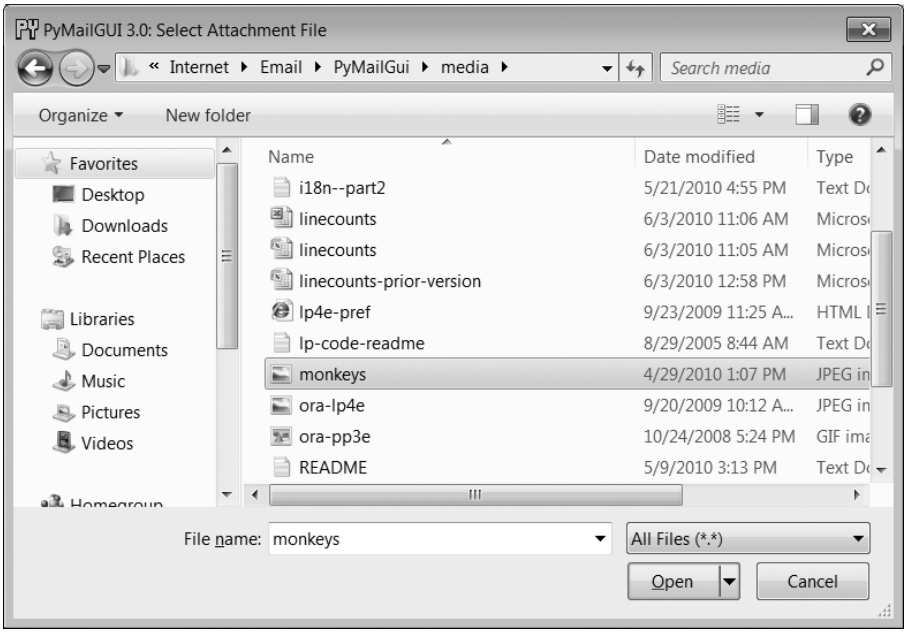


Рис. 14.15. *Диалог выбора файла вложения, открываемый кнопкой Attach (Вложить)*

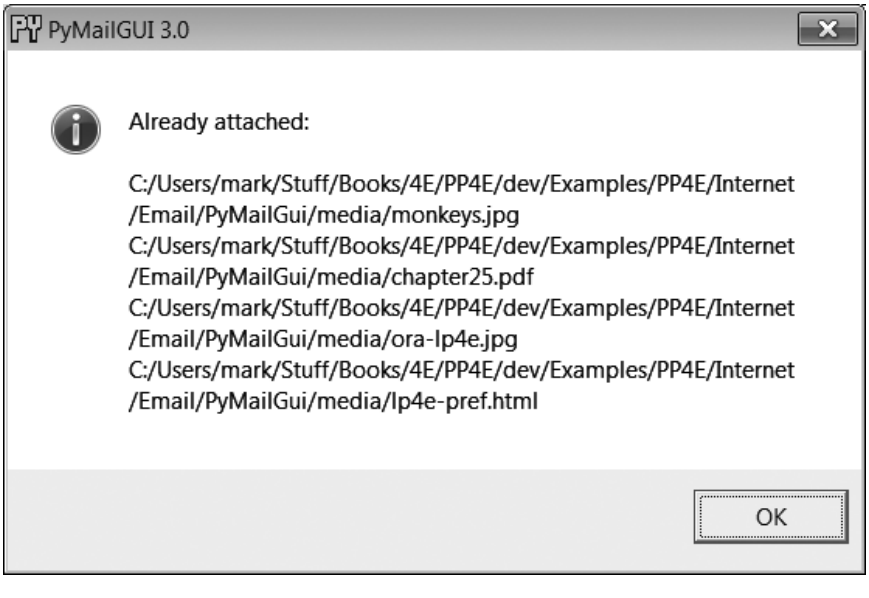


Рис. 14.16. *Диалог со списком вложений, выводимый кнопкой Parts (Части)*

Как мы видели, модуль smtplib в конечном итоге пересылает байты серверу через сокет. Это может оказаться длительной операцией, поэтому PyMailGUI передает ее выполнение дочернему потоку. Во время работы потока отправки на экране отображается немодальное окно, при этом графический интерфейс остается активным – события перерисовки и перемещения обрабатываются в главном потоке программы, в то время как поток отправки взаимодействует с сервером SMTP, и пользователь может выполнять другие операции, включая просмотр и отправку других сообщений.

Если по какой-либо причине Python не сможет отправить письмо по одному из указанных адресов получателей, будет выведено окно с сообщением об ошибке, а затем окно составления письма, чтобы дать вам возможность повторить попытку или сохранить письмо и вернуться к нему позднее. Если окно ошибки не было выведено, значит, все сработало правильно и ваше письмо появится в почтовых ящиках получателей на их почтовых серверах. Так как приведенное выше сообщение я послал себе самому, оно появится в моем ящике и будет загружено при следующем нажатии кнопки Load (Загрузить) в главном окне, как показано на рис. 14.17.

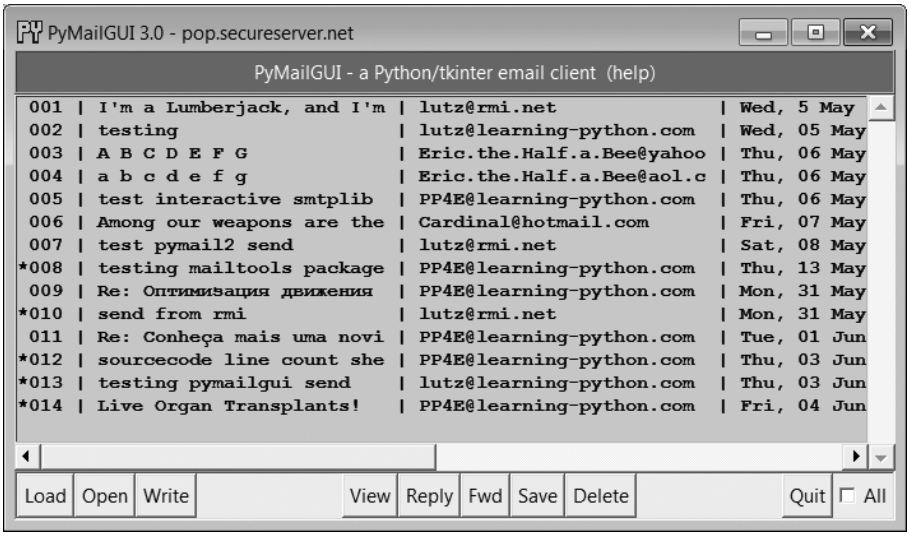


Рис. 14.17. *Главное окно PyMailGUI после загрузки отправленной почты*

Если взглянуть на последний снимок главного окна, можно заметить, что теперь в нем только одно новое письмо – программа PyMailGUI достаточно сообразительна, чтобы загрузить текст заголовка только одного нового письма и поместить его в конец списка загруженной почты. Операция отправки почты автоматически сохраняет отправленные сообщения в файле, имя которого указывается в модуле с настройками.

Для просмотра отправленных сообщений без подключения к Интернету используйте кнопку Open (Открыть), а кнопку Delete (Удалить) – для очистки файла с отправленными сообщениями, если он станет слишком большим (можно также сохранить письма из этого файла в других файлах, рассортировав их по некоторым критериям).

Просмотр электронных писем и вложений

Теперь просмотрим почтовое сообщение, которое было отправлено и получено. PyMailGUI позволяет просматривать почту в форматированном или исходном виде. Сначала выделим (щелчком) сообщение в главном окне, которое требуется просмотреть, а затем нажмем кнопку View (Просмотреть). После этого будет загружен полный текст сообщения (если он не был загружен в кэш ранее), и появится окно просмотра форматированного сообщения, как на рис. 14.18. Если было выбрано несколько сообщений, операция, вызываемая кнопкой View (Просмотреть), загрузит все сообщения, которые пока отсутствуют в кэше (то есть те, которые прежде еще не были загружены), и выведет их в отдельных окнах просмотра. Подобно любым продолжительным операциям загрузка полного текста сообщений производится в параллельных потоках выполнения, чтобы избежать блокирования.

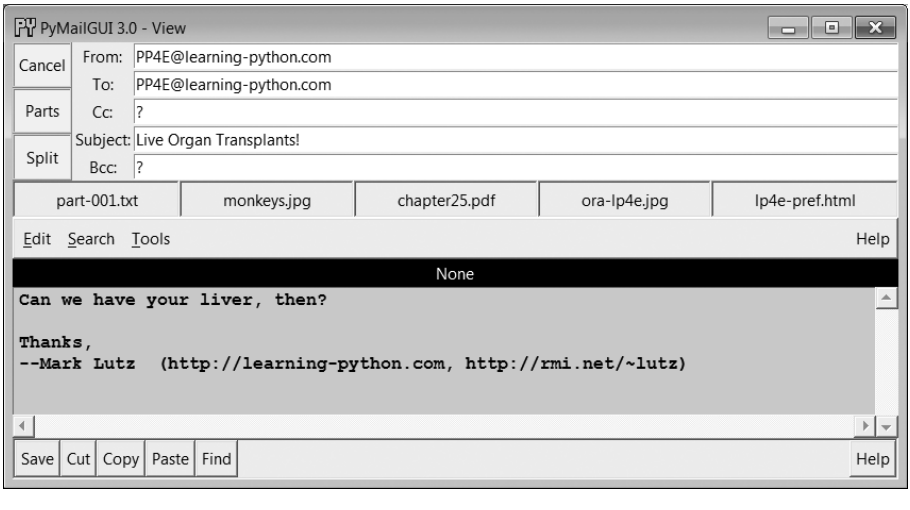


Рис. 14.18. *Окно PyMailGUI просмотра входящего сообщения*

С помощью модуля Python `email` из исходного текста почтового сообщения выделяются заголовки – их текст помещается в поля, справа в верхней части окна. Основной текст сообщения выделяется из его тела и вставляется в новый объект `TextEditor`, находящийся в нижней части окна. Для извлечения основного текста сообщения PyMailGUI использует эвристические алгоритмы – она не стремится вслепую вывести

весь исходный текст сообщения. Почта в формате HTML обрабатывается особым образом, но о подробностях я расскажу немного позже.

Любые другие части сообщения с вложениями открываются и отображаются с помощью кнопок быстрого доступа, находящихся в середине окна. Вложения также перечисляются в диалоге, вызываемом кнопкой Parts (Части), и могут сохраняться и открываться все сразу кнопкой Split (Разбить). На рис. 14.19 изображено диалоговое окно со списком вложений, вызываемое кнопкой Parts (Части), а на рис. 14.20 изображено диалоговое окно выбора каталога для сохранения вложений, вызываемое кнопкой Split (Разбить).

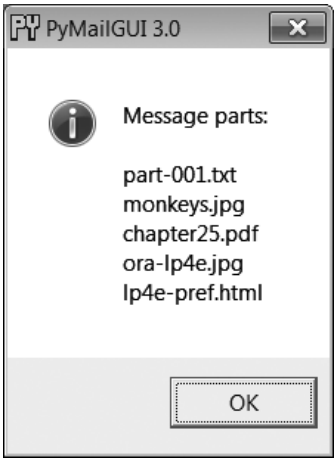


Рис. 14.19. *Щелчок на кнопке Parts (Части) выводит диалог со списком всех частей сообщения*

Когда в диалоговом окне, вызываемом кнопкой Split (Разбить) и изображенном на рис. 14.20, подтверждается выполнение операции, все части сообщения сохраняются в выбранный каталог, а части с известными типами автоматически открываются. Отдельные части после сохранения их во временном каталоге также автоматически открываются кнопками быстрого доступа в средней части окна, которые помечены надписями, соответствующими именам файлов частей; обычно это более удобно, особенно при большом количестве вложений.

Например, на рис. 14.21 показаны два изображения, вложенные в письмо, которое было отправлено на мой ноутбук с Windows, и открытые в стандартной программе просмотра изображений на этой платформе. На других платформах изображения могут открываться в веб-броузере. Щелкните на кнопках быстрого доступа с именами файлов изображений на них, расположенных ниже полей с заголовками, как показано на рис. 14.18, чтобы немедленно просмотреть их, или выполните операцию Split (Разбить), чтобы открыть все части сразу.

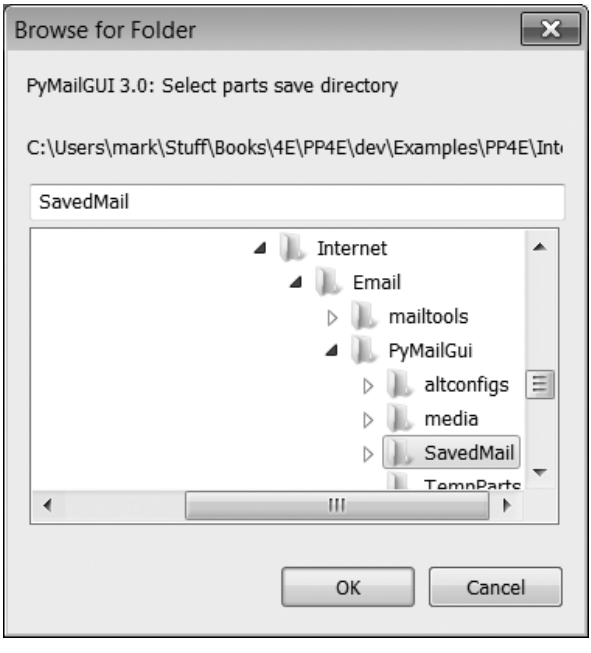


Рис. 14.20. *Окно диалога выбора каталога для сохранения вложений, вызываемого кнопкой Split (Разбить)*

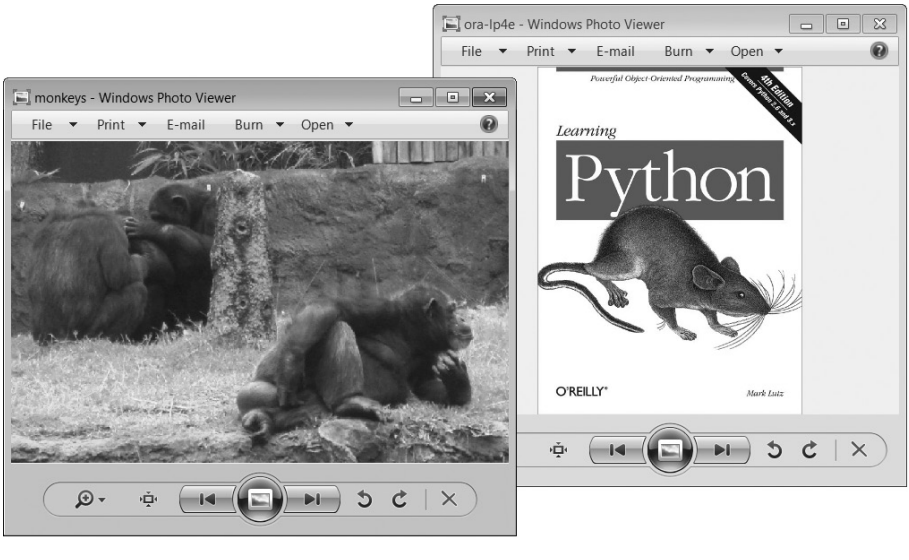


Рис. 14.21. *PyMailGUI открывает части с изображениями в программе просмотра или в броузере*

К этому моменту фотографии во вложениях, изображенные на рис. 14.21, совершили полный круг: они были преобразованы в формат MIME, вложены в письмо и отправлены, получены, проанализированы и преобразованы из формата MIME. Попутно они прошли через множество компьютеров – с компьютера клиента на сервер SMTP, затем на сервер POP и опять на компьютер клиента, покрыв произвольное расстояние.

С точки зрения взаимодействия с пользователем: перед отправкой почты мы вложили изображения в письмо с рис. 14.14, использовав диалог с рис. 14.15. Для доступа к ним позднее мы выбрали письмо для просмотра, как на рис. 14.17, и щелкнули на соответствующих кнопках быстрого доступа с рис. 14.18. Программа PyMailGUI преобразовала фотографии в формат Base64, вставила их в текст письма и позднее извлекла и преобразовала в оригинальные фотографии. Благодаря инструментам для работы с электронной почтой в языке Python и нашему программному коду, опирающемуся на них, все работает именно так, как ожидалось.

Обратите внимание: как показано на рис. 14.18 и 14.19, основной текст сообщения также считается одной из частей письма – при ее выборе открывается окно PyEdit, изображенное на рис. 14.22, где этот текст можно обработать и сохранить (основной текст сообщения можно также сохранить, щелкнув на кнопке Save (Сохранить) в окне просмотра сообщения). Основная часть включается в письмо, потому что не все письма имеют текстовую часть. Для сообщений, основная часть которых представлена разметкой HTML, программа PyMailGUI отображает простой текст, извлеченный из текста HTML в своем собственном окне, и открывает веб-броузер для просмотра письма в формате HTML. Подробнее о почтовых сообщениях в формате HTML мы поговорим ниже.

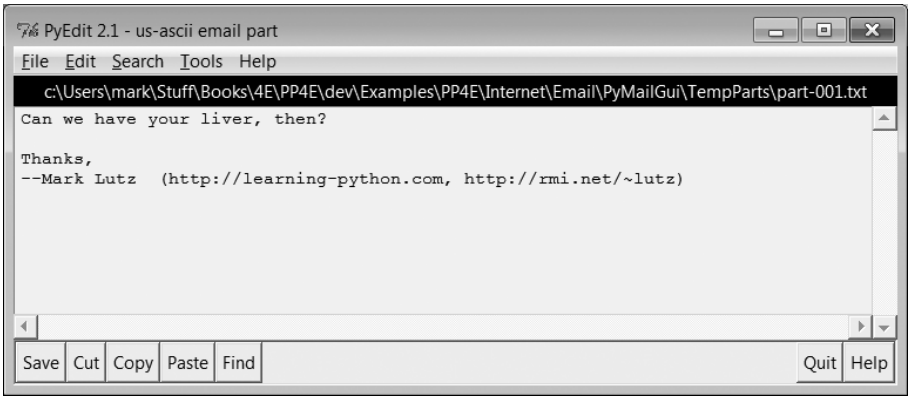


Рис. 14.22. *Основная текстовая часть, открытая в PyEdit*

Помимо изображений и простого текста PyMailGUI также открывает вложения в формате HTML и XML в веб-броузере и использует реестр Windows для открытия документов, типы которых хорошо известны Windows. Например, файлы *.doc* и *.docx*, *.xls* и *.xlsx*, и *.pdf* обычно открываются в Word, Excel и Adobe Reader соответственно. На рис. 14.23 изображен ответ на нажатие кнопки быстрого доступа к файлу *lp4e-pref.html* на рис. 14.18 на моем ноутбуке с Windows. Если вы внимательно рассмотрите этот снимок или сами выполните эту операцию, то сможете заметить, что вложение в формате HTML отображается одновременно и в веб-броузере, и в окне PyEdit – последнее можно запретить в `mailconfig`, но по умолчанию эта особенность включена, чтобы дать представление о преобразовании формата HTML.

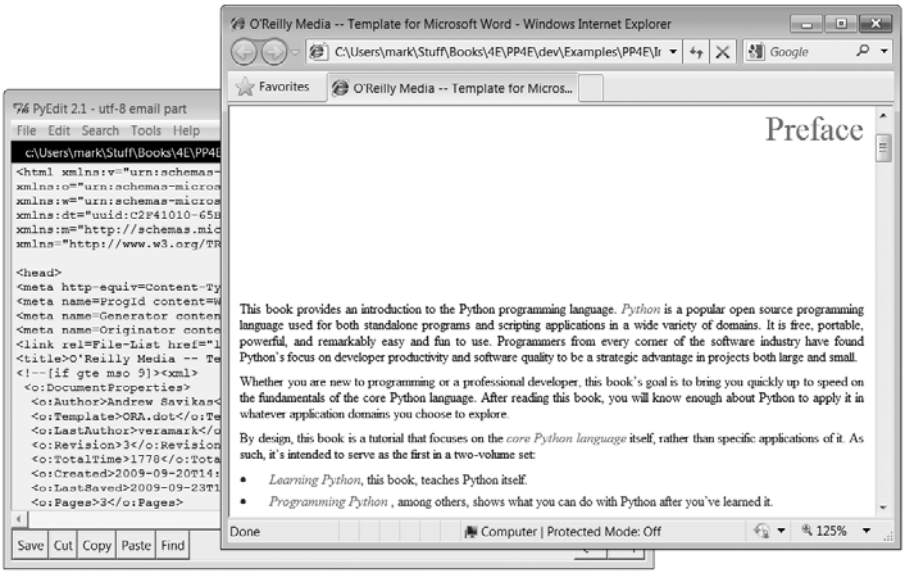


Рис. 14.23. *Вложение в формате HTML, открытое в веб-броузере*

Кнопки быстрого доступа в середине окна просмотра на рис. 14.18 предоставляют более непосредственный доступ к вложениям, чем кнопка Split (Разбить), – они не требуют выбирать каталог для сохранения и позволяют открывать только те вложения, которые требуется просмотреть. Однако кнопка Split (Разбить) позволяет открыть сразу все части за один шаг, выбрать каталог для сохранения и поддерживает произвольное количество частей. Файлы, которые не могут быть открыты автоматически из-за того, что их тип неизвестен операционной системе, можно исследовать в каталоге сохранения после щелчка на кнопке Split (Разбить) или на кнопке быстрого доступа (путь к каталогу указывается

в диалоге, который открывается при попытке обратиться к файлу вложения неизвестного типа).

Если количество частей больше заданного максимального числа, полоса кнопок быстрого доступа завершается кнопкой с меткой «...», щелчок на которой просто запускает операцию Split (Разбить) для сохранения и открытия непоместившихся частей. Одно из таких писем представлено на рис. 14.24 – его можно найти в каталоге *SavedMail*, в файле *version30-4E*, если вы захотите просмотреть его. Это достаточно сложное письмо, содержащее 11 частей различного типа.

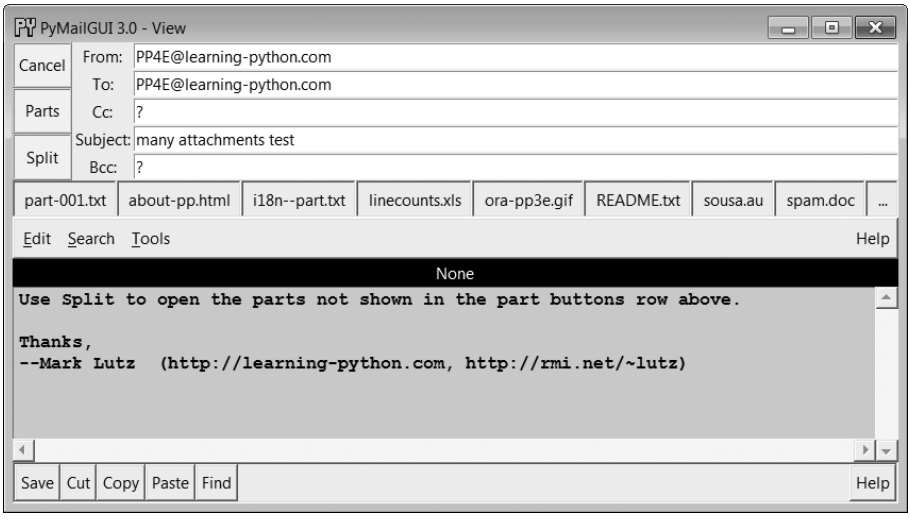


Рис. 14.24. *Окно просмотра сообщения, содержащего множество частей*

Как и многое в PyMailGUI, максимальное количество кнопок быстрого доступа к частям сообщения в окне просмотра можно настроить в модуле `mailconfig.py` с пользовательскими настройками. При создании снимка на рис. 14.24 этот параметр был равен восьми. На рис. 14.25 показано, как выглядит то же самое письмо после того, как параметру, определяющему максимальное число кнопок быстрого доступа, было присвоено число пять. Этому параметру можно присвоить число больше восьми, но в некоторый момент надписи на кнопках могут стать нечитаемыми (используйте кнопку Split (Разбить) вместо этого).

Как пример реакции на вложения разных типов, на рис. 14.26 и 14.27 показано, что произошло при выборе кнопок с надписями *sousa.au* и *chapter25.pdf* на рис. 14.24 и 14.18 на моем ноутбуке с Windows. Результат может отличаться для разных компьютеров – аудиофайл открывается в проигрывателе Windows Media Player, в то время как файлы MP3 открываются в проигрывателе iTunes, а некоторые платформы могут открывать такие файлы непосредственно в веб-броузере.

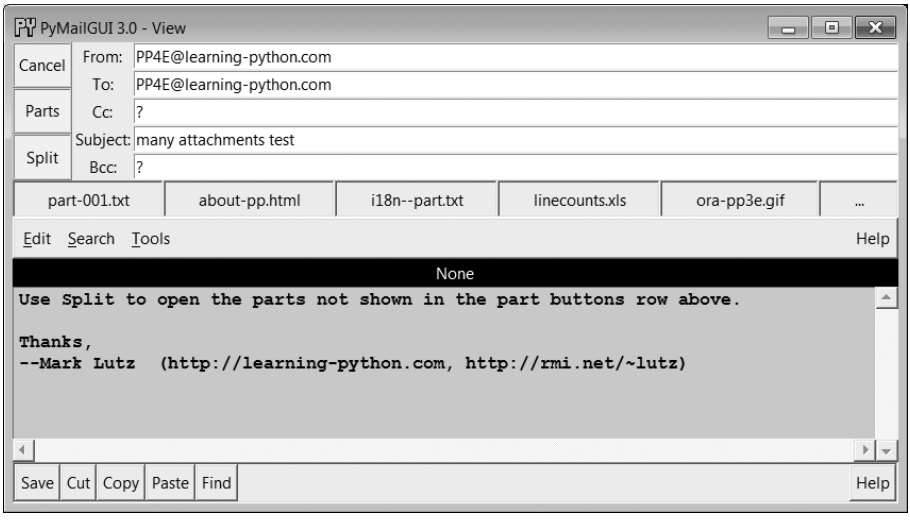


Рис. 14.25. *Окно просмотра после того, как максимальное количество кнопок частей было уменьшено*

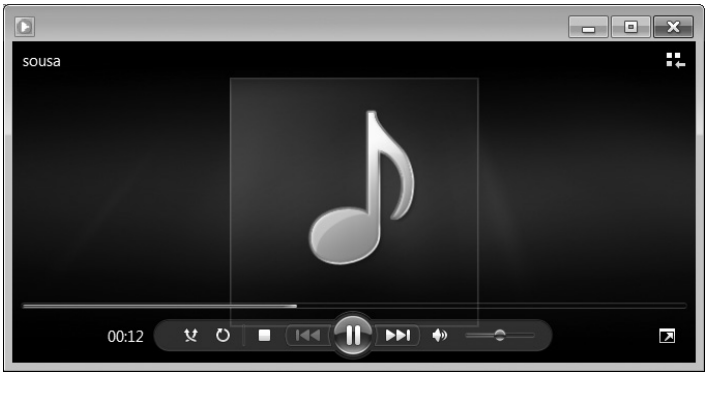


Рис. 14.26. *Вложенный аудиофайл, открытый программой PyMailGUI*

Помимо окна просмотра с красивым форматированием PyMailGUI позволяет увидеть необработанный текст почтового сообщения. Двойной щелчок в главном окне на строке сообщения открывает простой неформатированный текст сообщения (при этом полный текст сообщения загружается в отдельном потоке выполнения, если он не был загружен и помещен в кэш ранее). Исходный текст сообщения, которое я послал сам себе на рис. 14.18, показан на рис. 14.28 – в этом издании исходный текст сообщения отображается в отдельном окне редактора PyEdit (прежний инструмент просмотра текста в окне с прокруткой по-прежнему доступен, но PyEdit предоставляет дополнительные инструменты, такие как поиск, сохранение и так далее).

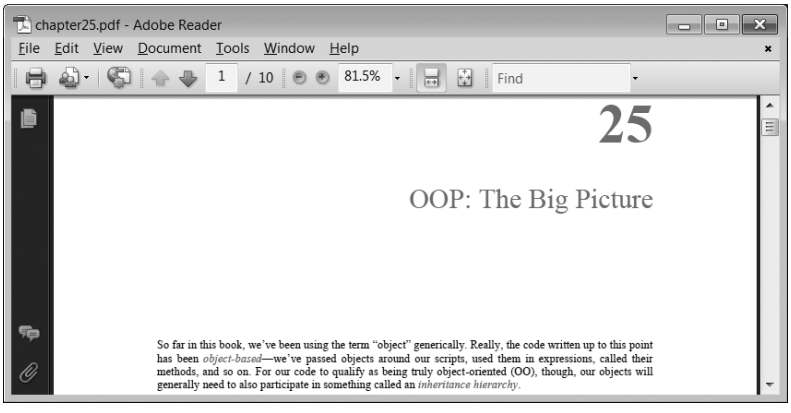
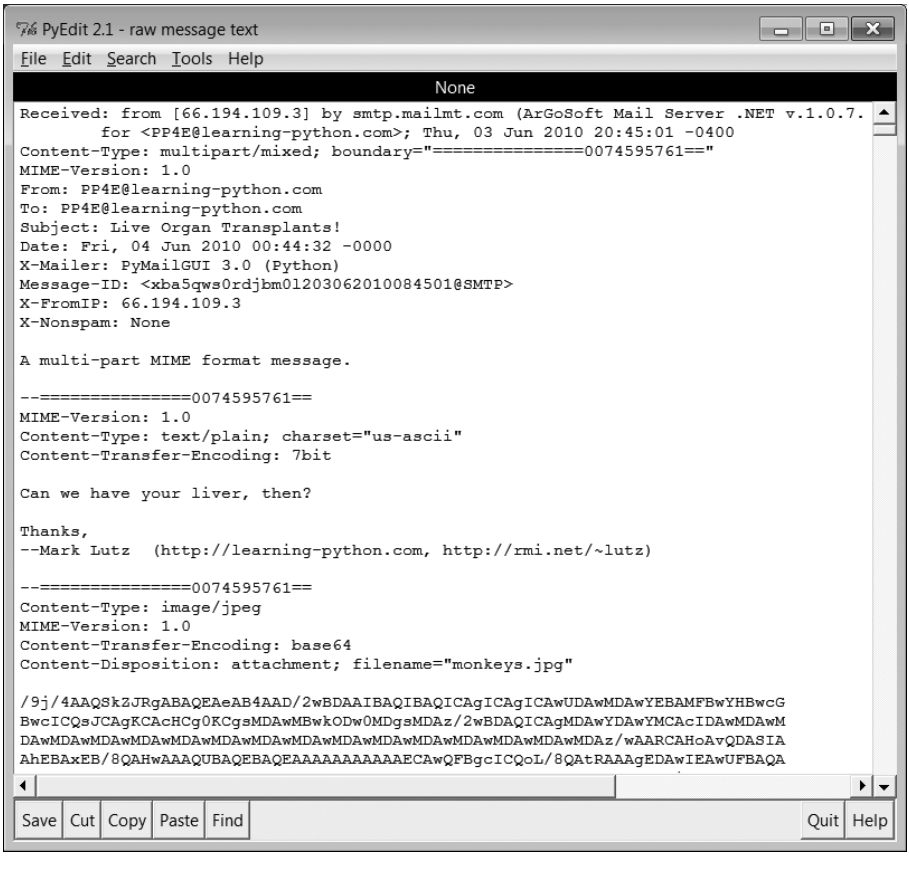


Рис. 14.27. *Вложенный файл PDF, открытый программой PyMailGUI*




```
Received: from [66.194.109.3] by smtp.mailmt.com (ArGoSoft Mail Server .NET v.1.0.7.
        for <PP4E@learning-python.com>; Thu, 03 Jun 2010 20:45:01 -0400
Content-Type: multipart/mixed; boundary="===============0074595761==="
MIME-Version: 1.0
From: PP4E@learning-python.com
To: PP4E@learning-python.com
Subject: Live Organ Transplants!
Date: Fri, 04 Jun 2010 00:44:32 -0000
X-Mailer: PyMailGUI 3.0 (Python)
Message-ID: <xba5qws0rdjbm01203062010084501@SMTP>
X-FromIP: 66.194.109.3
X-Nonspam: None

A multi-part MIME format message.

--===============0074595761==
MIME-Version: 1.0
Content-Type: text/plain; charset="us-ascii"
Content-Transfer-Encoding: 7bit

Can we have your liver, then?

Thanks,
--Mark Lutz  (http://learning-python.com, http://rmi.net/~lutz)

--===============0074595761==
Content-Type: image/jpeg
MIME-Version: 1.0
Content-Transfer-Encoding: base64
Content-Disposition: attachment; filename="monkeys.jpg"

/9j/4AAQSkZJRgABAQEAeAB4AAD/2wBDAAIBAQIBAQICAgICAgICAwUDAwMDAwYEBAMFBwYHBwcG
BwcICQsJCAgKCAcHCg0KCgsMDAwMBwkODw0MDgsMDAz/2wBDAQICAgMDAwYDAwYMCAcIDAwMDAwM
DAwMDAwMDAwMDAwMDAwMDAwMDAwMDAwMDAwMDAwMDAwMDAwMDAz/wAARCAHoAvQDASIA
AhEBAxEB/8QAHwAAAQUBAQEBAQEAAAAAAAAAAAECAwQFBgcICQoL/8QAtRAAAgEDAwIEAwUFBAQA
```

Рис. 14.28. *Окно PyMailGUI (PyEdit) с исходным текстом*

Такое отображение исходного текста может оказаться полезным для просмотра специальных заголовков, не отображаемых в окне форматированного просмотра. Например, необязательный заголовок «X-Mailer» в исходном тексте идентифицирует программу, отправившую сообщение. PyMailGUI автоматически добавляет этот заголовок в сообщения, наряду со стандартными заголовками, такими как «From» и «To». Другие заголовки добавляются в процессе передачи сообщения: заголовки «Received» указывают имена компьютеров, через которые прошло сообщение на пути к нашему почтовому серверу, а заголовок «Content-Type» добавляется и анализируется пакетом Python `email` в ответ на требование программы PyMailGUI

В действительности исходный текст представляет все, что относится к почтовому сообщению, – это то, что передается с компьютера на компьютер, когда письмо отправлено. Красиво отформатированное отображение в окнах просмотра графического интерфейса получается в результате анализа и декодирования компонентов исходного текста сообщения с помощью стандартных средств Python и помещения их в соответствующие поля на экране. Обратите, например, внимание на файл изображения в формате Base64 в конце сообщения на рис. 14.28 – это вложение было создано при отправке, передано через Интернет и преобразовано в оригинальное двоичное представление при извлечении. Огромный объем работы, значительная часть которой выполняется автоматически, благодаря программному коду и библиотекам.

Ответ на сообщения, пересылка и особенности адресации

PyMailGUI позволяет не только читать почту и создавать новые письма, но также пересылать входящую почту и отвечать на письма, отправленные другими. Обе эти операции являются обычными операциями составления сообщения, но включают оригинальный текст и предварительно заполняют заголовки. Чтобы ответить на электронное письмо, выделите его в списке главного окна и щелкните на кнопке Reply (Ответить). Если я отвечу на письмо, которое только что послал самому себе (отдает нарциссизмом, но служит целям демонстрации), появится окно составления письма, изображенное на рис. 14.29.

Формат этого окна идентичен тому, который мы видели для операции Write (Написать), за исключением того, что некоторые части автоматически заполняются программой PyMailGUI. Фактически единственное, что я руками добавил в этом окне, – это первая строка в текстовом редакторе, все остальное было автоматически заполнено программой PyMailGUI:

- Строка «From» устанавливается в соответствии с вашим почтовым адресом в модуле `mailconfig`.
- Строка «To» инициализируется адресом «From» исходного сообщения (мы ведь отвечаем тому, кто послал сообщение).

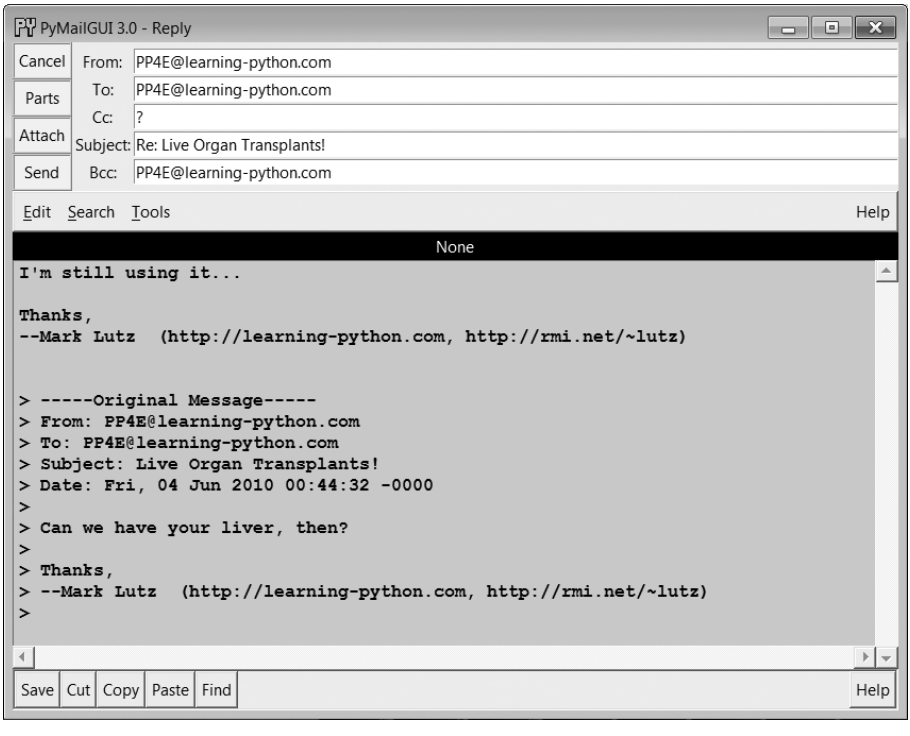


Рис. 14.29. *Окно PyMailGUI составления письма*

- В строку «Subject» записывается тема исходного сообщения, в начало которой добавляется последовательность «Re:» – стандартный формат продолжения строки темы (если только она была указана, при этом в любом регистре символов).
- Необязательная строка «Bcc», если разрешено в `mailconfig`, также заполняется адресом отправителя, так как этот способ часто используется для получения копии (новое в этой версии).
- В тело сообщения помещаются строка подписи из `mailconfig`, а также текст оригинального сообщения. Текст оригинального сообщения цитируется с помощью символов >, и перед ним помещается несколько строк заголовка исходного сообщения, чтобы дать представление о контексте.
- В данном примере не показана еще одна новинка в этой версии – заголовок «Cc» ответа также заполняется списком получателей оригинального письма, составляемым из содержимого заголовков «To» и «Cc» оригинального сообщения, из которого удаляются дубликаты и ваш адрес. Иными словами, операция «ответить» в действительности является по умолчанию операцией «ответить всем» – ответ посылается отправителю, а всем остальным получателям ориги-

нального сообщения отправляются копии. Поскольку последнее не всегда желательно, настройки в mailconfig можно изменить так, что при ответе будет заполняться только заголовок «To» с адресом отправителя. При необходимости можно также просто очищать предварительно заполненный заголовок «Cc», но если эта особенность запрещена, вам может потребоваться добавлять адреса в заголовок «Cc» вручную. Как действует предварительное заполнение заголовка «Cc», мы увидим ниже.

К счастью, все это реализуется значительно проще, чем может показаться. С помощью стандартного пакета Python email извлекаются все строки заголовков исходного сообщения, а все действия по добавлению в тело исходного сообщения символов цитирования > выполняет строковый метод replace. Я просто ввожу с клавиатуры все, что хочу сообщить в ответ (первый абзац в текстовой области письма), и нажимаю кнопку Send (Отправить), чтобы снова отправить ответное сообщение в почтовый ящик на моем почтовом сервере. Физически отправка ответа происходит точно так же, как отправка нового сообщения, – почта направляется на сервер SMTP в дочернем потоке отправки почты, при этом на экране до окончания выполнения операции отображается информационное окно.

Пересылка сообщения аналогична ответу: выделите сообщение в главном окне, нажмите кнопку Fwd (Переслать) и заполните поля и область текста появившегося окна составления сообщения. На рис. 14.30 изображено окно, создаваемое для пересылки письма, полученного ранее, после редактирования.

Как и при создании ответа, заголовок «From» заполняется адресом отправителя из mailconfig, первоначальный текст в теле сообщения автоматически цитируется, заголовок «Bcc» заполняется тем же адресом, что и заголовок «From», а в строку темы копируется тема исходного сообщения с добавлением в начало последовательности «Fwd:». При желании все эти строки можно изменить вручную перед отправкой. Однако строку «To» я должен заполнить вручную, поскольку это не прямой ответ – письмо не обязательно возвращается первоначальному отправителю. Далее, заголовок «Cc» не заполняется адресами получателей оригинального сообщения, как это делается при создании ответа, потому что пересылка письма не является продолжением дискуссии в группе.

Обратите внимание, что я пересылаю письмо в три разных адреса (два адреса в заголовке «To» и один вручную введен в заголовок «Bcc»). Кроме того, я использовал полный формат адресов «имя <адрес>». Адреса получателей в заголовках «To», «Cc» и «Bcc» разделяются запятой («,») и программа PyMailGUI благополучно распознает адреса в полной форме записи, где бы они ни встретились, включая настройки в модуле mailconfig. Запятая, которую можно увидеть в первом адресе в заголовке «To» на рис. 14.30, не вызывает конфликтов с запятой, разделяющей адреса получателей, потому что строки с адресами в этой версии подвергаются полному анализу. Кнопка Send (Отправить) в этом окне отправ-

ляет пересылаемое письмо по всем адресам, перечисленным в упомянутых заголовках, после удаления из них дубликатов, чтобы избежать отправки нескольких копий одному и тому же получателю.

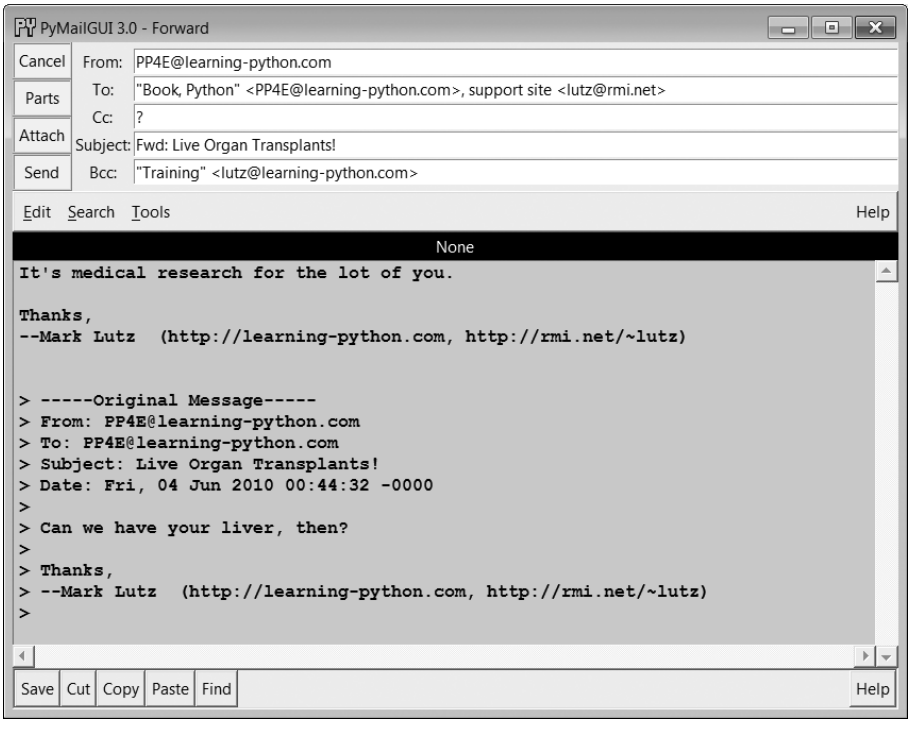


Рис. 14.30. *Окно PyMailGUI для пересылки письма*

Теперь я написал новое письмо, ответил на него и переслал его. Ответ и пересылаемое письмо также были отправлены на мой почтовый адрес. Если снова нажать кнопку Load (Загрузить), ответ и пересланное письмо должны показаться в списке главного окна. На рис. 14.31 они показаны как сообщения с номерами 15 и 16 (порядок их следования может зависеть от особенностей синхронизации вашего сервера, кроме того, я растянул окно по горизонтали, чтобы сделать видимым заголовок «To» в последнем из них).

Имейте в виду, что PyMailGUI выполняется на локальном компьютере, но сообщения, которые вы видите в главном окне, фактически находятся в почтовом ящике на сервере. Каждый раз, когда мы нажимаем кнопку Load (Загрузить), программа PyMailGUI загружает с сервера на ваш компьютер заголовки вновь поступивших сообщений, но не удаляет их с сервера. Три сообщения, которые мы только что написали (с 14 по 16), появятся и в любой другой почтовой программе, которой вы воспользуетесь со своей учетной записью (например, в Outlook или в системе

электронной почты с веб-интерфейсом). Программа PyMailGUI не удаляет сообщения при загрузке, а просто записывает их в память вашего компьютера для обработки. Если теперь выделить сообщение 16 и нажать кнопку View (Просмотреть), можно увидеть сообщение, которое мы переслали, как показано на рис. 14.32.

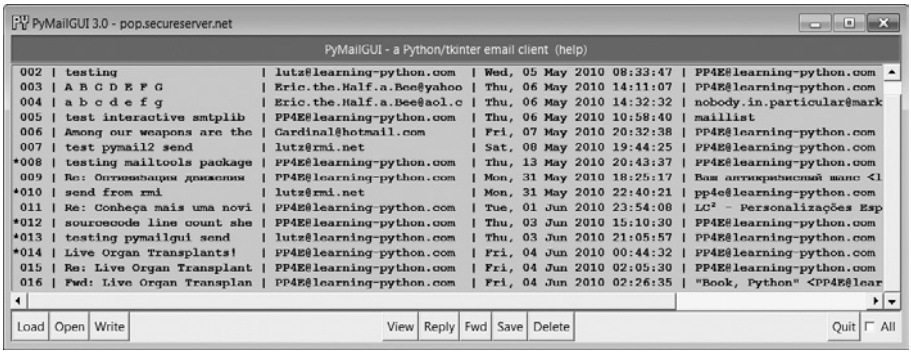


Рис. 14.31. *Список писем в PyMailGUI после отправок и загрузки*

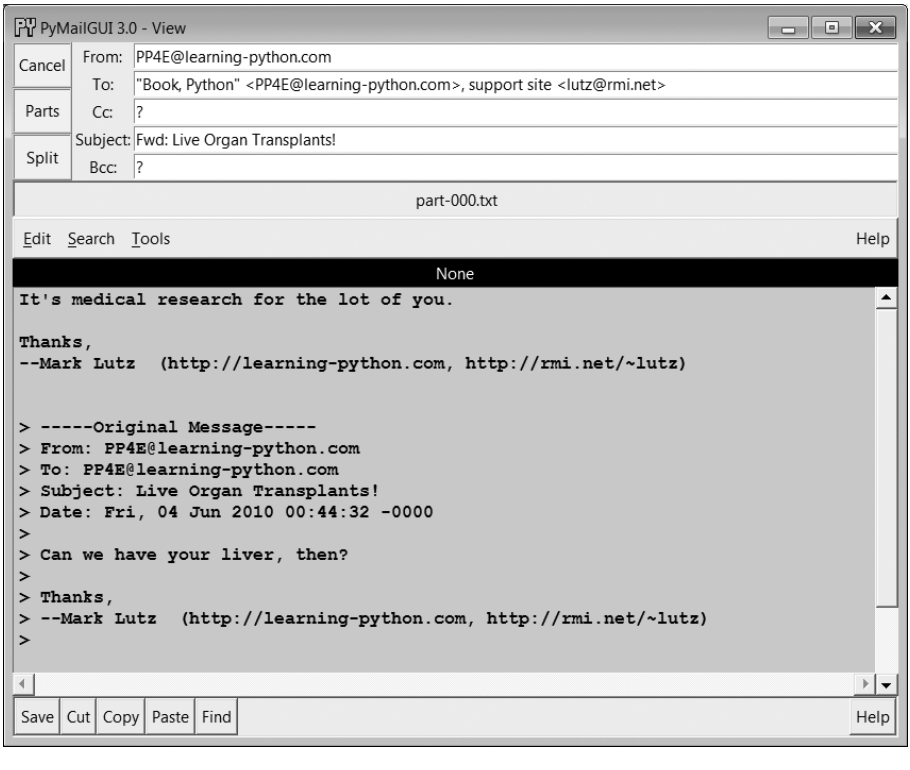


Рис. 14.32. *Окно PyMailGUI просмотра пересланного сообщения*

Это сообщение ушло с моего компьютера на удаленный почтовый сервер и было загружено с него в список Python, из которого и показано. Фактически оно поступило на три различные учетные записи электронной почты, имеющиеся у меня (две другие появятся ниже в этом разделе – см. рис. 14.45). Третий получатель не виден на рис. 14.32, потому что его адрес находился в заголовке «Bcc» (скрытая копия) – он примет сообщение, но его адрес не включается в строку заголовка сообщения.

На рис. 14.33 показано, как выглядит необработанный текст пересланного сообщения. Чтобы вывести это окно, нужно выполнить двойной щелчок по записи в главном окне. Форматированное отображение с рис. 14.32 – это просто результат извлечения элементов из текста, приведенного в окне с необработанным текстом.

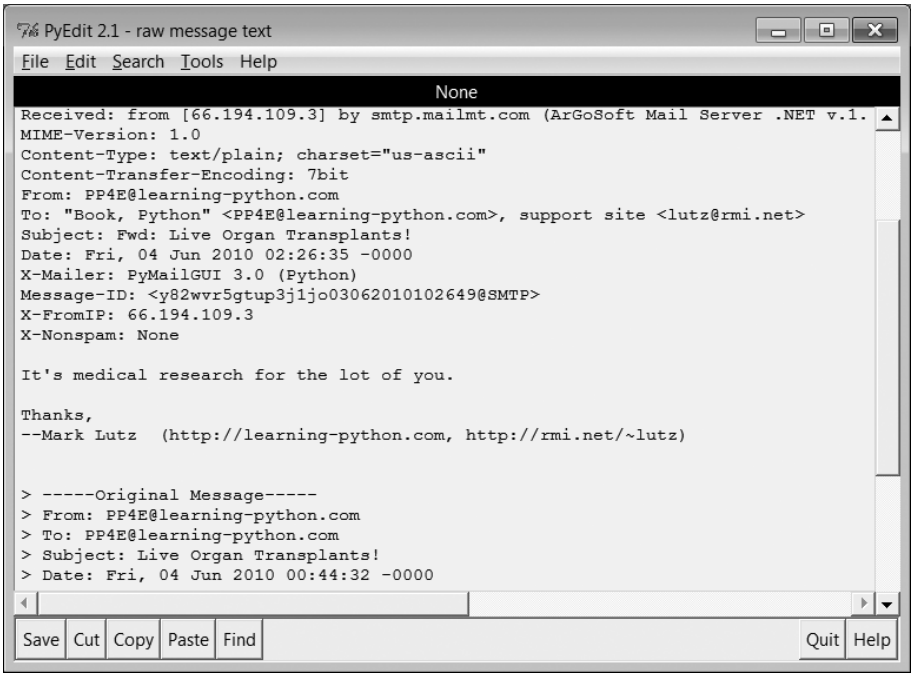


Рис. 14.33. *Окно PyMailGUI просмотра исходного текста пересланного сообщения*

И последнее замечание, касающееся ответов и пересылки: как уже упоминалось, в этой версии предусматривается отправка ответов всем получателям оригинального письма, исходя из предположения, что ответ означает продолжение дискуссии в группе. Для иллюстрации в верхней части рис. 14.34 показано оригинальное сообщение, ниже и левее – пересылаемое сообщение, а ниже и правее – ответ. При создании ответа

заголовок «Cc» автоматически заполняется адресами получателей оригинального сообщения, из которых удаляются дубликаты, и устанавливается новый адрес отправителя. Заголовок «Bcc» (здесь он разрешен) также в обоих случаях заполняется адресом отправителя. Это лишь начальные установки полей заголовков, которые можно редактировать и очищать перед отправкой. Кроме того, предварительное заполнение заголовка «Cc» в ответах можно полностью запретить в файле с настройками. Однако в этом случае вам может потребоваться вручную копировать и вставлять адреса в этот заголовок при ведении обсуждения в группе. Откройте файл с почтой, сохраненной в этой версии программы, поэкспериментируйте с этой особенностью программы и затем посмотрите предлагаемые усовершенствования, которые приводятся далее в этой главе.

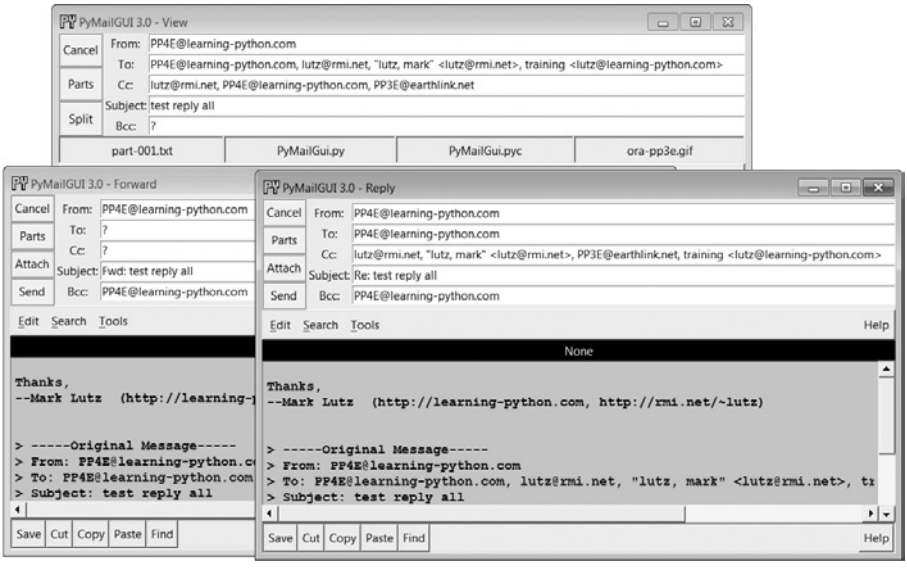


Рис. 14.34. *Операция «ответить всем» заполняет заголовок «Cc»*

Удаление сообщений

К настоящему времени мы осветили все, кроме кнопки Delete (Удалить) и флажка All (Все) в главном окне. Флажок All (Все) просто управляет выбором всех сообщений (операции, выполняемые при нажатии кнопок View (Просмотреть), Delete (Удалить), Reply (Ответить), Fwd (Переслать) и Save (Сохранить) применяются ко всем выделенным на этот момент сообщениям). Программа PyMailGUI позволяет также удалять сообщения с сервера, после чего они больше не будут появляться в списке при обращении к почтовому ящику.

Операция удаления действует точно так же, как операции просмотра и сохранения, только для ее выполнения нужно нажать кнопку Delete (Удалить). Обычно я время от времени удаляю почту, которая мне не интересна, и сохраняю с последующим удалением сообщения, которые важны для меня. Операцию сохранения мы уже рассматривали выше.

Подобно операциям просмотра, сохранения и другим операция удаления может применяться к одному или более сообщениям. Удаление производится немедленно и подобно другим операциям передачи почты выполняется в неблокирующем потоке, но только после подтверждения в диалоге, изображенном на рис. 14.35. В процессе выполнения операции удаления на экран выводится диалог, отображающий ход выполнения операции, подобный тем, что изображены на рис. 14.8 и 14.9.

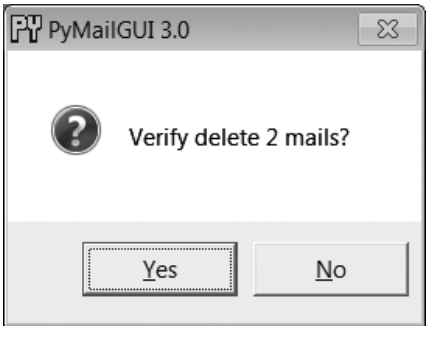


Рис. 14.35. *PyMailGUI запрашивает подтверждение операции удаления*

По умолчанию почта никогда не удаляется автоматически: те же самые сообщения будут показаны при очередном запуске PyMailGUI. Почта удаляется с сервера, только если потребовать этого и только при подтверждении в последнем показанном диалоге (это последняя возможность предотвратить удаление почты навсегда). После удаления производится обновление оглавления, и сеанс работы с программой продолжается.

До своего завершения операция удаления блокирует запуск операции загрузки почты и другие операции удаления и сама не может выполняться параллельно с другими операциями загрузки или удаления, так как эти операции могут изменять номера POP-сообщений и содержимое списка в главном окне (кроме того, они могут изменять содержимое внутреннего кэша электронной почты). Однако в процессе выполнения операции удаления сохраняется возможность составлять новые сообщения и обрабатывать файлы с сохраненной почтой.

Номера POP-сообщений и синхронизация

К настоящему моменту мы познакомились со всеми основными возможностями программы PyMailGUI; этого вполне достаточно, чтобы начать отправлять и получать простые текстовые сообщения. В оставшейся части этой демонстрации мы рассмотрим некоторые более глубинные концепции этой системы, включая синхронизацию почтового ящика, сообщения в формате HTML, интернационализацию и настройку нескольких учетных записей. Поскольку первая из этих тем перекликается с темой удаления почты, рассматривавшейся в предыдущей главе, рассмотрим ее в первую очередь.

С точки зрения конечного пользователя в операции удаления нет ничего сложного, но, как оказывается, удаление осложняется схемой POP нумерации сообщений. В главе 13 мы узнали о потенциальных ошибках синхронизации между почтовым ящиком на сервере и списком сообщений в локальном приложении, когда изучали пакет `mailtools`, используемый программой PyMailGUI (обсуждение приводится вокруг примера 13.24). В двух словах: в POP каждому сообщению присваивается последовательный номер, начиная с единицы. Эти номера и передаются серверу для получения и удаления сообщений. Почтовый ящик на сервере обычно блокируется на время соединения с сервером, чтобы серия операций удаления могла выполняться как единая операция – никакие другие изменения в ящике невозможны, пока соединение не будет закрыто.

Однако изменение номеров сообщений имеет определенное значение и для графического интерфейса. Нет ничего плохого, если в то время, когда выводится результат предыдущей загрузки, поступают новые письма – им будут присвоены более высокие номера по сравнению с теми, которые отображаются у клиента. Но если удалить сообщение, находящееся в середине почтового ящика, после того как оглавление будет загружено с сервера, номера всех сообщений, расположенных после удаленного, изменятся (они уменьшатся на единицу). Это означает, что некоторые сообщения могут иметь неверные номера, если удаление осуществляется во время просмотра ранее загруженной почты

Для решения этой проблемы PyMailGUI корректирует все отображаемые номера после операции удаления и просто удаляет ненужную запись из оглавления и из кэша. Однако этого недостаточно, чтобы обеспечить синхронизацию графического интерфейса с почтовым ящиком на сервере, так как содержимое почтового ящика может измениться в позиции, предшествующей позиции удаления, – выполнением операций удаления с помощью других почтовых клиентов (или даже в другом сеансе работы с PyMailGUI) или в результате удаления, выполняемого самим почтовым сервером (например, сообщения, которые определяются как недоставленные, автоматически удаляются из почтового ящика). Такие изменения, инициированные за рамками сеанса работы с PyMailGUI, редки, но возможны.

Для подобных случаев в PyMailGUI используется безопасное удаление и проверка синхронизации, реализованные в `mailtools`. Этот пакет выполняет сопоставление заголовков сообщений для определения ошибок синхронизации локального списка и почтового ящика на сервере. Например, если другой почтовый клиент удалит сообщение, предшествующее удаляемому в PyMailGUI, `mailtools` обнаружит проблему, отменит операцию удаления и выведет диалог, подобный тому, что изображен на рис. 14.36.

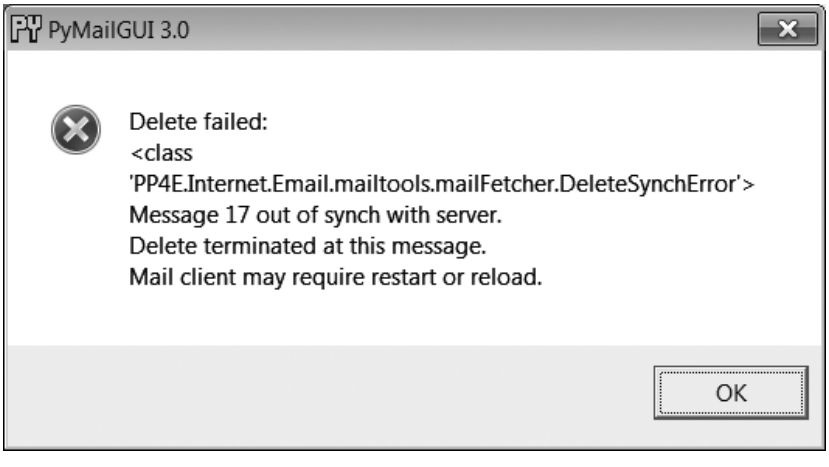


Рис. 14.36. *Операция безопасного удаления обнаружила различия с почтовым ящиком*

Аналогично, загрузка оглавления почтового ящика и получение отдельных сообщений также сопровождаются запуском процедуры проверки синхронизации в `mailtools`. На рис. 14.37 изображен диалог, сообщающий об ошибке, обнаруженной при получении сообщения, которое было удалено с помощью другого клиента после загрузки списка сообщений с сервера. Тот же диалог отображается, когда ошибка обнаруживается во время операции загрузки списка, только в первой строке выводится текст «Load failed» (Ошибка загрузки).

В обоих случаях, когда обнаруживается ошибка синхронизации, PyMailGUI автоматически выполняет повторную загрузку списка сообщений из почтового ящика на сервере сразу же после закрытия диалога. Такой подход гарантирует, что программа PyMailGUI не будет по ошибке удалять или отображать не те сообщения в тех редких случаях, когда содержимое почтового ящика на сервере будет изменяться без ее ведома. Подробнее о процедуре проверки синхронизации в пакете `mailtools` рассказывается в главе 13. Эти ошибки обнаруживаются, и исключения возбуждаются в пакете `mailtools`, но весь механизм задействуется вызовами в менеджере кэша почты в этой программе.

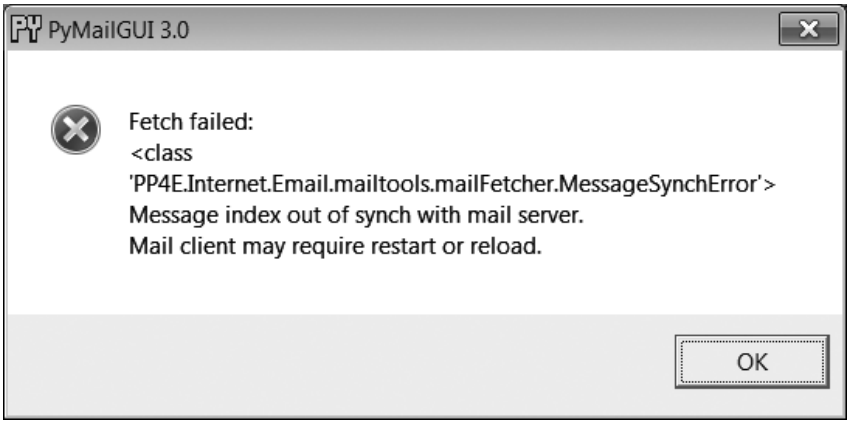


Рис. 14.37. *Ошибка синхронизации после удаления письма с помощью другого клиента*

Обработка содержимого электронной почты в формате HTML

До настоящего момента мы рассматривали базовые операции PyMailGUI в контексте простых текстовых сообщений. Мы также видели, что она может обрабатывать вложения в формате HTML, но мы еще не сталкивались со случаем, когда основной текст сообщения является разметкой HTML. В настоящее время использование формата HTML для представления основной части электронных писем стало обычным делом. Поскольку компонент редактора PyEdit, используемый программой PyMailGUI, опирается на виджет `Text` из библиотеки tkinter, ориентированный на простой текст, содержимое в формате HTML приходится обрабатывать особым образом:

- Для почтовых сообщений с альтернативной частью типа text/HTML PyMailGUI отображает в окне просмотра часть с простым текстом и добавляет кнопку, позволяющую по требованию открыть разметку HTML в веб-броузере.
- Для почтовых сообщений, содержащих только разметку HTML, в области отображения основного текста выводится простой текст, извлеченный из разметки HTML простым механизмом синтаксического анализа (не исходная разметка HTML), а сам текст в формате HTML автоматически открывается в веб-броузере.

Во всех случаях отображение интернациональных символов, присутствующих в разметке HTML, в веб-броузере зависит от информации о кодировках, имеющейся в тегах HTML, от механизма определения кодировки или от настроек пользователя. Корректно оформленные части в формате HTML уже содержат теги «<meta>» в разделах «<head>»,

определяющие имя кодировки, но они могут быть определены некорректно или вообще отсутствовать. Подробнее о поддержке интернационализации мы поговорим в следующем разделе.

На рис. 14.38 представлен случай просмотра альтернативной части типа text/HTML, а на рис. 14.39 показано, что происходит при просмотре сообщения, содержащего только текст в формате HTML. На рис. 14.38 веб-броузер был открыт щелчком на кнопке, соответствующей части в формате HTML, – этот случай ничем не отличается от примера с вложением в формате HTML, который мы видели выше.

Однако сообщения, содержащие только разметку HTML, обрабатываются в этой версии иначе: окно просмотра, на рис. 14.39 слева, отображает результат извлечения простого текста из разметки HTML, отображаемой в веб-броузере позади этого окна. Механизм синтаксического анализа HTML, используемый для этого, представляет собой первоначальный прототип, но даже те результаты, которые он способен воспроизвести, предпочтительнее, чем отображение исходной разметки HTML в окне просмотра. Для простейших почтовых сообщений в формате HTML, которые обычно отправляются отдельными лицами, а не компаниями, занимающимися массовыми рассылками рекламы (подобными той, что показана здесь), результаты при тестировании получаются в целом неплохие, хотя время покажет, как этот прототип будет чувствовать себя в суровых джунглях нестандартной разметки HTML – дальнейшее его улучшение весьма желательно.

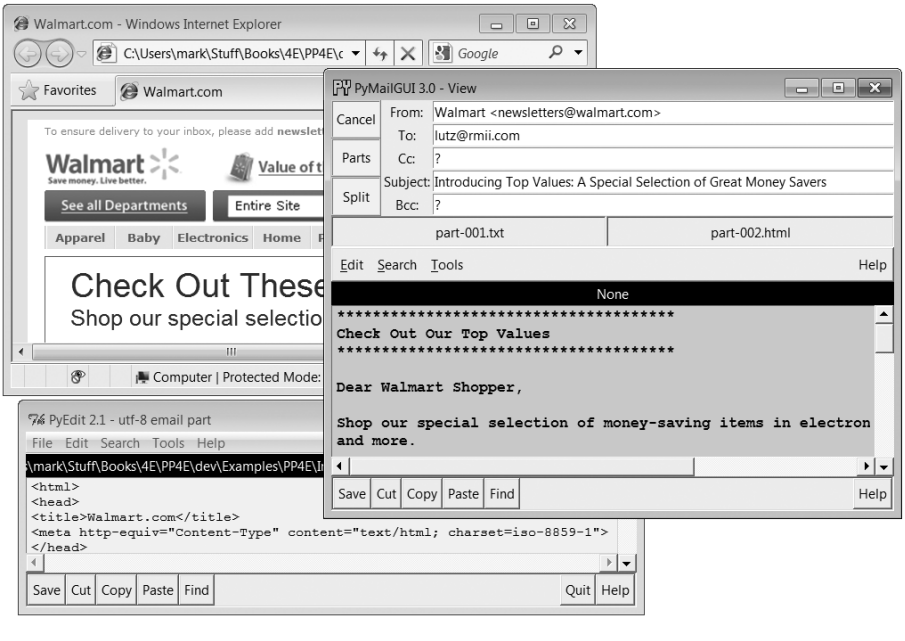


Рис. 14.38. *Просмотр сообщений с альтернативной частью типа text/HTML*

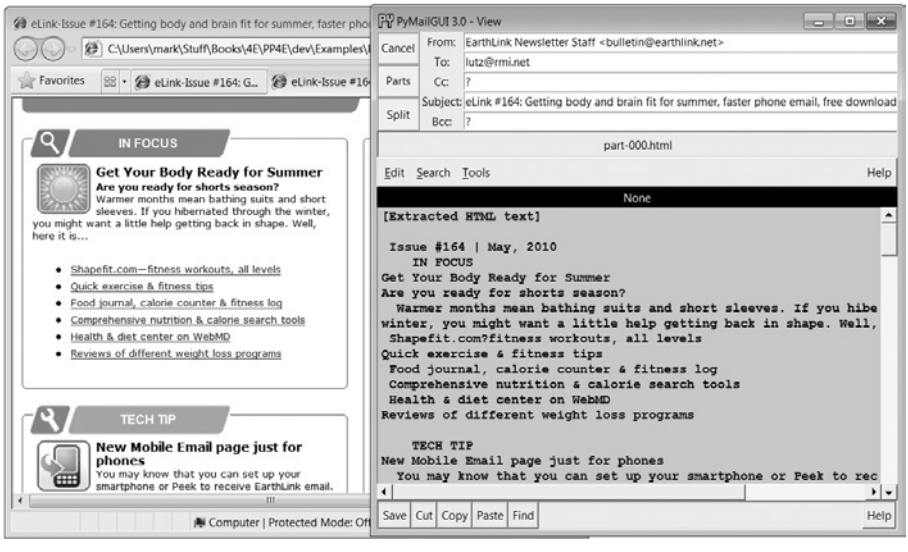


Рис. 14.39. *Просмотр сообщений, содержащих только разметку HTML*

Одно замечание: программа PyMailGUI в настоящее время может отображать разметку HTML в веб-броузере и извлекать из нее простой текст, но она не может отображать разметку HTML непосредственно в своем собственном окне и не поддерживает ее редактирование. Данная функциональность относится к разряду расширений, ожидающих, когда свое внимание им уделят другие программисты, которые посчитают это полезным.

Поддержка интернационализации содержимого

Наша следующая особенность является следствием неизбежного успеха Интернета. Как описывалось выше, в разделе, где перечислялись нововведения в версии 3.0, программа PyMailGUI полностью поддерживает интернациональные наборы символов в почтовых сообщениях – текстовое содержимое и заголовки декодируются перед отображением и кодируются перед отправкой в соответствии со стандартами электронной почты, MIME и Юникода. Это, пожалуй, самое заметное новшество в данной версии программы. К сожалению, это сложно отразить на снимках с экрана, но вы можете получить более полное представление об этой особенности, открыв следующий файл с сохраненными почтовыми сообщениями, включенный в состав примеров, и просмотрев сообщения из него в форматированном и в исходном виде, попробовав создавать на их основе ответы или пересылаемые сообщения, и так далее:

```
C:\...\PP4E\Internet\Email\PyMailGui\SavedMail\i18n-4E
```

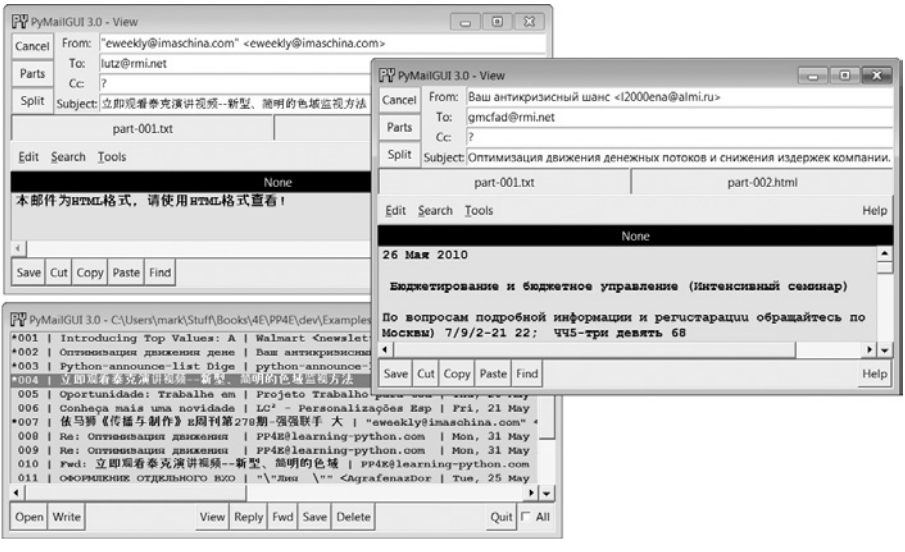


Рис. 14.40. *Поддержка интернационализации, заголовки и тело декодированы для отображения*

Для демонстрации особенностей этой поддержки на рис. 14.40 показана ситуация, когда данный файл был открыт для разнообразия с одной из альтернативных настроек учетной записи, описываемых в следующем разделе. На этом снимке изображено окно со списком сообщений и окна просмотра с сообщениями на русском и китайском языках, отправленными на мой почтовый адрес (эти письма рекламного характера не имеют какого-то особого значения, но вполне пригодны для тестирования). Обратите внимание, что и заголовки сообщений, и их содержимое было корректно декодировано для отображения как в окне со списком, так и в окнах просмотра сообщений.

На рис. 14.41 показаны фрагменты исходного текста двух сообщений, полученные двойным щелчком на соответствующих им элементах списка (вы можете просмотреть эти сообщения, открыв указанный выше файл, если не сможете разглядеть какие-то детали на снимке в книге). Обратите внимание, что тело обоих сообщений представлено в кодированном виде в соответствии с требованиями стандартов MIME и Юникода – заголовки вверху и текст внизу в этих окнах представлен в виде строк в формате Base64 и quoted-printable, которые необходимо декодировать, чтобы получить форматированный результат, изображенный на рис. 14.40.

Информация в заголовках текстовых частей описывает схемы кодирования их содержимого. Например, значение `charset="gb2312"` в заголовке «Content-type» идентифицирует набор символов Юникода китайского языка, а заголовок «Content-Transfer-Encoding» определяет формат MIME (например, base64).

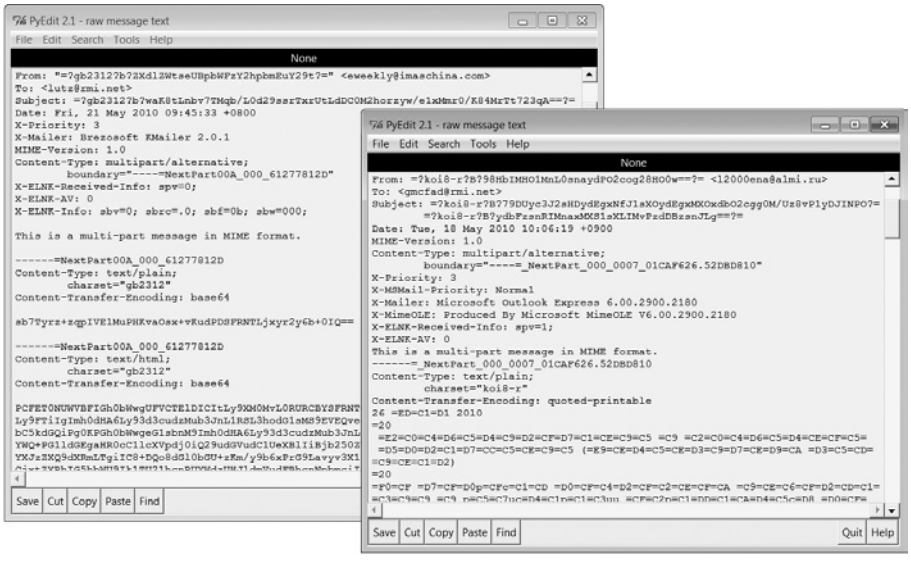


Рис. 14.41. *Исходный текст интернационализированных сообщений, закодированные заголовки и тело*

Здесь также присутствуют заголовки, кодированные в соответствии со стандартами поддержки интернационализации, – их содержимое само описывает формат MIME и кодировку Юникода. Например, префикс `=?koi8-r?B` в заголовке означает текст на русском языке в формате Base64. Программа PyMailGUI достаточно интеллектуальна, чтобы декодировать полные заголовки и компоненты имен в адресах независимо от того, закодированы они полностью (как показано здесь) или просто содержат кодированные подстроки (как в других почтовых сообщениях, хранящихся в файле *version30-4E* в каталоге *SavedMail*).

Дополнительно на рис. 14.42 показано, как выглядят основные части сообщений после их открытия с помощью соответствующих кнопок быстрого доступа. Содержимое частей было сохранено в файл в двоичном режиме после преобразования из формата MIME, а компоненты PyEdit в окнах просмотра выполняют их декодирование в соответствии с именами кодировок, полученными из заголовков сообщений. Как мы узнали в главах 9 и 11, стоящая за этим библиотека tkinter обычно отображает декодированные строки `str` лучше, чем двоичные строки `bytes`.

К настоящему моменту мы увидели, как PyMailGUI отображает интернационализированные сообщения, но эта программа позволяет также отправлять их и выполняет любые операции кодирования, предполагаемые текстовым содержимым. На рис. 14.43 демонстрируется результат запуска операций создания ответа и пересылки сообщения на русском языке, при этом, чтобы оградить случайного адресата от его получения, был изменен адрес в заголовке «To». Заголовки в окне просмотра

были декодированы перед отображением, закодированы перед отправкой и вновь декодированы перед отображением. Текстовые части в теле письма точно так же были декодированы, закодированы и вновь декодированы, кроме того, были декодированы цитированные заголовки и оригинальный текст, начинающиеся с символов «>» и добавленные в конец сообщения.

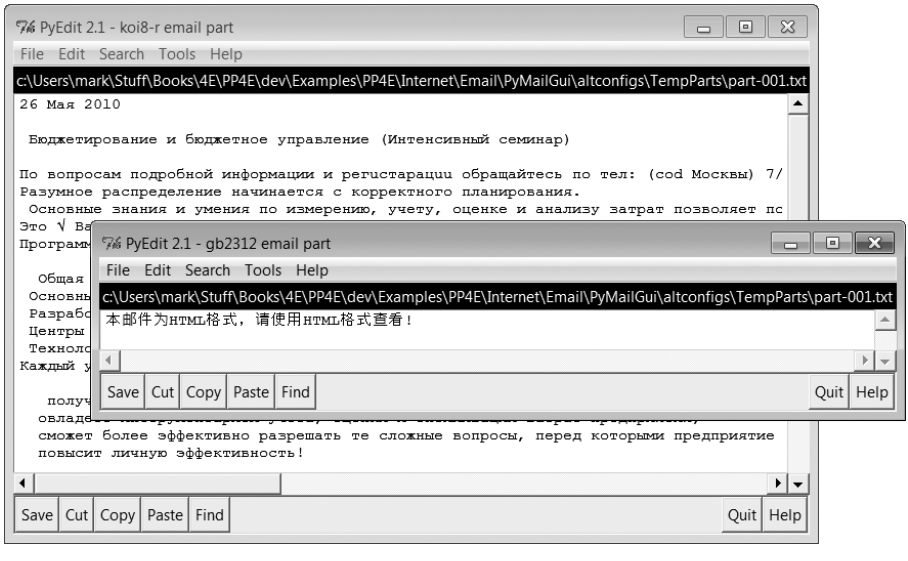


Рис. 14.42. *Основные текстовые части интернационализированных сообщений, декодированные компонентами PyEdit*

И наконец, на рис. 14.44 демонстрируется фрагмент исходного текста ответа на русском языке, форматированное представление которого можно видеть в правом нижнем окне на рис. 14.43. Щелкните дважды на соответствующем элементе списка, чтобы рассмотреть его во всех подробностях. Обратите внимание, что и заголовки, и текст тела сообщения были закодированы в соответствии со стандартами электронной почты и MIME.

Согласно действующим настройкам текст в теле сообщения перед передачей всегда кодируется с применением кодировки UTF-8 и преобразуется в формат MIME, если его не удается закодировать как текст ASCII – настройке по умолчанию, заданной в модуле `mailconfig`. При желании можно определить другие настройки по умолчанию, в соответствии с которыми текст будет кодироваться перед отправкой. Фактически, текст, который не может вставляться непосредственно внутрь полного текста сообщения, преобразуется в формат MIME точно так же, как двоичные вложения, такие как изображения.

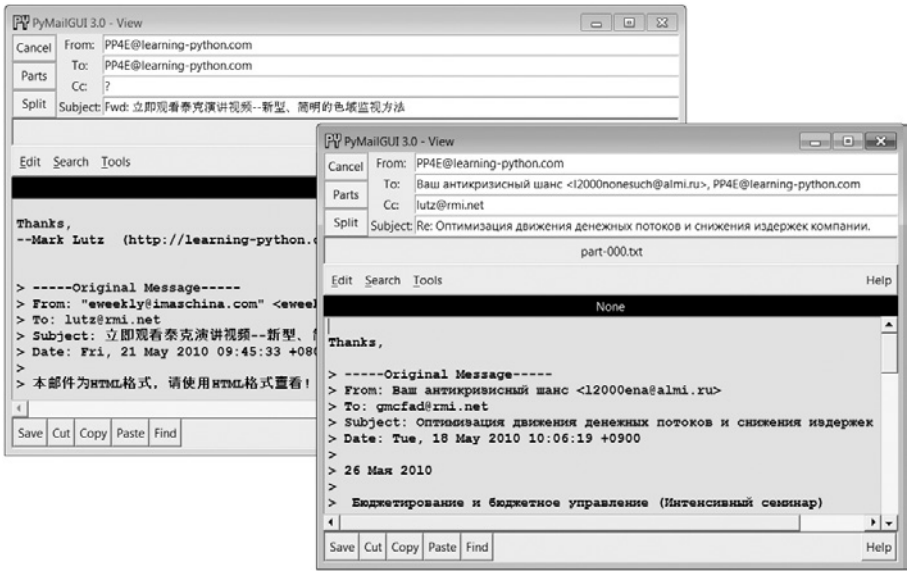


Рис. 14.43. *Результат выполнения операций ответа и пересылки сообщений с интернациональными символами, повторно декодированными*

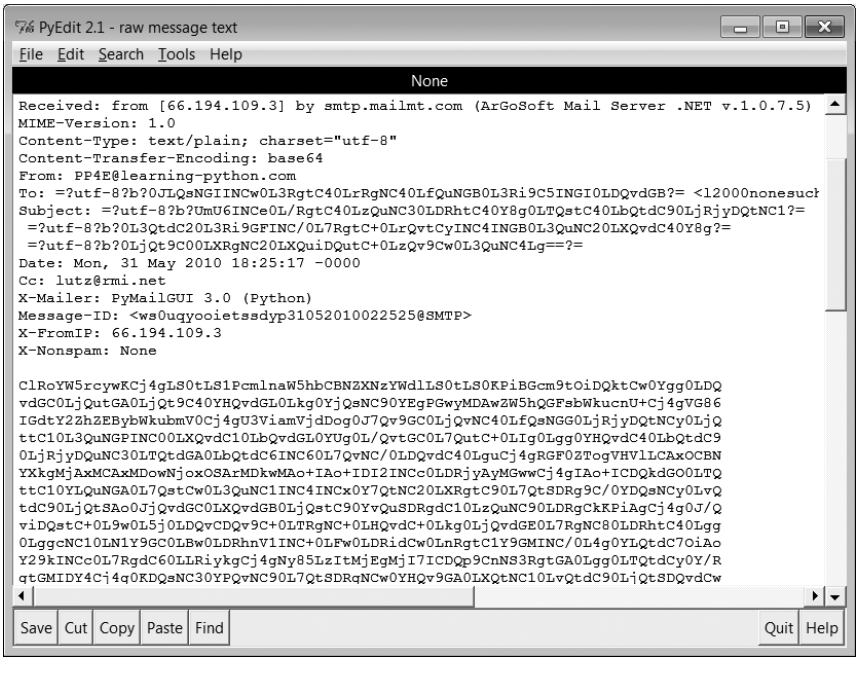


Рис. 14.44. *Исходный текст ответа на русском языке, заголовки и тело письма закодированы повторно*

То же относится и к неинтернационализированным наборам символов. Например, текстовая часть сообщения на английском языке, содержащая кавычки не из диапазона символов ASCII, будет закодирована с применением кодировки UTF-8 и преобразована в формат Base64, точно так же, как сообщение на рис. 14.44, при этом предполагается, что программа чтения почты получателя должна будет выполнить декодирование (все современные клиенты электронной почты поддерживают такую возможность). Это позволяет встраивать текст с символами вне диапазона ASCII в полный текст отправляемого сообщения.

Заголовки отправляемого сообщения тоже кодируются с применением кодировки UTF-8, если они содержат символы вне диапазона ASCII, благодаря чему они могут вставляться в полный текст сообщения. Фактически, если внимательно рассмотреть это сообщение, можно заметить, что заголовок «Subject» первоначально был закодирован с применением кириллической кодировки, но теперь он закодирован с применением кодировки UTF-8 – это новое представление воспроизводит те же самые символы (кодовые пункты) при декодировании для отображения.

Проще говоря, несмотря на то, что графический интерфейс по-прежнему остается англоязычным (метки и прочие надписи), операции отображения и отправки электронных писем поддерживают произвольные наборы символов. Декодирование для отображения, где это возможно, выполняется в зависимости от конкретного содержимого, с использованием информации в заголовках для текстового содержимого сообщения и содержимого в заголовках для самих заголовков. Кодирование для отправки выполняется в соответствии с пользовательскими настройками и правилами, с использованием пользовательских настроек или ввода или кодировки UTF-8 по умолчанию. Обязательное преобразование в формат MIME и кодирование заголовков электронной почты выполняется в значительной мере автоматически пакетом `email`, лежащим в основе программы.

Здесь не показаны диалоги, которые могут выводиться для запроса кодировок текстовых частей во время отправки, если это определено настройками в `mailconfig`. Подобные диалоги при определенных пользовательских настройках выводит и редактор PyEdit. Некоторые из этих пользовательских конфигураций задумывались для иллюстрации и для полноты картины. Предопределенные настройки прекрасно подходят для большинства ситуаций, с которыми мне приходилось сталкиваться, но у вас могут оказаться иные требования к поддержке интернационализации. Чтобы получить более полное представление, поэкспериментируйте с сообщениями, хранящимися в файле, на своем компьютере и рассмотрите исходный программный код системы.

Альтернативные конфигурации и учетные записи

До сих пор мы видели, как PyMailGUI действует с учетной записью электронной почты, созданной мной для демонстрации примеров в кни-

ге, но точно так же легко можно настроить в модуле `mailconfig` и другие учетные записи, а также другие визуальные эффекты. Например, на рис. 14.45 изображена ситуация, когда программа PyMailGUI работает с тремя различными учетными записями, используемыми мной для примеров в книгах и на учебных занятиях. Все три экземпляра здесь являются независимыми процессами. Все окна со списками отображают сообщения для разных учетных записей, и каждое из них имеет свои настройки внешнего вида и поведения. Окно просмотра сообщения вверху было открыто из окна со списком сообщений на сервере, находящегося внизу слева; к нему применены собственная цветовая схема и настройки отображения заголовков.

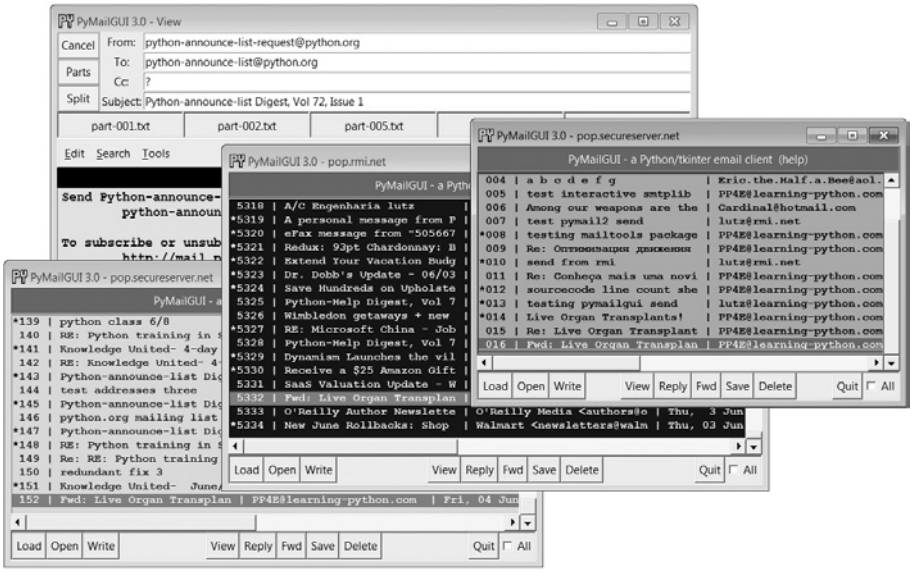


Рис. 14.45. *Альтернативные учетные записи и настройки*

Вы всегда можете изменить модуль `mailconfig` для определенной учетной записи, если она у вас одна, однако ниже будет показано одно из возможных решений на основе подкаталога `altconfigs`, позволяющее настраивать множество учетных записей, как в данном примере, не вмешиваясь в исходный код программы. Решение `altconfigs` отображает окна, как показано на рис. 14.45, и удовлетворительно работает в качестве интерфейса запуска – смотрите его реализацию далее.

Многооконный интерфейс и сообщения о состоянии

В довершение всего отметим, что программа PyMailGUI создана с поддержкой многооконного интерфейса – деталь, которую, возможно, трудно уловить по представленным снимкам экранов. Например, на рис. 14.46

показаны главное окно PyMailGUI со списком сообщений на сервере, два окна со списками сообщений, сохраненных в файлах, два окна просмотра писем и окно со справкой. Все эти окна являются немодальными, то есть действуют независимо и не мешают выбору других окон, хотя все они действуют в рамках одного процесса.

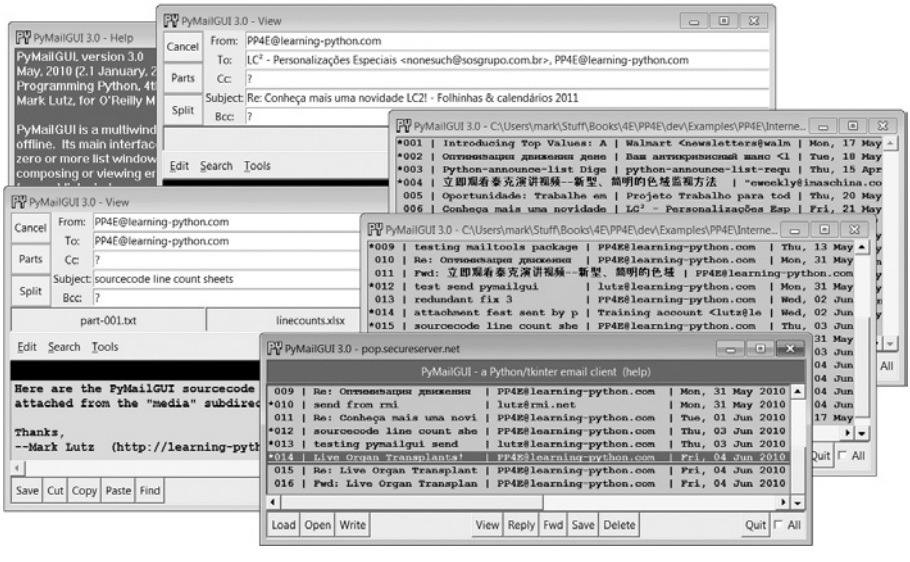


Рис. 14.46. *Многооконный интерфейс PyMailGUI и текстовые редакторы*

В целом одновременно может быть открыто любое количество окон для просмотра или составления сообщений, между которыми можно осуществлять операции копирования. Это важно, поскольку программа PyMailGUI должна обеспечить наличие в каждом окне отдельного объекта текстового редактора. Если бы объект текстового редактора был глобальным или использовал глобальные переменные, могло оказаться, что в каждом окне отображается один и тот же текст (а операции отправки могли бы привести к отправке текста из другого окна). Чтобы избежать этого, PyMailGUI создает и вставляет новые экземпляры `TextEditor` в каждое новое окно просмотра или составления сообщения и связывает новый редактор с обработчиком кнопки Send (Отправить), обеспечивая получение нужного текста. Это всего лишь действие обычного для ООП механизма сохранения состояния, но здесь он приносит ощутимую выгоду.

Во время своей работы PyMailGUI выводит множество различных сообщений о состоянии, но увидеть их можно, только если запустить программу из командной строки в консоли (например, в окне DOS в Windows или в xterm в Linux) или двойным щелчком на значке с именем файла (главным сценарием является файл с расширением *.py*, а не *.pyw*).

В Windows эти сообщения не будут видны при запуске PyMailGUI из других программ, таких как панели запуска PyDemos и PyGadgets. Эти сообщения о состоянии содержат информацию о сервере, показывают состояние загрузки почты и трассируют порождаемые попутно потоки загрузки, сохранения и удаления. Если вы хотите увидеть сообщения PyMailGUI, запустите ее из командной строки и наблюдайте:

```
C:\...\PP4E\Internet\Email\PyMailGui> PyMailGui.py
user: PP4E@learning-python.com
loading headers
Connecting...
b'+OK <24715.1275632750@pop05.mesa1.secureserver.net>'
load headers exit
synch check
Connecting...
b'+OK <26056.1275632770@pop19.prod.mesa1.secureserver.net>'
Same headers text
loading headers
Connecting...
b'+OK <18798.1275632771@pop04.mesa1.secureserver.net>'
load headers exit
synch check
Connecting...
b'+OK <28403.1275632790@pop19.prod.mesa1.secureserver.net>'
Same headers text
load 16
Connecting...
b'+OK <28472.1275632791@pop19.prod.mesa1.secureserver.net>'
Sending to...['lutz@rmi.net', 'PP4E@learning-python.com']
MIME-Version: 1.0
Content-Type: text/plain; charset="us-ascii"
Content-Transfer-Encoding: 7bit
From: PP4E@learning-python.com
To: lutz@rmi.net
Subject: Already got one...
Date: Fri, 04 Jun 2010 06:30:26 -0000
X-Mailer: PyMailGUI 3.0 (Python)

> -----Origin
Send exit
```

Можно также выполнить двойной щелчок на файле *PyMailGui.py* в окне менеджера файлов и наблюдать появившееся окно консоли DOS в Windows. Сообщения в консоли предназначены главным образом для отладки, но также они могут помочь понять, как действует система.

Подробности об использовании PyMailGUI вы найдете в окне со справкой (щелкните на полосе вызова справки в верхней части окна со списком сообщений на сервере) или в строке справки в модуле `PyMailGUIHelp.py`, описываемом в следующем разделе.

Реализация PyMailGUI

В конце концов, мы добрались до программного кода. Программа PyMailGUI состоит из нескольких новых модулей, перечисленных в начале главы, а также нескольких дополнительных файлов, описанных там же. Исходный программный код этих модулей приводится в этом разделе. Прежде чем двинуться дальше, следует сделать два небольших напоминания, которые помогут в изучении:

Повторное использование программного кода

Программа PyMailGUI получает существенный выигрыш от повторного использования модулей, которые мы написали раньше и не будем повторять здесь: `mailtools` – реализует операции загрузки, составления, анализа и удаления почты; `threadtools` – управляет потоками выполнения, взаимодействующими с сервером и локальными файлами; `TextEditor`, из части книги, посвященной графическим интерфейсам, – реализует отображение и редактирование текста почтовых сообщений; и так далее. Список номеров примеров приводится в начале этой главы

Кроме того, здесь используются стандартные модули и пакеты Python, такие как `poplib`, `smtplib` и `email`, скрывающие большую часть действий по пересылке байтов по Сети, извлечению и конструированию частей сообщений. Библиотека tkinter реализует компоненты графического интерфейса с поддержкой переносимости.

Организация программного кода

Как уже упоминалось выше, для повышения степени повторного использования программного кода в PyMailGUI применяются приемы выделения общего программного кода и ООП. Например, окна просмотра списков реализованы как общий суперкласс, в котором запрограммировано большинство операций, и два подкласса: один создает окно со списком входящей почты на сервере, а другой – окно со списком сообщений в локальном файле. Подклассы специализируют общий суперкласс под определенную среду хранения почты.

Такая архитектура отражает принцип действия самого графического интерфейса – окна со списками сообщений на сервере загружают почту с использованием протокола POP, а окна со списками сообщений в файлах загружают почту из локальных файлов. Однако компоновка графических элементов и базовые операции, общие для обоих видов окон, реализованы в суперклассе, чтобы избежать избыточности и упростить программный код. Подобный прием выделения общего программного кода применяется и к окнам просмотра сообщений: общий суперкласс окна просмотра повторно используется и специализируется подклассами окон создания нового сообщения, ответа и пересылки.

Чтобы упростить сопровождение программного кода, он был разделен на два основных модуля, отражающих структуру графического

интерфейса, – один из них реализует операции для окон со списками, а другой – для окон просмотра сообщений. Если, к примеру, требуется отыскать реализацию кнопки, которая находится в окне просмотра или редактирования сообщения, для этого следует заглянуть в модуль реализации окна просмотра и найти метод, имя которого начинается со слова *on*, в соответствии с соглашением, использовавшимся при выборе имен для методов обработчиков. Текст на кнопке также можно применять для поиска в таблицах соответствия надписей на кнопках и обработчиков, используемых при конструировании окон. Операции, выполняемые в окнах со списками, в свою очередь следует искать в модуле реализации окна со списком.

Кроме того, реализация кэша сообщений была выделена в отдельный объект и модуль, а потенциально пригодные для многократного использования инструменты оформлены в виде импортируемых модулей (например, утилита переноса длинных строк и диалоги общего пользования). Программа PyMailGUI также включает главный модуль, в котором определяются классы запускаемых окон; упрощенный механизм преобразования разметки HTML в простой текст; модуль, содержащий текст справки в виде строки; модуль `mailconfig` с пользовательскими настройками (здесь используется версия, характерная для PyMailGUI) и несколько дополнительных файлов.

В следующем разделе вашему вниманию будут представлены все файлы с исходными текстами PyMailGUI. По мере знакомства с ними возвращайтесь к демонстрации, описанной выше в этой главе, и запускайте программу, чтобы иметь возможность сопоставить ее поведение с реализацией.

Одно предварительное замечание: в этом разделе отсутствует только один из 18 файлов с исходными текстами PyMailGUI – это файл инициализации пакета `__init__.py`. Данный файл содержит единственную строку комментария и в настоящее время не используется системой. Он присутствует исключительно с целью расширения в будущем, когда программа PyMailGUI, возможно, будет использоваться как пакет – некоторые из ее модулей вполне могут пригодиться в других программах. Однако в текущей реализации используются операции импортирования из того же каталога, а не относительно пакета, то есть предполагается, что система либо выполняется как программа верхнего уровня (импортирование выполняется из текущего каталога «`.`») или путь к каталогу с ней находится в списке `sys.path` (импортирование выполняется по абсолютному пути). В Python 3.X каталог пакета не включается в `sys.path` автоматически, поэтому, чтобы в будущем программу можно было использовать как пакет, необходимо изменить внутренние операции импортирования (например, переместить главный сценарий в дереве каталогов на уровень выше и повсюду использовать конструкцию `from . import module`). За дополнительными подробностями о пакетах и импортировании пакетов обращайтесь к другим ресурсам, таким как книга «Изучаем Python».

PyMailGUI: главный модуль

В примере 14.1 приводится файл, который запускает программу PyMailGUI. Он реализует окна верхнего уровня со списками сообщений – комбинацию прикладной логики PyMailGUI и суперклассов оконного протокола, реализованных ранее в этой книге. Эти суперклассы определяют заголовки окон, значки и поведение при закрытии.

Кроме того, в этом модуле находится основная документация для разработчиков и логика обработки параметров командной строки – программа принимает из командной строки имена одного или более файлов с сохраненной в них почтой и автоматически открывает их при запуске графического интерфейса. Эта особенность, например, используется панелью запуска PyDemos, представленной в главе 10.

Пример 14.1. PP4E\Internet\Email\PyMailGui\PyMailGui.py

```
""""
################################################################################
PyMailGui 3.0 - почтовый клиент Python/tkinter.
Клиентский графический интерфейс на базе tkinter для отправки и приема
электронной почты.

Смотрите строку с текстом справки в PyMailGuiHelp.py, где приводится
информация об использовании, и список предлагаемых расширений этой версии.

Версия 2.0 была полностью переписана. Изменения между версиями 2.0
(июль 2005) и 2.1 (январь 2006): кнопки быстрого доступа в окнах просмотра,
выполнение в потоках операций загрузки и удаления сообщений в локальных
файлах, проверка и восстановление синхронизации с почтовым ящиком на сервере
по номерам сообщений при удалении, загрузке оглавления и загрузке сообщений.

Версия 3.0 (4E) является результатом переноса для работы под управлением
Python 3.X; использует компоновку по сетке вместо фреймов колонок
для улучшения размещения полей заголовков в окнах просмотра; вызывает
метод update() после вставки нового текстового редактора, чтобы обеспечить
более точное позиционирование текстового курсора (смотрите изменения
метода loadFirst в PyEdit, в главе 11); предоставляет HTML-версию текста
справки; извлекает простой текст из основных частей в формате HTML для
отображения и цитирования; поддерживает разделители в панелях инструментов;
решает проблему выбора кодировки для тела сообщения и заголовков при
получении, отправке и сохранении почты (см. главы 13 и 14); и многое другое
(см. главу 14, где приводится полный перечень изменений в версии 3.0);
декодирование полученного сообщения выполняется глубоко в недрах пакета
mailtools, при выполнении операций загрузки в кэш; кроме того, пакет
mailtools исправляет некоторые ошибки пакета email (см. главу 13);

Этот файл реализует окна верхнего уровня и интерфейс. PyMailGui использует
ряд модулей, которые ничего не подозревают об этом графическом интерфейсе,
но решают свои задачи, часть из которых была реализована в других разделах
книги. Модуль mailconfig расширен для использования в этой программе.
```

```
==Модули, определяемые в другом месте и используемые здесь:==

mailtools (пакет)
    глава, посвященная созданию клиентских сценариев
    отправляет на сервер и принимает с сервера, анализирует,
    конструирует (пример 13.21+)
threadtools.py
    глава об инструментах графического интерфейса
    управление очередью обработчиков графического интерфейса (пример 10.20)
windows.py
    глава об инструментах графического интерфейса
    настройка рамок окон верхнего уровня (пример 10.16)
textEditor.py
    глава с примерами программ с графическим интерфейсом
    текстовый виджет, используемый в окнах просмотра сообщений,
    некоторые диалоги (пример 11.4)

==В целом полезные модули, определяемые здесь:==

popuputil.py
    диалог с текстом справки и диалоги с информацией о ходе
    выполнения операций для общего использования
messagecache.py
    кэш для слежения за уже загруженными сообщениями
wraplines.py
    утилита переноса длинных строк в сообщениях
html2text.py
    упрощенный механизм преобразования HTML в простой текст
mailconfig.py
    пользовательские параметры: имена серверов, шрифты и так далее

==Модули программы, определяемые здесь:==

SharedNames.py
    объекты, совместно используемые классами окон и главным файлом
ViewWindows.py
    реализация окон просмотра, создания, ответа и пересылки
ListWindows.py
    реализация окон со списками содержимого почтового ящика
    на сервере и локальных файлов
PyMailGuiHelp.py (see also PyMailGuiHelp.html)
    текст справки для пользователя, который выводится щелчком
    на полосе вызова справки в главном окне
PyMailGui.py
    главный файл верхнего уровня (запускает программу) с типами главного окна
##############################################################################
"""

import mailconfig, sys
from SharedNames import appname, windows
from ListWindows import PyMailServer, PyMailFile
```

```
##############################################################################
# классы окон верхнего уровня
#
# Все окна просмотра, создания, ответа, пересылки, справки и диалоги
# с информацией о ходе выполнения операции являются прямыми наследниками
# класса PopupWindow: только используется; askPasswordWindow вызывает
# конструктор PopupWindow и прикрепляет виджеты к полученному окну;
# порядок здесь имеет большое значение! - классы PyMail переопределяют
# некоторые методы по умолчанию в классах окон, такие как destroy
# и okayToExit: они должны указываться первыми;
# чтобы использовать PyMailFileWindow отдельно, необходимо
# имитировать логику PyMailCommon.onOpenMailFile;
##############################################################################

# использует файл значка в cwd или в каталоге по умолчанию tools

srvrname = mailconfig.popservername or 'Server'

class PyMailServerWindow(PyMailServer, windows.MainWindow):
    "окно Tk с расширенным протоколом и подмешанными методами"

    def __init__(self):
        windows.MainWindow.__init__(self, appname, srvrname)
        PyMailServer.__init__(self)

class PyMailServerPopup(PyMailServer, windows.PopupWindow):
    "окно Toplevel с расширенным протоколом и подмешанными методами"

    def __init__(self):
        windows.PopupWindow.__init__(self, appname, srvrnane)
        PyMailServer.__init__(self)

class PyMailServerComponent(PyMailServer, windows.ComponentWindow):
    "фрейм Frame с расширенным протоколом и подмешанными методами"

    def __init__(self):
        windows.ComponentWindow.__init__(self)
        PyMailServer.__init__(self)

class PyMailFileWindow(PyMailFile, windows.PopupWindow):
    "окно Toplevel с расширенным протоколом и подмешанными методами"

    def __init__(self, filename):
        windows.PopupWindow.__init__(self, appname, filename)
        PyMailFile.__init__(self, filename)

##############################################################################
# когда выполняется как программа верхнего уровня: создает главное окно
# со списком сообщений на сервере
##############################################################################
```

```
if __name__ == '__main__':
    rootwin = PyMailServerWindow()       # открыть окно для сервера
    if len(sys.argv) > 1:                # 3.0: исправ. для добавл. len()
        for savename in sys.argv[1:]:
            rootwin.onOpenMailFile(savename)   # открыть окна
                                               # для файлов (демо)
        rootwin.lift()                   # сохранить файлы, загруженные
    rootwin.mainloop()                   # в потоках
```

SharedNames: глобальные переменные программы

Модуль в примере 14.2 реализует общее, глобальное для программы пространство имен, в котором собраны ресурсы, используемые в большинстве модулей программы, и определяются глобальные объекты, объединяющие модули. Позволяет другим модулям избежать избыточного повторения инструкций импортирования общих модулей и инкапсулирует инструкции импортирования пакета – это единственный модуль, куда придется вносить изменения, если пути к каталогам изменятся в будущем. Использование глобальных переменных в целом может сделать программу более сложной для понимания (местоположение некоторых имен может быть неочевидным), но это вполне разумный шаг, если все такие имена собраны в единственном модуле, как этот, потому что это единственное место, где придется искать неизвестные имена.

Пример 14.2. PP4E\Internet\Email\PyMailGui\SharedNames.py

```
……

##############################################################################
объекты, совместно используемые всеми классами окон и главным модулем:
глобальные переменные
##############################################################################
……


# используется во всех окнах, заголовки
appname = 'PyMailGUI 3.0'

# используется операциями сохранения, открытия, удаления из списка;
# а также для файла, куда сохраняются отправленные сообщения
saveMailSeparator = 'PyMailGUI' + ('-'*60) + 'PyMailGUI\n'

# просматриваемые в настоящее время файлы; а также для файла,
# куда сохраняются отправленные сообщения
openSaveFiles = {}                       # 1 окно на файл,{имя:окно}

# службы стандартной библиотеки
import sys, os, email.utils, email.message, webbrowser, mimetypes
from tkinter import *
from tkinter.simpledialog import askstring
from tkinter.filedialog   import SaveAs, Open, Directory
from tkinter.messagebox   import showinfo, showerror, askyesno
```

```
# повторно используемые примеры из книги
from PP4E.Gui.Tools      import windows     # рамка окна, протокол
                                            # завершения
from PP4E.Gui.Tools      import threadtools # очередь обработчиков в потоках
from PP4E.Internet.Email import mailtools   # утилиты загрузки, отправки,
                                            # анализа, создания
from PP4E.Gui.TextEditor import textEditor  # компонент и окно

# модули, определяемые здесь
import mailconfig  # пользовательские параметры: серверы, шрифты и т.д.
import popuputil   # диалоги вывода справки, инф. о ходе выполнения операции
import wraplines       # перенос длинных строк
import messagecache    # запоминает уже загруженную почту
import html2text       # упрощенный механизм преобразования html->текст
import PyMailGuiHelp   # документация пользователя

def printStack(exc_info):
    """
    отладка: выводит информацию об исключении и трассировку стека в stdout;
    3.0: выводит трассировочную информацию в файл журнала, если вывод
    в sys.stdout терпит неудачу: это происходит при запуске из другой
    программы в Windows; без этого обходного решения PyMailGUI аварийно
    завершает работу, поскольку вызывается из главного потока при появлении
    исключения в дочернем потоке; вероятно, ошибка в Python 3.1:
    эта проблема отсутствует в версиях 2.5 и 2.6, и объект с трассировочной
    информацией прекрасно работает, если вывод осуществляется в файл;
    по иронии, простой вызов print() здесь тоже работает (но вывод
    отправляется в никуда) при запуске из другой программы;
    """
    print(exc_info[0])
    print(exc_info[1])
    import traceback
    try:                                                # ok, если
        traceback.print_tb(exc_info[2], file=sys.stdout) # не запущена
    except:                                           # из другой программы!
        log = open('_pymailerrlog.txt', 'a')          # использовать файл
        log.write('-'*80)                             # иначе завершится
        traceback.print_tb(exc_info[2], file=log)     # в 3.X, но не в 2.5/6

# счетчики рабочих потоков выполнения, запущенных этим графич. интерфейсом
# sendingBusy используется всеми окнами отправки,
# используется операцией завершения главного окна

loadingHdrsBusy = threadtools.ThreadCounter()  # только 1
deletingBusy = threadtools.ThreadCounter()     # только 1
loadingMsgsBusy = threadtools.ThreadCounter()  # может быть множество
sendingBusy = threadtools.ThreadCounter()      # может быть множество
```

ListWindows: окна со списками сообщений

Пример 14.3 реализует окна со списками оглавления почты – окно для отображения содержимого почтового ящика на сервере и одно или более окон с содержимым локальных файлов. Окна этих двух типов в значительной степени выглядят и действуют одинаково и фактически выполняют общий программный код в суперклассе. Подклассы окон главным образом всего лишь специализируют суперкласс, отображая операции загрузки и удаления сообщений для работы с сервером или с локальным файлом.

Окна со списками создаются при запуске программы (начальное окно со списком сообщений на сервере и, возможно, окна со списками сообщений в файлах, имена которых передаются в виде аргументов командной строки), а также в ответ на нажатие кнопки Open (Открыть) в существующем окне со списком (чтобы открыть новое окно со списком сообщений в файле). Программный код инициализации окна можно найти в обработчике кнопки Open (Открыть) в этом примере.

Обратите внимание, что базовые операции обработки почты из пакета mailtools из главы 13 подмешиваются в PyMailGUI различными способами. Классы окон со списками в примере 14.3 наследуют из пакета mailtools класс MailParser, а окно со списком сообщений на сервере встраивает экземпляр объекта кэша сообщений, который в свою очередь наследует из пакета mailtools класс MailFetcher. Класс MailSender из пакета mailtools наследуется окнами просмотра и создания сообщений, но не наследуется окнами со списками. Окна просмотра также наследуют класс MailParser.

Это довольно длинный файл. В принципе, его можно было бы разбить на три файла, по одному на каждый класс, но эти классы настолько тесно связаны между собой, что для разработки было удобнее поместить их в один файл. По сути это один класс с двумя небольшими расширениями.

Пример 14.3. PP4E\Internet\Email\PyMailGui\ListWindows.py

```
……

##############################################################################
Реализация окон со списками сообщений на почтовом сервере и в локальных
файлах: по одному классу для каждого типа окон. Здесь использован прием
выделения общего программного кода для повторного использования: окна
со списками содержимого на сервере и в файлах являются специализированными
версиями класса PyMailCommon окон со списками; класс окна со списком почты
на сервере отображает свои операции на операции получения почты с сервера,
класс окна со списком почты в файле отображает свои операции на операции
с локальным файлом.

В ответ на действия пользователя окна со списками создают окна просмотра,
создания, ответа и пересылки писем. Окно со списком сообщений на сервере
играет роль главного окна и открывается при запуске сценарием верхнего
уровня; окна со списками сообщений в файлах открываются по требованию,
```

```
щелчком на кнопке "Open" в окне со списком сообщений на сервере или в файле.
Номера сообщений могут временно рассинхронизироваться с сервером, в случае
изменения содержимого почтового ящика входящих сообщений (вызывает полную
перезагрузку списка).

Изменения в этом модуле в версии 2.1:
-теперь проверяет синхронизации номеров сообщений при удалении и загрузке
-добавляет в окна просмотра до N кнопок прямого доступа к вложениям
-загрузка из файлов выполняется в отдельном потоке, чтобы избежать задержек
 при загрузке больших файлов
-удаление отправленных сообщений из файла также выполняется в потоке,
 благодаря чему исчезла задержка в работе графического интерфейса

Что сделать:
-сохранение отправленных сообщений в файл до сих пор выполняется в главном
 потоке: это может вызывать короткие паузы в работе графического интерфейса,
 хотя это маловероятно - в отличие от загрузки и удаления,
 сохранение/отправка просто записывает сообщение в конец файла.
-реализация операций с локальными файлами, как с текстовыми файлами
 с разделителями в значительной степени является лишь прототипом:
 она загружает полные тексты сообщений целиком в память, что ограничивает
 возможный размер этих файлов; лучшее решение: использовать 2 DBM-файла
 с доступом по ключу - для заголовков и для полного текста сообщений,
 плюс список для отображения ключей в позиции; в такой реализации
 файлы с сообщениями превращаются в каталоги и становятся нечитаемыми
 для человека.
##############################################################################
"""

from SharedNames import *      # глобальные объекты программы
from ViewWindows import ViewWindow, WriteWindow, ReplyWindow, ForwardWindow

##############################################################################
# основной фрейм - общая структура списков с сообщениями на сервере
# и в файлах
##############################################################################

class PyMailCommon(mailtools.MailParser):
    """
    абстрактный пакет виджетов с главным списком сообщений;
    смешивается с классами Tk, Toplevel или Frame окон верхнего уровня;
    должен специализироваться в подклассах с помощью метода actions()
    и других; создает окна просмотра и создания по требованию:
    они играют роль MailSenders;
    """
    # атрибуты класса, совместно используемые всеми окнами со списками
    threadLoopStarted = False                  # запускается первым окном
    queueChecksPerSecond = 20                  # изменит, в зависимости
                                               # от нагрузки на процессор
    queueDelay = 1000 // queueChecksPerSecond # минимальное число миллисекунд
                                               # между событиями таймера
```

```
queueBatch = 5                          # максимальное число вызовов
                                        # обработчиков на одно событие
                                        # от таймера

# все окна используют одни и те же диалоги: запоминают последний каталог
openDialog = Open(title=appname + ': Open Mail File')
saveDialog = SaveAs(title=appname + ': Append Mail File')

# 3.0: чтобы избежать загрузки одного и того же сообщения
# в разных потоках
beingFetched = set()

def __init__(self):
    self.makeWidgets()     # нарисовать содержимое окна: список, панель
    if not PyMailCommon.threadLoopStarted:

        # запустить цикл проверки завершения потоков
        # цикл событий от таймера, в котором производится вызов
        # обработчиков из очереди;
        # один цикл для всех окон: окна со списками с сервера
        # и из файла, окна просмотра могут выполнять операции в потоках;
        # self - экземпляр класса Tk, Toplevel или Frame: достаточно
        # будет виджета любого типа;
        # 3.0/4E: для увеличения скорости обработки увеличено количество
        # вызываемых обработчиков и количество событий в секунду:
        # ~100x/sec;

        PyMailCommon.threadLoopStarted = True
        threadtools.threadChecker(self, self.queueDelay, self.queueBatch)

def makeWidgets(self):
    # добавить флажок "All" внизу
    tools = Frame(self, relief=SUNKEN, bd=2, cursor='hand2') # 3.0: настр.
    tools.pack(side=BOTTOM, fill=X)
    self.allModeVar = IntVar()
    chk = Checkbutton(tools, text="All")
    chk.config(variable=self.allModeVar, command=self.onCheckAll)
    chk.pack(side=RIGHT)

    # добавить кнопки на панель инструментов внизу
    for (title, callback) in self.actions():
        if not callback:
            sep = Label(tools, text=title)        # 3.0: разделитель
            sep.pack(side=LEFT, expand=YES, fill=BOTH)  # растет с окном
        else:
            Button(tools, text=title, command=callback).pack(side=LEFT)

    # добавить список с возможностью выбора нескольких записей
    # и полосами прокрутки
    listwide = mailconfig.listWidth  or 74    # 3.0: настр. нач. размера
    listhigh = mailconfig.listHeight or 15 # ширина=символы, высота=строки
```

```
        mails    = Frame(self)
        vscroll  = Scrollbar(mails)
        hscroll  = Scrollbar(mails, orient='horizontal')
        fontsz   = (sys.platform[:3] == 'win' and 8) or 10 # по умолчанию
        listbg   = mailconfig.listbg or 'white'
        listfg   = mailconfig.listfg or 'black'
        listfont = mailconfig.listfont or ('courier', fontsz, 'normal')
        listbox  = Listbox(mails, bg=listbg, fg=listfg, font=listfont)
        listbox.config(selectmode=EXTENDED)
        listbox.config(width=listwide, height=listhigh) # 3.0: иниц. шире
        listbox.bind('<Double-1>', (lambda event: self.onViewRawMail()))

        # связать список и полосы прокрутки
        vscroll.config(command=listbox.yview, relief=SUNKEN)
        hscroll.config(command=listbox.xview, relief=SUNKEN)
        listbox.config(yscrollcommand=vscroll.set, relief=SUNKEN)
        listbox.config(xscrollcommand=hscroll.set)

        # присоединяется последним = усекается первым
        mails.pack(side=TOP, expand=YES, fill=BOTH)
        vscroll.pack(side=RIGHT, fill=BOTH)
        hscroll.pack(side=BOTTOM, fill=BOTH)
        listbox.pack(side=LEFT, expand=YES, fill=BOTH)
        self.listBox = listbox

    #####################
    # обработчики событий
    #####################

    def onCheckAll(self):
        # щелчок на флажке "All"
        if self.allModeVar.get():
            self.listBox.select_set(0, END)
        else:
            self.listBox.select_clear(0, END)

    def onViewRawMail(self):
        # может вызываться из потока: просмотр выбранных сообщений -
        # необработанный текст заголовков, тела
        msgnums = self.verifySelectedMsgs()
        if msgnums:
            self.getMessages(msgnums, after=lambda: self.contViewRaw(msgnums))

    def contViewRaw(self, msgnums, pyedit=True): # необх. полный TextEditor?
        for msgnum in msgnums:             # может быть вложенное определение
            fulltext = self.getMessage(msgnum) # полный текст - декодир. Юникод
            if not pyedit:
                # вывести в виджете scrolledtext
                from tkinter.scrolledtext import ScrolledText
                window = windows.QuietPopupWindow(appname,
                                                  'raw message viewer')
```

```
            browser = ScrolledText(window)
            browser.insert('0.0', fulltext)
            browser.pack(expand=YES, fill=BOTH)
        else:
            # 3.0/4E: более полноценный текстовый редактор PyEdit
            wintitle = ' - raw message text'
            browser  = textEditor.TextEditorMainPopup(self,
                                                      winTitle=wintitle)
            browser.update()
            browser.setAllText(fulltext)
            browser.clearModified()

def onViewFormatMail(self):
    """
    может вызываться из потока: просмотр отобранных сообщений -
    выводит форматированное отображение в отдельном окне; вызывается
    не из потока, если вызывается из списка содержимого файла или
    сообщения уже были загружены; действие after вызывается, только если
    предварительное получение в getMessages возможно и было выполнено
    """
    msgnums = self.verifySelectedMsgs()
    if msgnums:
        self.getMessages(msgnums, after=lambda: self.contViewFmt(msgnums))

def contViewFmt(self, msgnums):
    """
    завершение операции вывода окна просмотра: извлекает основной текст,
    выводит окно (окна) для отображения; если необходимо, извлекает
    простой текст из html, выполняет перенос строк; сообщения в формате
    html: выводит извлеченный текст, затем сохраняет во временном файле
    и открывает в веб-броузере; части сообщений могут также открываться
    вручную из окна просмотра с помощью кнопки Split (Разбить)
    или кнопок быстрого доступа к вложениям; в сообщении, состоящем
    из единственной части, иначе: часть должна открываться вручную
    кнопкой Split (Разбить) или кнопкой быстрого доступа к части;
    проверяет необходимость открытия html в mailconfig;

    3.0: для сообщений, содержащих только разметку html, основной текст
    здесь имеет тип str, но сохраняется он в двоичном режиме, чтобы
    обойти проблемы с кодировками; это необходимо, потому что
    значительная часть электронных писем в настоящее время
    отправляется в формате html; в первых версиях делалась попытка
    подобрать кодировку из N возможных (utf-8, latin-1, по умолчанию
    для платформы), но теперь тело сообщения получается и сохраняется
    в двоичном виде, чтобы минимизировать любые потери точности;
    если позднее часть будет открываться по требованию, она точно
    так же будет сохранена в файл в двоичном режиме;

    предупреждение: запускаемый веб-броузер не получает оригинальные
    заголовки письма: ему, вероятно, придется строить свои догадки
    о кодировке или вам придется явно сообщать ему о кодировке, если
```

```
        в разметке html отсутствуют собственные заголовки с информацией
        о кодировке (они принимают форму тегов <meta> в разделе <head>,
        если таковой имеются; здесь ничего не вставляется в разметку html,
        так как некоторые корректно оформленные части в формате html уже
        имеют все необходимое); IE, похоже, способен обработать большинство
        файлов html; кодирования частей html в utf-8 также может оказаться
        вполне достаточным: эта кодировка может с успехом применяться
        к большинству типов текста;
        """
        for msgnum in msgnums:
            fulltext = self.getMessage(msgnum)        # 3.0: str для анализа
            message  = self.parseMessage(fulltext)
            type, content = self.findMainText(message) # 3.0: декод. Юникод
            if type in ['text/html', 'text/xml']:      # 3.0: извлечь текст
                content = html2text.html2text(content)
            content = wraplines.wrapText1(content, mailconfig.wrapsz)
            ViewWindow(headermap   = message,
                       showtext    = content,
                       origmessage = message)    # 3.0: декодирует заголовки

            # единственная часть, content-type text/HTML (грубо, но верно!)
            if type == 'text/html':
                if ((not mailconfig.verifyHTMLTextOpen) or
                    askyesno(appname, 'Open message text in browser?')):

                    # 3.0: перед декодированием в Юникод
                    # преобразовать из формата MIME
                    type, asbytes = self.findMainText(message, asStr=False)
                    try:
                        from tempfile import gettempdir    # или виджет Tk
                        tempname = os.path.join(gettempdir(),
                                        'pymailgui.html')  # просмотра HTML?
                        tmp = open(tempname, 'wb')         # уже кодирован
                        tmp.close()
                        tmp.write(asbytes)
                        webbrowser.open_new('file://' + tempname)
                    except:
                        showerror(appname, 'Cannot open in browser')

    def onWriteMail(self):
        """
        окно составления нового письма на пустом месте, без получения других
        писем; здесь ничего не требуется цитировать, но следует добавить
        подпись и записать адрес отправителя в заголовок Bcc, если этот
        заголовок разрешен в mailconfig; значение для заголовка From
        в mailconfig может быть интернационализированным:
        окно просмотра декодирует его;
        """
        starttext = '\n'                    # использовать текст подписи
        if mailconfig.mysignature:
            starttext += '%s\n' % mailconfig.mysignature
```

```
        From = mailconfig.myaddress
        WriteWindow(starttext = starttext,                  # 3.0: заполнить
                    headermap = dict(From=From, Bcc=From))  # заголовок Bcc

    def onReplyMail(self):
        # может вызываться из потока: создание ответа на выбранные письма
        msgnums = self.verifySelectedMsgs()
        if msgnums:
            self.getMessages(msgnums, after=lambda: self.contReply(msgnums))

    def contReply(self, msgnums):
        """
        завершение операции составления ответа: отбрасывает вложения,
        цитирует текст оригинального сообщения с помощью символов '>',
        добавляет подпись; устанавливает начальные значения заголовков
        To/From, беря их из оригинального сообщения или из модуля
        с настройками; не использует значение оригинального заголовка To
        для From: может быть несколько адресов или название списка рассылки;
        заголовок To сохраняет формат имя+<адрес>, даже если в имени
        используется ','; для заголовка To используется значение
        оригинального заголовка From, игнорирует заголовок reply-to,
        если таковой имеется; 3.0: копия ответа по умолчанию также
        отправляется всем получателям оригинального письма;

        3.0: чтобы обеспечить возможность использования запятых
        в качестве разделителей адресов, теперь используются
        функции getaddresses/parseaddr, благодаря которым корректно
        обрабатываются запятые, присутствующие в компонентах имен адресов;
        в графическом интерфейсе адреса в заголовке также разделяются
        запятыми, мы копируем запятые при отображении заголовков
        и используем getaddresses для разделения адресов,
        когда это необходимо; почтовые серверы требуют, чтобы в качестве
        разделителей адресов использовался символ ',';
        больше не использует parseaddr для получения первой пары
        имя/адрес из результата, возвращаемого getaddresses:
        используйте полное значение заголовка From для заголовка To;

        3.0: здесь предусматривается декодирование заголовка Subject,
        потому что необходимо получить его текст, но класс окна просмотра,
        являющийся суперклассом окна редактирования, выполняет декодирование
        всех отображаемых заголовков (дополнительное декодирование
        заголовка Subject является пустой операцией); при отправке все
        заголовки и имена в заголовках с адресами, содержащие символы
        вне диапазона ASCII, в графическом интерфейсе присутствуют
        в декодированном виде, но внутри пакета mailtools обрабатываются
        в кодированном виде; quoteOrigText также декодирует первоначальные
        значения заголовков, вставляемые в цитируемый текст, и окна
        со списками декодируют заголовки для отображения;
        """
        for msgnum in msgnums:
            fulltext = self.getMessage(msgnum)
```

```
        message = self.parseMessage(fulltext) # при ошибке - объект ошибки
        maintext = self.formatQuotedMainText(message) # то же для пересыл.

        # заголовки From и To декодируются окном просмотра
        From = mailconfig.myaddress       # не оригинальный To
        To   = message.get('From', '')    # 3.0: разделитель ','
        Cc   = self.replyCopyTo(message)  # 3.0: всех получателей в cc?
        Subj = message.get('Subject', '(no subject)')
        Subj = self.decodeHeader(Subj)    # декодировать, чтобы
                                          # получить str
        if Subj[:4].lower() != 're: ':    # 3.0: объединить
            Subj = 'Re: ' + Subj
        ReplyWindow(starttext = maintext,
                    headermap =
                        dict(From=From, To=To, Cc=Cc,
                             Subject=Subj, Bcc=From))

def onFwdMail(self):
    # может вызываться из потока: пересылка выбранных писем
    msgnums = self.verifySelectedMsgs()
    if msgnums:
        self.getMessages(msgnums, after=lambda: self.contFwd(msgnums))

def contFwd(self, msgnums):
    """
    завершение операции пересылки письма: отбрасывает вложения, цитирует
    текст оригинального сообщения с помощью символов '>', добавляет
    подпись; смотрите примечания по поводу заголовков в методах
    реализации операции составления ответа; класс окна просмотра,
    являющийся суперклассом, выполнит декодирование заголовка From;
    """
    for msgnum in msgnums:
        fulltext = self.getMessage(msgnum)
        message = self.parseMessage(fulltext)
        maintext = self.formatQuotedMainText(message) # то же для ответа

        # начальное значение From берется из настроек,
        # а не из оригинального письма
        From = mailconfig.myaddress     # кодированный или нет
        Subj = message.get('Subject', '(no subject)')
        Subj = self.decodeHeader(Subj)  # 3.0: закодируется при отправке
        if Subj[:5].lower() != 'fwd: ': # 3.0: объединить
            Subj = 'Fwd: ' + Subj
        ForwardWindow(starttext = maintext,
                      headermap = dict(From=From, Subject=Subj, Bcc=From))

def onSaveMailFile(self):
    """
    сохраняет выбранные письма в файл для просмотра без подключения
    к Интернету; запрещается, если в текущий момент выполняются
    операции загрузки/удаления с целевым файлом; также запрещается
```

```
        методом getMessages, если self в текущий момент выполняет операции
        с другим файлом; метод contSave не поддерживает многопоточный
        режим выполнения: запрещает все остальные операции;
        """
        msgnums = self.selectedMsgs()
        if not msgnums:
            showerror(appname, 'No message selected')
        else:
            # предупреждение: диалог предупреждает о перезаписи
            # существующего файла
            filename = self.saveDialog.show()        # атрибут класса
            if filename:                             # не проверять номера
                filename = os.path.abspath(filename) # нормализовать / в \
                self.getMessages(msgnums,
                        after=lambda: self.contSave(msgnums, filename))

    def contSave(self, msgnums, filename):
        # проверяет возможность продолжения, после возможной загрузки
        # сообщений с сервера
        if (filename in openSaveFiles.keys() and  # этот файл просматривается?
            openSaveFiles[filename].openFileBusy): # операции загр./удал.?
            showerror(appname, 'Target file busy - cannot save')
        else:
            try:                              # предупр.: не многопоточн.
            fulltextlist = []                 # 3.0: использ. кодировку
            mailfile = open(filename, 'a', encoding=mailconfig.fetchEncoding)
            for msgnum in msgnums:            # < 1 сек. для N сообщ.
                fulltext = self.getMessage(msgnum)     # но сообщений может
                if fulltext[-1] != '\n': fulltext += '\n' # быть слишком много
                mailfile.write(saveMailSeparator)
                mailfile.write(fulltext)
                fulltextlist.append(fulltext)
            mailfile.close()
        except:
            showerror(appname, 'Error during save')
            printStack(sys.exc_info())
        else:                                  # почему .keys(): EIBTI[1]
            if filename in openSaveFiles.keys():   # этот файл уже открыт?
                window = openSaveFiles[filename]   # обновить список, чтобы
                window.addSavedMails(fulltextlist) # избежать повторного откр.
                #window.loadMailFileThread()       # это было очень медленно

    def onOpenMailFile(self, filename=None):
        # обработка сохраненной почты без подключения к Интернету
        filename = filename or self.openDialog.show()  # общий атрибут класса
        if filename:
            filename = os.path.abspath(filename)    # полное имя файла
            if filename in openSaveFiles.keys():    # только 1 окно на файл
```

[1] Аббревиатура от «Explicit Is Better Than Implicit» («явное лучше неявного») – один из девизов Python. – *Прим. перев.*

```
            openSaveFiles[filename].lift()      # поднять окно файла,
            showinfo(appname, 'File already open') # иначе будут возникать
        else:                                   # ошибки после удаления
            from PyMailGui import PyMailFileWindow # избежать дублирования
            popup = PyMailFileWindow(filename)  # новое окно со списком
            openSaveFiles[filename] = popup     # удаляется при закр.
            popup.loadMailFileThread()          # загрузить в потоке

def onDeleteMail(self):
    # удаляет выбранные письма с сервера или из файла
    msgnums = self.selectedMsgs()                # подкласс: fillIndex
    if not msgnums:                              # всегда проверять
        showerror(appname, 'No message selected')
    else:
        if askyesno(appname, 'Verify delete %d mails?' % len(msgnums)):
            self.doDelete(msgnums)

########################
# вспомогательные методы
########################

def selectedMsgs(self):
    # возвращает выбранные в списке сообщения
    selections = self.listBox.curselection() # кортеж строк-чисел, 0..N-1
    return [int(x)+1 for x in selections]   # преобр. в int, создает 1..N

warningLimit = 15
def verifySelectedMsgs(self):
    msgnums = self.selectedMsgs()
    if not msgnums:
        showerror(appname, 'No message selected')
    else:
        numselects = len(msgnums)
        if numselects > self.warningLimit:
            if not askyesno(appname, 'Open %d selections?' % numselects):
                msgnums = []
    return msgnums

def fillIndex(self, maxhdrsize=25):
    """
    заполняет запись в списке содержимым заголовков;
    3.0: декодирует заголовки в соответствии с email/mime/unicode,
         если необходимо;
    3.0: предупреждение: крупные символы из китайского алфавита могут
         нарушить выравнивание границ '|' колонок;
    """
    hdrmaps = self.headersMaps()                 # может быть пустым
    showhdrs = ('Subject', 'From', 'Date', 'To') # загол-ки по умолчанию
    if hasattr(mailconfig, 'listheaders'):       # загол-ки в mailconfig
        showhdrs = mailconfig.listheaders or showhdrs
    addrhdrs = ('From', 'To', 'Cc', 'Bcc')   # 3.0: декодируются особо
```

```
        # вычислить максимальный размер поля
        maxsize = {}
        for key in showhdrs:
            allLens = []                    # слишком большой для списка!
            for msg in hdrmaps:
                keyval = msg.get(key, ' ')
                if key not in addrhdrs:
                    allLens.append(len(self.decodeHeader(keyval)))
                else:
                    allLens.append(len(self.decodeAddrHeader(keyval)))
            if not allLens: allLens = [1]
            maxsize[key] = min(maxhdrsize, max(allLens))

        # заполнить окно списка полями фиксированной ширины с выравниванием
        # по левому краю
        self.listBox.delete(0, END)        # наличие неск. частей отметить *
        for (ix, msg) in enumerate(hdrmaps):     # по заг. content-type
            msgtype = msg.get_content_maintype() # нет метода is_multipart
            msgline = (msgtype == 'multipart' and '*') or ' '
            msgline += '%03d' % (ix+1)
            for key in showhdrs:
                mysize = maxsize[key]
                if key not in addrhdrs:
                    keytext = self.decodeHeader(msg.get(key, ' '))
                else:
                    keytext = self.decodeAddrHeader(msg.get(key, ' '))
                msgline += ' | %-*s' % (mysize, keytext[:mysize])
            msgline += '| %.1fK' % (self.mailSize(ix+1) / 1024)
                                                        # 3.0: .0 необяз.
            self.listBox.insert(END, msgline)
        self.listBox.see(END)   # самое свежее сообщение в последней строке

    def replyCopyTo(self, message):
        """
        3.0: при создании ответа все получатели оригинального письма
        копируются из заголовков To и Cc в заголовок Cc ответа после
        удаления дубликатов, и определяется новый отправитель;
        могло бы потребоваться декодировать интернационализированные адреса,
        но окно просмотра уже декодирует их для отображения (при отправке
        они повторно кодируются). Фильтрация уникальных значений здесь
        обеспечивается множеством в любом случае, хотя предполагается,
        что интернационализированный адрес отправителя в mailconfig будет
        представлен в кодированной форме (иначе здесь он не будет
        удаляться); здесь допускаются пустые заголовки To и Cc: split
        вернет пустой список;
        """
        if not mailconfig.repliesCopyToAll:
            # ответить только отправителю
            Cc = ''
        else:
            # скопировать всех получателей оригинального письма (3.0)
```

```
        allRecipients = (self.splitAddresses(message.get('To', '')) +
                         self.splitAddresses(message.get('Cc', '')))
        uniqueOthers = set(allRecipients) - set([mailconfig.myaddress])
        Cc = ', '.join(uniqueOthers)
    return Cc or '?'

def formatQuotedMainText(self, message):
    """
    3.0: общий программный код, используемый операциями создания ответа
    и пересылки: получает декодированный текст, извлекает текст из html,
    переносит длинные строки, добавляет символы цитирования '>'
    """
    type, maintext = self.findMainText(message) # 3.0: декод. строка str
    if type in ['text/html', 'text/xml']:       # 3.0: простой текст
        maintext = html2text.html2text(maintext)
    maintext = wraplines.wrapText1(maintext, mailconfig.wrapsz-2) # 2='> '
    maintext = self.quoteOrigText(maintext, message) # добавить загол. и >
    if mailconfig.mysignature:
        maintext = ('\n%s\n' % mailconfig.mysignature) + maintext
    return maintext

def quoteOrigText(self, maintext, message):
    """
    3.0: здесь необходимо декодировать все интернационализированные
    заголовки, иначе они будут отображаться в цитируемом тексте
    в кодированной форме email+MIME; decodeAddrHeader обрабатывает
    один адрес или список адресов, разделенных запятыми; при отправке
    это может вызвать кодирование полного текста сообщения, но основной
    текст уже в полностью декодированной форме: мог быть закодирован
    в любой кодировке;
    """
    quoted = '\n-----Original Message-----\n'
    for hdr in ('From', 'To', 'Subject', 'Date'):
        rawhdr = message.get(hdr, '?')
        if hdr not in ('From', 'To'):
            dechdr = self.decodeHeader(rawhdr)     # весь заголовок
        else:
            dechdr = self.decodeAddrHeader(rawhdr) # только имя в адресе
        quoted += '%s: %s\n' % (hdr, dechdr)
    quoted += '\n' + maintext
    quoted = '\n' + quoted.replace('\n', '\n> ')
    return quoted

######################
# требуются подклассам
######################

# используется операциями просмотра, сохранения,
# создания ответа, пересылки
def getMessages(self, msgnums, after): # переопределить, если имеется кэш,
    after()                            # проверка потока
```

```
    # плюс okayToQuit? и все уникальные операции
    def getMessage(self, msgnum): assert False  # исп. многими: полный текст
    def headersMaps(self): assert False       # fillIndex: список заголовков
    def mailSize(self, msgnum): assert False  # fillIndex: размер msgnum
    def doDelete(self): assert False          # onDeleteMail: кнопка Delete

##############################################################################
# главное окно - при просмотре сообщений в локальном файле
# (или в файле отправленных сообщений)
##############################################################################

class PyMailFile(PyMailCommon):
    """
    специализирует PyMailCommon для просмотра содержимого файла
    с сохраненной почтой; смешивается с классами Tk, Toplevel или Frame,
    добавляет окно списка; отображает операции загрузки,
    получения, удаления на операции с локальным текстовым файлом;
    операции открытия больших файлов и удаления
    писем здесь выполняются в потоках;

    операции сохранения и отправки выполняются в главном потоке, потому что
    вызывают лишь добавление в конец файла; сохранение запрещается, если
    исходный или целевой файл в текущий момент используются операциями
    загрузки/удаления; операция сохранения запрещает загрузку, удаление,
    сохранение только потому, что она не выполняется в параллельном потоке
    (блокирует графический интерфейс);

    Что сделать: может потребоваться использовать потоки выполнения
    и блокировки файлов на уровне операционной системы, если сохранение
    когда-либо будет выполняться в потоках; потоки выполнения: операции
    сохранения могли бы запрещать другие операции сохранения в этот же
    файл с помощью openFileBusy, но файл может открываться не только
    в графическом интерфейсе; блокировок файлов недостаточно, потому что
    графический интерфейс тоже обновляется; Что сделать: операция добавления
    в конец файла с сохраненными почтовыми сообщениями может потребовать
    использования блокировок на уровне операционной системы: в данной
    реализации при попытке выполнить отправку во время
    операций загрузки/удаления пользователь увидит диалог
    с сообщением об ошибке;

    3.0: сейчас файлы с сохраненными почтовыми сообщениями являются
    текстовыми, их кодировка определяется настройками в модуле mailconfig;
    вероятно, это не гарантирует поддержку в самом худшем случае
    применения необычных кодировок или смешивания разных кодировок,
    но в большинстве своем полный текст почтовых сообщений состоит только
    из символов ascii, и пакет email в Python 3.1 еще содержит ошибки;
    """
    def actions(self):
        return [ ('Open', self.onOpenMailFile),
                 ('Write',  self.onWriteMail),
                 (' ',      None),                # 3.0: разделители
```

```
                ('View',   self.onViewFormatMail),
                ('Reply',  self.onReplyMail),
                ('Fwd',    self.onFwdMail),
                ('Save',   self.onSaveMailFile),
                ('Delete', self.onDeleteMail),
                (' ',      None),
                ('Quit',   self.quit) ]

def __init__(self, filename):
    # вызывающий: выполняет затем loadMailFileThread
    PyMailCommon.__init__(self)
    self.filename = filename
    self.openFileBusy = threadtools.ThreadCounter() # 1 файл - 1 окно

def loadMailFileThread(self):
    """
    загружает или повторно загружает файл и обновляет окно со списком
    содержимого; вызывается операциями открытия, на этапе запуска
    программы и, возможно, операциями отправки, если файл
    с отправленными сообщениями открыт в настоящий момент;
    в файле всегда присутствует первый фиктивный элемент
    после разбиения текста;
    альтернатива: [self.parseHeaders(m) for m in self.msglist];
    можно было бы выводить диалог с информацией о ходе выполнения
    операции, но загрузка небольших файлов выполняется очень быстро;

    2.1: теперь поддерживает многопоточную модель выполнения - загрузка
    небольших файлов выполняется < 1 сек., но загрузка очень больших
    файлов может вызывать задержки в работе графического интерфейса;
    операция сохранения теперь использует addSavedMails для добавления
    списков сообщений, чтобы повысить скорость, но это не относится
    к повторной загрузке; все еще вызывается операцией отправки, просто
    потому что текст сообщения недоступен - требуется провести
    рефакторинг; удаление также производится в потоке выполнения:
    предусмотрено предотвращение возможности перекрытия операций
    открытия и удаления;
    """
    if self.openFileBusy:
        # не допускать параллельное выполнение операций
        # открытия/удаления
        errmsg = 'Cannot load, file is busy:\n"%s"' % self.filename
        showerror(appname, errmsg)
    else:
        #self.listBox.insert(END, 'loading...') # вызыв. ошибку при щелчке
        savetitle = self.title()          # устанавливается классом окна
        self.title(appname + ' - ' + 'Loading...')
        self.openFileBusy.incr()
        threadtools.startThread(
            action  = self.loadMailFile,
            args    = (),
            context = (savetitle,),
```

```
            onExit  = self.onLoadMailFileExit,
            onFail  = self.onLoadMailFileFail)

    def loadMailFile(self):
        # выполняется в потоке, оставляя графический интерфейс активным
        # операции открытия, чтения и механизм анализа могут возбуждать
        # исключения: перехватывается в утилитах работы с потоками
        file = open(self.filename, 'r',
                    encoding=mailconfig.fetchEncoding) # 3.0
        allmsgs = file.read()
        self.msglist = allmsgs.split(saveMailSeparator)[1:]  # полный текст
        self.hdrlist = list(map(self.parseHeaders,
                                self.msglist))          # объекты сообщений

    def onLoadMailFileExit(self, savetitle):
        # успешная загрузка в потоке
        self.title(savetitle) # записать имя файла в заголовок окна
        self.fillIndex()    # обновить граф. интерф.: вып-ся в главн. потоке
        self.lift()         # поднять окно
        self.openFileBusy.decr()

    def onLoadMailFileFail(self, exc_info, savetitle):
        # исключение в потоке
        showerror(appname, 'Error opening ¨%s¨\n%s\n%s' %
                           ((self.filename,) + exc_info[:2]))
        printStack(exc_info)
        self.destroy()           # всегда закрывать окно?
        self.openFileBusy.decr() # не требуется при уничтожении окна

    def addSavedMails(self, fulltextlist):
        """
        оптимизация: добавляет вновь сохраняемые сообщения в окна
        со списками содержимого, загруженного из файлов; в прошлом
        при сохранении выполнялась перезагрузка всего файла
        вызовом loadMailThread - это медленно;
        должен вызываться только в главном потоке выполнения:
        обновляет графический интерфейс; операция отправки по-прежнему
        вызывает перезагрузку файла с отправленными сообщениями, если он
        открыт: отсутствует текст сообщения;
        """
        self.msglist.extend(fulltextlist)
        self.hdrlist.extend(map(self.parseHeaders,
                                fulltextlist))             # 3.x итератор
        self.fillIndex()
        self.lift()

    def doDelete(self, msgnums):
        """
        бесхитростная реализация, но ее вполне достаточно: перезаписывает
        в файл все неудаленные сообщения; недостаточно просто удалить
        из self.msglist: изменяются индексы элементов списка;
        Py2.3 enumerate(L) - то же самое, что zip(range(len(L)), L)
```

```
        2.1: теперь поддерживает многопоточную модель выполнения, иначе
        может вызывать паузы в работе графического интерфейса при работе
        с очень большими файлами
        """
        if self.openFileBusy:
            # не допускать параллельное выполнение операций
            # открытия/удаления
            errmsg = 'Cannot delete, file is busy:\n"%s"' % self.filename
            showerror(appname, errmsg)
        else:
            savetitle = self.title()
            self.title(appname + ' - ' + 'Deleting...')
            self.openFileBusy.incr()
            threadtools.startThread(
                action  = self.deleteMailFile,
                args    = (msgnums,),
                context = (savetitle,),
                onExit  = self.onDeleteMailFileExit,
                onFail  = self.onDeleteMailFileFail)

    def deleteMailFile(self, msgnums):
        # выполняется в потоке, оставляя графический интерфейс активным
        indexed = enumerate(self.msglist)
        keepers = [msg for (ix, msg) in indexed if ix+1 not in msgnums]
        allmsgs = saveMailSeparator.join(['']  + keepers)
        file = open(self.filename, 'w',
                    encoding=mailconfig.fetchEncoding)   # 3.0
        file.write(allmsgs)
        self.msglist = keepers
        self.hdrlist = list(map(self.parseHeaders, self.msglist))

    def onDeleteMailFileExit(self, savetitle):
        self.title(savetitle)
        self.fillIndex() # обновляет граф. интерф.: выполняется в глав. потоке
        self.lift()     # сбросить заголовок, поднять окно
        self.openFileBusy.decr()

    def onDeleteMailFileFail(self, exc_info, savetitle):
        showerror(appname, 'Error deleting "%s"\n%s\n%s' %
         ((self.filename,) + exc_info[:2]))
        printStack(exc_info)
        self.destroy()            # всегда закрывать окно?
        self.openFileBusy.decr()  # не требуется при уничтожении окна

    def getMessages(self, msgnums, after):
        """
        используется операциями просмотра, сохранения, создания ответа,
        пересылки: потоки загрузки и удаления могут изменять
        списки сообщений и заголовков, поэтому следует
        запрещать выполнение других операций, безопасность которых
```

```
        зависит от них; этот метод реализует проверку для self:
        для операций сохранения также проверяется целевой файл;
        """
        if self.openFileBusy:
            errmsg = 'Cannot fetch, file is busy:\n"%s"' % self.filename
            showerror(appname, errmsg)
        else:
            after()                    # почта уже загружена

    def getMessage(self, msgnum):
        return self.msglist[msgnum-1]  # полный текст одного сообщения

    def headersMaps(self):
        return self.hdrlist            # объекты email.message.Message

    def mailSize(self, msgnum):
        return len(self.msglist[msgnum-1])

    def quit(self):
        # не выполняет завершение в ходе обновления:
        # следом вызывается fillIndex
        if self.openFileBusy:
            showerror(appname, 'Cannot quit during load or delete')
        else:
            if askyesno(appname, 'Verify Quit Window?'):
                # удалить файл из списка открытых файлов
                del openSaveFiles[self.filename]
                Toplevel.destroy(self)

##############################################################################
# главное окно - при просмотре сообщений на почтовом сервере
##############################################################################

class PyMailServer(PyMailCommon):
    """
    специализирует PyMailCommon для просмотра почты на сервере;
    смешивается с классами Tk, Toplevel или Frame, добавляет окно со списком
    сообщений; отображает операции загрузки, получения, удаления на операции
    с почтовым ящиком на сервере; встраивает объект класса MessageCache,
    который является наследником класса MailFetcher из пакета mailtools;
    """
    def actions(self):
        return [ ('Load',   self.onLoadServer),
                 ('Open',   self.onOpenMailFile),
                 ('Write',  self.onWriteMail),
                 (' ',      None),                 # 3.0: разделители
                 ('View',   self.onViewFormatMail),
                 ('Reply',  self.onReplyMail),
                 ('Fwd',    self.onFwdMail),
                 ('Save',   self.onSaveMailFile),
```

```
            ('Delete', self.onDeleteMail),
            (' ',      None),
            ('Quit',   self.quit) ]

def __init__(self):
    PyMailCommon.__init__(self)
    self.cache = messagecache.GuiMessageCache() # встраивается, не наслед.
   #self.listBox.insert(END, 'Press Load to fetch mail')

def makeWidgets(self):    # полоса вызова справки: только в главном окне
    self.addHelpBar()
    PyMailCommon.makeWidgets(self)

def addHelpBar(self):
    msg = 'PyMailGUI - a Python/tkinter email client (help)'
    title = Button(self, text=msg)
    title.config(bg='steelblue', fg='white', relief=RIDGE)
    title.config(command=self.onShowHelp)
    title.pack(fill=X)

def onShowHelp(self):
    """
    загружает и отображает блок текстовых строк
    3.0: теперь использует также HTML и модуль webbrowser
    выбор между текстом, HTML или отображением и в том, и в другом
    формате определяется настройками в mailconfig всегда
    отображает справку: если оба параметра имеют значение,
    отображается справка в формате html
    """
    if mailconfig.showHelpAsText:
        from PyMailGuiHelp import helptext
        popuputil.HelpPopup(appname, helptext,
                            showsource=self.onShowMySource)

    if mailconfig.showHelpAsHTML or (not mailconfig.showHelpAsText):
        from PyMailGuiHelp import showHtmlHelp
        showHtmlHelp() # 3.0: HTML-версия не предусматривает возможность
                       # вывести файлы с исходными текстами

def onShowMySource(self, showAsMail=False):
    """
    отображает файлы с исходными текстами плюс
    импортирует модули кое-где
    """
    import PyMailGui, ListWindows, ViewWindows, SharedNames, textConfig
    from PP4E.Internet.Email.mailtools import ( # mailtools теперь пакет
         mailSender, mailFetcher, mailParser)   # невозм. использовать *
    mymods = (                                  # в инструкции def
        PyMailGui, ListWindows, ViewWindows, SharedNames,
        PyMailGuiHelp, popuputil, messagecache, wraplines, html2text,
        mailtools, mailFetcher, mailSender, mailParser,
```

```
            mailconfig, textConfig, threadtools, windows, textEditor)
        for mod in mymods:
            source = mod.__file__
            if source.endswith('.pyc'):
                source = source[:-4] + '.py'  # предполагается присутствие
            if showAsMail:                    # файлов .py в том же каталоге
                # не очень элегантно...
                code   = open(source).read()  # 3.0: кодировка для платформы
                user   = mailconfig.myaddress
                hdrmap = {'From': appname, 'To': user,
                          'Subject': mod.__name__}
                ViewWindow(showtext=code,
                           headermap=hdrmap,
                           origmessage=email.message.Message())
            else:
                # более удобный текстовый редактор PyEdit
                # 4E: предполагается, что текст в кодировке UTF8
                # (иначе PyEdit может запросить кодировку!)
                wintitle = ' - ' + mod.__name__
                textEditor.TextEditorMainPopup(self, source, wintitle,
                                               'utf-8')

    def onLoadServer(self, forceReload=False):
        """
        может вызываться в потоках: загружает или повторно загружает
        список заголовков почты по требованию; операции Exit,Fail,Progress
        вызываются методом threadChecker посредством очереди
        в обработчике after; загрузка может перекрываться
        во времени с отправкой, но запрещает все остальные операции;
        можно было бы выполнять одновременно с oadingMsgs,
        но может изменять кэш сообщений; в случае ошибки операций
        удаления/синхронизации принудительно вызывается forceReload,
        иначе загружает только заголовки вновь прибывших сообщений;
        2.1: cache.loadHeaders можно использовать для быстрой проверки
        синхронизации номеров сообщений с сервером, когда загружаются
        только заголовки вновь прибывших сообщений;
        """
        if loadingHdrsBusy or deletingBusy or loadingMsgsBusy:
            showerror(appname, 'Cannot load headers during load or delete')
        else:
            loadingHdrsBusy.incr()
            self.cache.setPopPassword(appname) # не обновлять графический
                                               # интерфейс в потоке!
            popup = popuputil.BusyBoxNowait(appname,
                                            'Loading message headers')
            threadtools.startThread(
                action     = self.cache.loadHeaders,
                args       = (forceReload,),
                context    = (popup,),
                onExit     = self.onLoadHdrsExit,
                onFail     = self.onLoadHdrsFail,
```

```
            onProgress = self.onLoadHdrsProgress)

    def onLoadHdrsExit(self, popup):
        self.fillIndex()
        popup.quit()
        self.lift()
        loadingHdrsBusy.decr()       # разрешить выполнение других операций

    def onLoadHdrsFail(self, exc_info, popup):
        popup.quit()
        showerror(appname, 'Load failed: \n%s\n%s' % exc_info[:2])
        printStack(exc_info)          # вывести трассировку стека в stdout
        loadingHdrsBusy.decr()
        if exc_info[0] == mailtools.MessageSynchError: # синхронизировать
            self.onLoadServer(forceReload=True)      # нов. поток: перезагр.
        else:
            self.cache.popPassword = None  # заставить ввести в следующий раз

    def onLoadHdrsProgress(self, i, n, popup):
        popup.changeText('%d of %d' % (i, n))

    def doDelete(self, msgnumlist):
        """
        может вызываться в потоках: теперь удаляет почту
        на сервере - изменяет номера сообщений;
        может перекрываться во времени только с операцией
        отправки, запрещает все операции, кроме отправки;
        2.1: cache.deleteMessages теперь проверяет результат команды TOP,
        чтобы проверить соответствие выбранных сообщений на случай
        рассинхронизации номеров сообщений с сервером: это возможно в случае
        удаления почты другим клиентом или в результате автоматического
        удаления сообщений сервером - в случае ошибки загрузки некоторые
        провайдеры могут перемещать почту из папки входящих сообщений
        в папку недоставленных;
        """
        if loadingHdrsBusy or deletingBusy or loadingMsgsBusy:
            showerror(appname, 'Cannot delete during load or delete')
        else:
            deletingBusy.incr()
            popup = popuputil.BusyBoxNowait(appname,
                                            'Deleting selected mails')
            threadtools.startThread(
                action     = self.cache.deleteMessages,
                args       = (msgnumlist,),
                context    = (popup,),
                onExit     = self.onDeleteExit,
                onFail     = self.onDeleteFail,
                onProgress = self.onDeleteProgress)

    def onDeleteExit(self, popup):
        self.fillIndex() # не требуется повторно загружать с сервера
```

```
        popup.quit()     # заполнить оглавление обновленными данными из кэша
        self.lift()      # поднять окно с оглавлением, освободить блокировку
        deletingBusy.decr()

    def onDeleteFail(self, exc_info, popup):
        popup.quit()
        showerror(appname, 'Delete failed: \n%s\n%s' % exc_info[:2])
        printStack(exc_info)                # ошибка удаления
        deletingBusy.decr()                 # или проверки синхронизации
        self.onLoadServer(forceReload=True) # новый поток: номера изменились

    def onDeleteProgress(self, i, n, popup):
        popup.changeText('%d of %d' % (i, n))

    def getMessages(self, msgnums, after):
        """
        может вызываться в потоках: загружает все выбранные сообщения в кэш;
        используется операциями сохранения, просмотра, создания ответа
        и пересылки для предварительного заполнения кэша; может
        перекрываться во времени с другими операциями загрузки сообщений
        и отправки, запрещает выполнение операции удаления
        и загрузки заголовков; действие "after" выполняется, только если
        получение сообщений разрешено и было выполнено успешно;
        2.1: cache.getMessages проверяет синхронизацию оглавления
        с сервером,но здесь проверка выполняется только
        при необходимости взаимодействия с сервером,
        когда требуемые сообщения отсутствуют в кэше;

        3.0: смотрите примечания к messagecache: теперь предотвращаются
        попытки загрузить сообщения, которые в настоящий момент уже
        загружаются, если пользователь щелкнет мышью, чтобы повторно
        запустить операцию, когда она уже выполняется;
        если какое-либо сообщение из числа выбранных уже загружается
        другим запросом, необходимо запретить загрузку всего пакета
        сообщений toLoad: иначе необходимо ждать завершения N других
        операций загрузки; операции загрузки по-прежнему могут
        перекрываться во времени fetches при условии,
        что их ничто не объединяет;
        """
        if loadingHdrsBusy or deletingBusy:
            showerror(appname, 'Cannot fetch message during load or delete')
        else:
            toLoad = [num for num in msgnums if not self.cache.isLoaded(num)]
            if not toLoad:
                after()        # все уже загружено
                return         # обработать сейчас, не ждать диалога
            else:
                if set(toLoad) & self.beingFetched: # 3.0: хоть одно загруж.?
                    showerror(appname,
                              'Cannot fetch any message being fetched')
                else:
```

```
                self.beingFetched |= set(toLoad)
                loadingMsgsBusy.incr()
                from popuputil import BusyBoxNowait
                popup = BusyBoxNowait(appname,
                                      'Fetching message contents')
                threadtools.startThread(
                    action     = self.cache.getMessages,
                    args       = (toLoad,),
                    context    = (after, popup, toLoad),
                    onExit     = self.onLoadMsgsExit,
                    onFail     = self.onLoadMsgsFail,
                    onProgress = self.onLoadMsgsProgress)

def onLoadMsgsExit(self, after, popup, toLoad):
    self.beingFetched -= set(toLoad)
    popup.quit()
    after()
    loadingMsgsBusy.decr() # разрешить другие операции после onExit

def onLoadMsgsFail(self, exc_info, after, popup, toLoad):
    self.beingFetched -= set(toLoad)
    popup.quit()
    showerror(appname, 'Fetch failed: \n%s\n%s' % exc_info[:2])
    printStack(exc_info)
    loadingMsgsBusy.decr()
    if exc_info[0] == mailtools.MessageSynchError:  # синхр. с сервером
        self.onLoadServer(forceReload=True)    # новый поток: перезагр.

def onLoadMsgsProgress(self, i, n, after, popup, toLoad):
    popup.changeText('%d of %d' % (i, n))

def getMessage(self, msgnum):
    return self.cache.getMessage(msgnum) # полный текст

def headersMaps(self):
    # список объектов email.message.Message, в 3.x требуется вызвать
    # функцию list() при использовании map()
    # возвращает [self.parseHeaders(h) for h in self.cache.allHdrs()]
    return list(map(self.parseHeaders, self.cache.allHdrs()))

def mailSize(self, msgnum):
    return self.cache.getSize(msgnum)

def okayToQuit(self):
    # есть хоть один действующий поток?
    filesbusy = [win for win in openSaveFiles.values()
                     if win.openFileBusy]
    busy = loadingHdrsBusy or deletingBusy or \
           sendingBusy or loadingMsgsBusy
    busy = busy or filesbusy
    return not busy
```

ViewWindows: окна просмотра сообщений

В примере 14.4 приводится реализация окон просмотра и редактирования почтовых сообщений. Эти окна создаются в ответ на выполнение действий, запускаемых в окнах со списками кнопками View (Просмотреть), Write (Написать), Reply (Ответить) и Forward (Переслать). Порядок создания этих окон смотрите в обработчиках событий этих кнопок в модуле с реализацией окон со списками, представленном в примере 14.3.

Как и предыдущий модуль (пример 14.3), этот файл в действительности содержит один общий класс и несколько его расширений. Окно просмотра почтового сообщения практически идентично окну редактирования сообщения, используемому для выполнения операций Write (Написать), Reply (Ответить) и Forward (Переслать). В результате этот пример определяет общий суперкласс окна просмотра, реализующий внешний вид и поведение, и затем расширяет его подклассом окна редактирования.

Окна для создания ответа и пересылки здесь практически не отличаются от окна создания нового почтового сообщения, потому что все отличия (например, установка заголовков «From» и «To», вставка цитируемого текста) обрабатываются в реализации окна со списком, еще до того как будет создано окно редактирования.

Пример 14.4. PP4E\Internet\Email\PyMailGui\ViewWindows.py

```
……

##############################################################################
Реализация окон просмотра сообщения, создания нового, ответа и пересылки:
каждому типу соответствует свой класс. Здесь применяется прием выделения
общего программного кода для многократного использования: класс окна
создания нового сообщения является подклассом окна просмотра, а окна ответа
и пересылки являются подклассами окна создания нового сообщения.
Окна, определяемые в этом файле, создаются окнами со списками, в ответ
на действия пользователя.

Предупреждение: действие кнопки 'Split', открывающей части/вложения,
недостаточно очевидно. 2.1: эта проблема была устранена добавлением кнопок
быстрого доступа к вложениям.
Новое в версии 3.0: для размещения полей заголовков вместо фреймов колонок
задействован менеджер компоновки grid(), одинаково хорошо действующий
на всех платформах.
Новое в версии 3.0: добавлена поддержка кодировок Юникода для основного
текста сообщения + текстовых вложений при передаче.
Новое в версии 3.0: PyEdit поддерживает произвольные кодировки Юникода
для просматриваемых частей сообщений.
Новое в версии 3.0: реализована поддержка кодировок Юникода и форматов MIME
для заголовков в отправляемых сообщениях.

Что сделать: можно было бы не выводить запрос перед закрытием окна, если
текст сообщения не изменялся (как в PyEdit2.0), но эти окна несколько
больше, чем просто редактор, и не определяют факт изменения заголовков.
```

```
Что сделать: возможно, следует сделать диалог выбора файла в окнах создания
новых сообщений глобальным для всей программы? (сейчас каждое окно создает
собственный диалог).
##############################################################################
"""

from SharedNames import * # объекты, глобальные для всей программы

##############################################################################
# окно просмотра сообщения - также является суперклассом для окон создания
# нового сообщения, ответа и пересылки
##############################################################################

class ViewWindow(windows.PopupWindow, mailtools.MailParser):
    """
    подкласс класса Toplevel с дополнительными методами и встроенным
    объектом TextEditor; наследует saveParts, partsList
    из mailtools.MailParser; подмешивается в логику специализированных
    подклассов прямым наследованием;
    """
    # атрибуты класса
    modelabel = 'View'                      # используется в заголовках окон
    from mailconfig import okayToOpenParts  # открывать ли вложения?
    from mailconfig import verifyPartOpens  # выводить запрос перед
                                            # открытием каждой части?
    from mailconfig import maxPartButtons       # макс. число кнопок + '...'
    from mailconfig import skipTextOnHtmlPart   # 3.0: только в броузере,
                                                # не использовать PyEdit?
    tempPartDir = 'TempParts'                   # каталог для временного
                                                # сохранения вложений

    # все окна просмотра используют один и тот же диалог: запоминается
    # последний каталог
    partsDialog = Directory(title=appname + ': Select parts save directory')

    def __init__(self, headermap, showtext, origmessage=None):
        """
        карта заголовков - это origmessage или собственный словарь
        с заголовками (для операции создания нового письма);
        showtext - основная текстовая часть сообщения: извлеченная
        из сообщения или любой другой текст;
        origmessage - объект email.message.Message для просмотра
        """
        windows.PopupWindow.__init__(self, appname, self.modelabel)
        self.origMessage = origmessage
        self.makeWidgets(headermap, showtext)

    def makeWidgets(self, headermap, showtext):
        """
        добавляет поля заголовков, кнопки операций и доступа к вложениям,
        текстовый редактор
```

```
        3.0: предполагается, что в аргументе showtext передается
        декодированная строка Юникода str; она будет кодироваться
        при отправке или при сохранении;
        """
        actionsframe = self.makeHeaders(headermap)
        if self.origMessage and self.okayToOpenParts:
            self.makePartButtons()
        self.editor = textEditor.TextEditorComponentMinimal(self)
        myactions   = self.actionButtons()
        for (label, callback) in myactions:
            b = Button(actionsframe, text=label, command=callback)
            b.config(bg='beige', relief=RIDGE, bd=2)
            b.pack(side=TOP, expand=YES, fill=BOTH)

        # тело текста, присоединяется последним = усекается первым
        self.editor.pack(side=BOTTOM)    # может быть несколько редакторов
        self.update()                    # 3.0: иначе текстовый курсор
                                         # встанет в строке
        self.editor.setAllText(showtext) # каждый имеет собств. содержимое
        lines = len(showtext.splitlines())
        lines = min(lines + 3, mailconfig.viewheight or 20)
        self.editor.setHeight(lines)   # иначе высота=24, ширина=80
        self.editor.setWidth(80)       # или из textConfig редактора PyEdit
        if mailconfig.viewbg:
            self.editor.setBg(mailconfig.viewbg) # цвета, шрифты в mailconfig
        if mailconfig.viewfg:
            self.editor.setFg(mailconfig.viewfg)
        if mailconfig.viewfont:        # также через меню Tools редактора
            self.editor.setFont(mailconfig.viewfont)

    def makeHeaders(self, headermap):
        """
        добавляет поля ввода заголовков, возвращает фрейм
        с кнопками операций;
        3.0: для создания рядов метка/поле ввода использует менеджер
        компоновки grid(), одинаково хорошо действующий на всех платформах;
        также можно было бы использовать менеджер компоновки pack()
        с фреймами рядов и метками фиксированной ширины;

        3.0: декодирование интернационализированных заголовков
        (и компонентов имен в заголовках с адресами электронной почты)
        выполняется здесь, если это необходимо, так как они добавляются
        в графический интерфейс; некоторые заголовки, возможно, уже были
        декодированы перед созданием окон ответа/пересылки, где требуется
        использовать декодированный текст, но лишнее декодирование здесь
        не вредит им и оно необходимо для других заголовков и случаев, таких
        как просмотр полученных сообщений; при отображении в графическом
        интерфейсе заголовки всегда находятся в декодированной форме
        и будут кодироваться внутри пакета mailtools при передаче, если они
        содержат символы за пределами диапазона ASCII (смотрите реализацию
        класса Write); декодирование интернационализированных заголовков
```

```
        также происходит в окне с оглавлением почты и при добавлении
        заголовков в цитируемый текст; текстовое содержимое в теле письма
        также декодируется перед отображением и кодируется перед передачей
        в другом месте в системе (окна со списками, класс WriteWindow);

        3.0: при создании окна редактирования вызывающий программный код
        записывает адрес отправителя в заголовок Bcc, который подхватывается
        здесь для удобства в типичных ситуациях, если этот заголовок
        разрешен в mailconfig; при создании окна ответа также заполняется
        заголовок Cc, если разрешен, уникальными адресами получателей
        оригинального письма, включая адрес в заголовке From;
        """
        top    = Frame(self); top.pack   (side=TOP,  fill=X)
        left   = Frame(top);  left.pack  (side=LEFT, expand=NO,  fill=BOTH)
        middle = Frame(top);  middle.pack(side=LEFT, expand=YES, fill=X)

        # множество заголовков может быть расширено в mailconfig (Bcc и др.?)
        self.userHdrs = ()
        showhdrs = ('From', 'To', 'Cc', 'Subject')
        if hasattr(mailconfig, 'viewheaders') and mailconfig.viewheaders:
            self.userHdrs = mailconfig.viewheaders
            showhdrs += self.userHdrs
        addrhdrs = ('From', 'To', 'Cc', 'Bcc')  # 3.0: декодируются отдельно
        self.hdrFields = []
        for (i, header) in enumerate(showhdrs):
            lab = Label(middle, text=header+':', justify=LEFT)
            ent = Entry(middle)
            lab.grid(row=i, column=0, sticky=EW)
            ent.grid(row=i, column=1, sticky=EW)
            middle.rowconfigure(i, weight=1)
            hdrvalue = headermap.get(header, '?')  # может быть пустым
            # 3.0: если закодирован, декодировать с учетом email+mime+юникод
            if header not in addrhdrs:
                hdrvalue = self.decodeHeader(hdrvalue)
            else:
                hdrvalue = self.decodeAddrHeader(hdrvalue)
            ent.insert('0', hdrvalue)
            self.hdrFields.append(ent)    # порядок имеет значение в onSend
        middle.columnconfigure(1, weight=1)
        return left

    def actionButtons(self):       # должны быть методами для доступа к self
        return [('Cancel', self.destroy), # закрыть окно просмотра тихо
                ('Parts',  self.onParts), # список частей или тело
                ('Split', self.onSplit)]

    def makePartButtons(self):
        """
        добавляет до N кнопок быстрого доступа к частям/вложениям;
        альтернатива кнопкам Parts/Split (2.1); это нормально,
        когда временный каталог совместно используется всеми операциями:
```

```
        файл вложения не сохраняется, пока позднее не будет выбран и открыт;
        partname=partname требуется в lambda-выражениях в Py2.4;
        предупреждение: можно было бы попробовать пропустить главную
        текстовую часть;
        """
        def makeButton(parent, text, callback):
            link = Button(parent, text=text, command=callback, relief=SUNKEN)
            if mailconfig.partfg: link.config(fg=mailconfig.partfg)
            if mailconfig.partbg: link.config(bg=mailconfig.partbg)
            link.pack(side=LEFT, fill=X, expand=YES)

        parts = Frame(self)
        parts.pack(side=TOP, expand=NO, fill=X)
        for (count, partname) in enumerate(self.partsList(self.origMessage)):
            if count == self.maxPartButtons:
                makeButton(parts, '...', self.onSplit)
                break
            openpart = (lambda partname=partname: self.onOnePart(partname))
            makeButton(parts, partname, openpart)

    def onOnePart(self, partname):
        """
        отыскивает часть, соответствующую кнопке, сохраняет ее и открывает;
        допускается открывать несколько сообщений: каждый раз сохранение
        производится заново; вероятно, мы могли бы здесь просто
        использовать веб-броузер;
        предупреждение: tempPartDir содержит путь относительно cwd -
        может быть любым каталогом;
        предупреждение: tempPartDir никогда не очищается: может занимать
        много места на диске, можно было бы использовать модуль tempfile
        (как при отображении главной текстовой части в формате HTML
        в методе onView класса окна со списком);
        """
        try:
            savedir = self.tempPartDir
            message = self.origMessage
            (contype, savepath) = self.saveOnePart(savedir, partname, message)
        except:
            showerror(appname, 'Error while writing part file')
            printStack(sys.exc_info())
        else:
            self.openParts([(contype,
                             os.path.abspath(savepath))])   # повт. исп.

    def onParts(self):
        """
        отображает содержимое части/вложения сообщения в отдельном окне;
        использует ту же схему именования файлов, что и операция Split;
        если сообщение содержит единственную часть, она является
        текстовым телом
        """
```

```
        partnames = self.partsList(self.origMessage)
        msg = '\n'.join(['Message parts:\n'] + partnames)
        showinfo(appname, msg)

    def onSplit(self):
        """
        выводит диалог выбора каталога и сохраняет туда все части/вложения;
        при желании мультимедийные части и HTML открываются
        в веб-броузере, текст - в TextEditor, а документы
        известных типов - в соответствующих программах Windows;
        можно было бы отображать части в окнах View, где имеется
        встроенный текстовый редактор с функцией сохранения,
        но большинство частей являются нечитаемым текстом;
        """
        savedir = self.partsDialog.show() # атрибут класса: предыдущий каталог
        if savedir:                # диалог tk выбора каталога, не файла
            try:
                partfiles = self.saveParts(savedir, self.origMessage)
            except:
                showerror(appname, 'Error while writing part files')
                printStack(sys.exc_info())
            else:
                if self.okayToOpenParts: self.openParts(partfiles)

    def askOpen(self, appname, prompt):
        if not self.verifyPartOpens:
            return True
        else:
            return askyesno(appname, prompt) # диалог

    def openParts(self, partfiles):
        """
        автоматически открывает файлы известных и безопасных типов,
        но только после подтверждения пользователем; файлы других
        типов должны открываться вручную из каталога сохранения;
        к моменту вызова этого метода именованные части уже преобразованы
        из формата MIME и хранятся в двоичных файлах, однако текстовые
        файлы могут содержать текст в любой кодировке; редактору PyEdit
        требуется знать кодировку для декодирования, веб-броузеры могут
        пытаться сами определять кодировку или позволять сообщать ее им;

        предупреждение: не открывает вложения типа application/octet-stream,
        даже если имя файла имеет безопасное расширение, такое как .html;
        предупреждение: изображения/аудио/видео можно было бы открывать
        с помощью сценария playfile.py из этой книги; в случае ошибки
        средства просмотра текста: в Windows можно было бы также
        запускать Notepad Блокнот) с помощью startfile;
        (в большинстве случаев также можно было бы использовать
        модуль webbrowser, однако специализированный
        инструмент всегда лучше универсального;
        """
```

```
def textPartEncoding(fullfilename):
    """
    3.0: отображает имя файла текстовой части в содержимое
    параметра charset в заголовке content-type для данной части
    сообщения Message, которое затем передается конструктору PyEdit,
    чтобы обеспечить корректное отображение текста; для текстовых
    частей можно было бы возвращать параметр charset вместе
    с content-type из mailtools, однако проще обрабатывать эту
    ситуацию как особый случай здесь;

    содержимое части сохраняется пакетом mailtools в двоичном
    режиме, чтобы избежать проблем с кодировками, но здесь
    отсутствует прямой доступ к оригинальной части сообщения;
    необходимо выполнить это отображение, чтобы получить
    имя кодировки, если оно присутствует;
    редактор PyEdit в 4 издании теперь позволяет явно указывать
    кодировку открываемого файла и определяет кодировку
    при сохранении; смотрите главу 11, где описываются особенности
    поведения PyEdit: он запрашивает кодировку у пользователя,
    если кодировка не указана или оказывается неприменимой;
    предупреждение: перейти на mailtools.mailParser в PyMailCGI,
    чтобы повторно использовать для тега <meta>?
    """
    partname = os.path.basename(fullfilename)
    for (filename, contype, part) in \
        self.walkNamedParts(self.origMessage):
        if filename == partname:
                return part.get_content_charset()  # None, если нет
                                                   # в заг.
    assert False, 'Text part not found'   # никогда не должна
                                          # выполняться

for (contype, fullfilename) in partfiles:
    maintype  = contype.split('/')[0]             # левая часть
    extension = os.path.splitext(fullfilename)[1] # не [-4:]
    basename  = os.path.basename(fullfilename)    # отбросить путь

    # текст в формате HTML и XML, веб-страницы, некоторые
    # мультимедийные файлы
    if contype in ['text/html', 'text/xml']:
        browserOpened = False
        if self.askOpen(appname, 'Open "%s" in browser?' % basename):
            try:
                webbrowser.open_new('file://' + fullfilename)
                browserOpened = True
            except:
                showerror(appname, 'Browser failed: trying editor')
        if not browserOpened or not self.skipTextOnHtmlPart:
            try:
                # попробовать передать редактору PyEdit имя кодировки
                encoding = textPartEncoding(fullfilename)
```

```
                        textEditor.TextEditorMainPopup(parent=self,
                            winTitle=' - %s email part' % (encoding or '?'),
                            loadFirst=fullfilename, loadEncode=encoding)
                    except:
                        showerror(appname, 'Error opening text viewer')

            # text/plain, text/x-python и др.; 4E: кодировка может не подойти
            elif maintype == 'text':
                if self.askOpen(appname, 'Open text part "%s"?' % basename):
                    try:
                        encoding = textPartEncoding(fullfilename)
                        textEditor.TextEditorMainPopup(parent=self,
                            winTitle=' - %s email part' % (encoding or '?'),
                            loadFirst=fullfilename, loadEncode=encoding)
                    except:
                        showerror(appname, 'Error opening text viewer')

            # мультимедийные файлы: Windows открывает
            # mediaplayer, imageviewer и так далее
            elif maintype in ['image', 'audio', 'video']:
                if self.askOpen(appname, 'Open media part "%s"?' % basename):
                    try:
                        webbrowser.open_new('file://' + fullfilename)
                    except:
                        showerror(appname, 'Error opening browser')

            # типичные документы Windows: Word, Excel, Adobe, архивы и др.
            elif (sys.platform[:3] == 'win' and
                maintype == 'application' and                # 3.0: +x типы
                extension in ['.doc', '.docx', '.xls', '.xlsx',# обобщить
                              '.pdf', '.zip', '.tar', '.wmv']):
                    if self.askOpen(appname, 'Open part "%s"?' % basename):
                        os.startfile(fullfilename)
            else:     # пропустить!
                msg = 'Cannot open part: "%s"\nOpen manually in: "%s"'
                msg = msg % (basename, os.path.dirname(fullfilename))
                showinfo(appname, msg)

##############################################################################
# окна редактирования сообщений - операции создания нового сообщения,
# ответа и пересылки
##############################################################################

if mailconfig.smtpuser:         # пользователь определен в mailconfig?
    MailSenderClass = mailtools.MailSenderAuth   # требуется имя/пароль
else:
    MailSenderClass = mailtools.MailSender

class WriteWindow(ViewWindow, MailSenderClass):
    """
    специализирует окно просмотра для составления нового сообщения
```

```
наследует sendMessage из mailtools.MailSender
"""
modelabel = 'Write'

def __init__(self, headermap, starttext):
    ViewWindow.__init__(self, headermap, starttext)
    MailSenderClass.__init__(self)
    self.attaches = []       # каждое окно имеет свой диалог открытия
    self.openDialog = None   # диалог запоминает последний каталог

def actionButtons(self):     # должны быть методами для доступа к self
    return [('Cancel', self.quit),
            ('Parts',  self.onParts), # PopupWindow проверяет отмену
            ('Attach', self.onAttach),
            ('Send',   self.onSend)]  # 4E: без отступов: по центру

def onParts(self):
    # предупреждение: удаление в настоящее время не поддерживается
    if not self.attaches:
        showinfo(appname, 'Nothing attached')
    else:
        msg = '\n'.join(['Already attached:\n'] + self.attaches)
        showinfo(appname, msg)

def onAttach(self):
    """
    вкладывает файл в письмо: имя, добавляемое здесь, будет добавлено
    как часть в операции Send, внутри пакета mailtools;
    4E: имя кодировки Юникода можно было бы запрашивать здесь,
    а не при отправке
    """
    if not self.openDialog:
        self.openDialog = Open(title=appname + ': Select Attachment File')
    filename = self.openDialog.show()  # запомнить каталог
    if filename:
        self.attaches.append(filename) # для открытия в методе отправки

def resolveUnicodeEncodings(self):
    """
    3.0/4E: в качестве подготовки к отправке определяет кодировку Юникода
    для текстовых частей: для основной текстовой части и для любых
    текстовых вложений; кодировка для основной текстовой части может
    быть уже известна, если это ответ или пересылка, но она не известна
    при создании нового письма, к тому же в результате редактирования
    кодировка может измениться; модуль smtplib в 3.1 требует, чтобы
    полный текст отправляемого сообщения содержал только символы ASCII
    (если это str), поэтому так важно определить кодировку прямо здесь;
    иначе будет возникать ошибка при попытке отправить
    ответ/пересылаемое письмо с текстом в кодировке UTF8, когда
    установлен параметр config=ascii, а текст содержит символы
    вне диапазона ASCII; пытается использовать настройки пользователя
```

```
и выполнить ответ, а в случае неудачи возвращается к универсальной
кодировке UTF8 как к последней возможности;
"""

def isTextKind(filename):
    contype, encoding = mimetypes.guess_type(filename)
    if contype is None or encoding is not None:  # утилита 4E
        return False          # не определяется, сжатый файл?
    maintype, subtype = contype.split('/', 1)  # проверить на text/?
    return maintype == 'text'

# выяснить кодировку основного текста
bodytextEncoding = mailconfig.mainTextEncoding
if bodytextEncoding == None:
    asknow = askstring('PyMailGUI',
                       'Enter main text Unicode encoding name')
    bodytextEncoding = asknow or 'latin-1' # или
                                         # sys.getdefaultencoding()?

# последний шанс: использовать utf-8, если кодировку
# так и не удалось определить выше
if bodytextEncoding != 'utf-8':
    try:
        bodytext = self.editor.getAllText()
        bodytext.encode(bodytextEncoding)
    except (UnicodeError, LookupError): # Lookup: неверная кодировка
        bodytextEncoding = 'utf-8'      # универсальная кодировка

# определить кодировки текстовых вложений
attachesEncodings = []
config = mailconfig.attachmentTextEncoding
for filename in self.attaches:
    if not isTextKind(filename):
        attachesEncodings.append(None)   # не текст: не спрашивать
    elif config != None:
        attachesEncodings.append(config) # для всех текстовых
    else:                                # частей, если установлена
        prompt = 'Enter Unicode encoding name for %' % filename
        asknow = askstring('PyMailGUI', prompt)
        attachesEncodings.append(asknow or 'latin-1')

# последний шанс: использовать utf-8, если кодировку
# так и не удалось определить выше
choice = attachesEncodings[-1]
if choice != None and choice != 'utf-8':
    try:
        attachbytes = open(filename, 'rb').read()
        attachbytes.decode(choice)
    except (UnicodeError, LookupError, IOError):
        attachesEncodings[-1] = 'utf-8'
return bodytextEncoding, attachesEncodings
```

```
def onSend(self):
    """
    может вызываться из потока: обработчик кнопки
    Send (Отправить) в окне редактирования;
    может перекрываться во времени с любыми другими потоками
    выполнения, не запрещает никаких операций, кроме завершения;
    обработчики Exit, Fail выполняются методом threadChecker
    посредством очереди в обработчике after;
    предупреждение: здесь не предусматривается вывод информации о ходе
    выполнения, потому что операция отправки почты является атомарной;
    допускается указывать несколько адресов получателей,
    разделенных ',';
    пакет mailtools решает проблемы с кодировками, обрабатывает
    вложения, форматирует строку даты и так далее; кроме того,
    пакет mailtools сохраняет текст отправленных сообщений
    в локальном файле

    3.0: теперь выполняется полный разбор заголовков To,Cc,Bcc
    (в mailtools) вместо простого разбиения по символу-разделителю;
    вместо простых полей ввода можно было бы использовать
    многострочные виджеты;
    содержимое заголовка Bcc добавляется на "конверт",
    а сам заголовок удаляется;

    3.0: кодировка Юникода для текстовых частей определяется здесь,
    потому что она может потребоваться для вывода подсказок
    в графическом интерфейсе; фактическое кодирование
    текстовых частей выполняется внутри
    пакета mailtools, если это необходимо;

    3.0: интернационализированные заголовки уже декодируются в полях
    ввода графического интерфейса; кодирование любых
    интернационализированных заголовков с символами вне диапазона ASCII
    выполняется не здесь, а в пакете mailtools, потому что эта операция
    не требуется для работы графического интерфейса;
    """

    # определить кодировку для текстовых частей;
    bodytextEncoding, attachesEncodings = self.resolveUnicodeEncodings()

    # получить компоненты графического интерфейса;
    # 3.0: интернационализированные заголовки уже декодированы
    fieldvalues = [entry.get() for entry in self.hdrFields]
    From, To, Cc, Subj = fieldvalues[:4]
    extraHdrs = [('Cc', Cc), ('X-Mailer', appname + ' (Python)')]
    extraHdrs += list(zip(self.userHdrs, fieldvalues[4:]))
    bodytext = self.editor.getAllText()

    # разбить список получателей на адреса по ',', исправить пустые поля
    Tos = self.splitAddresses(To)
    for (ix, (name, value)) in enumerate(extraHdrs):
```

```
            if value:                               # игнорировать, если ''
                if value == '?':                    # ? не заменяется
                    extraHdrs[ix] = (name, '')
                elif name.lower() in ['cc', 'bcc']: # разбить по ','
                    extraHdrs[ix] = (name, self.splitAddresses(value))

        # метод withdraw вызывается, чтобы предотвратить повторный запуск
        # передачи во время передачи
        # предупреждение: не устраняет вероятность ошибки полностью -
        # пользователь может восстановить окно, если значок
        # останется видимым
        self.withdraw()
        self.getPassword() # если необходимо; не запускайте диалог в потоке!
        popup = popuputil.BusyBoxNowait(appname, 'Sending message')
        sendingBusy.incr()
        threadtools.startThread(
            action = self.sendMessage,
            args   = (From, Tos, Subj, extraHdrs, bodytext, self.attaches,
                                       saveMailSeparator,
                                       bodytextEncoding,
                                       attachesEncodings),
            context = (popup,),
            onExit  = self.onSendExit,
            onFail  = self.onSendFail)

    def onSendExit(self, popup):
        """
        стирает окно ожидания, стирает окно просмотра, уменьшает счетчик
        операций отправки; метод sendMessage автоматически сохраняет
        отправленное сообщение в локальном файле; нельзя использовать
        window.addSavedMails: текст почтового сообщения недоступен;
        """
        popup.quit()
        self.destroy()
        sendingBusy.decr()

        # может быть \ при открытии, в mailconfig используется /
        sentname = os.path.abspath(mailconfig.sentmailfile) # расширяет '.'
        if sentname in openSaveFiles.keys():              # файл открыт?
            window = openSaveFiles[sentname]              # обновить список
            window.loadMailFileThread()                   # и поднять окно

    def onSendFail(self, exc_info, popup):
        # выводит диалог с сообщением об ошибке, оставляет нетронутым окно
        # с сообщением, чтобы имелась возможность сохранить или повторить
        # попытку, перерисовывает фрейм
        popup.quit()
        self.deiconify()
        self.lift()
        showerror(appname, 'Send failed: \n%s\n%s' % exc_info[:2])
```

```
            printStack(exc_info)
            MailSenderClass.smtpPassword = None # повт. попытку; 3.0/4E: не в self
            sendingBusy.decr()

        def askSmtpPassword(self):
            """
            получает пароль, если он необходим графическому интерфейсу,
            вызывается из главного потока выполнения;
            предупреждение: чтобы не возникла необходимость запрашивать пароль
            в потоке выполнения, если он не был введен в первый раз, выполняет
            цикл, пока пользователь не введет пароль; смотрите логику получения
            пароля доступа к POP-серверу, где приводится альтернативный
            вариант без цикла
            """
            password = ''
            while not password:
                prompt = ('Password for %s on %s?' %
                          (self.smtpUser, self.smtpServerName))
                password = popuputil.askPasswordWindow(appname, prompt)
            return password

class ReplyWindow(WriteWindow):
    """
    специализированная версия окна создания сообщения для ответа
    текст и заголовки поставляются окном со списком
    """
    modelabel = 'Reply'

class ForwardWindow(WriteWindow):
    """
    специализированная версия окна создания сообщения для пересылки
    текст и заголовки поставляются окном со списком
    """
    modelabel = 'Forward'
```

messagecache: менеджер кэша сообщений

Класс в примере 14.5 реализует кэш для хранения загруженных сообщений. Его логика была выделена в отдельный файл, чтобы не загромождать реализацию окон со списками. Окно со списком сообщений на сервере создает и встраивает экземпляр этого класса для обеспечения взаимодействий с почтовым сервером и сохранения загруженных заголовков и полного текста сообщений. В этой версии окно со списком сообщений на сервере также запоминает, какие письма загружаются в текущий момент, чтобы избежать попытки загрузить одно и то же письмо несколько раз в параллельных потоках. Данная задача не была реализована здесь лишь потому, что она может потребовать операций с графическим интерфейсом.

Пример 14.5. PP4E\Internet\Email\PyMailGui\messagecache.py

```
"""
##############################################################################
управляет загрузкой сообщений с заголовками и контекстом, но не графическим
интерфейсом; подкласс класса MailFetcher, со списком загруженных заголовков
и сообщений; вызывающая программа сама должна заботиться о поддержке потоков
выполнения и графического интерфейса;

изменения в версии 3.0: использует кодировку для полного текста сообщений
из локального модуля mailconfig; декодирование выполняется глубоко в недрах
mailtools, после загрузки текст сообщения всегда возвращается в виде
строки Юникода str; это может измениться в будущих версиях Python/email:
подробности смотрите в главе 13;

изменения в версии 3.0: поддерживает новую особенность mailconfig.fetchlimit
в mailtools, которая может использоваться для ограничения максимального
числа самых свежих заголовков или сообщений (если не поддерживается команда
TOP), загружаемых при каждом запросе на загрузку; обратите внимание, что эта
особенность является независимой от параметра loadfrom, используемого здесь,
чтобы ограничить загрузку только самыми новыми сообщениями, хотя они
и используются одновременно: загружается не больше чем fetchlimit вновь
поступивших сообщений;

изменения в версии 3.0: есть вероятность, что пользователь запросит загрузку
сообщения, которое в текущий момент уже загружается в параллельном потоке,
просто щелкнув на сообщении еще раз (операции загрузки сообщений, в отличие
от полной загрузки оглавления, могут перекрываться во времени с другими
операциями загрузки и отправки); в этом нет никакой опасности, но это может
привести к излишней и, возможно, параллельной загрузке одного и того же
письма, что бессмысленно и ненужно (если выбрать все сообщения в списке
и дважды нажать кнопку View, это может вызвать загрузку большинства
сообщений дважды!); в главном потоке графического интерфейса слежение
за загружаемыми сообщениями, чтобы  такое перекрытие во времени не было
возможным: загружаемое сообщение препятствует выполнению операций загрузки
любых наборов сообщений, в которых оно присутствует, параллельная загрузка
непересекающихся множеств сообщений по-прежнему возможна;
##############################################################################
"""

from PP4E.Internet.Email import mailtools
from popuputil import askPasswordWindow

class MessageInfo:
    """
    элемент списка в кэше
    """
    def __init__(self, hdrtext, size):
        self.hdrtext = hdrtext # fulltext - кэшированное сообщение
        self.fullsize = size   # hdrtext - только заголовки
        self.fulltext = None   # fulltext=hdrtext если не работает
                               # команда TOP
```

```
class MessageCache(mailtools.MailFetcher):
    """
    следит за уже загруженными заголовками и сообщениями;
    наследует от MailFetcher методы взаимодействия с сервером;
    может использоваться в других приложениях: ничего не знает о графическом
    интерфейсе или поддержке многопоточной модели выполнения;

    3.0: байты исходного полного текста сообщения декодируются в str, чтобы
    обеспечить возможность анализа пакетом email в Py3.1 и сохранения
    в файлах; использует настройки определения кодировок из локального
    модуля mailconfig; декодирование выполняется автоматически
    в пакете mailtools при получении;
    """
    def __init__(self):
        mailtools.MailFetcher.__init__(self) # 3.0: наследует fetchEncoding
        self.msglist = []                    # 3.0: наследует fetchlimit

    def loadHeaders(self, forceReloads, progress=None):
        """
        здесь обрабатываются три случая: первоначальная загрузка всего
        списка, загрузка вновь поступившей почты и принудительная
        перезагрузка после удаления; не получает повторно просмотренные
        сообщения, если список заголовков остался прежним или был дополнен;
        сохраняет кэшированные сообщения после удаления, если операция
        удаления завершилась успешно;
        2.1: выполняет быструю проверку синхронизации номеров сообщений
        3.0: теперь учитывает максимум mailconfig.fetchlimit;
        """
        if forceReloads:
            loadfrom = 1
            self.msglist = []              # номера сообщений изменились
        else:
            loadfrom = len(self.msglist)+1 # продолжить с места посл. загрузки

        # только если загружается вновь поступившая почта
        if loadfrom != 1:
            self.checkSynchError(self.allHdrs()) # возб. искл. при рассинхр.

        # получить все или только новые сообщения
        reply = self.downloadAllHeaders(progress, loadfrom)
        headersList, msgSizes, loadedFull = reply

        for (hdrs, size) in zip(headersList, msgSizes):
            newmsg = MessageInfo(hdrs, size)
            if loadedFull:              # zip может вернуть пустой результат
                newmsg.fulltext = hdrs  # получить полные сообщения, если
            self.msglist.append(newmsg) # не поддерживается команда TOP

    def getMessage(self, msgnum):       # получает исходный текст сообщения
        cacheobj = self.msglist[msgnum-1] # добавляет в кэш, если получено
        if not cacheobj.fulltext:         # безопасно использовать в потоках
```

```
            fulltext = self.downloadMessage(msgnum)  # 3.0: более простое
            cacheobj.fulltext = fulltext             # кодирование
        return cacheobj.fulltext

    def getMessages(self, msgnums, progress=None):
        """
        получает полный текст нескольких сообщений, может
        вызываться в потоках выполнения;
        2.1: выполняет быструю проверку синхронизации номеров сообщений;
        нельзя получить сообщения здесь, если не было загружено оглавление;
        """
        self.checkSynchError(self.allHdrs())    # возб. искл. при рассинхр.
        nummsgs = len(msgnums)                  # добавляет сообщения в кэш
        for (ix, msgnum) in enumerate(msgnums): # некоторые возм. уже в кэше
            if progress: progress(ix+1, nummsgs) # подключ. только при необх.
            self.getMessage(msgnum)         # но может выполнять подключ.
                                            # более одного раза

    def getSize(self, msgnum):              # инкапсулирует структуру кэша
        return self.msglist[msgnum-1].fullsize  # уже изменялось однажды!

    def isLoaded(self, msgnum):
        return self.msglist[msgnum-1].fulltext

    def allHdrs(self):
        return [msg.hdrtext for msg in self.msglist]

    def deleteMessages(self, msgnums, progress=None):
        """
        если удаление всех номеров сообщений возможно, изымает удаленные
        элементы из кэша, но не перезагружает ни список заголовков,
        ни текст уже просмотренных сообщений: список в кэше будет отражать
        изменение номеров сообщений на сервере; если удаление завершилось
        неудачей по каким-либо причинам, вызывающая программа должна
        принудительно перезагрузить все заголовки, расположенные выше,
        потому что номера некоторых сообщений на сервере могут измениться
        непредсказуемым образом;
        2.1: теперь проверяет синхронизацию номеров сообщений, если команда
        TOP поддерживается сервером; может вызываться в потоках выполнения
        """
        try:
            self.deleteMessagesSafely(msgnums, self.allHdrs(), progress)
        except mailtools.TopNotSupported:
            mailtools.MailFetcher.deleteMessages(self, msgnums, progress)

        # ошибок не обнаружено: обновить список оглавления
        indexed = enumerate(self.msglist)
        self.msglist = [msg for (ix, msg) in indexed if ix+1 not in msgnums]

class GuiMessageCache(MessageCache):
    """
```

```
    вызовы графического интерфейса добавляются здесь, благодаря чему
    кэш можно использовать в приложениях без графического интерфейса
    """


    def setPopPassword(self, appname):
        """
        получает пароль с помощью графического интерфейса, вызывается
        в главном потоке; принудительно вызывается из графического
        интерфейса, чтобы избежать вызова из дочерних потоков выполнения
        """
        if not self.popPassword:
            prompt = 'Password for %s on %s?' % (self.popUser, self.popServer)
            self.popPassword = askPasswordWindow(appname, prompt)

    def askPopPassword(self):
        """
        но здесь не использует графический интерфейс: я вызываю его
        из потоков!; попытка вывести диалог в дочернем потоке выполнения
        подвесит графический интерфейс; может вызываться суперклассом
        MailFetcher, но только если пароль остается пустой строкой
        из-за закрытия окна диалога
        """
        return self.popPassword
```

popuputil: диалоги общего назначения

В примере 14.6 реализовано несколько удобных вспомогательных диалогов, которые могут пригодиться в разных программах. Обратите внимание, что здесь импортируется уже знакомый нам вспомогательный модуль `windows`, обеспечивающий единство стиля диалогов (значки, заголовки и так далее).

Пример 14.6. PP4E\Internet\Email\PyMailGui\popuputil.py

```
"""
##############################################################################
вспомогательные окна - могут пригодиться в других программах
##############################################################################
"""


from tkinter import *
from PP4E.Gui.Tools.windows import PopupWindow

class HelpPopup(PopupWindow):
    """
    специализированная версия Toplevel, отображающая
    справочный текст в области с прокруткой
    кнопка Source вызывает указанный обработчик обратного вызова
    альтернатива в версии 3.0: использовать файл HTML и модуль webbrowser
    """
    myfont = 'system' # настраивается
```

```
    mywidth = 78      # 3.0: начальная ширина

    def __init__(self, appname, helptext, iconfile=None, showsource=lambda:0):
        PopupWindow.__init__(self, appname, 'Help', iconfile)
        from tkinter.scrolledtext import ScrolledText # немодальный диалог
        bar = Frame(self)         # присоедин-ся первым - усекается последним
        bar.pack(side=BOTTOM, fill=X)
        code = Button(bar, bg='beige', text="Source", command=showsource)
        quit = Button(bar, bg='beige', text="Cancel", command=self.destroy)
        code.pack(pady=1, side=LEFT)
        quit.pack(pady=1, side=LEFT)
        text = ScrolledText(self)          # добавить Text + полосы прокр.
        text.config(font=self.myfont)
        text.config(width=self.mywidth)    # слишком большой для showinfo
        text.config(bg='steelblue', fg='white')  # закрыть при нажатии
                                                 # на кнопку
        text.insert('0.0', helptext)       # или на клавишу Return
        text.pack(expand=YES, fill=BOTH)
        self.bind("<Return>", (lambda event: self.destroy()))

    def askPasswordWindow(appname, prompt):
        """
        модальный диалог для ввода строки пароля
        функция getpass.getpass использует stdin, а не графический интерфейс
        tkSimpleDialog.askstring выводит ввод эхом
        """
        win = PopupWindow(appname, 'Prompt') # настроенный экземпляр Toplevel
        Label(win, text=prompt).pack(side=LEFT)
        entvar = StringVar(win)
        ent = Entry(win, textvariable=entvar, show='*') # показывать *
        ent.pack(side=RIGHT, expand=YES, fill=X)
        ent.bind('<Return>', lambda event: win.destroy())
        ent.focus_set(); win.grab_set(); win.wait_window()
        win.update()          # update вызывает принудительную перерисовку
        return entvar.get()   # виджет ent к этому моменту уже уничтожен

class BusyBoxWait(PopupWindow):
    """
    блокирующее окно с сообщением: выполнение потока приостанавливается
    цикл обработки событий главного потока графического интерфейса
    продолжает выполняться, но сам графический интерфейс
    не действует, пока открыто это окно;
    используется переопределенная версия метода quit, потому что в дереве
    наследования он находится ниже, а не левее;
    """
    def __init__(self, appname, message):
        PopupWindow.__init__(self, appname, 'Busy')
        self.protocol('WM_DELETE_WINDOW', lambda:0) # игнор. попытку закрыть
        label = Label(self, text=message + '...') # win.quit(), чтобы закрыть
        label.config(height=10, width=40, cursor='watch') # курсор занятости
        label.pack()
```

```
        self.makeModal()
        self.message, self.label = message, label

    def makeModal(self):
        self.focus_set() # захватить фокус ввода
        self.grab_set()  # ждать вызова threadexit

    def changeText(self, newtext):
        self.label.config(text=self.message + ': ' + newtext)

    def quit(self):
        self.destroy()       # не запрашивать подтверждение

class BusyBoxNowait(BusyBoxWait):
    """
    неблокирующее окно
    вызывайте changeText, чтобы отобразить ход выполнения операции,
    quit - чтобы закрыть окно
    """
    def makeModal(self):
        pass

if __name__ == '__main__':
    HelpPopup('spam', 'See figure 1...\n')
    print(askPasswordWindow('spam', 'enter password'))
    input('Enter to exit') # пауза, если сценарий был запущен щелчком мыши
```

wraplines: инструменты разбиения строк

Модуль в примере 14.7 реализует универсальные инструменты, используемые для переноса длинных строк в фиксированной позиции или по первому разделителю перед фиксированной позицией. Программа PyMailGUI использует функцию `wrapText1` из этого модуля для оформления текста в окнах просмотра, создания ответа и пересылки, однако этот модуль может пригодиться и в других программах. Запустите этот файл как самостоятельный сценарий, чтобы увидеть, как действует его программный код самотестирования, и изучите его функции, чтобы понять логику обработки текста.

Пример 14.7. PP4E\Internet\Email\PyMailGui\wraplines.py

```
"""
##############################################################################
разбивает строки по фиксированной позиции или по первому разделителю перед
фиксированной позицией; смотрите также: иной, но похожий модуль textwrap
в стандартной библиотеке (2.3+);
4E предупреждение: предполагается работа со строками str; поддержка строк
типа bytes могла бы помочь избежать некоторых сложностей с декодированием;
##############################################################################
"""
```

```
defaultsize = 80

def wrapLinesSimple(lineslist, size=defaultsize):
    "разбивает по фиксированной позиции"
    wraplines = []
    for line in lineslist:
        while True:
            wraplines.append(line[:size]) # OK, если длина строки < size
            line = line[size:]            # разбить без анализа
            if not line: break
    return wraplines

def wrapLinesSmart(lineslist, size=defaultsize, delimiters='.,:\t '):
    "выполняет перенос по первому разделителю левее позиции size"
    wraplines = []
    for line in lineslist:
        while True:
            if len(line) <= size:
                wraplines += [line]
                break
            else:
                for look in range(size-1, size // 2, -1): # 3.0: // а не /
                    if line[look] in delimiters:
                        front, line = line[:look+1], line[look+1:]
                        break
                else:
                    front, line = line[:size], line[size:]
                wraplines += [front]
    return wraplines

##############################################################################
# утилиты для типичных случаев использования
##############################################################################

def wrapText1(text, size=defaultsize): # лучше для текста из строк: почта
    "когда текст читается целиком "    # сохраняет первонач. структуру строк
    lines = text.split('\n')           # разбить по '\n'
    lines = wrapLinesSmart(lines, size)  # перенести по разделителям
    return '\n'.join(lines)              # объединить все вместе

def wrapText2(text, size=defaultsize): # более равномерное разбиение
    "то же, но текст - одна строка"    # но теряется первонач. структ. строк
    text = text.replace('\n', ' ')       # отбросить '\n', если имеются
    lines = wrapLinesSmart([text], size) # перенести единую строку по разд.
    return lines                         # объединение выполняет вызывающий

def wrapText3(text, size=defaultsize):
    "то же, но выполняет объединение"
    lines = wrapText2(text, size)        # перенос как одной длинную строку
    return '\n'.join(lines) + '\n'       # объединить, добавляя '\n'
```

```
def wrapLines1(lines, size=defaultsize):
    "когда символ перевода строки добавляется в конец"
    lines = [line[:-1] for line in lines]    # отбросить '\n' (или .rstrip)
    lines = wrapLinesSmart(lines, size)      # перенести по разделителям
    return [(line + '\n') for line in lines] # объединить

def wrapLines2(lines, size=defaultsize):   # более равномерное разбиение
    "то же, но объединяет в одну строку "  # но теряется первонач. структура
    text = ''.join(lines)                    # объединить в 1 строку
    lines = wrapText2(text)                  # перенести по разделителям
    return [(line + '\n') for line in lines] # добавить '\n' в концы строк

##############################################################################
# самотестирование
##############################################################################

if __name__ == '__main__':
    lines = ['spam ham ' * 20 + 'spam,ni' * 20,
             'spam ham ' * 20,
             'spam,ni' * 20,
             'spam ham.ni' * 20,
             '',
             'spam'*80,
             ' ',
             'spam ham eggs']

    sep = '-' * 30
    print('all', sep)
    for line in lines: print(repr(line))
    print('simple', sep)
    for line in wrapLinesSimple(lines): print(repr(line))
    print('smart', sep)
    for line in wrapLinesSmart(lines): print(repr(line))

    print('single1', sep)
    for line in wrapLinesSimple([lines[0]], 60): print(repr(line))
    print('single2', sep)
    for line in wrapLinesSmart([lines[0]], 60): print(repr(line))
    print('combined text', sep)
    for line in wrapLines2(lines): print(repr(line))
    print('combined lines', sep)
    print(wrapText1('\n'.join(lines)))

    assert ''.join(lines) == ''.join(wrapLinesSimple(lines, 60))
    assert ''.join(lines) == ''.join(wrapLinesSmart(lines, 60))
    print(len(''.join(lines)), end=' ')
    print(len(''.join(wrapLinesSimple(lines))), end=' ')
    print(len(''.join(wrapLinesSmart(lines))), end=' ')
    print(len(''.join(wrapLinesSmart(lines, 60))), end=' ')
    input('Press enter') # пауза, если сценарий был запущен щелчком мыши
```

html2text: извлечение текста из разметки HTML (прототип, предварительное знакомство)

В примере 14.8 приводится реализация простейшего механизма анализа разметки HTML, который используется программой PyMailGUI для извлечения простого текста из почтовых сообщений, основная (или единственная) текстовая часть которых представлена в виде разметки HTML. Извлеченный текст используется для отображения при просмотре и в качестве начального содержимого при создании ответов и пересылке. Оригинальная разметка HTML также отображается во всей своей красе в веб-броузере, как и прежде.

Это только *прототип*. Поскольку в настоящее время программа PyMailGUI главным образом ориентирована на работу с простым текстом, этот механизм разрабатывался в качестве временного решения – до создания виджета просмотра/редактирования HTML. По этой причине он в лучшем случае может считаться лишь прототипом, который не подвергался сколько-нибудь существенной доводке. Реализация надежного механизма синтаксического анализа разметки HTML выходит далеко за рамки этой главы и книги. Если этот механизм окажется не в состоянии отобразить правильно простой текст (а такое обязательно будет случаться!), у пользователя останется возможность просмотреть и вырезать хорошо отформатированный текст из окна веб-броузера.

Этот раздел также является *предварительным знакомством*. Синтаксический анализ HTML не будет рассматриваться до главы 19 этой книги, поэтому до того момента многое в этой реализации вам придется принять на веру. К сожалению, эта функциональная особенность была добавлена в PyMailGUI на последних этапах создания книги, и, чтобы не отказываться от включения этого улучшения в книгу, пришлось пожертвовать правилом не использовать опережающих ссылок. А пока для получения более полной информации о синтаксическом анализе HTML наберитесь терпения (или перелистайте вперед) до главы 19.

Если говорить в двух словах, реализованный здесь класс предоставляет методы-обработчики, которые вызываются стандартной реализацией механизма анализа HTML по мере обнаружения тегов и содержимого – эта модель используется здесь для сохранения интересующего нас текста. Помимо класса мы могли бы также использовать модуль Python `html.entities`, позволяющий отобразить большее количество типов элементов, чем предусмотрено здесь, – еще один инструмент, с которым мы познакомимся в главе 19.

Несмотря на его ограничения, данный пример может служить черновым руководством, помогающим как-то начать работать, и какими бы ни были результаты, которые он предоставляет, они определенно лучше, чем отображение и цитирование исходного кода разметки HTML, как это было в предыдущих изданиях.

Пример 14.8. PP4E\Internet\Email\PyMailGui\html2text.py

```
"""
################################################################
*ОЧЕНЬ* простой механизм преобразования html-в-текст для получения
цитируемого текста в операциях создания ответа и пересылки и отображения
при просмотре основного содержимого письма. Используется, только когда
основная текстовая часть представлена разметкой HTML (то есть отсутствует
альтернативная или другие текстовые части для отображения). Нам также
необходимо знать, является текст разметкой HTML или нет, однако уже
findMainText возвращает тип содержимого для основной текстовой части.

Данная реализация является лишь прототипом, который поможет вам создать
более законченное решение. Она не подвергалась сколько-нибудь существенной
доводке, но любой результат лучше, чем отображение исходного кода разметки
HTML, и гораздо более удачной выглядит идея перехода к использованию виджета
просмотра/редактирования HTML в будущем. Однако на данный момент PyMailGUI
все еще ориентирована на работу с простым текстом.

Если (в действительности - когда) этот механизм не сможет воспроизвести
простой текст, пользователи смогут просмотреть и скопировать текст в окне
веб-броузера, запускаемого для просмотра HTML. Подробнее об анализе HTML
рассказывается в главе 19.
################################################################
"""

from html.parser import HTMLParser    # стандартный парсер
                                      # (SAX-подобная модель)

class Parser(HTMLParser): # наследует станд. парсер, определяет обработчики
    def __init__(self):   # текст - строка str, может быть в любой кодировке
        HTMLParser.__init__(self)
        self.text = '[Extracted HTML text]'
        self.save = 0
        self.last = ''

    def addtext(self, new):
        if self.save > 0:
            self.text += new
            self.last = new

    def addeoln(self, force=False):
        if force or self.last != '\n':
            self.addtext('\n')

    def handle_starttag(self, tag, attrs):                # + другие,
        if tag in ('p', 'div', 'table', 'h1', 'h2', 'li'): # с которых может
            self.save += 1                          # начинаться содержимое?
            self.addeoln()
        elif tag == 'td':
            self.addeoln()
```

```
        elif tag == 'style':                # + другие, которые могут
            self.save -= 1                  #   завершать содержимое?
        elif tag == 'br':
            self.addeoln(True)
        elif tag == 'a':
            alts = [pair for pair in attrs if pair[0] == 'alt']
            if alts:
                name, value = alts[0]
                self.addtext('[' + value.replace('\n', '') + ']')

    def handle_endtag(self, tag):
        if tag in ('p', 'div', 'table', 'h1', 'h2', 'li'):
            self.save -= 1
            self.addeoln()
        elif tag == 'style':
            self.save += 1

    def handle_data(self, data):
        data = data.replace('\n', '')  # а как быть с тегом <PRE>?
        data = data.replace('\t', ' ')
        if data != ' ' * len(data):
            self.addtext(data)

    def handle_entityref(self, name):
        xlate = dict(lt='<', gt='>', amp='&', nbsp='').get(name, '?')
        if xlate:
            self.addtext(xlate) # плюс множество других: показать ? как есть

def html2text(text):
    try:
        hp = Parser()
        hp.feed(text)
        return(hp.text)
    except:
        return text

if __name__ == '__main__':

    # для тестирования: html2text.py media\html2text-test\htmlmail1.html
    # получает имя файла из командной строки, отображает результат в Text
    # файл должен содержать текст в кодировке по умолчанию для данной
    # платформы, но это требование не применяется к аргументу text

    import sys, tkinter
    text = open(sys.argv[1], 'r').read()
    text = html2text(text)
    t = tkinter.Text()
    t.insert('1.0', text)
    t.pack()
    t.mainloop()
```

После того как я написал этот пример и закончил главу, я попробовал поискать в Интернете инструменты преобразования HTML в текст с целью найти более удачное решение и обнаружил намного более полноценное и надежное решение на языке Python, чем этот простой прототип. Кроме того, я обнаружил, что эта система называется так же, как мой сценарий!

Это совершенно случайное совпадение (увы, разработчики склонны думать одинаково). Чтобы поближе познакомится с этой более удачной и проверенной альтернативой, выполните поиск в Интернете по слову *html2text*. Описываемое решение распространяется с открытыми исходными текстами, но на условиях лицензии GPL, и на момент написания этих строк существовало только в версии для Python 2.X (например, в нем используется модуль `sgmllib`, имеющийся в 2.X, который был убран из версии 3.X в пользу нового модуля `html.parser`). Первым минусом можно считать тот факт, что лицензия GPL могла бы стать источником проблем с авторскими правами при распространении этого улучшенного решения с программой PyMailGUI в составе пакета примеров к книге. Но еще хуже то, что существует только версия для Python 2.X, а это означает, что она вообще не может использоваться в примерах из этой книги, выполняющихся под управлением Python 3.X.

В Интернете можно найти и другие инструменты, позволяющие извлекать простой текст, на которые стоит обратить внимание, включая BeautifulSoup и еще один инструмент с именем *html2text.py* (нет, правда!). Они также пока доступны только для версии 2.X, хотя все может измениться к тому времени, когда вы будете читать это примечание. Нет других причин изобретать колесо, кроме той, что имеющиеся колеса не подходят к вашей тележке!

mailconfig: настройки пользователя

В примере 14.9 приводится модуль `mailconfig` с пользовательскими настройками для программы PyMailGUI. Эта программа имеет собственную версию данного модуля, потому что многие из настроек являются уникальными для PyMailGUI. Чтобы использовать программу для чтения собственной электронной почты, необходимо изменить значения переменных так, чтобы они отражали имена ваших POP и SMTP-серверов и параметры учетной записи. Переменные в этом модуле позволяют также настраивать внешний вид и поведение программы без необходимости отыскивать и изменять логику ее работы.

В текущем виде модуль представляет конфигурацию с единственной учетной записью. Можно было бы обобщить реализацию этого модуля для поддержки нескольких учетных записей, запрашивая необходимую информацию у пользователя в консоли при первом импортировании. Однако в одном из следующих разделов будет представлено другое решение, позволяющее расширять этот модуль внешними средствами.

Пример 14.9. PP4E\Internet\Email\PyMailGui\mailconfig.py

```
"""
###############################################################################
Пользовательские настройки для PyMailGUI.

Сценарии для работы с электронной почтой получают имена серверов и другие
параметры из этого модуля: измените модуль так, чтобы он отражал имена ваших
серверов, вашу подпись и предпочтения. Этот модуль также определяет
некоторые параметры внешнего вида виджетов в графическом интерфейсе,
политику выбора кодировок Юникода и многие другие особенности версии 3.0.
Смотрите также: локальный файл textConfig.py, где хранятся настройки
редактора PyEdit, используемого программой PyMailGUI.

Внимание: программа PyMailGUI не работает без большинства переменных в этом
модуле: создайте резервную копию! Предупреждение: с некоторого момента
в этом файле непоследовательно используется смешивание регистров символов
в именах переменных...; Что сделать: некоторые настройки можно было бы
получать из командной строки и неплохо было бы реализовать возможность
настройки параметров в виде диалогов графического интерфейса, но и этот
общий модуль тоже неплохо справляется с задачей.
###############################################################################
"""


#------------------------------------------------------------------------------
# (обязательные параметры для загрузки, удаления) сервер POP3, пользователь;
#------------------------------------------------------------------------------

#popservername = '?Please set your mailconfig.py attributes?'

popservername = 'pop.secureserver.net' # другие примеры в каталоге altconfigs/
popusername   = 'PP4E@learning-python.com'

#------------------------------------------------------------------------------
# (обязательные параметры отправки) имя сервера SMTP;
# смотрите модуль Python smtpd, где имеется класс сервера SMTP,
# выполняющийся локально ('localhost');
# примечание: провайдер Интернета может требовать, чтобы вы напрямую
# подключались к его системе:
# у меня был случай, когда я мог отправлять почту через Earthlink, используя
# коммутируемое подключение, но не мог по кабельному соединению Comcast;
#------------------------------------------------------------------------------

smtpservername = 'smtpout.secureserver.net'

#------------------------------------------------------------------------------
# (необязательные параметры) личная информация, используемая PyMailGUI
# для заполнения полей в формах редактирования;
# если не установлено, не будет вставлять начальные значения;
# mysignature - может быть блоком текста в тройных кавычках,
# пустая строка игнорируется;
# myaddress   - используется как начальное значение поля "From",
```

```
# если не пустое, больше не пытается определить значение From для ответов -
# с переменным успехом;
#-----------------------------------------------------------------------------

myaddress   = 'PP4E@learning-python.com'
mysignature = ('Thanks,\n'
             '--Mark Lutz (http://learning-python.com, http://rmi.net/~lutz)')

#-----------------------------------------------------------------------------
# (может потребоваться при отправке) пользователь/пароль SMTP,
# если необходима аутентификация;
# если аутентификация не требуется, установите переменную smtpuser
# в значение None или '' и присвойте переменной smtppasswdfile имя файла
# с паролем SMTP, или пустую строку, если желательно заставить программу
# запрашивать пароль  (в консоли или в графическом интерфейсе)
#-----------------------------------------------------------------------------

smtpuser = None     # зависит от провайдера
smtppasswdfile = '' # присвойте '', чтобы заставить запрашивать пароль

#smtpuser = popusername

#-----------------------------------------------------------------------------
# (необязательные параметры) PyMailGUI: имя локального однострочного
# текстового файла с паролем к POP-серверу; если пустая строка
# или файл не может быть прочитан, пароль будет запрашиваться
# при первой попытке подключения;
# пароль не шифруется: оставьте эту переменную пустой при работе
# на общем компьютере; PyMailCGI всегда запрашивает пароль
# (выполняется, возможно, на удаленном сервере);
#-----------------------------------------------------------------------------

poppasswdfile = r'c:\temp\pymailgui.txt' # присвойте '', чтобы
                                         # запрашивать пароль

#-----------------------------------------------------------------------------
# (обязательные параметры) локальный файл для сохранения
# отправленных сообщений;этот файл можно открыть и просмотреть,
# щелкнув на кнопке 'Open' в PyMailGUI; не используйте форму '.',
# если программа может запускаться из другого каталога: например, pp4e demos
#-----------------------------------------------------------------------------

#sentmailfile = r'.\sentmail.txt'    # . означает текущий рабочий каталог

#sourcedir    = r'C:\...\PP4E\Internet\Email\PyMailGui\'
#sentmailfile = sourcedir + 'sentmail.txt'

# определить автоматически по одному из файлов с исходными текстами
import wraplines, os
mysourcedir  = os.path.dirname(os.path.abspath(wraplines.__file__))
sentmailfile = os.path.join(mysourcedir, 'sentmail.txt')
```

```
#-------------------------------------------------------------------------
# (более не используется) локальный файл, куда pymail сохраняет принятую
# почту (полный текст); PyMailGUI запрашивает имя файла с помощью диалога
# в графическом интерфейсе; Кроме того, операция Split запрашивает каталог,
# а кнопки быстрого доступа к частям сохраняют их в ./TempParts;
#-------------------------------------------------------------------------

#savemailfile = r'c:\temp\savemail.txt' # не используется в PyMailGUI: диалог

#-------------------------------------------------------------------------
# (необязательные параметры) списки заголовков, отображаемых в окнах
# со списками и в окнах просмотра в PyMailGUI; listheaders замещает список
# по умолчанию, viewheaders расширяет его; оба должны быть кортежами строк
# или None, чтобы использовать зн. по умолчанию;
#-------------------------------------------------------------------------

listheaders = ('Subject', 'From', 'Date', 'To', 'X-Mailer')
viewheaders = ('Bcc',)

#-------------------------------------------------------------------------
# (необязательные параметры) шрифты и цвета для текста в окнах со списками,
# содержимого в окнах просмотра и кнопок быстрого доступа к вложениям;
# шрифты определяются кортежами ('семейство', размер, 'стиль'); цвет (фона
# и переднего плана) определяется строкой 'имя_цвета' или шестнадцатеричным
# значением '#RRGGBB'; None означает значение по умолчанию; шрифты и цвета
# окон просмотра могут также устанавливаться интерактивно, с помощью
# меню Tools текстового редактора; смотрите также пример setcolor.py
# в части книги, посвященной графическим интерфейсам (глава 8);
#-------------------------------------------------------------------------

listbg   = 'indianred'            # None, 'white', '#RRGGBB'
listfg   = 'black'
listfont = ('courier', 9, 'bold') # None, ('courier', 12, 'bold italic')
                                  # для колонок в окнах со списками
                                  # использовать моноширинный шрифт
viewbg     = 'light blue'         # было '#dbbedc'
viewfg     = 'black'
viewfont   = ('courier', 10, 'bold')
viewheight = 18                   # макс. число строк при открытии (20)

partfg = None
partbg = None

# смотрите имена цветов в Tk: aquamarine paleturquoise powderblue
# goldenrod burgundy ....
#listbg = listfg = listfont = None
#viewbg = viewfg = viewfont = viewheight = None # использовать по умолчанию
#partbg = partfg = None

#-------------------------------------------------------------------------
# (необязательные параметры) позиция переноса строк оригинального текста
# письма при просмотре, создании ответа и пересылке; перенос выполняется
```

```
# по первому разделителю левее данной позиции;
# при редактировании не выполняется автоматический перенос строк: перенос
# должен выполняться самим пользователем или инструментами электронной
# почты получателя; чтобы запретить перенос строк, установите в этом
# параметре большое значение (1024?);
#-----------------------------------------------------------------------

wrapsz = 90

#-----------------------------------------------------------------------
# (необязательные параметры) управляют порядком открытия вложений
# в графическом интерфейсе PyMailGUI; используются при выполнении
# операции Split в окнах просмотра и при нажатии кнопок быстрого доступа
# к вложениям; если параметр okayToOpenParts имеет значение False, кнопки
# быстрого доступа не будут отображаться в графическом интерфейсе,
# а операция Split будет сохранять вложения в каталоге, но не будет
# открывать их; параметр verifyPartOpens используется кнопкой Split
# и кнопками быстрого доступа: если этот параметр имеет значение False, все
# вложения с известными типами автоматически будут открываться кнопкой Split;
# параметр verifyHTMLTextOpen определяет, следует ли использовать
# веб-броузер для открытия основной текстовой части в формате HTML:
#-----------------------------------------------------------------------

okayToOpenParts    = True  # открывать ли части/вложения вообще?
verifyPartOpens    = False # спрашивать ли разрешения перед открытием
                           # каждой части?
verifyHTMLTextOpen = False # спрашивать ли разрешение перед открытием основной
                           # текстовой части, если она в формате HTML?

#-----------------------------------------------------------------------
# (необязательные параметры) максимальное число кнопок быстрого доступа
# к частям сообщения, отображаемым в середине окон просмотра;
# если количество частей окажется больше этого числа, после кнопки
# с порядковым номером, соответствующим ему, будет отображаться
# кнопка "...", нажатие на которую будет запускать операцию "Split"
# для извлечения дополнительных частей;
#-----------------------------------------------------------------------

maxPartButtons = 8             # количество кнопок в окнах просмотра

# *** дополнительные параметры для версии 3.0 ***

#-----------------------------------------------------------------------
# (обязательные параметры, для получения сообщений) кодировка Юникода,
# используемая для декодирования полного текста сообщения и для кодирования
# и декодирования текста сообщения при сохранении в текстовых файлах;
# подробности смотрите в главе 13: это ограниченное и временное решение
# пока в пакете email не будет реализован новый механизм анализа сообщений,
# способный работать со строками bytes, способный более точно обрабатывать
# кодировки на уровне отдельных сообщений;
# обратите внимание: для декодирования сообщений в некоторых старых файлах
```

```
# я был вынужден использовать не utf-8, а 'latin1' (8-битовую кодировку,
# которая является надмножеством 7-битовой кодировки ascii);
#-----------------------------------------------------------------------------

fetchEncoding = 'latin-1' # как декодировать и сохранять полный текст
                          # сообщения (ascii?)

#-----------------------------------------------------------------------------
# (необязательные параметры, для отправки) кодировки для текста новых писем
# плюс всех текстовых вложений; установите эти параметры равными None, чтобы
# программа запрашивала кодировки при отправке письма, в противном случае
# эти значения будут действовать на протяжении всего сеанса;
# по умолчанию = 'latin-1', если в диалоге запроса кодировки щелкнуть
# на кнопке Cancel; в любом случае, если кодировки, указанные здесь
# или введенные в диалоге, окажутся неприменимыми, будет использоваться
# кодировка UTF-8 (например, при выборе ascii для отправки письма с текстом
# или вложением, содержащим символы вне диапазона ascii; пакет email более
# привередлив к именам кодировок, чем сам язык Python: ему известна
# кодировка latin-1 (будет использовать формат MIME quoted-printable), но
# кодировка latin1 ему не известна (будет использовать формат MIME base64);
# чтобы задействовать кодировку по умолчанию для текущей платформы,
# присвойте этим параметрам значение, возвращаемое функцией
# sys.getdefaultencoding(); кодировки текстовых частей полученных
# сообщений определяются автоматически из заголовков;
#-----------------------------------------------------------------------------

mainTextEncoding       = 'ascii' # основной текст сообщения (None=спрашивать)
attachmentTextEncoding = 'ascii' # все текстовые вложения (utf-8, latin-1)

#-----------------------------------------------------------------------------
# (необязательные параметры, для отправки) определите в этом параметре имя
# кодировки, которая будет применяться к заголовкам и к именам в адресах,
# содержащим символы вне диапазона ASCII, при отправке новых писем;
# None означает использование кодировки UTF-8 по умолчанию, которая подходит
# в большинстве случаев; имена в адресах электронной почты, которые
# не удастся декодировать, будут отбрасываться (будут использоваться только
# сами адреса);обратите внимание, что декодирование заголовков выполняется
# автоматически перед отображением note, в соответствии с их содержимым,
# без учета пользовательских настроек;
#-----------------------------------------------------------------------------

headersEncodeTo = None # как кодировать не-ASCII заголовки при передаче
                       # (None=UTF-8)

#-----------------------------------------------------------------------------
# (необязательные параметры) выбирают текстовую, HTML или обе версии справки;
# всегда отображается та или иная версия: если оба параметра отключены,
# отображается версия HTML
#-----------------------------------------------------------------------------

showHelpAsText = True # прокручиваемый текст, с кнопками открытия файлов
                      # с исходным программным кодом
```

```
showHelpAsHTML = True # HTML в веб-броузере, без ссылок на файлы с исходным
                      # программным кодом

#-----------------------------------------------------------------------
# (необязательный параметр) если имеет значение True, выбранная текстовая
# часть сообщения в формате HTML не будет отображаться в окне PyEdit,
# если она отображается в веб-броузере; если имеет значение False, будут
# отображаться обе версии и будет выполнено определение кодировки
# для передачи текстовому виджету (которая может быть неизвестна броузеру);
#-----------------------------------------------------------------------

skipTextOnHtmlPart = False # не показывать текст из разметки html в PyEdit

#-----------------------------------------------------------------------
# (необязательный параметр) максимальное количество заголовков или сообщений,
# которые будут загружаться по каждому запросу; если установить в этом
# параметре значение N, PyMailGUI будет загружать не более N самых свежих
# почтовых сообщений; более старые сообщения, не вошедшие в это число,
# не будут загружаться с сервера, но будут отображаться как пустые/фиктивные
# сообщения; если этот параметр будет иметь значение None (или 0),
# загрузка никак не будет ограничиваться; используйте этот параметр,
# если у вас слишком много сообщений в почтовом ящике и скорость подключения
# к Интернету или почтового сервера делает процедуру загрузки всех писем
# слишком длительной; кроме того, PyMailGUI загружает заголовки только самых
# новых сообщений, но данный параметр никак не зависит от этой особенности;
#-----------------------------------------------------------------------

fetchlimit = 50 # максимальное число загружаемых заголовков/сообщений

#-----------------------------------------------------------------------
# (необязательные параметры) начальные ширина и высота списка с оглавлением
# (символы x строки); исключительно для удобства, поскольку окно легко может
# растягиваться после открытия;
#-----------------------------------------------------------------------

listWidth  = None      # None = используется значение 74 по умолчанию
listHeight = None      # None = используется значение 15 по умолчанию

#-----------------------------------------------------------------------
# (необязательный параметр, для ответов) если имеет значение True,
# операция Reply будет заполнять заголовок Cc ответа адресами всех
# получателей оригинального сообщения, после удаления повторяющихся адресов,
# и адресом нового отправителя; если имеет значение False, заголовок CC
# заполняться не будет и ответ будет настроен на отправку только отправителю
# оригинального сообщения; в любом случае, заголовок Cc всегда можно
# отредактировать позднее.
#-----------------------------------------------------------------------

repliesCopyToAll = True # True=ответить отправителю + всем получателям,
                        # иначе - только отправителю

#end
```

textConfig: настройка окон редактора PyEdit

Модуль mailconfig, представленный в предыдущем разделе, содержит некоторые параметры настройки компонента PyEdit, используемого для просмотра и редактирования основного текста сообщения, но PyMailGUI также использует PyEdit для отображения других разновидностей текста в отдельных окнах, включая необработанный текст письма, некоторые текстовые вложения и исходный программный код в окне справки. Для настройки параметров отображения для этих окон PyMailGUI полагается на собственные возможности редактора PyEdit, который пытается загрузить модуль, такой как в примере 14.10, из каталога приложения. Настройки кодировок для редактора PyEdit загружаются из единого модуля textConfig, находящегося в каталоге пакета, поскольку, как предполагается, они не должны изменяться на одной и той же платформе (подробности смотрите в главе 11).

Пример 14.10. PP4E\Internet\Email\PyMailGui\textConfig.py

```
"""
настройки для окон редактора PyEdit, отличных от компонента отображения
основного текста сообщения; этот модуль (не его пакет), как предполагается,
должен находиться в пути поиска; настройки кодировок Юникода для PyEdit
берутся не из этого файла, а из файла из textConfig.py, находящегося
в пакете;
"""


bg = 'beige'     # отсутствует=white; имя или шестнадцатеричное значение RGB
fg = 'black'     # отсутствует=black; например, 'beige', '#690f96'

# другие настройки - смотрите PP4E\Gui\TextEditor\textConfig.py
# font = ('courier', 9, 'normal')
# height = 20    # Tk по умолчанию: 24 строки
# width = 80     # Tk по умолчанию: 80 символов
```

PyMailGUIHelp: текст справки и ее отображение

Наконец, в примере 14.11 приводится модуль, где в виде строки в тройных кавычках определяется текст, отображаемый в окне справки PyMailGUI, а также функция, отображающая текст справки в формате HTML. HTML-версия справки находится в отдельном файле, который не приводится здесь, но включен в пакет примеров для этой книги.

Кроме того, я также опустил большую часть справочного текста ради экономии места в книге (в предыдущем издании текстом справки было занято 11 страниц и еще больше потребовалось бы в этом издании!). Полный текст вы найдете в этом модуле, находящемся в пакете примеров, или вы можете запустить PyMailGUI и щелкнуть на полосе вызова справки в главном окне с оглавлением почтового ящика на сервере, чтобы узнать больше о том, как действует интерфейс PyMailGUI. В справке вы найдете описание некоторых особенностей PyMailGUI, которые не

были представлены в демонстрационном разделе, выше в этой главе, и другие сведения.

Справка в формате HTML содержит ссылки для перехода между разделами и отображается в веб-броузере. Поскольку при отображении текстовой версии справки предоставляется возможность открыть файлы с исходными текстами и минимизируются внешние зависимости (попытка просмотреть HTML-версию будет терпеть неудачу, если программе не удастся отыскать веб-броузер), с программой поставляются обе версии, текстовая и в формате HTML, и пользователю предоставляется возможность выбора версии для просмотра в модуле `mailconfig`. Возможны и другие схемы (например, преобразование HTML в текст, в случае невозможности отобразить HTML-версию), но мы оставим их для самостоятельной реализации.

Пример 14.11. PP4E\Internet\PyMailGui\PyMailGuiHelp.py (частично)

```
"""
##########################################################################
строка с текстом справки для PyMailGUI и функция отображения справки
в формате HTML;

Предыстория: первоначально справка отображалась в окне информационного
диалога, который был узким местом для Linux; позднее стал использоваться
текстовый компонент с прокруткой и кнопки; теперь реализована возможность
отображения справки в формате HTML в веб-броузере;

2.1/3E: строка с текстом справки помещена в отдельный модуль, чтобы
не отвлекать внимание от выполняемого программного кода. В данном случае
этот текст помещается в простое окно с прокруткой; в будущем можно было бы
использовать файл HTML и открывать его в веб-броузере (с помощью модуля
webbrowser или выполняя команду "browser help.html" или "start help.html"
с применением функции os.system);

3.0/4E: теперь также имеется возможность выводить справочный текст
в веб-броузере, в формате HTML, с оформлением списков, ссылок на разделы
и с разделителями; представление справочного текста в формате HTML,
который приводится в виде простой строки ниже и отображается в веб-броузере,
вы найдете в файле PyMailGuiHelp.html, входящем в состав пакета с примерами;
в настоящее время поддерживаются обе версии, текстовая и HTML: измените
параметры в mailconfig.py, чтобы выбрать наиболее предпочтительную для вас;
##########################################################################
"""


# новая справка в формате HTML для 3.0/4E
helpfile = 'PyMailGuiHelp.html' # смотрите пакет с примерами к книге
def showHtmlHelp(helpfile=helpfile):
    """
    3.0: открывает HTML-версию справки в локальном веб-броузере с помощью
    модуля webbrowser; этот модуль доступен для импорта, но html-файл может
    оказаться за пределами текущего рабочего каталога
    """
```

```
    import os, webbrowser
    mydir = os.path.dirname(__file__) # каталог из имени файла этого модуля
    mydir = os.path.abspath(mydir)   # получ. абс. путь: мог быть .., или иным
    webbrowser.open_new('file://' + os.path.join(mydir, helpfile))

##########################################################################
# строка с текстовой версией справки: за создание графического интерфейса
# отвечает клиент
##########################################################################

helptext = """PyMailGUI, version 3.0
May, 2010 (2.1 January, 2006)
Programming Python, 4th Edition
Mark Lutz, for O'Reilly Media, Inc.

PyMailGUI - это многооконный интерфейс для работы с электронной почтой
в автономном режиме и с подключением к Интернету. Основной интерфейс
этой программы состоит из одного окна со списком писем на почтовом
сервере и нуля или более окон со списками содержимого файлов, куда ранее
сохранялась почта, а также множества окон составления новых сообщений
или просмотра содержимого писем, выбранных в окне со списком.
При запуске первым открывается окно с главным (на сервере) списком
сообщений, но соединение с сервером не устанавливается, пока явно
не будет запущена операция загрузки или отправки сообщения.
Все окна PyMailGUI могут изменяться в размерах, что особенно удобно
для окон со списками, так как это дает возможность отобразить
дополнительные колонки.

Примечание: Чтобы использовать PyMailGUI для чтения и отправки вашей
собственной почты, необходимо изменить имена POP и SMTP серверов и параметры
учетной записи в файле mailconfig.py, находящемся в каталоге с исходными
текстами программы PyMailGUI. Подробности смотрите в разделе 11.

Содержание:
0)  РАСШИРЕНИЯ В ЭТОЙ ВЕРСИИ
1)  ОПЕРАЦИИ, ДОСТУПНЫЕ В ОКНЕ СО СПИСКОМ
2)  ОПЕРАЦИИ, ДОСТУПНЫЕ В ОКНЕ ПРОСМОТРА
3)  РАБОТА В АВТОНОМНОМ РЕЖИМЕ
4)  ПРОСМОТР ТЕКСТА И ВЛОЖЕНИЙ
5)  ОТПРАВКА ТЕКСТА И ВЛОЖЕНИЙ
6)  ОДНОВРЕМЕННОЕ ВЫПОЛНЕНИЕ НЕСКОЛЬКИХ ОПЕРАЦИЙ С СЕРВЕРОМ
7)  УДАЛЕНИЕ ПОЧТЫ
8)  СИНХРОНИЗАЦИЯ НОМЕРОВ ВХОДЯЩИХ СООБЩЕНИЙ
9)  ЗАГРУЗКА ПОЧТЫ
10) ПОДДЕРЖКА ЮНИКОДА И ИНТЕРНАЦИОНАЛИЗАЦИИ
11) МОДУЛЬ mailconfig С НАСТРОЙКАМИ
12) ЗАВИСИМОСТИ
13) ПРОЧИЕ ПОДСКАЗКИ ("Шпаргалка")
```

...остальная часть файла...

```
13) ПРОЧИЕ ПОДСКАЗКИ ("Шпаргалка")
- Используйте ',' для разделения адресов в заголовках To, Cc и Bcc.
- Адреса могут указываться в полной форме '"имя" <адрес>'.
- Содержимое и заголовки декодируются при получении и кодируются
  при отправке.
- Сообщения в формате HTML отображаются как простой текст, плюс как HTML
  в веб-броузере.
- Заголовки To, Cc и Bcc получают значения, указанные при составлении
  письма, но заголовок Bcc не отправляется.
- Если разрешено в mailconfig, заголовок Bcc заполняется адресом отправителя.

- При составлении ответа и при пересылке в письмо автоматически вставляется
  текст оригинального сообщения.
- Если разрешено, при составлении ответа заголовок Cc заполняется адресами
  получателей оригинального письма.
- При отправке письма в него можно добавлять вложения, которые будут
  кодироваться по мере необходимости.
- Вложения могут открываться в окне просмотра, щелчком на кнопке Split
  или на кнопках быстрого доступа.
- Двойной щелчок на сообщении в окне со списком открывает окно с исходным
  текстом сообщения.
- Можно выбрать сразу несколько сообщений в списке, чтобы обработать их
  как единое множество: Ctrl|Shift + щелчок, или флажок All.

- Отправляемая почта сохраняется в файле, имя которого указано в mailconfig:
  открыть его для просмотра можно с помощью кнопки Open.
- Операция сохранения открывает диалог выбора файла, куда будет
  сохранено сообщение.
- Операция сохранения никогда не удаляет выбранный файл, а добавляет
  сообщение в конец.
- Операция извлечения вложений (Split) запрашивает каталог сохранения;
  кнопки быстрого доступа сохраняют вложения в каталоге ./TempParts.
- Диалоги открытия и сохранения файла всегда запоминают последний каталог.
- Для сохранения черновика письма во время его составления используйте
  операцию сохранения редактора.

- Пароли запрашиваются программой PyMailGUI по мере необходимости
  и не сохраняются.
- Вы можете сохранить свои пароли в файл и указать его имя в mailconfig.py.
- Чтобы напечатать текст сообщения, сохраните его в текстовый файл
  и напечатайте с помощью другого инструмента.
- В каталоге altconfigs вы увидите примеры использования нескольких
  учетных записей.

- Письма никогда не удаляются с почтового сервера автоматически.
- При успешном завершении операция удаления не вызывает перезагрузку
  заголовков.
- Операция удаления предварительно проверяет почтовый ящик,
  чтобы гарантировать удаление выбранной почты.
- Операция получения письма определяет изменения в ящике входящей почты
  и может автоматически перезагрузить оглавление.
```

```
- Одновременно может выполняться любое количество операций отправки
  и получения непересекающихся множеств писем.

- Щелкните на кнопке Source в этом окне, чтобы увидеть
  исходные тексты PyMailGUI.
- Смотрите обновления и изменения на сайте http://www.rmi.net/~lutz
- Эта система распространяется с открытыми исходными текстами: изменяйте
  ее программный код по своему усмотрению.
"""


if __name__ == '__main__':
    print(helptext)          # вывести в stdout, если запущен как сценарий
    input('Press Enter key') # пауза, если был запущен щелчком мыши
```

Полный текст файла справки в формате HTML, первые строки которого показаны в примере 14.12, смотрите в пакете примеров. Он представляет собой простой перевод строки с текстом справки (добавление оформления этой страницы мы оставим в качестве самостоятельного упражнения).

Пример 14.12. PP4E\Internet\PyMailGui\PyMailGuiHelp.html (частично)

```
<HTML>
<TITLE>PyMailGUI 3.0 Help</TITLE>
<!-- ЧТО СДЕЛАТЬ: добавить рисунки, снимки экрана и прочее --!>
<BODY>

<H1 align=center>PyMailGUI, Version 3.0</H1>
<P align=center>
<B><I>May, 2010 (2.1 January, 2006)</I></B><BR>
<B><I>Programming Python, 4th Edition</I></B><BR>
<B><I>Mark Lutz, for O'Reilly Media, Inc.</I></B>
<P>
<I>PyMailGUI</I> это многооконный интерфейс для работы с электронной почтой
в автономном режиме и с подключением к Интернету.
...остальная часть файла опущена...
```

altconfigs: настройка нескольких учетных записей

Имеется еще несколько коротких файлов, которые не являются «официальной» частью системы, но я пользуюсь ими для запуска и тестирования системы. Для тех, у кого имеется несколько учетных записей электронной почты, может быть неудобным изменять файл с настройками каждый раз, когда требуется использовать какую-то конкретную учетную запись. Кроме того, когда одновременно открывается несколько сеансов PyMailGUI для разных учетных записей, желательно, чтобы графический интерфейс имел различный внешний вид, чтобы их можно было отличать.

Для решения этой проблемы в пакете примеров был создан каталог `altconfigs`, предоставляющий простой способ выбора учетной записи

и настроек для нее на этапе запуска. В нем находится новый сценарий верхнего уровня, настраивающий путь поиска модулей, а также модуль mailconfig, запрашивающий и загружающий модуль с параметрами настройки электронной почты, суффикс в имени которого запрашивается в консоли. Также предоставляется сценарий запуска, не выполняющий настройку пути поиска, – для запуска из PyGadgets или для создания ярлыка на рабочем столе, например, не требующий, чтобы переменная окружения PYTHONPATH содержала путь к корневому каталогу PP4E. Все эти файлы приводятся в примерах с 14.13 по 14.17.

Пример 14.13. PP4E\Internet\PyMailGui\altconfigs\PyMailGui.py

```
import sys               # ..\PyMailGui.py или 'book' для уч. записи книги
sys.path.insert(1, '..')             # добавить действительный каталог
exec(open('../PyMailGui.py').read()) # запустить этот сценарий,
                                     # но mailconfig взять здесь
```

Пример 14.14. PP4E\Internet\PyMailGui\altconfigs\mailconfig.py

```
above = open('../mailconfig.py').read() # скопировать версию из каталога
                                        # выше (банально?)
open('mailconfig_book.py', 'w').write(above) # используется для учетной
                                     # записи 'book' и как основа для других
acct = input('Account name?')        # book, rmi, train
exec('from mailconfig_%s import *' % acct)   # . первый элемент в sys.path
```

Пример 14.15. PP4E\Internet\PyMailGui\altconfigs\mailconfig_rmi.py

```
from mailconfig_book import * # базовые настройки в . (копируются из ..)
popservername = 'pop.rmi.net' # это огромный почтовый ящик: 4800 сообщений!
popusername   = 'lutz'
myaddress     = 'lutz@rmi.net'
listbg = 'navy'
listfg = 'white'
listHeight = 20        # выше изначально
viewbg = '#dbbedc'
viewfg = 'black'
wrapsz = 80            # переносить по позиции 80 символов в строке
fetchlimit = 300       # загружать больше заголовков
```

Пример 14.16. PP4E\Internet\PyMailGui\altconfigs\mailconfig_train.py

```
from mailconfig_book import *   # базовые настройки в . (копируются из ..)
popusername = 'lutz@learning-python.com'
myaddress   = 'lutz@learning-python.com'
listbg = 'wheat'       # красновато-золотистый, темно-зеленый, бежевый
listfg = 'navy'        # шоколадный, коричневый,...
viewbg = 'aquamarine'
viewfg = 'black'
wrapsz = 80
viewheaders = None     # без Bcc
fetchlimit = 100       # загружать больше заголовков
```

Пример 14.17. PP4E\Internet\PyMailGui\altconfigs\launch_PyMailGui.py

```
# для запуска без настройки PYTHONPATH (например, с помощью ярлыка)
import os                          # Launcher.py слишком мощный инструмент
os.environ['PYTHONPATH'] = r'..\..\..\..\..' # хм-м; нужно обобщить
os.system('PyMailGui.py')          # получает переменную с настройками путей
```

Файлы с настройками учетных записей, подобные тем, что представлены в примерах 14.15 и 14.16, могут импортировать базовый модуль с настройками учетной записи «book» (чтобы затем расширить его), а могут не импортировать (целиком заменить его). Чтобы использовать эти альтернативные настройки, выполните следующую команду или запустите из любого каталога сценарий самонастройки, представленный в примере 14.17. В любом случае вы сможете открыть окна для этих учетных записей и просмотреть включенные в примеры сохраненные сообщения, но если вы собираетесь использовать эти клиенты для получения и отправки своей почты, не забудьте указать параметры своих учетных записей и свои настройки:

```
C:\...\PP4E\Internet\Email\PyMailGui\altconfigs> PyMailGui.py
Account name?rmi
```

Добавьте «start» в начало этой команды, чтобы окно консоли оставалось доступным для ввода и открытия других учетных записей в Windows (в Unix попробуйте добавить «&» в конец команды). На рис. 14.45 выше показана ситуация, когда были открыты все три мои учетные записи. Я держу их постоянно открытыми на своем настольном компьютере, поскольку операция загрузки извлекает только вновь поступившую почту независимо от того, как долго графический интерфейс находился в бездействии, а операция отправки вообще не требует, чтобы что-то было загружено. Различные цветовые схемы позволяют отличать открытые учетные записи. Ярлык сценария запуска на рабочем столе делает открытие моих учетных записей еще проще.

В данной реализации имена учетных записей запрашиваются, только когда запускается этот специальный файл *PyMailGui.py*, и не запрашиваются, когда запускается оригинальный сценарий непосредственно или программами запуска примеров (в этом случае поток ввода `stdin` может оказаться недоступным для чтения). Расширение модуля, такого как `mailconfig`, который может импортироваться из разных мест, представляет собой весьма интересную задачу (в значительно мере именно поэтому я не рассматриваю это быстрое решение как официальную особенность для конечного пользователя). Например, существуют другие способы обеспечить поддержку нескольких учетных записей:

- Изменять единственный модуль `mailconfig` на месте.
- Импортировать альтернативные модули и сохранять их как ключ «mailconfig» в `sys.modules`.
- Копировать переменные из альтернативных модулей в атрибуты модуля `mailconfig` с использованием `__dict__` и `setattr`.

- Использовать для хранения настроек класс, что позволит обеспечить поддержку изменение настроек с помощью подклассов.
- Открывать диалог в графическом интерфейсе, запрашивающий имя учетной записи до или после появления главного окна.

И так далее. Схема с отдельным подкаталогом, использованная здесь, была выбрана, потому что она оказывает минимальное влияние на существующий программный код. В частности, она позволяет избежать необходимости вносить изменения в существующий модуль `mailconfig` (который прекрасно подходит для случая использования единственной учетной записи); избежать необходимости заставлять пользователя вводить дополнительные сведения в случае использования единственной учетной записи; и учитывает тот факт, что инструкция «import module1 as module2» не лишает возможности импортировать модуль «module1» напрямую позднее. Этот последний пункт чреват более неприятными последствиями, чем можно было бы предположить; импортирование специализированной версии модуля – это не просто вопрос использования расширения «as» для переименования:

```
import m1 as m2 # импорт специализированной версии: загрузит m1 как m2
print(m2.attr)  # выведет attr из m1.py

import m2       # позднее: загрузит модуль m2.py!
print(m2.attr)  # выведет attr из m2.py
```

Иными словами, это решение на скорую руку, которое первоначально предназначалось для тестирования, похоже, является первым кандидатом на усовершенствование – наряду с другими идеями, изложенными в следующем разделе, которым завершается эта глава.

Идеи по усовершенствованию

Я постоянно пользуюсь версией 3.0 программы PyMailGUI как для личного, так и для делового общения, тем не менее, в любом программном обеспечении всегда найдется, что усовершенствовать, и эта система не исключение. Если у вас появится желание поэкспериментировать с программным кодом, ниже перечислены некоторые возможные направления улучшений, которыми я завершаю эту главу:

Компоновка списка и сортировка по столбцам

Было бы неплохо реализовать возможность сортировки по столбцам в окнах со списками. Это может потребовать усложнить структуру окна со списком, обратив внимание на столбцы. В настоящее время окно со списком почты бесспорно выглядит первым кандидатом на внесение косметических улучшений, и любое решение по реализации сортировки по столбцам наверняка будет способствовать этому. Определенные надежды дают расширения для библиотеки tkinter, такие как виджет Tix `HList`, кроме того, имеется сторонний виджет `TkinterTreectrl`, поддерживающий многоколоночные списки с сорти-

ровкой, но в настоящее время он доступен только в версии для Python 2.X – ищите дополнительную информацию по этой теме в Интернете и в других ресурсах.

Размер файла с сохраненной почтой (принятой и отправленной)

Реализация *сохранения почты в файлы* ограничивает размеры файлов объемом доступной памяти, куда они загружаются полностью. Избавиться от этого ограничения помогла бы реализация сохранения почты в файлы DBM с доступом по ключу. Смотрите дополнительные примечания в комментариях внутри модуля `windows`. То же относится и к файлам, куда сохраняется *отправленная почта*, хотя пользователь может сам регулировать их размеры, периодически удаляя ненужные письма. Кроме того, было бы полезно реализовать вывод диалога, предлагающего удалить почту по достижении файлом значительного размера.

Встроенные ссылки

Адреса URL внутри сообщений можно было бы выделять визуально и при щелчке на них автоматически открывать в веб-броузере с помощью инструментов запуска, с которыми мы встречались в частях книги, посвященных графическим интерфейсам и системным инструментам (текстовый виджет в библиотеке tkinter имеет непосредственную поддержку гиперссылок).

Устранение избыточности текста справки

В этой версии объем текста справки вырос настолько, что был также реализован в виде страницы HTML для отображения в веб-броузере при помощи модуля `webbrowser` (вместо или в дополнение к отображению справки в текстовом виде, в соответствии с настройками в `mailconfig`). Это означает, что в настоящее время имеется две копии текста справки: в виде простого текста и в виде HTML. Это не лучший вариант с точки зрения сопровождения.

Можно было бы вообще исключить отображение справки в текстовом виде либо реализовать извлечение простого текста из HTML с помощью модуля Python `html.parser` и тем самым избежать избыточности. О синтаксическом анализе разметки HTML в целом подробнее рассказывается в главе 19. Обратите также внимание на новый модуль `html2text` в PyMailGUI, реализующий прототип инструмента извлечения простого текста из HTML. Кроме того, версия справки в формате HTML не содержит ссылок на файлы с исходными текстами; их можно было бы добавлять в HTML автоматически с помощью строковых методов форматирования, хотя не совсем ясно, что будут делать броузеры с исходным программным кодом на языке Python (некоторые могут попытаться запустить его).

Расширение области применения многопоточной модели выполнения

Выполнение операций сохранения сообщений и извлечения вложений также можно было бы организовать в параллельных потоках,

чтобы учесть наиболее тяжелые случаи. Дополнительно о параллельном выполнении операций сохранения говорится в комментариях в файле *ListWindows.py* – здесь могут возникать некоторые тонкие проблемы, требующие использования блокировок в потоках и системных блокировок файлов из-за возможных параллельных попыток обновления файлов. Заполнение списка с оглавлением также можно было бы организовать в потоках, что особенно полезно при работе с почтовыми ящиками, содержащими огромное количество писем, и на медлительных компьютерах (здесь также может помочь оптимизация, позволяющая избегать повторного анализа заголовков).

Удаление вложений из списка

В настоящее время отсутствует возможность исключить вложение после его добавления в окне составления сообщения. Эту поддержку можно было бы реализовать, добавив кнопки быстрого доступа в окно составления письма, по щелчку на которых можно было бы запрашивать подтверждение и удалять соответствующие им вложения.

Фильтрация спама

В дополнение к спам-фильтрам, действующим на почтовых серверах или предоставляемых интернет-провайдерами, мы могли бы добавить свой автоматический спам-фильтр, обрабатывающий получаемую почту. В этом может помочь пакет *SpamBayes* на языке Python. Эту функцию лучше реализовывать на серверах, а не на клиентах, но не все провайдеры фильтруют спам.

Усовершенствование поддержки нескольких учетных записей

Как описывалось в предыдущем разделе, в настоящее время система выбирает одну учетную запись электронной почты и использует соответствующий модуль с настройками, выполняя специальный программный код в подкаталоге `altconfigs`. Это решение вполне пригодно для книжного примера, но просто само напрашивается на усовершенствование для широкого круга пользователей.

Упрощение доступа к файлу с отправленной почтой

Было бы неплохо добавить кнопку, открывающую файл с отправленной почтой. Программа PyMailGUI автоматически сохраняет отправленную почту в отдельный файл, который в настоящее время можно открыть щелчком на кнопке Open (Открыть) в окне со списком. Однако, откровенно говоря, об этой возможности очень легко забыть – я сам забыл о ней, когда готовил обновленную версию программы для этого издания! Возможно, также было бы полезно предусмотреть параметр в `mailconfig`, запрещающий сохранять почту, особенно для тех, кто не собирается удалять почту из этого файла (он может расти очень быстро; смотрите также предложение, касающееся удаления сохраненных сообщений, выше).

Настройка скорости обработки очереди с функциями обратного вызова в потоках

Как уже упоминалось в описании изменений в версии 3.0, в ней в 10 раз была увеличена скорость обработки очереди с функциями обратного вызова в потоках с целью ускорить начальную загрузку заголовков. Это было достигнуто за счет увеличения количества обработчиков, вызываемых по каждому событию от таймера, и двукратного увеличения частоты срабатывания таймера. Однако слишком частая проверка очереди может повысить потребление процессора до неприемлемого уровня на некоторых компьютерах. На моем ноутбуке с Windows нагрузка на процессор практически незаметна (использование процессора программой при работе в холостом режиме составляет 0%), но у вас может возникнуть потребность настроить эти параметры на своей платформе.

Смотрите параметры настройки скорости в реализации окон со списками и программный код в файле *threadtools.py*, представленном в главе 10. В целом, увеличение количества обработчиков, вызываемых при каждом событии от таймера, и уменьшение частоты срабатывания таймера должно уменьшить нагрузку на процессор без потери отзывчивости интерфейса. (Эх, если бы всякий раз, когда повторяю эти слова, я получал бы по пять центов...)

Списки рассылки

Мы могли бы добавить поддержку списков рассылки, позволив пользователям ассоциировать несколько адресов с сохраняемым названием списка рассылки. При отправке по названию списка рассылки сообщение можно было бы отправлять по всем адресам в списке (адреса «To», передаваемые модулю `smtplib`), а название списка указывать в заголовке «To». Примеры, связанные со списками рассылки, вы найдете в описании инструментов для работы с протоколом SMTP в главе 13.

Просмотр и редактирование основного текста сообщения в формате HTML

Программа PyMailGUI по-прежнему ориентирована на поддержку простого текста в основной текстовой части сообщения, несмотря на то, что многие почтовые клиенты в настоящее время в большей степени ориентированы на поддержку HTML. Отчасти это объясняется тем, что в качестве инструмента редактирования основного текста в PyMailGUI используется простой виджет `Text`. PyMailGUI способна отображать сообщения в формате HTML, открывая их в веб-броузере, и пытается извлекать текст из разметки HTML для отображения, как отмечается в следующем предложении по усовершенствованию, но в ней отсутствует собственный редактор HTML. Обеспечение полной поддержки HTML в качестве основного текста сообщения, вероятно, потребует расширения библиотеки tkinter (или, как это ни прискорбно, перехода на использование другого набора инструментов

создания графических интерфейсов с имеющейся поддержкой этой особенности).

Улучшение механизма синтаксического анализа разметки HTML

Как отмечалось выше, эта версия включает простейший механизм синтаксического анализа разметки HTML, применяемый с целью извлечения текста из основной (или единственной) текстовой части в формате HTML для отображения или цитирования в ответах и в пересылаемых письмах. Также выше отмечалось, что этот механизм нельзя считать законченным или достаточно надежным – чтобы довести его до уровня, пригодного для нормальной эксплуатации, этот механизм необходимо усовершенствовать и протестировать на большом количестве электронных писем в формате HTML. Возможно, было бы лучше поискать более полные и надежные альтернативы для Python 3.X с открытыми исходными текстами, подобные сторонней утилите с тем же названием *html2text.py*, которая была описана в примечании выше. Еще один гибкий механизм анализа разметки HTML предоставляет система *BeautifulSoup* с открытыми исходными текстами, но она опирается на инструменты из модуля `SGMLParser`, доступного только в Python 2.X (исключен из Python 3.X).

Поддержка альтернативных частей с содержимым в виде простого текста и HTML

Кроме того, что касается разметки HTML, в настоящее время отсутствует поддержка возможности отправлять текстовую и HTML версии сообщения с применением популярной схемы MIME multipart/alternative, обеспечивающей поддержку простых текстовых клиентов и клиентов с поддержкой HTML и позволяющей пользователям выбирать, какую из частей использовать для отображения. Такие сообщения могут просматриваться (в графическом интерфейсе предлагаются обе части), но не могут составляться. Но, опять же, поскольку в программе отсутствует возможность редактирования текста в формате HTML, это весьма спорный момент. Если такой редактор когда-нибудь будет добавлен, нам необходимо будет обеспечить поддержку писем подобного рода в программном коде, создающем объект сообщения в пакете `mailtools`, и переписать текущую логику отправки письма так, чтобы совместно использовать эту поддержку.

Интернационализированные заголовки нарушают стройность колонок в списках

Как часто бывает в мире программного обеспечения, добавление новой особенности в этой версии нарушает работу другой, существовавшей прежде: шрифты, используемые для отображения некоторых символов Юникода в заголовках, занимают слишком много места, из-за чего нарушается стройность колонок фиксированной ширины в окнах со списками. Реализация отображения списков опирается на предположение, что строка из N символов на экране всегда имеет

одну и ту же ширину для всех писем, однако это оказывается не так для некоторых китайских и других наборов символов.

Это обстоятельство не препятствует работе программы – оно возникает только при отображении некоторых интернационализированных заголовков и означает, что разделитель колонок «|» будет смещаться для некоторых писем, и, тем не менее, эту проблему желательно решить. Одно из решений заключается в том, чтобы перейти к более сложной реализации отображения списка, побочным эффектом которой может стать решение проблемы сортировки по столбцам, описанной выше.

Адресные книги

В программе PyMailGUI отсутствует механизм автоматического заполнения полей с адресами электронной почты из адресной книги, имеющийся во многих современных клиентах. Добавление этой возможности могло бы стать интересным расширением – с помощью низкоуровневого обработчика событий от клавиатуры можно было бы организовать поиск адреса по мере ввода, а для реализации хранения содержимого адресной книги можно было бы использовать модули `pickle` и `shelve`, описываемые в главах 1 и 17.

Проверка орфографии

В настоящее время в PyMailGUI отсутствует проверка орфографии, имеющаяся в большинстве современных программ для работы с электронной почтой. Ее можно было бы добавить в PyMailGUI, но, пожалуй, еще более уместным было бы добавить ее в текстовый редактор PyEdit, используемый в PyMailGUI, чтобы эта проверка могла быть унаследована всеми клиентами PyEdit. Запуск поиска в Интернете дает множество вариантов, включая интересный сторонний пакет `PyEnchant`, для исследования которого у нас просто нет места.

Поиск в электронных письмах

Аналогично в PyMailGUI отсутствует поддержка поиска в электронных письмах (в заголовках и содержимом) указанной строки. Не совсем ясно, как этот поиск должен осуществляться, учитывая, что система извлекает и кэширует только заголовки сообщений, а собственно содержимое загружается только по требованию пользователя, но, тем не менее, возможность выполнять поиск по входящей почте могла бы оказаться совсем нелишней. В настоящее время эту операцию можно выполнить вручную, сохранив полученную почту в текстовый файл и выполнив поиск в этом файле с помощью внешних инструментов.

Создание двоичного дистрибутива

В качестве настольного приложения программа PyMailGUI представляется идеальным кандидатом на создание самостоятельного двоичного выполняемого файла с помощью таких инструментов, как PyInstaller, Py2Exe и других. При распространении программы в та-

ком виде пользователям не требуется устанавливать Python, потому что интерпретатор Python и библиотеки времени выполнения включаются в выполняемый файл.

Выбор операции «Ответить» вместо «Ответить всем» в графическом интерфейсе

Как описывалось выше, в обзоре изменений в версии 3.0, операция «Ответить» в этой версии по умолчанию копирует в заголовок «Cc» адреса всех получателей оригинального сообщения вдобавок к адресу отправителя оригинального сообщения. Такое заполнение заголовка «Cc» можно отключить в `mailconfig`, потому что это не во всех случаях желательно. Однако в идеале выбор того или иного типа ответа можно было бы реализовать в графическом интерфейсе отдельно для каждого письма, а не для всего сеанса в целом. Для этого вполне достаточно добавить в окно со списком еще одну кнопку Reply All (Ответить всем). Из-за того что эта особенность появилась слишком поздно, чтобы добавить ее в графический интерфейс, она была оставлена как упражнение для самостоятельной реализации.

Распространение вложений

При ответе или пересылке электронной почты PyMailGUI отбрасывает все вложения, имеющиеся в оригинальном сообщении. Это сделано умышленно, отчасти потому, что в настоящее время в графическом интерфейсе отсутствует возможность удаления вложений из писем перед отправкой (вы не сможете удалить вложения ни выборочно, ни все сразу), а отчасти потому, что единственный пользователь этой системы предпочитает поступать именно так.

Пользователи могут обойти это ограничение, сохранив в некотором каталоге все вложения с помощью кнопки Split (Разбить), и затем, отправляя письмо оттуда, добавить любые нужные вложения. Однако лучше было бы предоставить пользователю возможность самому выбирать, как должна действовать система по умолчанию при создании ответов и при пересылке писем. Аналогично, пересылка писем с содержимым в формате HTML в настоящее время требует сохранения и вложения части с разметкой HTML, чтобы избежать цитирования текста – эта задача похожа на задачу распространения вложений в целом.

Запретить редактирование при просмотре письма?

Текст сообщения в окне просмотра доступен для редактирования, хотя новое письмо в этом случае не создается. Это сделано умышленно – пользователи могут вносить свои комментарии в текст сообщения и сохранять его в текстовом файле, щелкнув на кнопке Save (Сохранить) внизу окна или просто скопировав фрагменты текста в другое окно. Однако такая возможность может вводить в заблуждение и является избыточной (точно так же можно отредактировать и сохранить основной текст, щелкнув на соответствующей ему кнопке быстрого доступа). Возможность удаления инструментов редактирования

повлекла бы за собой необходимость расширения редактора PyEdit. Использование редактора PyEdit для отображения в целом является удачным решением – при работе с текстом сообщения пользователи получают доступ ко всем инструментам PyEdit, включая сохранение, поиск, переход по номеру строки, поиск во внешних файлах, замену, отмену/возврат ввода и так далее, однако в данном контексте возможность редактирования может оказаться лишней.

Автоматическая периодическая проверка электронной почты

Было бы совсем несложно добавить автоматическую периодическую проверку и получение новой входящей почты, зарегистрировав обработчик событий от таймера с большим интервалом с помощью метода `after` любого виджета или задействовав объект таймера из модуля `threading`. Я не реализовал эту возможность, потому что у меня есть некоторые предубеждения против неожиданностей, преподносимых программным обеспечением, однако ваш опыт может подсказывать иное.

Добавить кнопки Reply (Ответить) и Forward (Переслать) в окна просмотра?

Небольшое эргономическое усовершенствование: мы могли бы включить кнопки Reply (Ответить) и Forward (Переслать) в окна просмотра сообщений, не ограничивая доступ к этим операциям только из окон со списками. Как единственный пользователь этой системы я предпочитаю не нарушать внешний вид и концептуальную простоту установленного подхода – графические интерфейсы стремятся выйти из-под контроля при углублении иерархии всплывающих окон. Однако добавить кнопки Reply (Ответить) и Forward (Переслать) в окна просмотра совсем несложно, и они могли бы просто извлекать содержимое текущего сообщения, вместо того чтобы заново анализировать его.

Опустить заголовок Bcc в окнах просмотра?

Незначительное улучшение: в окнах просмотра сообщений, вероятно, лучше будет опустить заголовок «Bcc», даже если он разрешен в файле с настройками. Так как он должен отбрасываться после отправки сообщения, единственное место, где он необходим, – это окно составления нового сообщения. Он отображается в любом случае, чтобы позволить убедиться, что заголовок «Bcc» отбрасывается при передаче (в предыдущем издании этого не делалось), чтобы обеспечить единообразие внешнего вида всех окон для работы с отдельными сообщениями, чтобы избежать необходимости обработки специальных случаев и чтобы избежать принятия таких эргономических решений без согласия пользователей.

Проверка пустого заголовка «Subject»

Небольшая проблема, связанная с удобством использования: было бы несложно добавить проверку пустого поля заголовка «Subject» перед передачей и выводить диалог с запросом на подтверждение, что-

бы дать пользователю второй шанс заполнить заголовок. Чаще всего заголовок с темой оставляется пустым непреднамеренно. Такую же проверку можно было бы выполнять для поля заголовка «To», однако существуют вполне допустимые ситуации, когда этот заголовок остается пустым (сообщение отправляется по адресам в заголовках «Cc» и «Bcc»).

Более точное удаление дубликатов адресов получателей

В текущей реализации операция отправки пытается удалить повторяющиеся адреса получателей, используя операции с множествами. Этот прием работает, но он может быть неточным, когда один и тот же адрес электронной почты встречается дважды с разными компонентами имен (например, «*name1* <eml>, *name2* <eml>»). Для достижения более высокой точности можно было выполнить полный анализ адресов получателей, извлекая и сравнивая только компоненты адресов. Однако не совсем ясно, что *следует* делать, когда один и тот же адрес электронной почты появляется с разными именами. Может быть, несколько человек совместно используют одну и ту же учетную запись? Если нет, тогда какое имя использовать?

Сейчас в редких случаях может потребоваться вмешательство конечного пользователя или почтового сервера. В большинстве случаев другие клиенты электронной почты наверняка сумеют обработать имена непротиворечивыми способами. Кроме того, операция создания ответа удаляет повторяющиеся адреса перед заполнением поля заголовка «Cc» точно таким же упрощенным способом, и обе операции, отправки и создания ответа, могли бы сравнивать строки без учета регистра символов при фильтрации повторяющихся адресов.

Обработка сообщений телеконференций

Поскольку сообщения телеконференций в Интернете по своей структуре напоминают сообщения электронной почты (строки заголовков плюс текстовое тело – смотрите пример использования модуля `nntplib` в главе 13), этот сценарий можно было бы дополнить возможностью отображения не только электронных писем, но и статей из телеконференций. Право отнести это усовершенствование к искусным обобщениям или к чертовски нудной работе я оставляю за вами.

Отправка по протоколу SMTP может не выполняться в некоторых сетях?

В моей домашней сети и в офисе отправка писем по протоколу SMTP выполняется без каких-либо проблем, но иногда мне приходилось быть свидетелем проблем с отправкой писем в общедоступных сетях, имеющихся в отелях и аэропортах. В некоторых случаях операция отправки завершалась возбуждением исключения и выводом сообщения об ошибке в графическом интерфейсе; в более тяжелых случаях отправка могла завершаться без исключения и вообще без вывода сообщения об ошибке. Почта просто отправлялась в никуда,

что, как вы понимаете, неприемлемо, когда доставка содержимого имеет большое значение.

Не совсем понятно, связаны ли эти проблемы с ограничениями используемой сети, с реализацией модуля Python `smtplib` или с сервером SMTP моего интернет-провайдера. К сожалению, у меня не хватает времени смоделировать эту проблему и исследовать ее (так как я единственный пользователь системы и единственный, кто ее тестирует).

Поиск решений таких проблем я оставляю читателям в качестве самостоятельного упражнения, однако отмечу: если вы собираетесь использовать систему для отправки важной почты, вам сначала следует протестировать отправку в новом сетевом окружении, чтобы убедиться, что почта отправляется корректно. Для этого достаточно отправить почту самому себе и проверить полученное.

Оптимизация производительности

Практически все работы над этой системой до настоящего времени были посвящены ее функциональности. Система действительно позволяет выполнять некоторые операции параллельно и оптимизирует загрузку почты, первоначально получая только заголовки и кэшируя уже полученную почту, чтобы избежать необходимости повторной ее загрузки. Однако такие параметры, как нагрузка на процессор и потребление памяти вообще никак не исследовались. Вообще, в первую очередь мы стараемся писать функциональный и ясный программный код на языке Python и оставляем решение вопросов производительности на потом, только если в этом возникнет необходимость. С другой стороны, широкий круг пользователей этой программы мог бы взять на себя анализ производительности и поиск путей дальнейшего ее повышения.

Например, полный текст сообщения загружается в кэш только один раз, однако при каждом открытии его для просмотра в памяти создается копия разобранного сообщения. Для больших сообщений это может приводить к потреблению значительных объемов памяти. Снизить потребление памяти можно было бы за счет кэширования уже проанализированных писем, хотя большие письма при этом так и останутся большими, и кэш может храниться в памяти дольше, чем требуется, если не предусмотреть специальных решений. Хранение сообщений или их частей в файлах (возможно в виде объектов, преобразованных в последовательную форму), а не в памяти могло бы сдержать потребление памяти, однако для этого может потребоваться реализовать механизм обработки временных файлов. В данной реализации все сообщения, размещенные в памяти, в конечном итоге будут утилизированы сборщиком мусора Python, как только окна будут закрыты, но действия сборщика мусора могут зависеть от того, где и как мы сохраняем ссылки на объекты. Смотрите

также модули `gc` в стандартной библиотеке, где можно найти указания по управлению выборочной сборкой мусора.

Оптимизация модели поддержки Юникода

Как уже коротко обсуждалось в начале этой главы и подробно – в главе 13, в PyMailGUI поддержка кодировок Юникода текста сообщений и компонентов заголовков является достаточно широкой, но не настолько общей и универсальной, как могла бы быть. Некоторые ограничения, имеющиеся здесь, обусловлены ограничениями пакета `email` в Python 3.1, от которого сильно зависит программа PyMailGUI. Довольно сложно в почтовых клиентах на языке Python реализовать поддержку некоторых особенностей лучше, чем позволяют используемые библиотеки.

Кроме того, поддержка Юникода, присутствующая в этой программе, не подвергалась сколько-нибудь широкому или строгому тестированию. Так же, как и текстовый редактор PyEdit, представленный в главе 11, PyMailGUI в настоящее время является системой с единственным пользователем, разрабатывавшейся с целью служить книжным примером, а не проектом с открытыми исходными текстами. Вследствие этого некоторые текущие алгоритмы использования Юникода возникли отчасти из соображений здравого смысла и могут быть улучшены со временем и по мере накопления опыта.

Например, в конечном итоге при отправке может оказаться лучше повсеместно использовать кодировку UTF-8 (или вообще не использовать никакой кодировки) вместо поддержки некоторого набора пользовательских настроек, который включен в книгу для иллюстрации. Поскольку кодировку UTF-8 можно использовать для представления большинства кодовых пунктов Юникода, она является наиболее широко применимой.

Еще одна тонкая особенность: мы могли бы также предусмотреть передачу кодировки основной текстовой части компоненту PyEdit, встроенному в окна просмотра и редактирования, чтобы она могла использоваться операцией сохранения в PyEdit. В данной реализации пользователи могут открыть основную текстовую часть сообщения в окне просмотра и сохранить ее в кодировке, которая становится известной автоматически, но при сохранении черновиков редактируемых писем происходит откат к использованию политики поддержки Юникода в PyEdit и диалогах графического интерфейса. Однако невозможность однозначного задания кодировки, используемой при сохранении черновиков, может быть неизбежной – пользователи могут вводить символы из любого набора как при создании новых сообщений, так и при редактировании ответов или пересылаемых писем (как и в случае с заголовками в ответах и пересылаемых письмах, первоначальная кодировка оригинального сообщения может оказаться неприменимой после редактирования текста).

Кроме того, в программе отсутствует поддержка символов вне диапазона ASCII в полном тексте сообщения, однако в редких случаях интернационализированный текст может появляться в других контекстах (например, в именах вложенных файлов, недекодированная форма которых может оказаться недопустимой для файловой системы на платформе получателя, что может потребовать переименования, если это разрешено). И хотя программа обеспечивает поддержку интернационализированного содержимого сообщений, сам графический интерфейс по-прежнему использует английский текст для кнопок, меток и заголовков окон, что обычно не допускается в по-настоящему интернационализированных программах.

Иными словами, если эта программа когда-нибудь достигнет уровня коммерческого или широко используемого продукта, реализованную в ней поддержку Юникода наверняка придется пересмотреть. Кроме того, как уже отмечалось в главе 13, будущая версия пакета `email`, возможно, будет автоматически решать некоторые проблемы, связанные с Юникодом, однако при этом может потребоваться обновить реализацию программы PyMailGUI и исправить проблемы несовместимости, которые могут возникнуть вследствие этого. А пока она остается полезным наглядным примером: плохо ли, хорошо ли, но такие изменения всегда будут непреложным фактом в постоянно развивающемся мире разработки программного обеспечения.

И так далее – поскольку это программное обеспечение является открытым, продолжительность работы над ним не ограничивается какими-то временными рамками. В конечном счете, создание законченного клиента электронной почты является серьезным предприятием, и мы развили этот пример, насколько это возможно в данной книге. Чтобы продолжить дальнейшее развитие PyMailGUI, нам, вероятно, следует принять во внимание пригодность для этих целей как пакета `email` в версии Python 3.1, так и библиотеки tkinter создания графического интерфейса. Этих инструментов вполне достаточно для создания утилиты, которую мы реализовали здесь, но они могут тормозить дальнейшее ее развитие.

Например, отсутствие виджета просмотра HTML в базовом наборе инструментов tkinter препятствует реализации в графическом интерфейсе возможности просмотра и составления сообщений в формате HTML. Кроме того, несмотря на широкую поддержку интернационализации в PyMailGUI, приходится полагаться на обходные решения, чтобы обеспечить возможность работы с электронной почтой. Справедливости ради следует заметить, что некоторые проблемы в пакете `email`, описанные в этой книге, наверняка будут решены к тому времени, когда вы будете читать о них, и сейчас программа электронной почты, вероятно, представлена к своем худшем виде – из-за проблем интернационализации, возникающих вследствие особого отношения к Юникоду в Python 3.X. Подобные ограничения инструмента могут препятствовать дальнейшему развитию системы.

С другой стороны, несмотря на все ограничения в используемых инструментах, программа PyMailGUI успешно решает поставленные задачи. Этот в общем-то полнофункциональный и быстрый клиент электронной почты на удивление хорошо справляется с моей почтой, с моими настройками и прекрасно зарекомендовал себя в большинстве случаев, с которыми я сталкивался до настоящего времени. Возможно, он не совсем соответствует вашим вкусам и вашим условиям, но он полностью открыт для настройки и доработки. Предложенные упражнения и идеи по дальнейшему улучшению я официально передаю в ведение вашего воображения – в конце концов это Python.

На этом мы завершаем наш тур по работе с сетевыми протоколами на стороне клиента. В следующей главе мы перемахнем по ту сторону Интернета и рассмотрим сценарии, выполняющиеся на серверах. Такие программы лежат в основе важного понятия приложений, целиком живущих в Веб и запускаемых веб-броузерами. Мы делаем этот скачок в структуре, но следует понимать, что средств, с которыми мы ознакомились в этой и предыдущей главах, часто достаточно для полной реализации распределенной обработки, требуемой во многих приложениях, и работать они могут в согласии со сценариями, выполняемыми на сервере. Однако для полного понимания картины мира Веб необходимо также исследовать и царство серверов.

15

Сценарии на стороне сервера

«До чего же запутанную паутину мы плетем...»

Эта глава – уже четвертая в нашем обзоре интернет-программирования на языке Python. В предыдущих трех главах мы изучили сокеты и основные интерфейсы программирования клиентов, такие как FTP и электронная почта. В этой главе наше внимание будет сосредоточено на создании сценариев, выполняемых на стороне сервера, – программ, обычно называемых *CGI-сценариями*. Хотя эти сценарии относятся к самому нижнему уровню веб-разработки, тем не менее, они предоставляют простой способ, позволяющий начать разработку интерактивных сайтов на языке Python.

Сценарии, выполняемые на серверах, и производные от них лежат в основе большинства взаимодействий в Сети. Это справедливо как для CGI-сценариев, создаваемых вручную, так и для высокоуровневых фреймворков, автоматизирующих выполнение некоторых задач. Вследствие этого знание фундаментальной модели устройства Веб, которую мы будем исследовать здесь в контексте программирования CGI-сценариев, является совершенно необходимым условием для успешной разработки веб-приложений независимо от используемых инструментов.

Как будет показано ниже, Python является идеальным языком для создания сценариев, реализующих и настраивающих сайты, благодаря простоте применения и разнообразию библиотек. В следующей главе мы будем использовать знания, полученные здесь, для реализации полнофункциональных сайтов. А сейчас наша цель состоит в том, чтобы понять основные принципы разработки серверных сценариев, прежде

чем перейти к исследованию систем, строящихся и развертывающихся на основе этой простой модели.

Замок на песке

При чтении последующих двух глав этой книги следует иметь в виду, что они фокусируются на основах разработки серверных сценариев и могут служить лишь введением в программирование на языке Python в этой области. Сфера веб-программирования обширна и сложна, быстро и непрерывно меняется, и часто существуют различные способы достижения заданной цели, некоторые из которых могут разниться от броузера к броузеру и от сервера к серверу.

Например, схема шифрования паролей, представленная в следующей главе, может в некоторых ситуациях оказаться излишней (при наличии подходящего сервера можно использовать защищенную версию протокола HTTP). Кроме того, фрагменты HTML, которые мы будем использовать, вероятно, не отражают всей мощи этого языка разметки, и иногда могут даже не соответствовать стандартам HTML. На практике, большая часть нового материала, добавлявшегося при переизданиях этой книги, появилась в ответ на развитие технологий в области веб-программирования.

Учитывая, сколь велико и изменчиво это поле деятельности, данная часть книги не претендует на полноту рассмотрения области создания серверных сценариев. То есть вы не должны рассматривать эти главы, как последнее слово по данной теме. Чтобы стать настоящим специалистом в этой сфере, нужно изучить другие книги, излагающие дополнительные детали и приемы веб-мастерства, – например, книгу Чака Муссиано (Chuck Musciano) и Билла Кеннеди (Bill Kennedy) «HTML & XHTML: The Definitive Guide», издательство O'Reilly)[1].

Но не все так мрачно! Здесь вы познакомитесь с основными идеями, лежащими в основе разработки серверных сценариев, встретитесь с набором инструментов для CGI-программирования, имеющихся в языке Python, и получите знания, достаточные, чтобы начать создавать на языке Python собственные веб-сайты. Эти знания вы сможете применять повсюду в Веб.

[1] Чак Муссиано, Билл Кеннеди «HTML и XHTML. Подробное руководство», 6-е издание. – СПб: Символ-Плюс, 2008.

Что такое серверный CGI-сценарий?

Попросту говоря, CGI-сценарии реализуют значительную часть того взаимодействия, с которым вы сталкиваетесь в Веб. Это стандартный и широко используемый способ программирования веб-систем и взаимодействия с веб-сайтами и основа большинства моделей разработки веб-приложений.

Существуют другие способы сделать поведение сайтов интерактивным с помощью Python как на стороне клиента, так и на стороне сервера. Мы уже познакомились с некоторыми вариантами в начале главы 12. Например, в число клиентских решений входят апплеты Jython; инструменты разработки полнофункциональных интернет-приложений, такие как Silverlight и pyjamas; технология Active Scripting в Windows; и грядущий стандарт HTML 5. На стороне сервера используются различные дополнительные технологии, построенные на основе модели CGI, такие как Python Server Pages, и веб-фреймворки, такие как Django, App Engine, CherryPy и Zope, многие из которых используют модель программирования MVC (Model-View-Controller – модель-представление-поведение).

Однако в целом значительная доля операций в Веб реализована через серверные CGI-сценарии, независимо от того, созданы они вручную или частично автоматизированы применением фреймворков и инструментов. Разработка CGI-сценариев является, пожалуй, наиболее простым подходом к реализации веб-сайтов, не предоставляющим инструментов, которые часто встраиваются в большие фреймворки, таких как сохранение состояния, интерфейсы к базам данных и шаблоны. Однако разработка CGI-сценариев во многих отношениях является простейшим приемом создания сценариев, выполняющихся на стороне сервера. Как результат, они обеспечивают идеальный способ, позволяющий приступить к веб-программированию на стороне сервера. CGI-сценариев вполне достаточно для разработки простых сайтов, не требующих применения инструментов уровня предприятия, и при необходимости они могут быть усилены дополнительными библиотеками.

Притаившийся сценарий

Формально, CGI-сценарии являются программами, выполняющимися на сервере и придерживающимися общего шлюзового интерфейса (Common Gateway Interface) – модели связи между броузером и сервером, от которой CGI-сценарии берут свое название. CGI – это прикладной протокол, который используется веб-серверами для транспортировки входных данных и результатов между веб-броузерами или другими клиентами и серверными сценариями. Разобраться с протоколом CGI, видимо, удобнее исходя из предполагаемого им взаимодействия.

Большинство тех, кто путешествует по Сети и нажимает кнопки на веб-страницах, воспринимает такое взаимодействие как данность, но за ку-

лисами каждой операции в Сети происходит масса вещей. С точки зрения пользователя, это достаточно знакомые и простые операции:

Передача

Когда вы приходите на веб-сайт с целью выполнить поиск, приобрести товар или передать данные, то обычно заполняете форму в веб-броузере, нажимаете кнопку, чтобы передать информацию, и ждете ответа.

Ответ

Предполагая, что с вашим соединением с Интернетом и компьютером, с которым вы взаимодействуете, все в порядке, вы в итоге получает ответ в виде новой веб-страницы. Это может быть просто подтверждение (например, «Спасибо за сделанный вами заказ») или новая форма, которую опять нужно заполнить и отправить.

И верите вы или нет, но эта простая модель лежит в основе всей кипучей жизни в Сети. Но внутри все несколько сложнее. В действительности здесь функционирует изощренная архитектура клиент/сервер, базирующаяся на сокетах, – веб-броузер, выполняющийся на вашем компьютере, является *клиентом*, а компьютер, с которым вы связаны через Интернет, является *сервером*. Рассмотрим снова схему взаимодействия со всеми внутренними деталями, обычно невидимыми для пользователей.

Передача

После заполнения страницы с формой в веб-броузере и нажатия кнопки передачи веб-броузер незаметно для вас пересылает информацию через Интернет на сервер, указанный в качестве получателя. Сервер обычно является удаленным компьютером, расположенным в другом месте как кибернетического, так и реального пространства. Он указан в адресе URL, к которому происходит обращение, – строка адреса в Интернете, находящаяся в верхней части окна броузера. Целевые сервер и файл могут быть явно указаны в адресе URL, но чаще они указываются в разметке HTML, описывающей саму страницу для передачи, – в гиперссылке или в параметре `action` тега формы HTML.

Каким бы образом ни был указан сервер, выполняющийся на вашем компьютере, веб-броузер в конечном счете посылает ему вашу информацию через сокет с помощью технологий, рассмотренных нами в предыдущих трех главах. На компьютере сервера постоянно выполняется программа, называемая *HTTP-сервером*, которая ждет на сокете поступления данных от броузеров или других клиентов, обычно на порту с номером 80.

Обработка

Когда ваша информация оказывается на компьютере сервера, программа HTTP-сервера сначала должна обнаружить ее и решить, как обработать запрос. Если запрашиваемый адрес URL указывает про-

сто на веб-страницу (например, URL, оканчивающийся на *.html*), HTTP-сервер открывает указанный HTML-файл на компьютере сервера и возвращает его содержимое броузеру через сокет. На стороне клиента броузер читает разметку HTML и строит на ее основе страницу, которую вы видите.

Но если запрашиваемый броузером адрес URL указывает на *выполняемую программу* (например, если URL оканчивается на *.cgi* или *.py*), HTTP-сервер для обработки запроса запускает ее на компьютере сервера, переадресует данные, поступающие от броузера, в поток ввода `stdin` указанной программы и устанавливает для нее переменные окружения и аргументы командной строки. Обычно программа, запускаемая сервером, является CGI-сценарием – она выполняется на удаленном компьютере сервера, находящегося где-то в киберпространстве и, как правило, не на вашем компьютере. С этого момента вся ответственность за обработку поступивших данных ложится на CGI-сценарий – он может сохранить ваши данные в базе, выполнить поиск, списать средства с вашей кредитной карточки и так далее.

Ответ

В конечном итоге CGI-сценарий выводит разметку HTML, а также несколько строк заголовков, чтобы создать в вашем броузере новую страницу ответа. При запуске CGI-сценария HTTP-сервер должен обеспечить соединение стандартного выходного потока `stdout` сценария с сокетом, на котором броузер ждет данных. Благодаря этому разметка HTML, выводимая CGI-сценарием, передается через Интернет вашему броузеру и создает новую страницу. Разметка HTML, выводимая CGI-сценарием, действует точно так же, как если бы она хранилась и считывалась из файла, – она может описывать простую страницу ответа или совершенно новую форму для получения дополнительной информации. Поскольку разметка генерируется сценарием, она может включать информацию, определяемую динамически, на основе запроса.

Иными словами, CGI-сценарии представляют собой нечто вроде *обработчиков обратного вызова* для генерируемых веб-броузерами запросов, которые обуславливают динамическое выполнение программ. Они автоматически запускаются на компьютере сервера в ответ на действия в броузере. Хотя CGI-сценарии получают и посылают стандартные структурированные сообщения через сокеты, CGI более похож на процедурное соглашение высокого уровня для обмена информацией между броузером и сервером.

Создание CGI-сценариев на языке Python

Хотя все описанное выше может показаться сложным, не волнуйтесь – большую часть трудностей берут на себя Python и HTTP-сервер, выполняющий сценарии. Сценарии CGI пишутся как вполне автономные программы в предположении, что задачи начального запуска уже решены.

Серверная часть протокола HTTP реализуется веб-сервером, а не CGI-сценарием. Кроме того, библиотечные модули Python автоматически препарируют информацию, переданную броузером, и передают ее сценарию CGI в легко усвояемой форме. В результате сценарии могут сосредоточиться на решении прикладных задач, таких как обработка полученных данных и создание конечной страницы.

Как отмечалось выше, в контексте сценариев CGI стандартные потоки ввода-вывода, `stdin` и `stdout`, автоматически подключаются к сокетам, соединенным с броузером. Кроме того, HTTP-сервер передает часть информации от броузера CGI-сценарию в виде переменных окружения и, возможно, в виде аргументов командной строки. Для программистов CGI это означает:

- *Входные данные*, передаваемые броузером серверу, становятся потоком байтов в стандартном потоке ввода `stdin`, а также попадают в переменные окружения.
- *Выходные данные* посылаются сервером клиенту просто в результате вывода корректно сформированной разметки HTML в стандартный поток вывода `stdout`.

Наиболее сложными частями в этой схеме оказываются синтаксический анализ всех входных данных, посылаемых броузером, и форматирование данных в возвращаемом броузеру ответе. К счастью, стандартная библиотека Python в значительной мере автоматизирует обе задачи:

Ввод

С помощью модуля Python `cgi` входные данные, введенные в форму в окне веб-броузера или прикрепленные к строке URL, становятся для CGI-сценариев на языке Python значениями в объекте типа словаря. Python сам анализирует данные и совершенно независимо от стиля передачи (форма или URL) возвращает объект, в котором каждому значению, посланному броузером, соответствует пара *`key:value`*.

Вывод

В модуле `cgi` имеются также средства автоматического преобразования строк к виду, допустимому для использования в разметке HTML (например, замещающие встроенные символы `<`, `>` и `&` соответствующими экранированными последовательностями HTML). Модуль `urllib.parse` предоставляет дополнительные средства форматирования текста, вставляемого в строки URL (например, добавляющие экранирующие последовательности `%XX` и `+`).

Оба эти интерфейса мы подробно изучим далее этой главе. Сейчас же просто имейте в виду, что, хотя сценарии CGI можно писать на любом языке, эта задача решается очень просто с помощью стандартных модулей Python и атрибутов языка.

Пожалуй, картину несколько портит то, что сценарии CGI тесно связаны с синтаксисом языка разметки HTML, поскольку он должен генерироваться для создания страницы ответа. То есть можно сказать, что

в CGI-сценарии на языке Python приходится встраивать разметку на языке HTML, который является совершенно отдельным самостоятельным языком.[1] Как будет показано, то обстоятельство, что сценарии CGI создают интерфейс пользователя путем вывода разметки HTML, означает необходимость особенно тщательно следить за тем, какой текст будет помещен в код веб-страницы (например, придется экранировать операторы HTML). Еще хуже то, что для создания сценариев CGI требуется хотя бы беглое знание форм HTML, поскольку в них обычно и задаются входные данные и адрес целевого сценария.

Эта книга не ставит перед собой задачу обучения языку HTML – если вас озадачит таинственный синтаксис HTML, генерируемый приводимыми здесь сценариями, загляните в какое-нибудь введение в HTML, например, в книгу «HTML и XHTML. Подробное руководство». Имейте также в виду, что высокоуровневые инструменты и фреймворки могут иногда скрывать детали конструирования форм от программистов на языке Python, хотя и ценой новых усложнений, обычно свойственных фреймворкам. Например, с помощью HTMLgen и других подобных пакетов можно вообще не иметь дело с разметкой HTML и оперировать привычными объектами Python, однако для этого вам придется изучить прикладной интерфейс этой системы.

Запуск примеров серверных сценариев

Как и в случае с графическими интерфейсами, трудно оценить веб-системы, не учитывая их высокую интерактивность, поэтому лучшим способом прочувствовать работу приводимых примеров будет опробовать их на практике. Прежде чем перейти к программному коду, давайте подготовим среду, в которой они будут выполняться.

Для запуска программ CGI требуется наличие трех компонентов:

- Клиент для отправки запросов: броузер или сценарий
- Веб-сервер, принимающий запросы
- Сценарий CGI, который будет запускаться сервером для обработки запроса

---

[1] Интересно отметить: в главе 12 мы коротко представили другие системы, которые пошли противоположным путем, – они встраивают программный код на языке Python или обращения к нему в разметку HTML. Такую модель используют механизмы шаблонов в веб-фреймворках Zope, PSP и других, запуская программный код на языке Python, чтобы получить ту или иную часть страницы ответа. Для встраивания программного кода на языке Python эти системы должны обеспечивать поддержку встроенных специальных тегов. Сценарии CGI на языке Python, наоборот, встраивают разметку HTML внутрь программного кода, поэтому они могут выполняться непосредственно как автономные программы, хотя при этом должны запускаться веб-сервером с поддержкой CGI.

Написание сценариев CGI и есть наша дальнейшая задача, а в качестве клиента можно использовать любой веб-броузер (например, Firefox, Safari, Chrome или Internet Explorer). Как мы увидим далее, роль веб-клиента в сценариях может также играть модуль Python `urllib.request`. Единственный отсутствующий компонент – промежуточный веб-сервер.

Выбор веб-сервера

Существуют различные варианты выбора веб-серверов для проведения экспериментов. Например, открытый веб-сервер Apache представляет собой законченный веб-сервер промышленного уровня, а его модуль `mod_python`, который будет обсуждаться позднее, позволяет быстро запускать сценарии на языке Python. При желании вы можете установить и настроить его, получив полноценное решение, которое может выполняться на вашем собственном компьютере. Однако описание порядка использования веб-сервера Apache выходит далеко за рамки этой книги.

Если у вас имеется учетная запись на действующем веб-сервере, где установлен Python 3.X, можете установить туда файлы HTML и сценарии, которые встретите здесь. Для второго издания этой книги, например, все веб-примеры были выгружены в папку моей учетной записи на сервере «starship» с поддержкой Python и доступны по адресам URL вида:

```
http://starship.python.net/~lutz/PyInternetDemos.html
```

Если вы пойдете тем же путем, замените часть адреса `starship.python.net/~lutz` именем своего сервера и путем к каталогу своей учетной записи. Недостаток использования учетной записи на удаленном сервере состоит в том, что в этом случае изменять сценарии будет сложнее – вам придется либо работать непосредственно на сервере, либо копировать сценарии после каждого изменения. Кроме того, вам сначала нужно будет получить доступ к такому серверу и выяснить подробности его конфигурации, которые могут изменяться в весьма широких пределах. На сервере starship, например, файлы CGI-сценариев на языке Python должны были иметь расширение *.cgi*, право на выполнение и строку `#!`, характерную для Unix, в начале сценария, указывающую оболочке путь к интерпретатору Python.

Поиск сервера с поддержкой Python 3.X для опробования примеров из этой книги может стать сложной проблемой и занять значительное время – ни один из моих интернет-провайдеров не обеспечивал такую поддержку в середине 2010 года, когда я работал над этой главой, однако вполне возможно, что существуют коммерческие провайдеры, готовые ее предложить. Естественно, такое положение дел может измениться со временем.

Использование локального веб-сервера

Ради простоты в этом издании предпринят иной подход. Все примеры будут запускаться под управлением простого веб-сервера, написанного

на языке Python. Кроме того, веб-сервер будет выполняться на том же локальном компьютере, что и веб-броузер. Таким образом, теперь все, что осталось сделать, чтобы получить возможность опробовать примеры серверных сценариев, – это запустить сценарий веб-сервера и использовать «localhost» в качестве имени во всех адресах URL (вернитесь к главе 12, если забыли, почему это имя соответствует локальному компьютеру). Например, чтобы просмотреть веб-страницу, введите в адресную строку своего веб-броузера адрес URL следующего вида:

```
http://localhost/tutor0.html
```

Этот подход также позволяет избежать некоторых сложностей, связанных с различиями между серверами, и упрощает процедуру изменения программного кода – его можно редактировать непосредственно на локальном компьютере.

Для опробования примеров из этой книги мы будем использовать веб-сервер из примера 15.1. Это по сути тот же самый сценарий, который был представлен в главе 1, дополненный возможностью передачи ему пути к рабочему каталогу и номера порта в виде аргументов командной строки (этот же сценарий мы будем запускать в корневом каталоге большого примера в следующей главе). Мы не будем погружаться в особенности всех модулей и классов, использованных в примере 15.1, – за дополнительной информацией обращайтесь к руководству по стандартной библиотеке Python. Но, как описывалось в главе 1, этот сценарий реализует веб-сервер, который:

- Ожидает появления входящих запросов от клиентов, выполняющихся на том же компьютере, прослушивая порт, номер которого указывается в командной строке (по умолчанию используется порт с номером 80 – стандартный порт HTTP).
- Обслуживает страницы HTML из веб-каталога, путь к которому указывается в командной строке (по умолчанию используется каталог, из которого запускается сценарий веб-сервера).
- Запускает CGI-сценарии на языке Python с расширением *.py* в именах файлов из подкаталога *cgi-bin* (или *htbin*), находящегося в веб-каталоге.

Дополнительную информацию о работе веб-сервера можно найти в главе 1.

Пример 15.1. PP4E\Internet\Web\webserver.py

```
"""
Реализует веб-сервер на языке Python, способный обслуживать
страницы HTML и запускать серверные CGI-сценарии на языке Python;
этот сервер непригоден для промышленной эксплуатации (например, он
не поддерживает протокол HTTPS, медленно запускает/выполняет сценарии
на некоторых платформах), но его вполне достаточно для тестирования,
особенно на локальном компьютере;
```

```
По умолчанию обслуживает файлы и сценарии в текущем рабочем каталоге
и принимает соединения на порту 80, если не определены иные значения
для этих параметров с помощью аргументов командной строки; CGI-сценарии
на языке Python должны сохраняться в подкаталоге cgi-bin или htbin
в веб-каталоге; на одном и том же компьютере может быть запущено несколько
серверов для обслуживания различных каталогов при условии, что они
прослушивают разные порты;
"""

import os, sys
from http.server import HTTPServer, CGIHTTPRequestHandler

webdir = '.' # каталог с файлами HTML и подкаталогом cgi-bin для сценариев
port   = 80  # http://servername/ если 80, иначе http://servername:xxxx/

if len(sys.argv) > 1: webdir = sys.argv[1]    # аргументы командной строки
if len(sys.argv) > 2: port = int(sys.argv[2]) # иначе по умолчанию ., 80
print('webdir "%s", port %s' % (webdir, port))

os.chdir(webdir)        # перейти в корневой веб-каталог
srvraddr = ('', port)   # имя хоста, номер порта
srvrobj = HTTPServer(srvraddr, CGIHTTPRequestHandler)
srvrobj.serve_forever() # обслуживать клиентов до завершения
```

Чтобы запустить сервер, обслуживающий примеры сценариев из этой главы, просто выполните этот сценарий из каталога, где находится файл сценария, без аргументов командной строки. Например, из командной строки DOS:

```
C:\...\PP4E\Internet\Web> webserver.py
webdir ".", port 80
```

В Windows можно просто щелкнуть на значке этого сценария и оставить окно консоли открытым или запустить его из командной строки DOS. В Unix его можно запустить из командной строки в фоновом режиме или в отдельном окне терминала. Некоторые платформы могут потребовать от вас наличия привилегий администратора, чтобы запускать серверы, использующие порты из зарезервированного диапазона, такие как порт HTTP с номером 80. Если это ваш случай, либо запускайте сервер с необходимыми привилегиями, либо используйте иной номер порта (подробнее о портах рассказывается ниже в этой главе).

По умолчанию, когда сценарий выполняется локально, он обслуживает страницы HTML, которые запрашиваются с сервера «localhost», из каталога, где он находится или откуда запускается, и выполняет CGI-сценарии на языке Python, находящиеся в подкаталоге *cgi-bin*. Измените значение переменной `webdir` в сценарии или передайте путь к другому каталогу в командной строке. Вследствие такой организации файлы HTML в пакете примеров находятся в том же каталоге, что и сценарий веб-сервера, а CGI-сценарии находятся в подкаталоге *cgi-bin*. Иными словами, для посещения веб-страниц и запуска сценариев мы будем использовать адреса URL следующего вида соответственно:

```
http://localhost/somepage.html
http://localhost/cgi-bin/somescript.py
```

По умолчанию оба адреса отображаются в каталог, содержащий сценарий веб-сервера (*PP4E\Internet\Web*). Напомню: чтобы опробовать примеры, находящиеся на другом компьютере, просто замените части «localhost» и «localhost/cgi-bin» этих адресов именем своего сервера и строкой пути к каталогу (подробнее об адресах URL рассказывается ниже в этой главе). Сами примеры при этом будут действовать точно так же, но запросы будут передаваться уже не между программами, выполняющимися на локальном компьютере, а через сеть удаленному серверу.

Сервер в примере 15.1 не предназначен для промышленной эксплуатации, но его можно использовать для экспериментов с примерами из этой книги. Он вполне пригоден для тестирования CGI-сценариев локально, с использованием имени сервера «localhost», прежде чем вы будете развертывать их на действующем удаленном сервере. Если у вас появится желание опробовать примеры под управлением другого веб-сервера, вам потребуется экстраполировать примеры с учетом своих реалий. Такие параметры, как имена серверов и пути к каталогам в адресах URL, расширения в именах файлов CGI-сценариев и другие соглашения могут изменяться в широких пределах. За дополнительными подробностями обращайтесь к документации для своего сервера. В этой и следующей главе будет предполагаться, что используется сценарий *webserver.py*, запущенный на локальном компьютере.

Корневая страница с примерами на стороне сервера

Чтобы убедиться, что все готово к опробованию примеров, запустите сценарий веб-сервера из примера 15.1 и введите следующий адрес URL в адресную строку веб-броузера:

```
http://localhost/PyInternetDemos.html
```

По этому адресу загружается страница для запуска программ со ссылками на все файлы примеров (исходный код разметки HTML этой страницы смотрите в пакете с примерами). Сама страница для запуска выглядит в веб-броузере Internet Explorer в Windows 7, как показано на рис. 15.1 (в других броузерах и на других платформах она выглядит аналогично). Для каждого крупного примера на этой странице имеется ссылка, щелчок на которой запускает соответствующий сценарий.

Некоторые примеры можно открыть, щелкнув на соответствующих файлах HTML в окне менеджера файлов. Однако сценарии CGI, которые вызываются ссылками из примеров, должны запускаться веб-сервером. Если просматривать такие страницы непосредственно, то броузер, скорее всего, просто отобразит исходный программный код сценариев вместо того, чтобы выполнить его. Чтобы запускать сценарии CGI, необходимо открывать страницы HTML, вводя соответствующие им адреса URL с именем сервера «localhost» в адресной строке броузера.

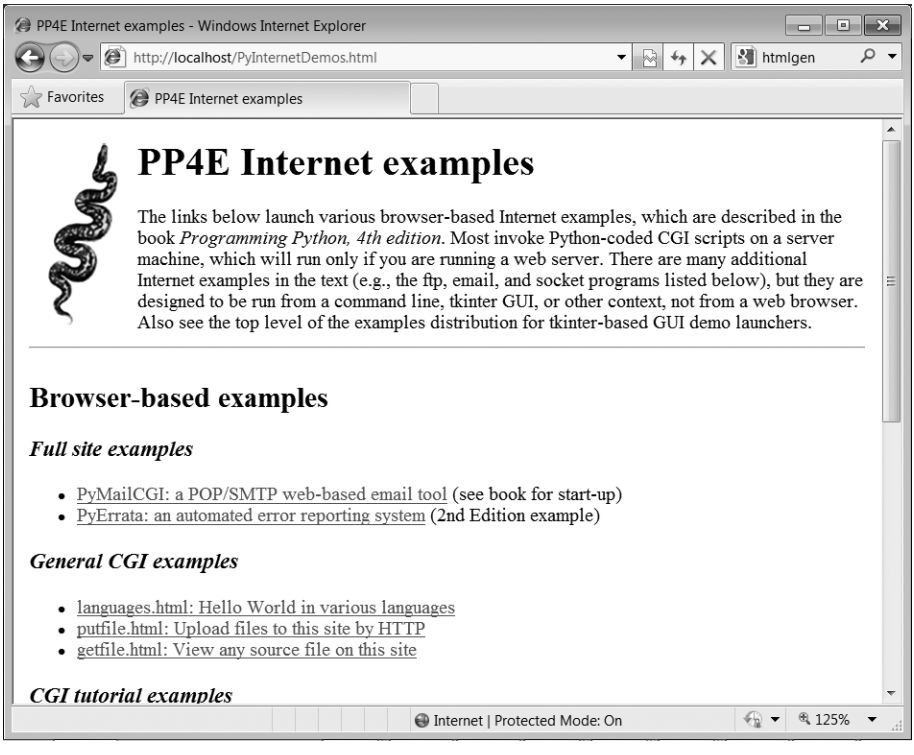


Рис. 15.1. *PyInternetDemos – страница запуска сценариев*

Рано или поздно у вас может появиться желание начать использовать более мощный веб-сервер, поэтому мы рассмотрим дополнительные подробности установки CGI ниже в этой главе. Возможно, также у вас появится желание еще раз просмотреть обзор доступных вариантов серверов в главе 12 (немалой популярностью пользуются Apache и mod_python). Эти подробности можно просто пропустить или прочитать вскользь, если вы не собираетесь устанавливать другой сервер прямо сейчас. А пока будем использовать локальный сервер.

Просмотр примеров серверных сценариев и их вывода

Исходный программный код примеров, относящихся к данной части книги, приведен в тексте и помещен в пакет примеров. Всегда, когда вам потребуется просмотреть исходный код разметки файла HTML или код разметки HTML, созданный CGI-сценарием Python, можно просто выбрать пункт меню броузера View Source (Исходный код страницы или Просмотр HTML-кода) во время отображения соответствующей веб-страницы.

Имейте, однако, в виду, что функция просмотра исходного кода страницы позволяет увидеть *вывод*, сгенерированный серверным сценарием

в процессе выполнения, а не исходный программный код самого сценария. Не существует автоматического способа получить исходный программный код Python самих CGI-сценариев, более короткого, чем отыскать его в книге или в пакете с примерами.

Для решения этой проблемы ниже в этой главе мы напишем CGI-программу с именем `getfile`, которая позволит загружать и просматривать содержимое любого файла (HTML, CGI-сценария и так далее), находящегося на веб-сайте этой книги. Достаточно будет просто ввести имя нужного файла в форму веб-страницы, на которую указывает ссылка *getfile.html* со страницы запуска демонстрационных программ, изображенной на рис. 15.1, или добавить его в качестве параметра в конец явно вводимого адреса URL, как показано ниже, – замените `tutor5.py` в конце на имя сценария, программный код которого вам необходимо увидеть, и опустите подстроку `cgi-bin` в конце, если хотите просмотреть содержимое файла HTML:

```
http://localhost/cgi-bin/getfile.py?filename=cgi-bin\tutor5.py
```

В ответ сервер возвратит текст указанного файла вашему броузеру. Этот процесс требует, однако, явного обращения к интерфейсу и значительно больших знаний об адресах URL, чем у нас пока есть. Чтобы понять, как и почему действует эта таинственная строчка, перейдем к следующему разделу.

Вверх к познанию CGI

Теперь, когда мы рассмотрели вопросы подготовки, пора заняться конкретными деталями программирования. В этом разделе мы поэтапно познакомимся с особенностями разработки CGI-сценариев – от простых неинтерактивных сценариев до больших программ, использующих все стандартные инструменты веб-страниц для ввода данных пользователем (которые были названы виджетами в третьей части книги, посвященной разработке графических интерфейсов на основе библиотеки tkinter).

Попутно мы исследуем основные идеи разработки серверных сценариев. Сначала мы будем продвигаться неспешно, чтобы методично изучить основы, а в следующей главе перейдем к использованию идей, с которыми познакомимся здесь, для создания примеров более крупных и практических сайтов. А пока давайте пройдем простой начальный учебный курс по CGI, с объемом разметки HTML, достаточным для написания базовых серверных сценариев.

Первая веб-страница

Как уже говорилось, сценарии CGI тесно связаны с HTML, поэтому начнем с простой страницы HTML. Файл *tutor0.html*, представленный в примере 15.2, определяет добротную, полнофункциональную веб-

страницу. Это текстовый файл, содержащий код разметки HTML, который определяет структуру и содержимое простой веб-страницы.

Пример 15.2. PP4E\Internet\Web\tutor0.html

```
<HTML>
<TITLE>HTML 101</TITLE>
<BODY>
<H1>A First HTML Page</H1>
<P>Hello, HTML World!</P>
</BODY></HTML>
```

Если ввести в веб-броузере адрес этого файла в Интернете, то должна появиться страница, изображенная на рис. 15.2. Здесь мы видим, как выглядит в броузере Internet Explorer страница, находящаяся по адресу *http://localhost/tutor0.html* (введите этот адрес в адресную строку своего броузера); при этом предполагается, что локальный веб-сервер, описанный в предыдущем разделе, был запущен и работает. В других броузерах страница будет отображена сходным образом. Поскольку это статическая страница HTML, вы получите тот же результат, просто щелкнув на значке файла в большинстве платформ, хотя в таком режиме его содержимое не будет получено от веб-сервера.

Рис. 15.2. *Простая веб-страница из файла HTML*

Чтобы действительно разобраться в том, как действует этот маленький файл, требуется некоторое представление о синтаксисе HTML, адресах Интернета и правах доступа к файлам. Рассмотрим кратко каждую из названных тем, а потом перейдем к следующему примеру.

Основы HTML

Я обещал, что не стану много рассказывать в этой книге о языке разметки HTML, но для понимания примеров некоторые знания все-таки потребуются. Вкратце, HTML – это описательный язык разметки, основанный на *тегах* – элементах, заключенных в пару символов <>. Одни теги являются самостоятельными (например, <HR> определяет горизон-

тальную линию). Другие действуют парно, обозначая начало и конец области действия тега; при этом в концевой тег входит дополнительный символ слэша.

Например, строка заголовка уровня 1 определяется на языке HTML как `<H1>text</H1>`. Текст *text* между тегами будет выведен на веб-странице. Некоторые теги позволяют также указывать дополнительные параметры (иногда их называют атрибутами). Например, пара тегов вида `<A  href="address">text</A>` определяет *гиперссылку*: если щелкнуть на тексте ссылки на странице, броузер выполнит переход по интернет-адресу (URL), указанному в параметре `href`.

Важно помнить, что язык разметки HTML используется только для описания страниц: веб-броузер читает код HTML и на основе содержащегося в нем описания строит веб-страницу с заголовками, абзацами, ссылками и тому подобным. Примечательно, что в разметке отсутствует *информация о расположении* (за размещение компонентов на странице отвечает броузер) и *программная логика* – в разметке нет операторов `if`, циклов и так далее. Кроме того, в тексте примера 15.2 отсутствует программный код на языке Python – исходный код HTML предназначен только для описания страниц, а не для составления программ или определения всех деталей интерфейса пользователя.

Отсутствие в HTML элементов управления интерфейсом пользователя и возможности определять логику работы является одновременной сильной и слабой его чертой. Он хорошо приспособлен для описания страниц и простых интерфейсов пользователя на высоком уровне. Броузер, а не вы, занимается физическим размещением компонентов страницы на экране. С другой стороны, HTML непосредственно не поддерживает развитые возможности графических интерфейсов, что вызывает необходимость применения на веб-сайтах CGI-сценариев (или других технологий, таких как полнофункциональные интернет-приложения, RIA) для введения динамической составляющей в статичный по сути HTML.

Интернет-адреса (URL)

После создания файла HTML его необходимо поместить в таком месте, где он будет доступен веб-броузеру. Если вы используете локальный веб-сервер на языке Python, как описано выше, это совсем просто: адреса URL вида *http://localhost/file.html* открывают доступ к веб-страницам, а *http://localhost/cgi-bin/file.py* – к CGI-сценариям. Такое задание адресов обусловлено тем, что сценарий веб-сервера по умолчанию обслуживает страницы и сценарии, находящиеся в том каталоге, откуда он был запущен.

На других серверах адреса URL могут иметь более сложный вид. Как и все файлы HTML, *tutor0.html* должен быть записан в каталог на сервере, из которого выполняющаяся программа веб-сервера разрешает броузерам получать страницы. Например, на сервере, использовавшемся в примерах ко второму изданию этой книги, файл страницы должен

был храниться внутри или ниже каталога *public_html* в моем личном домашнем каталоге, то есть где-то в дереве каталогов, растущем из */home/lutz/public_html*. Полный путь к этому файлу на сервере Unix будет:

```
/home/lutz/public_html/tutor0.html
```

Этот путь отличается от адреса *PP4E\Internet\Web* в пакете примеров, прилагаемом к книге, указанного в заголовке примера 15.2. Однако при ссылке на этот файл вместо полного пути к файлу в дереве каталогов клиент должен указывать его адрес в Интернете, также называемый URL. Ниже приводится адрес URL, который использовался для загрузки страницы с удаленного сервера:

```
http://starship.python.net/~lutz/tutor0.html
```

Удаленные серверы автоматически отображают такие адреса URL в путь к файлу в файловой системе, практически так же, как начальная часть адреса *http://localhost* отображается в каталог с примерами, содержащий сценарий веб-сервера, действующий на локальном компьютере. Вообще говоря, строки URL, подобные этой, состоят из нескольких частей:

Имя протокола: http

Часть этого адреса URL, содержащая протокол, сообщает броузеру, что он должен связаться с программой HTTP-сервера, выполняющейся на компьютере сервера, с помощью протокола сообщений HTTP. В адресах URL, используемых в броузерах, могут указываться различные протоколы. Например, *ftp://* – для ссылки на файл, обслуживаемый сервером FTP и действующий по протоколу FTP, *file://* – для ссылки на локальный файл, *telnet* – для начала сеанса Telnet клиента и так далее.

Имя сервера и порт: starship.python.net

Адрес URL также определяют доменное имя или IP-адрес (Internet Protocol – протокол Интернета) целевого сервера. В данном случае это доменное имя сервера, где установлены примеры. Заданное имя используется при открытии сокета для связи с сервером. Как обычно, имя *localhost* (или эквивалентный ему IP-адрес *127.0.0.1*) обозначает веб-сервер, выполняющийся на том же компьютере, что и клиент.

При необходимости данная часть адреса URL может также явно определять номер порта, на котором сервер принимает запросы на соединение, следующий за двоеточием (например, *starship.python.net:8000* или *127.0.0.1:80*). В случае использования протокола HTTP сокет обычно соединяется с портом номер 80, который используется по умолчанию, если номер порта не указан в адресе. Более подробную информацию об именах и портах можно найти в главе 12.

Путь к файлу: ~lutz/tutor0.html

Наконец, адрес URL определяет путь к нужному файлу на удаленной машине. Веб-сервер HTTP автоматически транслирует путь к файлу из URL в действительный путь к файлу: на моем сервере starship

часть адреса `~lutz` автоматически транслируется в каталог *public_html* в моем домашнем каталоге. При использовании сценария веб-сервера на языке Python, представленного в примере 15.1, пути к файлам отображаются в текущий рабочий каталог сервера. Обычно адреса URL отображаются в такие файлы, но могут ссылаться и на элементы других видов и, как будет показано чуть ниже, даже на выполняемые CGI-сценарии, которые автоматически запускаются сервером при обращении к ним.

Параметры запроса (в более поздних примерах)

За адресом URL могут также следовать дополнительные входные параметры для программ CGI. Параметры указываются после знака `?` и отделяются один от другого символом `&`. Например, строка в конце URL вида `?name=bob&job=hacker` передает CGI-сценарию, указанному в URL, параметры `name` и `job` со значениями `bob` и `hacker` соответственно. Как будет обсуждаться ниже в этой главе, когда мы будем исследовать правила экранирования, иногда параметры могут отделяться символами `;`, как в строке `?name=bob;job=hacker`, хотя такая форма встречается значительно реже.

Эти значения иногда называются *параметрами строки запроса* URL, они обрабатываются так же, как получаемые сценариями данные формы. Технически параметры запроса могут иметь различную структуру (например, можно передавать неименованные значения, разделяемые символом `+`), однако в этой книге мы дополнительные возможности рассматривать не будем; подробнее о формах и параметрах рассказывается ниже в этой главе.

Чтобы убедиться, что мы ухватили суть синтаксиса адресов URL, разберем еще один пример, который мы будем использовать ниже в этой главе. Компоненты следующего адреса URL для протокола HTTP:

```
http://localhost:80/cgi-bin/languages.py?language=All
```

однозначно идентифицируют сценарий на сервере, как описывается ниже:

- Имя сервера `localhost` обозначает веб-сервер, выполняющийся на том же компьютере, что и клиент, – как объяснялось выше, именно такую конфигурацию мы будем использовать в наших примерах.
- Номер порта 80 определяет порт сокета, на котором веб-сервер принимает запросы на соединение (порт 80 используется по умолчанию, если номер порта отсутствует в адресе URL, поэтому мы обычно опускаем его).
- Путь к файлу `cgi-bin/languages.py` определяет местоположение файла, который будет выполняться на сервере, внутри каталога, откуда сервер начинает поиск запрашиваемых файлов.
- Строка запроса `?language=All` содержит входной параметр для указанного сценария `languages.py` и является альтернативой полям ввода в форме (описываются ниже).

Несмотря на то, что под представленное описание подпадает большинство реальных адресов URL, тем не менее, полная форма представления адресов URL выглядит так:

```
protocol://networklocation/path;parameters?querystring#fragment
```

Например, часть `fragment` адреса определяет имя раздела внутри страницы (например, `#part1`). Кроме того, каждая часть адреса может иметь свой собственный формат, а некоторые части поддерживаются не для всех протоколов. Например, часть `;parameters` не поддерживается для протокола HTTP (она явно определяет тип файла при использовании протокола FTP), а часть `networklocation` может также определять необязательные параметры аутентификации пользователя для некоторых протоколов (полный формат представления этой части для протоколов FTP и Telnet имеет вид `user:password@host:port`, а для протокола HTTP – `host:port`). Например, в главе 13 мы уже использовали сложные адреса URL для протокола FTP, включающие имя пользователя и пароль, а также определяющие двоичный тип файла (если тип файла не указан, он может быть определен сервером автоматически):

```
ftp://lutz:password@ftp.rmi.net/filename;type=i
```

Мы не будем здесь отвлекаться на дополнительные правила форматирования адресов URL. Если вам нужны подробности, можно начать с чтения описания модуля `urllib.parse` в руководстве по стандартной библиотеке Python, а также просмотреть его программный код в стандартной библиотеке Python. Можно также обратить внимание, что URL, который вводится для доступа к некоторой странице, выглядит несколько иначе после того, как страница получена (пробелы превращаются в символы `+`, добавляются `%` и так далее). Это связано с тем, что броузеры, как правило, должны следовать соглашениям по экранированию (то есть трансляции) адресов URL, которые будут изучены далее в этой главе.

Использование минимальных адресов URL

Поскольку броузеры запоминают интернет-адрес предыдущей страницы, адреса URL, встроенные в HTML-файлы, часто могут опускать названия протоколов и имена серверов, а также путь к каталогу файла. При отсутствии каких-то частей броузер просто использует их значения, взятые из адреса предшествующей страницы. Такой минимальный синтаксис действует как для адресов URL, встроенных в гиперссылки, так и для действий форм (с формами мы познакомимся ниже). Например, внутри страницы, полученной из каталога *dirpath* на сервере *http://www.server.com*, минимальные адреса в гиперссылках и действиях форм, такие как:

```
<A HREF="more.html">
<FORM ACTION="next.py" ...>
```

обрабатываются в точности так же, как если бы был задан полный адрес URL с явными составными частями, содержащими имя сервера и путь к каталогу:

```
<A HREF="http://www.server.com/dirpath/more.html">
<FORM ACTION="http://www.server.com/dirpath/next.py" ...>
```

Первый минимальный URL ссылается на файл *more.html* на том же сервере и в том же каталоге, откуда получена страница, содержащая эту гиперссылку. Броузер расширяет его до полного URL. В компоненте адреса URL пути к файлу может также применяться синтаксис относительного пути в стиле Unix. Например, тег гиперссылки типа `<A HREF="../spam.gif">` указывает на GIF-файл на сервере в родительском каталоге файла, содержащего URL этой ссылки.

К чему нужна вся эта возня с укороченными адресами URL? Помимо увеличения срока службы клавиатуры и сохранения зрения главным преимуществом таких минимальных адресов URL является отсутствие необходимости изменять их при перемещении страниц в новый каталог или на другой сервер – сервер и путь логически определяются при использовании страницы, а не задаются жестко в ее разметке HTML. В противном случае последствия могут оказаться достаточно болезненными: примеры, содержащие явные ссылки на сайты и имена путей в адресах URL, находящихся в коде HTML, нельзя копировать на другие серверы без внесения изменений в исходный код разметки. Помощь в этом могут оказать специальные сценарии, но редактирование исходного кода разметки чревато появлением ошибок.

Недостаток минимальных адресов URL заключается в том, что при переходе по ним не происходит автоматического соединения с Интернетом. Это становится заметным только при загрузке страниц из локальных файлов на вашем компьютере. Например, обычно можно открывать страницы HTML вообще без соединения с Интернетом, открывая локальные файлы страниц в веб-броузере (например, щелкнув на значке файла). При таком локальном просмотре страницы переход по полностью заданному адресу URL заставляет броузер автоматически соединиться с Интернетом, чтобы получить нужную страницу или сценарий. Однако минимальные URL открываются снова на локальном компьютере – обычно в этом случае броузер просто выводит исходный код страницы или сценария.

В итоге оказывается, что минимальные адреса URL лучше переносимы, но успешнее работают, когда все страницы действительно загружаются из Интернета. Чтобы упростить работу с примерами, приведенными в этой книге, в адресах URL, содержащихся в них, часто опускаются составляющие с именем сервера и путем к каталогу. Для данной книги превращение минимального адреса URL страницы или сценария в истинный происходит добавлением перед именем файла в адресе URL строки:

```
http://localhost
```

Ваш броузер сделает это, если этого не сделаете вы.

Ограничения прав доступа к файлам HTML

Одно дополнительное указание, прежде чем двинуться дальше: если вам потребуется установить примеры на другой сервер, на некоторых платформах может потребоваться предоставить всем пользователям право чтения файлов веб-страниц и их каталогов. Это требуется потому, что их будет загружать из Сети произвольный пользователь (часто некто с именем «nobody», с которым мы познакомимся чуть ниже).

В Unix-подобных системах изменение прав доступа может осуществляться соответствующей командой `chmod`. Например, обычно достаточно команды оболочки `chmod 755 filename`; она дает всем права на чтение и выполнение файла *filename*, а запись разрешается только вам.[1] Такие права доступа к файлам и каталогам являются типичными, но на некоторых серверах могут быть отличия. Загружая файлы HTML на свой сайт, поинтересуйтесь тем, какие соглашения приняты на вашем сервере.

Первый CGI-сценарий

HTML-файл, который мы видели в предыдущем разделе, – это всего лишь HTML-файл, а не сценарий CGI. Когда броузер обращается к нему, удаленный веб-сервер просто отправляет обратно текст файла, с помощью которого в броузере создается новая страница. Чтобы проиллюстрировать природу CGI-сценариев, перепишем этот пример в виде CGI-программы на языке Python, как показано в примере 15.3.

Пример 15.3. PP4E\Internet\Web\cgi-bin\tutor0.py

```
#!/usr/bin/python
"""
выполняется на сервере, выводит разметку HTML для создания новой страницы;
url=http://localhost/cgi-bin/tutor0.py
"""


print('Content-type: text/html\n')
print('<TITLE>CGI 101</TITLE>')
print('<H1>A First CGI Script</H1>')
print('<P>Hello, CGI World!</P>')
```

Файл *tutor0.py* создает такую же страницу, как показано на рис. 15.2, если обратиться к нему из броузера, – просто замените расширение файла в адресе URL с **.html** на **.py** и добавьте в путь имя подкаталога

[1] Это не какие-то магические числа. В системах Unix режим 755 является битовой маской. Первая цифра 7 означает, что вы (владелец файла) можете читать, записывать и выполнять файл (7 в двоичном представлении имеет вид 111, где каждый бит соответствует определенному режиму доступа). Две цифры 5 (101 в двоичном представлении) указывают, что все остальные (ваша группа и те, кто остался) могут осуществлять чтение и выполнение (но не запись) файла. Детали смотрите в страницах справочного руководства по команде `chmod`.

cgi-bin, чтобы в результате в адресной строке броузера получился адрес *http://localhost/cgi-bin/tutor0.py*.

Но это совсем другой зверь – это *выполняемая программа*, запускаемая на сервере в ответ на попытку обратиться к ней. Это также совершенно законная программа на языке Python, в которой разметка HTML страницы выводится динамически, а не заготовлена заранее в статичном файле. На самом деле в этой программе вообще мало чего специфического для CGI – при запуске из командной строки она просто выведет разметку HTML:

```
C:\...\PP4E\Internet\Web\cgi-bin> python tutor0.py
Content-type: text/html

<TITLE>CGI 101</TITLE>
<H1>A First CGI Script</H1>
<P>Hello, CGI World!</P>
```

Однако при выполнении этой программы HTTP-сервером на компьютере веб-сервера стандартный поток вывода подключается к сокету, из которого осуществляет чтение броузер на компьютере клиента. Поэтому весь вывод пересылается через Интернет веб-броузеру. Вследствие этого выводимый сценарием текст должен иметь формат, соответствующий ожиданиям броузера.

В частности, когда вывод сценария достигает броузера, первая полученная им строка интерпретируется как заголовок, описывающий последующий текст. В выводимом ответе может быть несколько строк заголовков, но между заголовками и началом разметки HTML (или другими данными) всегда должна быть одна пустая строка. Как будет показано ниже, директивы «cookies», управляющие механизмом сохранения информации, также передаются в области заголовков – перед пустой строкой.

В данном сценарии первая строка заголовка сообщает броузеру, что дальнейшая передача представляет собой текст в формате HTML (`text/html`), а символ новой строки (`\n`) в конце первого вызова функции `print` генерирует дополнительный перевод строки помимо того, который создает сама функция `print`. В результате сразу после заголовка добавляется пустая строка. Оставшаяся часть вывода программы представляет собой стандартную разметку HTML, которая используется броузером для создания веб-страницы, в точности так, как если бы разметка HTML находилась в статическом HTML-файле на сервере.[1]

---

[1] Обратите внимание, что этот сценарий не генерирует охватывающие теги `<HEAD>` и `<BODY>`, имеющиеся в статическом файле HTML, представленном в предыдущем разделе. Строго говоря, он должен был бы сделать это – разметка HTML без таких тегов является недействительной. Но все стандартные броузеры не обращают внимания на такое упущение. Если от вас требуется заботиться о подобных тонкостях, обращайтесь за официальными подробностями к справочникам по языку HTML.

Обращение к сценариям CGI производится в точности как к файлам HTML: вам необходимо ввести в адресную строку броузера полный адрес URL этого сценария либо выполнить щелчок на ссылке *tutor0.py* в корневой странице примеров, изображенной на рис. 15.1 (осуществляющий переход по минимальной гиперссылке, преобразуемой в полный адрес URL сценария). На рис. 15.3 показана страница, которая будет сгенерирована сценарием, если обратиться к нему из броузера.

Рис. 15.3. *Простая веб-страница, созданная сценарием CGI*

Установка сценариев CGI

Если вы используете локальный веб-сервер, описанный в начале этой главы, то, чтобы заставить этот пример работать, никаких дополнительных действий по установке выполнять не требуется, и вы можете просто пропустить большую часть этого раздела. Однако, если сценарий CGI необходимо задействовать на другом сервере, вам потребуется знание некоторых практических деталей. В этом разделе для справки дается краткий обзор типичных особенностей настройки CGI.

Как и файлы HTML, сценарии CGI являются простыми текстовыми файлами, которые можно создать на локальном компьютере и загрузить на сервер по FTP либо написать в тестовом редакторе, выполняемом непосредственно на сервере (возможно, с помощью клиента Telnet или SSH). Однако из-за того что сценарии CGI выполняются как программы, для них существуют некоторые особые требования к установке, отличные от требований к обычным файлам HTML. В частности, их, как правило, требуется хранить и именовать специальным образом, и они должны быть настроены как программы, выполнять которые разрешено любым пользователям. В зависимости от потребностей после выгрузки CGI-сценариев на сервер может потребоваться обеспечить им возможность отыскивать импортируемые модули, а также преобразование их в формат текстовых файлов, соответствующий платформе сервера. Рассмотрим каждое ограничение установки более подробно:

Соглашения по каталогам и именам файлов

Прежде всего, сценарии CGI должны быть помещены в каталог, который веб-сервер распознает как каталог программ, и им должны даваться имена, которые сервер распознает как имена сценариев CGI. Для использования с локальным веб-сервером, задействованным в этой главе, сценарии должны помещаться в специальный подкаталог *cgi-bin* и иметь расширение *.py*. На сервере, использовавшемся для опробования примеров во втором издании книги, напротив, CGI-сценарии должны были помещаться в каталог *public_html*, как обычные файлы HTML, а имена файлов сценариев должны были оканчиваться расширением *.cgi*, а не *.py*. Некоторые серверы допускают расширения *.py* имен файлов и могут распознавать другие каталоги программ, но в этом они могут существенно отличаться друг от друга, иногда в зависимости от настроек, установленных для сервера или для пользователя.

Соглашения по выполнению

Поскольку сценарии CGI должны выполняться веб-сервером от имени произвольных пользователей в Сети, их файлам также требуется дать право на выполнение, чтобы пометить их как программы, и разрешение на выполнение должно быть дано остальным пользователям. На большинстве серверов это можно сделать командой оболочки `chmod 0755 filename`.

На некоторых серверах сценарии CGI также должны начинаться со строки `#!`, определяющей интерпретатор Python, который будет выполнять программный код в файле. Текст после `#!` в первой строке просто указывает путь к выполняемому файлу Python на сервере. Подробнее об этой специальной первой строке читайте в главе 3 и обязательно ознакомьтесь с соглашениями, действующими на вашем сервере, если он работает не под управлением Unix.

Некоторые серверы могут требовать наличие этой строк, даже если они работают не под управлением Unix. Большинство сценариев CGI в этой книге содержат строку `#!` на тот случай, если когда-нибудь они будут выполняться в Unix-подобных системах, – наш локальный веб-сервер при выполнении в Windows просто игнорирует первую строку, интерпретируя ее как комментарий Python.

Следует отметить одну тонкость: как было показано выше в книге, специальная первая строка в выполняемых текстовых файлах обычно может содержать либо жестко определенный путь к интерпретатору Python (например, *#!/usr/bin/python*), либо вызов программы `env` (например, *#!/usr/bin/env python*), которая определяет местонахождение Python по значениям переменных окружения (например, `$PATH`). Однако прием с `env` менее полезен в сценариях CGI, так как значения их переменных окружения будут соответствовать пользователю «nobody» (а не вашим настройкам), что разъясняется в следующем абзаце.

Настройка пути поиска модулей (необязательная)

Некоторые HTTP-серверы по соображениям безопасности обычно выполняют CGI-сценарии с привилегиями пользователя «nobody» (это ограничивает доступ пользователя к серверу). Поэтому для файлов, публикуемых в Сети, должны быть установлены специальные права, делающие их доступными другим пользователям. Это также означает, что сценарии CGI не могут рассчитывать, что путь поиска модулей Python будет установлен каким-либо особым образом. Как мы видели, путь к модулям обычно инициализируется согласно значению переменной окружения `PYTHONPATH` для пользователя и содержимому файлов *.pth* плюс значениям по умолчанию, которые обычно включают текущий рабочий каталог. Но так как сценарии CGI выполняются с привилегиями пользователем «nobody», переменная `PYTHONPATH` может иметь произвольное значение при выполнении CGI-сценария.

Не ломайте себе над этим голову, поскольку на практике проблем с этим обычно не возникает. Благодаря тому, что по умолчанию Python обычно ищет импортируемые модули в текущем каталоге, проблем вообще не будет, если все сценарии, а также любые используемые ими модули и пакеты будут храниться в вашем веб-каталоге, а веб-сервер будет запускать сценарии CGI из каталога, где они размещаются. Но если модуль находится в другом месте, может потребоваться изменять список `sys.path` в сценариях, чтобы вручную настроить путь поиска перед импортом – например, с помощью вызова функции `sys.path.append(`*`dirname`*`)`, присваиваний по индексам и так далее.

Соглашения по символам конца строки (необязательные)

На некоторых серверах Unix (и Linux) может также потребоваться обеспечить соответствие текстовых файлов сценариев соглашениям Unix по концу строки (`\n`), а не DOS (`\r\n`). Проблем не возникнет, если редактирование и отладку производить прямо на сервере (или на другом компьютере, работающем под управлением Unix) или один за другим передавать файлы по FTP в текстовом режиме. Но если редактировать сценарии на PC и загружать файлы на сервер Unix в виде *tar*-архива (или в двоичном режиме FTP), то после загрузки может потребоваться преобразовать символы конца строки. Например, сервер, который использовался при работе над вторым изданием этой книги, возвращал установленную по умолчанию страницу сообщения об ошибке для сценариев, в которых конец строки имел формат DOS. Приемы и примечания по созданию сценариев, автоматизирующих преобразование символов конца строки, вы найдете в главе 6.

Небуферизованный режим работы потока вывода (необязательно)

На некоторых серверах функция `print` может буферизовать вывод. Если ваш сценарий CGI выполняется продолжительное время, что-

бы не заставлять пользователя ждать начала появления результатов, можно вручную выталкивать буферы (вызывая метод `sys.stdout.flush()`) или запустить сценарии Python в небуферизованном режиме. Напомню, что в главе 5 рассказывалось о возможности перевода потока вывода в небуферизованный режим за счет использования флага `-u` командной строки интерпретатора или установки переменной окружения `PYTHONUNBUFFERED` в любое непустое значение.

Чтобы передать интерпретатору флаг `-u` в мире CGI, попробуйте указать его в первой строке сценария на Unix-подобных платформах, например: `#!/usr/bin/python -u`. Однако обычно буферизация не оказывает существенного влияния. Тем не менее, для некоторых серверов и клиентов отключение ее может помочь решить проблему получения пустой страницы ответа или преждевременной отправки заголовка ошибки завершения сценария – на стороне клиента может истечь предельное время ожидания до того, как сервер отправит буферизованный поток вывода (хотя обычно подобные проблемы являются отражением действительных программных ошибок в сценариях).

На первый взгляд, этот процесс установки может показаться несколько сложным, но во многом он зависит от особенностей сервера, и когда вы разберетесь с ним самостоятельно, все окажется не так страшно. Установка занимает некоторое время и обычно до некоторой степени может быть автоматизирована с помощью сценариев Python, выполняемых на сервере. Подведем итоги. Большинство сценариев CGI на языке Python являются текстовыми файлами, содержащими программный код Python, которые:

- Именуются согласно соглашениям веб-сервера (например, *file.cgi*).
- Хранятся в каталоге, распознаваемом веб-сервером (например, *cgi-bin/*).
- Имеют права доступа исполняемых файлов, если это необходимо (например, `chmod 755 file.py`).
- На некоторых серверах могут потребовать добавить в начало особую строку `#!`*`pythonpath`*.
- Настраивают `sys.path`, только если это необходимо, чтобы увидеть модули в других каталогах.
- Используют соглашения Unix по концу строки, только если сервер не принимает формат DOS.
- При необходимости выталкивают выходные буферы или отправляют ответ порциями.

Даже если приходится использовать сервер, который настраивает кто-то другой, большая часть соглашений легко выявится еще на этапе отладки. Как обычно, необходимо проконсультироваться по поводу соглашений, действующих на каждом из тех серверов, куда вы планируете скопировать эти файлы примеров.

Поиск Python на удаленном сервере

Последняя подсказка по установке: в контексте веб-приложения на стороне сервера не требуется установка Python у *клиентов*, но он должен присутствовать на компьютере *сервера*, где должны выполняться сценарии CGI. Если вы пользуетесь своим собственным веб-сервером – сценарием *webserver.py*, представленным выше, или другим свободно распространяемым сервером, таким как Apache, то это не представляет проблемы.

Но если вы пользуетесь сервером, который настраивали не вы сами, необходимо убедиться в наличии интерпретатора Python на этом компьютере. Более того, необходимо узнать каталог его установки, чтобы указать путь к нему в строке `#!` в начале своего сценария. Если вы не уверены, установлен ли интерпретатор Python на сервере и где он находится, вам помогут несколько советов:

- Особенно в системах Unix следует сначала предположить, что Python находится в стандартном месте (например, */usr/local/bin/python*): введите команду `python` (или `which python`) в окне командной оболочки и проверьте его работоспособность. Весьма вероятно, что Python уже установлен в системе. Если у вас есть доступ к серверу посредством Telnet или SSH, можно воспользоваться командой Unix `find`, выполнив с ее помощью поиск, начиная с каталога */usr*.
- Если ваш сервер работает под управлением Linux, то, вероятно, все готово к работе. В настоящее время интерпретатор Python является стандартной частью дистрибутивов Linux, а под управлением операционной системы Linux работают многие веб-сайты и серверы провайдеров услуг Интернета. На таких сайтах Python обычно устанавливается в каталог */usr/bin/python*.
- В других средах, где нет возможности самостоятельно контролировать компьютер сервера, получить доступ к уже установленному интерпретатору Python может оказаться труднее. В этих случаях можно переместить свой сайт на сервер, где Python точно установлен, или убедить провайдера установить Python на сервере, который вы хотите использовать, или установить Python на сервере самому.

Если ваш провайдер без понимания относится к вашей потребности в Python и вы хотите перенести свой сайт на сервер провайдера, который предоставляет возможность его использования, поищите в Интернете списки провайдеров, дружественно относящихся к Python. А если вы предпочтете установить Python на сервере самостоятельно, обязательно рассмотрите возможность применения инструментов создания *скомпилированных файлов* программ на языке Python – с их помощью можно создать единственный файл выполняемой программы, который целиком содержит интерпретатор Python вместе со всеми стандартными библиотечными модулями. Такой скомпилированный интерпретатор можно по FTP за один шаг выгрузить в каталог вашей учетной записи на сервере, при этом не потребуется полной инсталляции Python на сервере.

Создать двоичный скомпилированный файл с программой на языке можно с помощью свободно распространяемых программ PyInstaller и Py2Exe.

Наконец, обратите внимание, что для опробования примеров из этой книги необходима версия Python 3.X, а не Python 2.X. Как уже упоминалось выше, на момент, когда я работал над четвертым изданием, многие коммерческие провайдеры Интернета поддерживали версию 2.X, а не 3.X, но такое положение вещей, как ожидается, должно измениться с течением времени. Если вам удалось найти провайдера, обеспечивающего поддержку Python 3.X, выгрузите свои файлы по FTP и используйте для работы SSH или Telnet. Вы можете также запустить на удаленном сервере сценарий *webserver.py* из этой главы, хотя при этом вам, возможно, придется отказаться от идеи использовать стандартный порт 80 – все зависит от того, какой уровень управления будет предоставлен вашей учетной записи.

Добавление картинок и создание таблиц

Вернемся к написанию серверных сценариев. Каждому, когда-либо бродившему по Сети, известно, что веб-страницы обычно содержат не только обычный текст. В примере 15.4 представлен CGI-сценарий на языке Python, который выводит разметку HTML, содержащую тег `<IMG>` для включения графического изображения в страницу, отображаемую броузером клиента. Чего-либо особенно характерного для Python в этом примере нет, но обратите внимание, что как и простые файлы HTML, графический файл (*ppsmall.gif*) также располагается на сервере и загружается с него в процессе интерпретации броузером вывода этого сценария.

Пример 15.4. PP4E\Internet\Web\cgi-bin\tutor1.py

```
#!/usr/bin/python

text = """Content-type: text/html

<TITLE>CGI 101</TITLE>
<H1>A Second CGI Script</H1>
<HR>
<P>Hello, CGI World!</P>
<IMG src="../ppsmall.gif" BORDER=1 ALT=[image]>
<HR>
"""
print(text)
```

Обратите внимание на использование блока строк в тройных кавычках – вся строка с разметкой HTML посылается броузеру в один прием с помощью находящегося в конце вызова функции `print`. Обязательно убедитесь в том, что пустая строка между заголовком «Content-type» и первой строкой разметки HTML действительно является пустой (некоторые броузеры могут допускать ошибки при интерпретации, если в этой

строке будут присутствовать пробелы или символы табуляции). Если и клиент, и сервер функционируют нормально, при обращении к этому сценарию будет сгенерирована страница, изображенная на рис. 15.4.

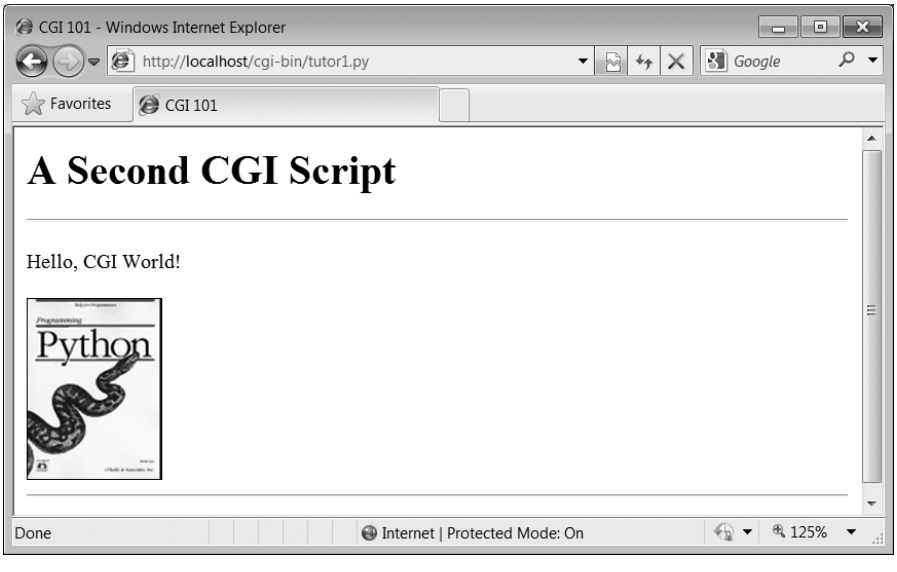


Рис. 15.4. *Страница с изображением, созданная сценарием tutor1.py*

До сих пор наши сценарии CGI выводили готовую разметку HTML, которую с таким же успехом можно было бы хранить в файле HTML. Но поскольку CGI-сценарии являются выполняемыми программами, они также способны воспроизводить разметку HTML динамически – даже, например, в ответ на определенный набор данных, введенных пользователем и переданных сценарию. В конце концов, в этом и состоит назначение сценариев CGI. Давайте начнем использовать это качество в своих целях и напишем сценарий Python, который программно конструирует разметку HTML, как показано в примере 15.5.

Пример 15.5. PP4E\Internet\Web\cgi-bin\tutor2.py

```
#!/usr/bin/python

print("""Content-type: text/html

<TITLE>CGI 101</TITLE>
<H1>A Third CGI Script</H1>
<HR>
<P>Hello, CGI World!</P>

<table border=1>
""")
```

```
for i in range(5):
    print('<tr>')
    for j in range(4):
        print('<td>%d.%d</td>' % (i, j))
    print('</tr>')

print("""
</table>
<HR>
""")
```

Несмотря на все теги, это действительно программный код на языке Python – сценарий *tutor2.py* также встраивает блоки HTML с помощью строк в тройных кавычках. Но на этот раз вложенные циклы `for` динамически генерируют часть разметки HTML, посылаемой броузеру. В частности, сценарий создает разметку HTML двумерной таблицы, размещающейся в середине страницы, как показано на рис. 15.5.

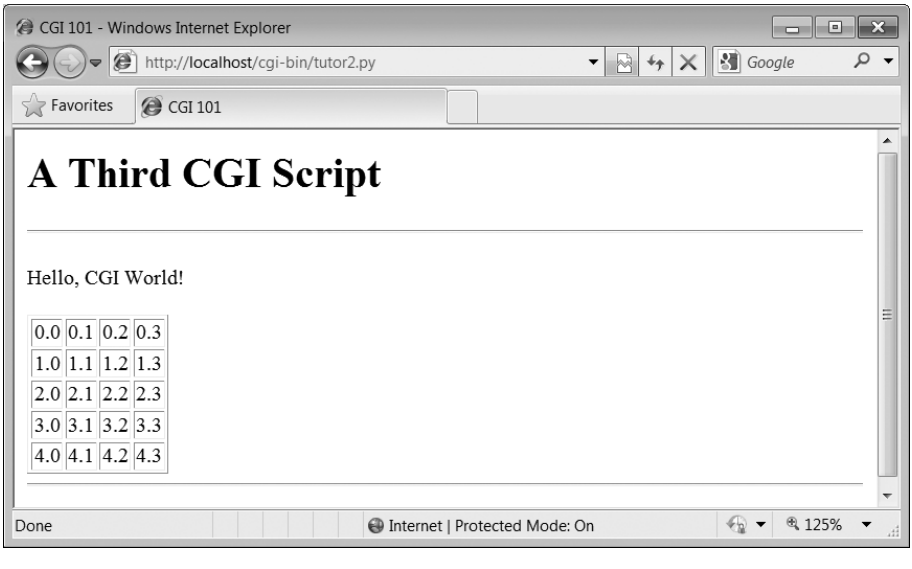


Рис. 15.5. *Страница с таблицей, генерируемая tutor2.py*

В каждой строке таблицы выводятся пары чисел «номер_строки.номер_колонки», полученные в процессе выполнения сценария Python. Если вам интересно узнать, как выглядит сгенерированная разметка HTML, откройте страницу в броузере и выберите пункт меню View Source (Исходный код страницы или Просмотр HTML-кода). Это единая страница HTML, сгенерированная первым вызовом функции `print` сценария, последующими циклами `for` и, наконец, последним вызовом `print`. Иными словами, объединение вывода этого сценария представляет собой HTML-документ с заголовками.

Теги таблиц

Сценарий в примере 15.5 генерирует теги таблиц HTML. Опять-таки, мы не собираемся здесь изучать язык HTML, но представим его в объеме, достаточном для понимания примеров, приводимых в этой книге. Таблицы HTML объявляются как текст между тегами `<table>` и `</table>`. Текст таблицы, в свою очередь, обычно состоит из строк, объявляемых с помощью тегов `<tr>` и `</tr>`, и колонок в каждой из строк, заключенных в теги `<td>` и `</td>`. Циклы в нашем сценарии конструируют разметку HTML, объявляющую пять строк по четыре колонки в каждой, путем вывода соответствующих тегов, при этом значениями ячеек являются номера текущих строки и колонки.

Например, ниже приводится часть вывода сценария, определяющая первые две строки (чтобы увидеть полный вывод, запустите этот сценарий как обычную программу из командной строки или воспользуйтесь пунктом меню View Source (Исходный код страницы или Просмотр HTML-кода) в вашем броузере):

```
<table border=1>
<tr>
<td>0.0</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>1.0</td>
<td>1.1</td>
<td>1.2</td>
<td>1.3</td>
</tr>
. . .
</table>
```

Другие теги и параметры таблиц позволяют определять заголовки строк (`<th>`), тип границ и так далее. В одном из следующих разделов мы приведем расширенный синтаксис таблиц для придания организованной структуры формам.

Добавление взаимодействия с пользователем

Сценарии CGI прекрасно умеют генерировать разметку HTML на лету, как было показано выше, но их также часто используют для реализации взаимодействий с пользователями, вводящими данные в веб-броузере. Как отмечалось ранее в этой главе, веб-взаимодействия обычно представляют собой двухэтапный процесс и реализуются с использованием двух разных веб-страниц: вы заполняете форму ввода на странице и нажимаете кнопку Submit Query (Отправить запрос), а в ответ возвращается новая страница. В промежутке данные формы обрабатывает сценарий CGI.

Передача данных

Это описание выглядит достаточно просто, но процедура получения данных, введенных пользователем, требует понимания специального тега HTML, `<form>`. Рассмотрим реализацию простого веб-взаимодействия, чтобы посмотреть, как действуют формы. Сначала нужно определить страницу с формой, заполняемой пользователем, как показано в примере 15.6.

Пример 15.6. PP4E\Internet\Web\tutor3.html

```
<html>
<title>CGI 101</title>
<body>
<H1>A first user interaction: forms</H1>
<hr>
<form method=POST action="http://localhost/cgi-bin/tutor3.py">
    <P><B>Enter your name:</B>
    <P><input type=text name=user>
    <P><input type=submit>
</form>
</body></html>
```

tutor3.html является простым файлом HTML, а не сценарием CGI (хотя его содержимое можно было бы вывести и с помощью сценария). При обращении к этому файлу текст между тегами `<form>` и `</form>` генерирует поля ввода и кнопку Submit Query (Отправить запрос), как показано на рис. 15.6.

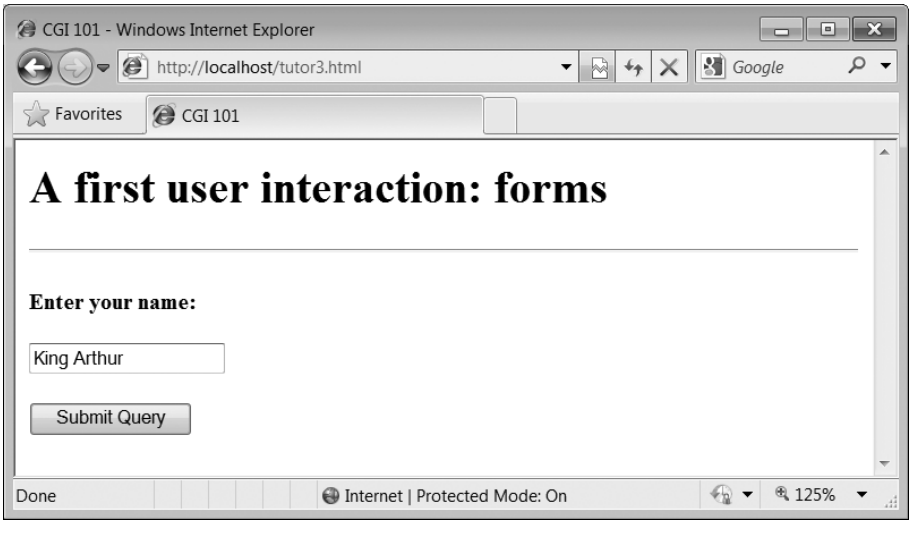


Рис. 15.6. *Страница простой формы, генерируемая файлом tutor3.html*

Дополнительно о тегах формы

Мы не станем подробно обсуждать основы создания форм HTML, но подчеркнем некоторые места. Внутри формы HTML:

Обработчик action формы

Параметр формы `action` определяет адрес URL CGI-сценария, который будет вызван для обработки отправленных данных формы. Это ссылка из формы на обрабатывающую ее программу – в данном случае программу *tutor3.py* в подкаталоге *cgi-bin*, находящемся в текущем рабочем каталоге, откуда был запущен локальный веб-сервер. Параметр `action` близок по духу параметру `command` в кнопках tkinter – это место, где регистрируется обработчик обратного вызова (в данном случае удаленный сценарий-обработчик).

Поля ввода

Управляющие элементы ввода определяются вложенными тегами `<input>`. В данном примере теги ввода обладают двумя ключевыми параметрами. Параметр `type` в текстовых полях ввода принимает значение `text`, а в кнопках отправки формы (которые отправляют данные серверу и по умолчанию имеют метку «Submit Query») – значение `submit`. Параметр `name` служит для идентификации введенного значения по имени, когда данные формы будут получены сервером. Например, серверный сценарий CGI, который мы увидим чуть ниже, использует строку `user` как ключ к получению данных, введенных в текстовое поле этой формы.

Как будет показано в последующих примерах, другие параметры тега `<input>` могут определять начальные значения (`value=X`), режим «только для чтения» (`readonly`) и так далее. Также ниже мы увидим, что параметр `type` может принимать другие значения – для передачи скрытых данных (`type=hidden`), повторной инициализации полей (`type=reset`) или создания кнопки для выбора нескольких вариантов (`type=checkbox`).

Метод отправки: `get` *и* `post`

Формы имеют также параметр `method`, определяющий стиль кодирования, используемый при передаче данных через сокет на целевой сервер. В данном случае используется стиль `post`, при котором устанавливается связь с сервером, а затем ему в отдельной передаче посылается поток данных, введенных пользователем.

Другим возможным вариантом является стиль `get`, при котором входная информация передается серверу за один шаг путем добавления данных пользователя в конец адреса URL, вызывающего сценарий, обычно после символа `?`. Параметры запроса были уже представлены выше, когда мы знакомились с адресами URL, – мы еще будем использовать их далее в этом разделе.

При передаче методом get входные данные обычно оказываются на сервере в виде переменных окружения или аргументов командной строки, используемой для запуска сценария. При передаче методом post данные принимаются со стандартного ввода и декодируются. Поскольку метод get добавляет входные параметры в конец адреса URL, он дает пользователям возможность делать закладки с параметрами для отправки данных позднее (например, ссылки на страницы интернет-магазинов с названиями товаров). Метод post в большей степени приспособлен для однократной отправки данных (например, для отправки комментария).

Метод get обычно считается более эффективным, но он ограничивается конечной длиной строки запроса в операционной системе и менее безопасен (параметры могут сохраняться в файлах журналов на сервере, например). Метод post позволяет обрабатывать большие объемы входных данных и в некоторых случаях обеспечивает более высокий уровень безопасности данных, но он требует выполнить дополнительную передачу. К счастью, модуль Python cgi прозрачным образом обрабатывает оба стиля кодирования данных, поэтому нашим сценариям CGI не приходится беспокоиться о том, который из стилей использован.

Обратите внимание, что для иллюстрации адрес URL в параметре action в данном примере записан в виде полного адреса. Так как броузер помнит, откуда пришла страница HTML, содержащая адрес URL, он будет действовать так же, если указать лишь имя файла сценария, как показано в примере 15.7.

Пример 15.7. PP4E\Internet\Web\tutor3-minimal.html

```
<html>
<title>CGI 101</title>
<body>
<H1>A first user interaction: forms</H1>
<hr>
<form method=POST action="cgi-bin/tutor3.py">
    <P><B>Enter your name:</B>
    <P><input type=text name=user>
    <P><input type=submit>
</form>
</body></html>
```

Полезно помнить, что адреса URL, указываемые в параметре action тегов форм и в гиперссылках, служат адресами в первую очередь для броузера, а не для сценария. Сам сценарий *tutor3.py* безразличен к тому, какого вида URL его запустил – минимальный или полный. На самом деле все части URL, включая имя файла сценария (и вплоть до параметров запроса URL) участвуют в диалоге между броузером и HTTP-сервером до того, как будет запущен сценарий CGI.

С другой стороны, адреса URL, поступающие из-за пределов страницы (например, введенные в поле адреса броузера или отправляемые модулю Python `urllib.request`), обычно должны иметь полную форму, потому что в этом случае к ним не применимо понятие предыдущей страницы.

Сценарий, возвращающий ответ

Пока мы создали только одну статическую страницу с полем ввода. Но кнопка отправки на этой странице творит чудеса. При нажатии на нее запускается удаленная программа, адрес URL которой указан в параметре `action` формы, и этой программе передаются данные, введенные пользователем, в соответствии с параметром `method` формы, определяющим стиль кодирования. Пока пользователь на клиентской машине ждет ответа, на сервере запускается сценарий Python, представленный в примере 15.8, который обрабатывает данные, введенные в форму.

Пример 15.8. PP4E\Internet\Web\cgi-bin\tutor3.py

```
#!/usr/bin/python
"""
выполняется на стороне сервера, читает данные формы, выводит разметку HTML;
url=http://server-name/cgi-bin/tutor3.py
"""


import cgi
form = cgi.FieldStorage()          # извлечь данные из формы
print('Content-type: text/html')   # плюс пустая строка

html = """
<TITLE>tutor3.py</TITLE>
<H1>Greetings</H1>
<HR>
<P>%s</P>
<HR>"""

if not 'user' in form:
    print(html % 'Who are you?')
else:
    print(html % ('Hello, %s.' % form['user'].value))
```

Как и прежде, этот CGI-сценарий на языке Python выводит разметку HTML, воспроизводящую страницу ответа в броузере клиента. Но этот сценарий делает еще кое-что: он использует стандартный модуль `cgi` для извлечения данных, введенных пользователем на предыдущей веб-странице (см. рис. 15.6).

К счастью, в Python это происходит автоматически: обращение к конструктору класса `FieldStorage` модуля `cgi` автоматически выполняет всю работу по извлечению данных формы из входного потока и переменных окружения независимо от способа передачи этих данных – в виде потока в стиле `post` или в виде параметров, добавляемых в URL в стиле `get`.

Вводимые данные, пересылаемые в любом из стилей, выглядят для сценариев Python одинаково.

Сценарии должны вызвать конструктор `cgi.FieldStoreage` только один раз, перед обращением к значениям полей. В результате этого вызова возвращается объект, имеющий вид словаря, – поля для ввода данных пользователем из формы (или параметры из адреса URL) представляются в виде значений ключей этого объекта. Например, `form['user']` в сценарии является объектом, атрибут `value` которого представляет собой строку, содержащую текст, введенный в текстовое поле формы. Если вы перелистаете книгу назад, к примеру с разметкой HTML страницы формы, то заметите, что параметр `name` поля ввода имел значение `user` – имя в форме HTML стало ключом, по которому введенное значение извлекается из словаря. Объект, возвращаемый конструктором `FieldStorage`, поддерживает и другие операции со словарями, например с помощью метода `in` можно проверить, есть ли некоторое поле во входных данных.

Перед завершением этот сценарий выводит разметку HTML, создающую страницу результата, на которой повторяются данные, введенные пользователем в форму. Два выражения форматирования строк (%) применяются для вставки введенного текста в строку ответа, а эта строка ответа – в заключенный в тройные кавычки блок строк HTML. Вывод сценария выглядит так:

```
<TITLE>tutor3.py</TITLE>
<H1>Greetings</H1>
<HR>
<P>Hello, King Arthur.</P>
<HR>
```

В броузере этот вывод превращается в страницу, изображенную на рис. 15.7.

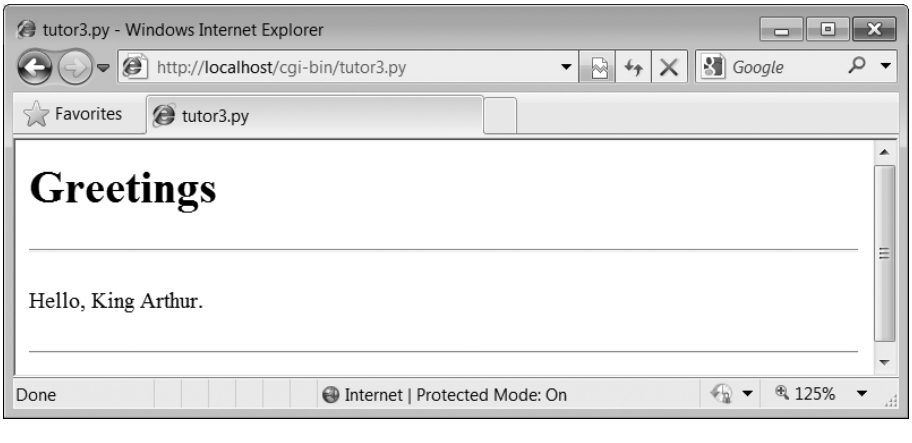


Рис. 15.7. *Результат обработки параметров формы сценарием tutor3.py*

Передача параметров в адресах URL

Обратите внимание, что вверху броузера отображается адрес URL сценария, сгенерировавшего эту страницу. Сам адрес URL мы не вводили – он появился из атрибута `action` тега `<form>` в разметке HTML предыдущей страницы. Однако ничто не мешает ввести адрес URL сценария вручную в адресную строку броузера, чтобы запустить сценарий, как мы это делали ранее, в примерах сценария CGI и файла HTML.

Но здесь есть одна ловушка: откуда возьмется значение поля ввода, если страницы с формой нет? То есть если ввести адрес URL сценария CGI самостоятельно, как будет заполнено поле ввода? Ранее, говоря о форматах адресов URL, я упомянул, что в схеме кодирования `get` входные параметры помещаются в конец URL. При явном вводе адресов сценариев тоже можно добавлять входные значения в конец URL, где они служат той же цели, что и поля `<input>` в формах. Кроме того, модуль Python `cgi` обеспечивает одинаковое представление в сценариях данных, полученных из адресов URL и форм.

Например, можно вообще пропустить заполнение страницы с формой и прямо вызвать сценарий *tutor3.py*, обратившись к URL вида (введя его вручную в адресной строке броузера):

```
http://localhost/cgi-bin/tutor3.py?user=Brian
```

В этом адресе URL значение поля ввода с именем `user` задано явно, как если бы пользователь заполнил форму ввода. При вызове подобным образом единственным ограничением является соответствие имени параметра `user` имени, ожидаемому сценарием (и жестко определенному в разметке HTML формы). Мы используем здесь лишь один параметр, но в целом перечень параметров URL обычно предваряется символом `?`, за которым следует одно или более присваиваний вида `name=value`, разделяемых символами `&`, если их больше одного. На рис. 15.8 представлена страница ответа, получаемая после ввода URL с явно указанными данными.

Фактически формы HTML, определяющие стиль кодирования `get`, тоже вызывают добавление входных параметров в адреса URL подобным образом. Попробуйте изменить пример 15.6 так, чтобы он использовал параметр `method=GET`, и отправьте форму – содержимое поля ввода имени из формы появится в адресе страницы ответа в виде параметра запроса, как и в адресе URL, который был введен вручную на рис. 15.8. Формы могут использовать любой стиль, `post` или `get`. При вводе адресов URL вручную можно использовать только метод `get`.

В целом, любой сценарий CGI можно вызвать, заполнив и передав страницу формы или передав входные данные в конце адреса URL. Вводить вручную параметры в адреса URL может оказаться довольно трудным делом, особенно если сценарий ожидает получить множество сложных параметров, однако процесс конструирования строк запроса можно автоматизировать с помощью других программ.

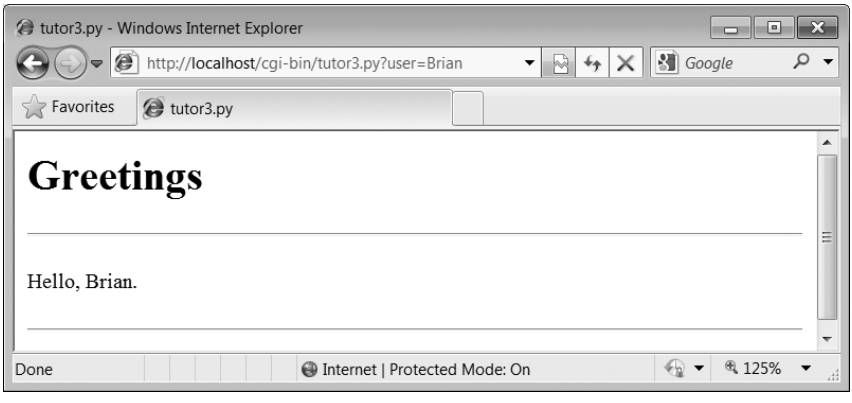


Рис. 15.8. *Результат обработки параметров URL сценарием tutor3.py*

Нетрудно заметить сходство между приведенными вызовами сценариев CGI с передачей им явно входных параметров и *функциями*, использующимися удаленно в Сети. Передача данных сценариям через URL аналогична передаче именованных аргументов функциям Python как по действию, так и синтаксически. На самом деле некоторые продвинутые веб-фреймворки, такие как Zope, делают связь между адресами URL и вызовами функций в Python еще более явной (адреса URL более тесно связаны с вызовами функций Python).

Между прочим, если очистить поле ввода имени на форме ввода (то есть сделать его пустым) и нажать кнопку отправки, то поле `user` с именем окажется пустым. Точнее, броузер может вообще не послать это поле вместе с другими данными формы, несмотря на то, что оно присутствует в разметке HTML, описывающей структуру формы. Сценарий CGI обнаруживает такое отсутствующее поле с помощью метода `in` словаря и порождает в ответ страницу, изображенную на рис. 15.9.

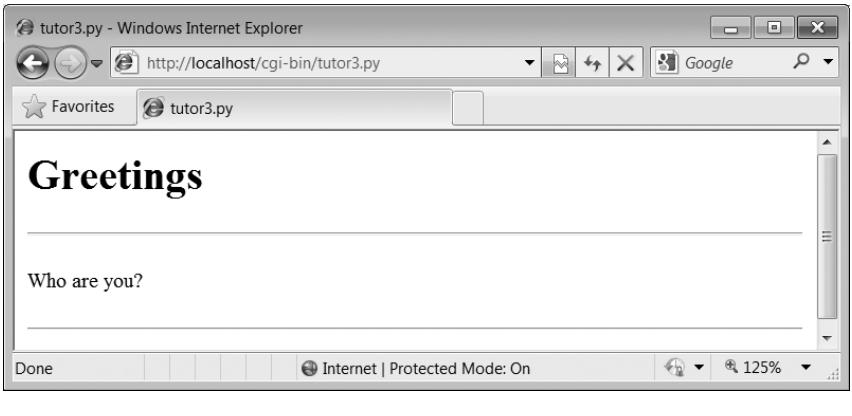


Рис. 15.9. *Пустое поле с именем порождает страницу сообщения об ошибке*

Вообще говоря, сценарии CGI должны проверять отсутствие каких-либо полей, поскольку пользователь мог не ввести их в форму или формы вообще нет, – поля ввода могут быть не добавлены в конец явно введенного адреса URL. Например, если ввести URL сценария вообще без параметров – то есть отбросить текст, начиная с ?, и оставить только адрес *http://localhost/cgi-bin/tutor3.py* – будет получена та же самая страница ответа с сообщением об ошибке. Так как любой сценарий CGI может быть вызван через форму или через URL, сценарии должны быть готовы к обеим ситуациям.

Тестирование с помощью модуля urllib.request без использования броузера

После того как мы разобрались, как отправлять поля ввода форм в виде параметров строки запроса в конце адреса URL, модуль Python `urllib.request`, о котором говорилось в главах 1 и 13, оказывается для нас еще более полезным. Напомню, что этот модуль позволяет получать ответ, сгенерированный для любого адреса URL. Когда адрес URL ссылается на простой файл HTML, мы просто загрузим его содержимое. Но когда адрес ссылается на сценарий CGI, действие будет состоять в запуске удаленного сценария и получении его вывода. Это понятие открывает путь к *веб-службам*, которые генерируют разметку XML в ответ на входные параметры. В простейшем случае это позволяет *тестировать* удаленные сценарии.

Например, мы можем напрямую запустить сценарий из примера 15.8, без использования веб-страницы *tutor3.html* или ввода адреса URL в адресную строку броузера:

```
C:\...\PP4E\Internet\Web> python
>>> from urllib.request import urlopen
>>> reply = urlopen('http://localhost/cgi-bin/tutor3.py?user=Brian').read()
>>> reply
b'<TITLE>tutor3.py</TITLE>\n<H1>Greetings</H1>\n<HR>\n<P>Hello, Brian.</P>\
n<HR>\n'

>>> print(reply.decode())
<TITLE>tutor3.py</TITLE>
<H1>Greetings</H1>
<HR>
<P>Hello, Brian.</P>
<HR>

>>> url = 'http://localhost/cgi-bin/tutor3.py'
>>> conn = urlopen(url)
>>> reply = conn.read()
>>> print(reply.decode())
<TITLE>tutor3.py</TITLE>
<H1>Greetings</H1>
<HR>
```

```
<P>Who are you?</P>
<HR>
```

В главе 13 рассказывалось, что метод `urllib.request.urlopen` возвращает объект файла, подключенный к потоку сгенерированного ответа. Операция чтения из этого файла возвращает разметку HTML, которую обычно получают броузеры и отображают ее в виде страницы. В версии 3.X ответ принимается из сокета, в виде строки байтов `bytes`, но его можно декодировать в строку `str`, если это необходимо.

Когда подобным способом выполняется непосредственное получение разметки HTML ответа, ее можно проанализировать с помощью инструментов Python обработки текста, включая строковые методы, такие как `split` и `find`, модуль `re` сопоставления с регулярными выражениями или модуль `html.parser` разбора HTML – все эти инструменты мы будем рассматривать в главе 19. Извлечение текста из ответа, подобное этому, иногда неформально называют *чтением с экрана* (*screen scraping*) – способ использования содержимого веб-сайта в других программах. Прием чтения с экрана является альтернативой использованию более сложных фреймворков веб-служб, хотя и ненадежной: незначительные изменения в формате страницы могут нарушать работу программ, опирающихся на этот метод. Наконец, текст ответа можно просто просмотреть – модуль `urllib.request` позволяет тестировать сценарии CGI из интерактивной оболочки Python или из других сценариев без использования броузера.

В более общем случае эта методика позволяет использовать серверный сценарий как своеобразную функцию. Например, графический интерфейс на стороне клиента может вызывать сценарий CGI и анализировать сгенерированную страницу ответа. Аналогично, сценарий CGI, сохраняющий изменения в базе данных, можно вызывать программно, с помощью модуля `urllib.request`, без использования страницы с формой ввода. Это также открывает путь к автоматизации регрессионного тестирования сценариев CGI – мы можем вызывать сценарии на любом удаленном компьютере и сравнивать их ответы с ожидаемыми результатами.[1] Мы еще встретимся с модулем `urllib.request` в последующих примерах.

Прежде чем двинуться дальше, необходимо сделать несколько замечаний, касающихся особенностей использования модуля `urllib.request`. Во-первых, этот модуль также поддерживает прокси-серверы, альтернативные режимы передачи данных, поддержку защищенного протокола HTTPS на стороне клиента, cookies, перенаправление и многое другое. Например, он может обеспечить прозрачную поддержку прокси-серверов с применением переменных окружения или настроек системы или

[1] Если в ваши обязанности входит тестирование серверных сценариев, обратите внимание на систему Twill, написанную на языке Python, которая предоставляет простой язык для описания взаимодействий клиента с веб-приложениями. Подробности ищите в Сети.

с помощью объектов ProxyHandler в этом модуле (подробности и примеры их использования смотрите в документации).

Кроме того, хотя это обычно и не имеет значения для сценариев Python, модуль urllib.request предоставляет возможность отправлять параметры методами get и post, описанными выше. По умолчанию используется метод get, и тогда параметры передаются в виде строки запроса в конце адреса URL, как было показано выше. Чтобы использовать метод post, необходимо передать параметры в виде отдельного аргумента:

```
>>> from urllib.request import urlopen
>>> from urllib.parse import urlencode
>>> params = urlencode({'user': 'Brian'})
>>> params
'user=Brian'
>>>
>>> print(urlopen('http://localhost/cgi-bin/tutor3.py', params).read().
decode())
<TITLE>tutor3.py</TITLE>
<H1>Greetings</H1>
<HR>
<P>Hello, Brian.</P>
<HR>
```

Наконец, если ваше веб-приложение опирается на использование cookies на стороне клиента (обсуждаются ниже), их поддержка обеспечивается модулем urllib.request автоматически, с использованием имеющейся поддержки в стандартной библиотеке Python, позволяющей сохранять cookies локально и позднее возвращать их серверу. Также поддерживаются перенаправление, аутентификация и многое другое. Кроме того, поддерживается передача данных по защищенному протоколу HTTP (HTTPS), если ваш компьютер обеспечивает работу с защищенными сокетами (такая поддержка имеется в большинстве систем). Подробности смотрите в руководстве по стандартной библиотеке Python. Поддержку работы с cookies мы будем рассматривать далее в этой главе, а в следующей познакомимся с защищенным протоколом HTTPS.

Табличная верстка форм

Теперь возьмем что-нибудь более реальное. В большинстве CGI-приложений на страницах с формами ввода присутствуют несколько полей. Когда полей больше одного, метки и поля для ввода обычно располагаются в виде таблицы, чтобы придать форме структурированный внешний вид. Файл HTML в примере 15.9 определяет форму с двумя полями ввода.

Пример 15.9. PP4E\Internet\Web\tutor4.html

```
<html>
<title>CGI 101</title>
<body>
```

```
<H1>A second user interaction: tables
</H1>
<hr>
<form method=POST action="cgi-bin/tutor4.py">
    <table>
        <TR>
            <TH align=right>Enter your name:
            <TD><input type=text name=user>
        <TR>
            <TH align=right>Enter your age:
            <TD><input type=text name=age>
        <TR>
            <TD colspan=2 align=center>
            <input type=submit value="Send">
    </table>
</form>
</body></html>
```

Тег <TH> определяет колонку, как и <TD>, но также помечает ее как колонку заголовка, что обычно означает вывод ее полужирным шрифтом. Размещая поля ввода и метки в такой таблице, получаем страницу ввода, изображенную на рис. 15.10. Метки и поля ввода автоматически выравниваются по вертикали в колонках подобно тому, как это обеспечивалось менеджерами компоновки графического интерфейса из библиотеки tkinter, с которыми мы встречались ранее в этой книге.

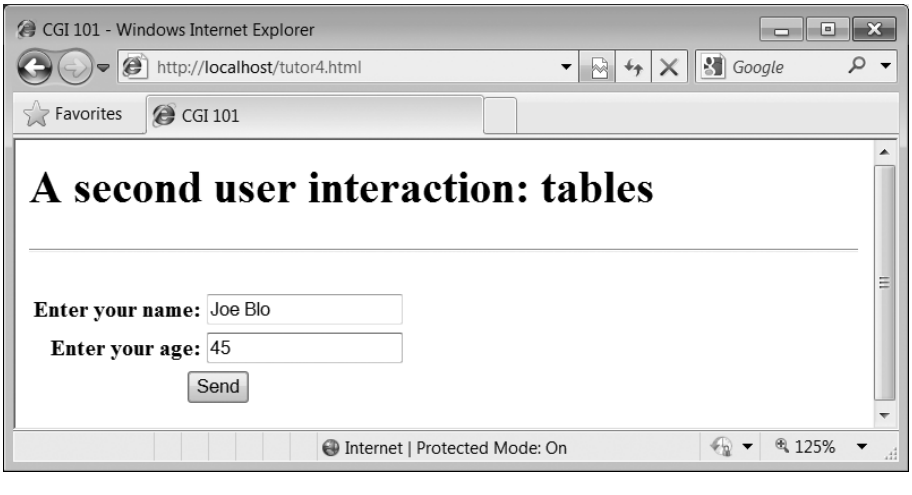


Рис. 15.10. *Форма, скомпонованная с помощью тегов таблиц*

При нажатии кнопки отправки на этой форме (с надписью Send (Передать)) на сервере будет запущен сценарий из примера 15.10 со входными данными, введенными пользователем.

Пример 15.10. PP4E\Internet\Web\cgi-bin\tutor4.py

```
#!/usr/bin/python
"""
выполняется на стороне сервера, читает данные формы, выводит разметку HTML;
URL http://server-name/cgi-bin/tutor4.py
"""

import cgi, sys
sys.stderr = sys.stdout              # для вывода сообщений
                                     # об ошибках в броузере
form = cgi.FieldStorage()            # извлечь данные из формы
print('Content-type: text/html\n')   # плюс пустая строка

# class dummy:
#     def __init__(self, s): self.value = s
# form = {'user': dummy('bob'), 'age':dummy('10')}

html = """
<TITLE>tutor4.py</TITLE>
<H1>Greetings</H1>
<HR>
<H4>%s</H4>
<H4>%s</H4>
<H4>%s</H4>
<HR>"""

if not 'user' in form:
    line1 = 'Who are you?'
else:
    line1 = 'Hello, %s.' % form['user'].value

line2 = "You're talking to a %s server." % sys.platform

line3 = ""
if 'age' in form:
    try:
        line3 = "Your age squared is %d!" % (int(form['age'].value) ** 2)
    except:
        line3 = "Sorry, I can't compute %s ** 2." % form['age'].value

print(html % (line1, line2, line3))
```

Табличная структура определяется файлом HTML, а не этим CGI-сценарием на языке Python. На самом деле в этом сценарии не так много нового – как и прежде, с помощью операторов форматирования входные значения вставляются в строку шаблона с разметкой HTML ответа, заключенную в тройные кавычки, но на этот раз для каждого поля ввода выводится отдельная строка. Когда этот сценарий будет запущен в результате операции отправки формы ввода, он выведет новую страницу ответа, изображенную на рис. 15.11.

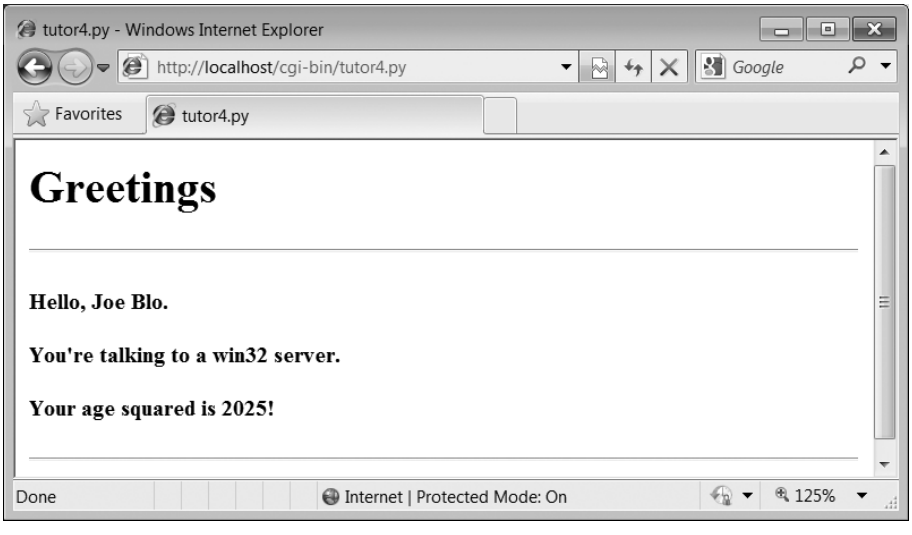


Рис. 15.11. *Страница ответа, сгенерированная сценарием tutor4.py*

Как обычно, мы можем передавать параметры этому CGI-сценарию в конце URL. На рис. 15.12 изображена страница, полученная при явной передаче имени и возраста пользователя в следующем адресе URL:

```
http://localhost/cgi-bin/tutor4.py?user=Joe+Blow&age=30
```

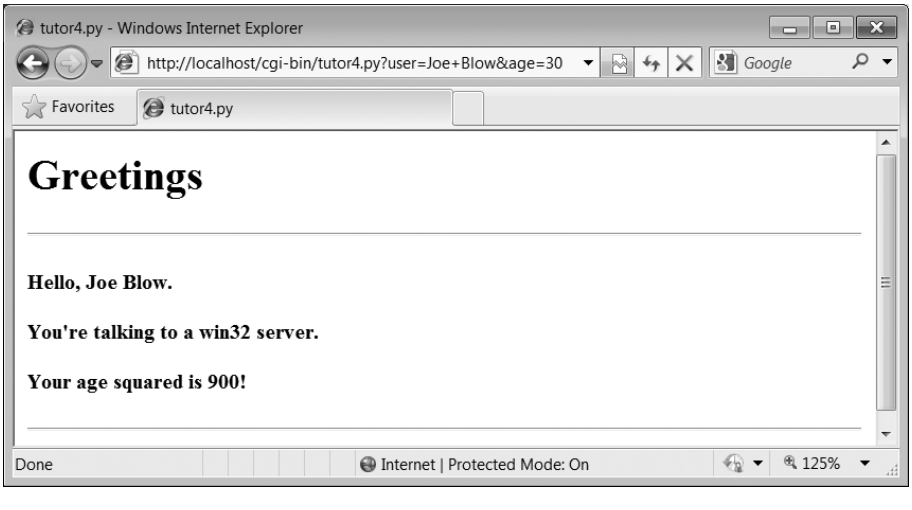


Рис. 15.12. *Страница, сгенерированная при обращении к сценарию tutor4.py с параметрами в URL*

Обратите внимание, что на этот раз за символом ? следуют два параметра; они отделяются друг от друга символом &. Кроме того, обратите внимание, что в значении параметра user мы определили пробел с помощью символа +. Это обычное соглашение по кодированию адресов URL. На стороне сервера символ + автоматически будет замещен пробелом. Кроме того, это одно из стандартных правил экранирования строк URL, которые мы рассмотрим позднее. И хотя пример 15.10 не содержит ничего особенно нового для сценариев CGI, тем не менее, он высвечивает несколько новых приемов программирования, имеющих отношение к отладке и безопасности сценариев CGI. Давайте кратко рассмотрим их.

Преобразование строк в сценариях CGI

Просто для красоты этот сценарий возвращает название платформы сервера, полученное вызовом функции sys.platform, а также квадрат значения поля ввода age. Обратите внимание на необходимость преобразования значения age в целое число с помощью встроенной функции int — в сценарии CGI все входные значения поступают в виде строк. Преобразование в целое число можно также выполнить с помощью встроенной функции eval. Ошибки преобразования (и прочие) элегантно перехватываются инструкцией try, чтобы вывести строку ошибки и не дать сценарию завершиться раньше времени.

Никогда нельзя использовать функцию eval для преобразования строк, полученных из Интернета, таких как значение поля age в данном примере, если только нет абсолютной уверенности, что строка не может содержать злонамеренного программного кода. Например, если бы этот пример был размещен в Интернете, то кто-нибудь мог бы ввести в поле age строку, содержащую вызов команды оболочки (или дописать параметр age с таким значением в конец URL). В зависимости от контекста и при наличии соответствующих прав у процесса при передаче в eval такая строка может уничтожить все файлы в каталоге сценариев сервера или сделать еще что похуже!

Только если сценарии CGI запускаются в процессе с ограниченными правами и с ограниченным доступом к серверу, можно не опасаться запуска в сценарии CGI на выполнение строк, получаемых из Сети. Никогда не передавайте такие строки инструментам динамического программирования, таким как ev al или exec, а также средствам, выполняющим произвольные команды оболочки, например os.popen и os.system, если только вы не уверены в их безвредности. Для числовых преобразований всегда используйте простые инструменты, такие как int и float, которые могут распознавать только числа, а не произвольный программный код на языке Python.

Отладка сценариев CGI

Ошибки встречаются даже в прекрасном новом мире Интернета. Вообще говоря, отладка сценариев CGI может оказаться значительно более

трудной, чем отладка программ, выполняемых локально. Ошибки происходят на удаленном компьютере, да и вообще сценарии не будут выполняться вне контекста, предполагаемого моделью CGI. Сценарий из примера 15.10 демонстрирует следующие два распространенных приема отладки:

Перехват сообщений об ошибках

Данный сценарий перенаправляет стандартный поток вывода `sys.stderr` в `sys.stdout`, чтобы сообщения интерпретатора об ошибках в конечном счете выводились на странице ответа в броузере. Обычно сообщения интерпретатора об ошибках выводятся в поток вывода `stderr`, в результате чего они выводятся в окно консоли на стороне сервера или в файл журнала. Чтобы направить их в броузер, нужно заставить `stderr` ссылаться на тот же самый объект файла, что и `stdout` (который в сценариях CGI соединяется с броузером). Если не сделать такого переназначения, сообщения об ошибках, в том числе сообщения о программных ошибках в сценарии, не появятся в броузере.

Имитация контрольных данных

Определение класса `dummy`, закомментированное в этой окончательной версии, использовалось для отладки сценария, перед тем как он был размещен в Сети. Помимо того, что сообщения, выводимые в поток вывода `stderr`, по умолчанию не видны, сценарии CGI также предполагают наличие окружающего контекста, которого нет при отладке их вне окружения CGI. Например, при запуске из командной строки системы этот сценарий не получает входные данные из формы. Раскомментируйте этот фрагмент для тестирования сценария из командной строки. Класс `dummy` маскируется под объект разобранных полей формы, и `form` получает значения словаря, содержащего два объекта полей формы. В итоге `form` подменяет результат вызова `cgi.FieldStorage`. Как обычно в Python, необходимо придерживаться лишь интерфейсов объектов, а не типов данных.

Ниже приводится несколько общих рекомендаций по отладке CGI-сценариев на сервере:

Запускайте сценарий из командной строки

Возможно, при этом не будет сгенерирована разметка HTML, но при автономном выполнении выявятся синтаксические ошибки в программном коде. Не забывайте, что в командной строке Python можно запускать файлы с исходным программным кодом независимо от их расширений: например, команда `python somescript.cgi` будет выполнена как обычно.

Перенаправляйте `sys.stderr` *в* `sys.stdout` *как можно ближе к началу своего сценария*

В результате текст сообщений об ошибках и дамп стека будут показываться в броузере клиента, а не в окне консоли или файле журнала на сервере. Если не считать утомительный поиск в журналах сервера

или обработку исключительных ситуаций вручную, это может оказаться единственным способом увидеть текст сообщения об ошибке при аварийном завершении сценария.

Имитируйте передачу входных данных для моделирования окружающего контекста CGI

Например, определите классы, имитирующие интерфейс входных данных CGI (как это делает класс `dummy` в данном сценарии), чтобы посмотреть вывод сценария для разных контрольных примеров при запуске его из командной строки.[1] Иногда также бывает полезно установить переменные окружения так, чтобы они имитировали данные формы или URL (далее в этой главе будет показан пример такого приема).

Вызывайте утилиты вывода контекста CGI в броузере

В модуле CGI имеются вспомогательные функции, посылающие в броузер форматированный дамп переменных окружения CGI и входных значений. Например, вызов `cgi.print_form(form)` выведет все параметры, отправленные клиентом, а вызов `cgi.test()` выведет значения переменных окружения, полей ввода формы, имя каталога и многое другое. Иногда этого оказывается достаточно для решения проблем, связанных с соединением или с входными данными. Некоторые из этих функций будут использованы в примере почтовой программы с веб-интерфейсом в следующей главе.

Показывайте перехватываемые исключительные ситуации

При перехвате возбужденной исключительной ситуации сообщение Python об ошибке по умолчанию не выводится в `stderr`. В таких случаях ваш сценарий решает, выводить ли имя исключения и его значение на странице ответа. Подробности, касающиеся исключения, могут быть получены с помощью функции `sys.exc_info()` из встроенного модуля `sys`. Кроме того, для получения трассировки стека вручную и вывода ее на страницу можно воспользоваться модулем `traceback` — трассировочная информация содержит номера строк в исходном программном коде, выполнявшихся в момент появления исключения. Это также будет использовано в реализации страницы с ошибками в примере PyMailCGI (глава 16).

---

[1] Этот прием применим не только к сценариям CGI. В главе 12 мы познакомились с системами, встраивающими программный код на языке Python в разметку HTML, такими как Python Server Pages. К сожалению, нет достаточно хорошего способа протестировать такой код вне контекста объемлющей системы, кроме как извлечь его (возможно, с помощью модуля `htm.parser`, поставляемого в составе стандартной библиотеки Python и описываемого в главе 19) и запустить его, передав ему имитацию API, который будет использован в конечном итоге.

Добавляйте вывод отладочной информации

Вы всегда можете вставить в программный код CGI-сценария вызовы функции `print` для вывода отладочной информации, как в любую другую программу на языке Python. Однако вам следует убедиться, что строка заголовка «Content-type» выводится первой, иначе выводимая информация может не отображаться на странице. В самом тяжелом случае можно также выводить отладочные и трассировочные сообщения в локальный текстовый файл на сервере – при выводе в файл отпадает необходимость форматировать трассировочные сообщения в соответствии с соглашениями о выводе разметки HTML в поток ответа.

Опробуйте сценарий в действии

Конечно, если ваш сценарий хоть как-то работает, лучше всего дать ему работать на сервере и пусть обрабатывает реальные данные от броузера. Запуск локального веб-сервера на вашем компьютере, как показано в данной главе, позволит быстрее вносить изменения и тестировать их.

Добавление стандартных инструментов ввода

До сих пор мы вводили данные в текстовые поля. Но формы HTML поддерживают ряд элементов ввода (которые в традиционных графических интерфейсах мы называли виджетами) для получения данных, вводимых пользователем. Рассмотрим программу CGI, которая демонстрирует сразу все стандартные элементы для ввода. Как обычно, определим файл HTML для страницы формы и CGI-сценарий на языке Python для обработки введенных данных и создания ответной страницы. Файл HTML представлен в примере 15.11.

Пример 15.11. PP4E\Internet\Web\tutor5a.html

```
<HTML><TITLE>CGI 101</TITLE>
<BODY>
<H1>Common input devices</H1>
<HR>
<FORM method=POST action="cgi-bin/tutor5.py">
    <H3>Please complete the following form and click Send</H3>
    <P><TABLE>
        <TR>
            <TH align=right>Name:
            <TD><input type=text name=name>
        <TR>
            <TH align=right>Shoe size:
            <TD><table>
            <td><input type=radio name=shoesize value=small>Small
            <td><input type=radio name=shoesize value=medium>Medium
            <td><input type=radio name=shoesize value=large>Large
            </table>
```

```
        <TR>
            <TH align=right>Occupation:
            <TD><select name=job>
                <option>Developer
                <option>Manager
                <option>Student
                <option>Evangelist
                <option>Other
            </select>
        <TR>
            <TH align=right>Political affiliations:
            <TD><table>
            <td><input type=checkbox name=language value=Python>Pythonista
            <td><input type=checkbox name=language value=Perl>Perlmonger
            <td><input type=checkbox name=language value=Tcl>Tcler
            </table>
        <TR>
            <TH align=right>Comments:
            <TD><textarea name=comment cols=30 rows=2>
                Enter text here</textarea>
        <TR>
            <TD colspan=2 align=center>
            <input type=submit value="Send">
    </TABLE>
</FORM>
<HR>
</BODY></HTML>
```

При отображении в броузере появляется страница, изображенная на рис. 15.13

Как и прежде, эта страница содержит обычное текстовое поле, а также переключатели, окно раскрывающегося списка, группу флажков для выбора нескольких вариантов и область ввода многострочного текста. Для всех них в файле HTML задан параметр `name`, идентифицирующий выбранное ими значение в данных, отправляемых клиентом серверу. Если заполнить эту форму и щелкнуть на кнопке отправки Send (Передать), на сервере будет запущен сценарий из примера 15.12, который обработает все входные данных, введенные с клавиатуры или выбранные на форме.

Пример 15.12. PP4E\Internet\Web\cgi-bin\tutor5.py

```
#!/usr/bin/python
"""
выполняется на стороне сервера, читает данные формы, выводит разметку HTML
"""


import cgi, sys
form = cgi.FieldStorage()         # извлечь данные из формы
print("Content-type: text/html")  # плюс пустая строка
```

```
html = """
<TITLE>tutor5.py</TITLE>
<H1>Greetings</H1>
<HR>
<H4>Your name is %(name)s</H4>
<H4>You wear rather %(shoesize)s shoes</H4>
<H4>Your current job: %(job)s</H4>
<H4>You program in %(language)s</H4>
<H4>You also said:</H4>
<P>%(comment)s</P>
<HR>"""

data = {}
for field in ('name', 'shoesize', 'job', 'language', 'comment'):
    if not field in form:
        data[field] = '(unknown)'
    else:
        if not isinstance(form[field], list):
            data[field] = form[field].value
        else:
            values = [x.value for x in form[field]]
            data[field] = ' and '.join(values)
print(html % data)
```

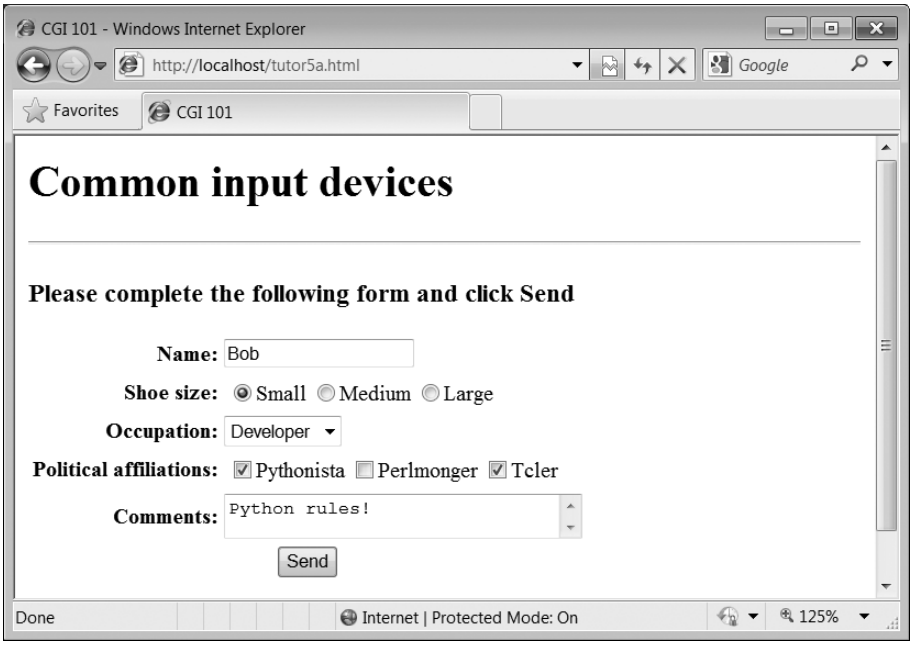


Рис. 15.13. Страница формы ввода, генерируемая файлом tutor5a.html

Этот сценарий Python не очень сложен; в основном он просто копирует данные из полей формы в словарь с именем `data`, чтобы их можно было легко вставить в заключенную в тройные кавычки строку ответа. Стоит пояснить некоторые использованные приемы:

Проверка правильности полей

Как обычно, необходимо проверить все ожидаемые поля ввода и убедиться в их присутствии во входных данных с помощью оператора `in` словарей. Некоторые или все поля ввода могут отсутствовать, если на форме в них не были введены данные или они не были добавлены в строку адреса URL.

Форматирование строк

На этот раз в строке формата мы использовали ссылки на ключи словаря – напомню, что `%(name)s` означает необходимость извлечь значение для ключа `name` из словаря данных и преобразовать его в строку.

Поля с несколькими вариантами выбора

Проверяется также тип значений всех ожидаемых полей, чтобы определить, не получен ли список вместо обычной строки. Значения элементов ввода с несколькими вариантами выбора, таких как поле выбора `language` на этой странице ввода, возвращаются из `cgi.FieldStorage` в виде списка объектов с атрибутами `value`, а не простого одиночного объекта с атрибутом `value`.

Этот сценарий дословно копирует в словарь значения простых полей, а для извлечения значений полей с несколькими вариантами выбора использует генератор списков и с помощью строкового метода `join` конструирует единую строку, вставляя `and` между выбранными значениями (например, `Python and Tcl`). Генератор списков, использованный в сценарии, эквивалентен вызову `map(lambda x: x.value, form[field])`.

Здесь это не показано, однако объекты `FieldStorage` имеют альтернативные методы `getfirst` и `getlist`, которые позволяют интерпретировать поля ввода, как значения из одного или из нескольких элементов независимо от того, чем в действительности они являются (подробности ищите в руководстве по стандартной библиотеке Python). И как будет показано ниже, помимо простых строк и списков может возвращаться еще один, третий тип объектов – для полей ввода, определяющих выгруженные файлы. Для надежности сценарий в действительности должен также экранировать текст, полученный из полей ввода и вставляемый в разметку HTML ответа – на случай, если он содержит операторы языка HTML. Подробнее об экранировании мы поговорим ниже.

Когда эта страница с формой будет заполнена и отправлена на сервер, сценарий создаст ответ, как показано на рис. 15.14, – в сущности, лишь форматированное отражение того, что было передано.

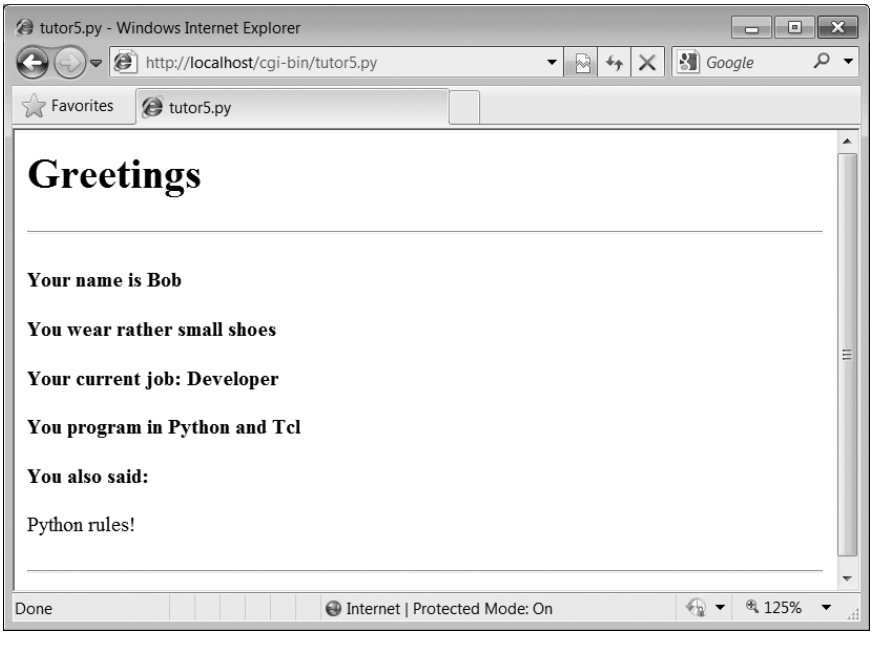


Рис. 15.14. *Страница ответа, созданная сценарием tutor5.py (1)*

Изменение размещения элементов формы ввода

Предположим, что вы написали систему, подобную той, что была представлена в предыдущем разделе, а ваши пользователи, клиенты и «вторая половина» начали жаловаться, что форму ввода трудно читать. Не волнуйтесь. Поскольку модель CGI естественным образом отделяет *интерфейс* пользователя (описание страницы HTML) от *логики обработки* (сценария CGI), можно совершенно безболезненно изменить структуру формы. Достаточно просто модифицировать файл HTML, при этом изменять программный код CGI не понадобится. Так, в примере 15.13 приводится новое определение формы ввода, где таблицы используются несколько иначе, обеспечивая более красивую разметку с границами.

Пример 15.13. PP4E\Internet\Web\tutor5b.html

```
<HTML><TITLE>CGI 101</TITLE>
<BODY>
<H1>Common input devices: alternative layout</H1>
<P>Use the same tutor5.py server side script, but change the
layout of the form itself. Notice the separation of user interface
and processing logic here; the CGI script is independent of the
HTML used to interact with the user/client.</P><HR>
<FORM method=POST action="cgi-bin/tutor5.py">
    <H3>Please complete the following form and click Submit</H3>
    <P><TABLE border cellpadding=3>
```

```
            <TR>
                <TH align=right>Name:
                <TD><input type=text name=name>
            <TR>
                <TH align=right>Shoe size:
                <TD><input type=radio name=shoesize value=small>Small
                <input type=radio name=shoesize value=medium>Medium
                <input type=radio name=shoesize value=large>Large
            <TR>
                <TH align=right>Occupation:
                <TD><select name=job>
                    <option>Developer
                    <option>Manager
                    <option>Student
                    <option>Evangelist
                    <option>Other
                </select>
            <TR>
                <TH align=right>Political affiliations:
                <TD><P><input type=checkbox name=language value=Python>Pythonista
                <P><input type=checkbox name=language value=Perl>Perlmonger
                <P><input type=checkbox name=language value=Tcl>Tcler
            <TR>
                <TH align=right>Comments:
                <TD><textarea name=comment cols=30 rows=2>
                    Enter spam here</textarea>
            <TR>
                <TD colspan=2 align=center>
                <input type=submit value="Submit">
                <input type=reset value="Reset">
        </TABLE>
    </FORM>
    </BODY></HTML>
```

Если открыть в броузере эту альтернативную страницу, мы получим интерфейс, показанный на рис. 15.15.

Теперь, прежде чем вы попытаетесь найти отличия между этим файлом HTML и предыдущим, я должен отметить, что отличия в HTML, порождающем эту страницу, значительно менее важны, чем то обстоятельство, что атрибуты `action` форм этих двух страниц ссылаются на одинаковые адреса URL. Нажатие кнопки Submit (Отправить) в этой версии вызывает запуск того же самого и совершенно не изменившегося CGI-сценария Python *tutor5.cgi* (пример 15.12).

Это означает, что сценарии совершенно не зависят от структуры интерфейса пользователя, с помощью которого им отправляется информация. Изменения страницы ответа требуют, конечно, изменения сценария, но код HTML разметки страницы с формой ввода можно изменять по своему вкусу, и это не оказывает влияния на программный код Python, выполняемый на сервере. На рис. 15.16 показана страница ответа, порождаемая сценарием на этот раз.

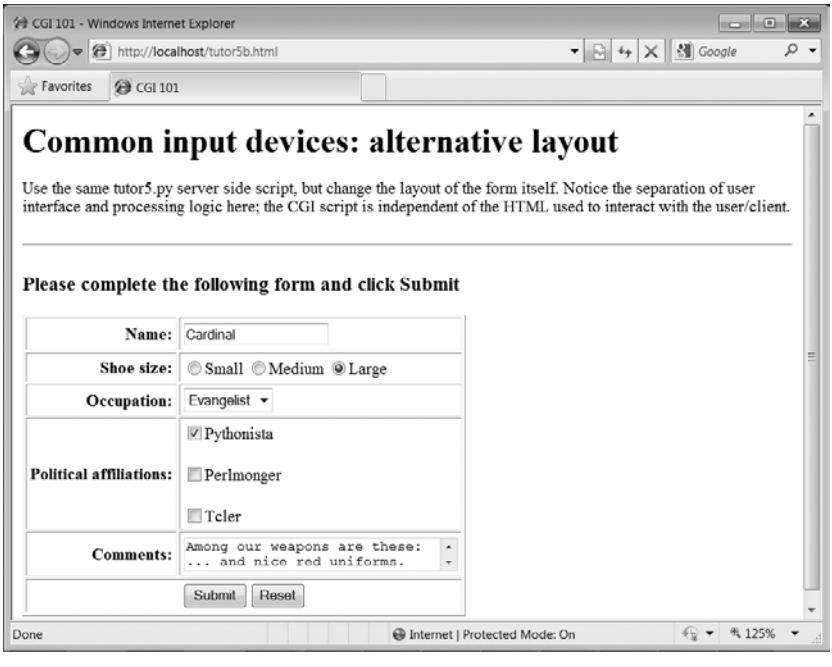


Рис. 15.15. Страница формы, созданная файлом tutor5b.html

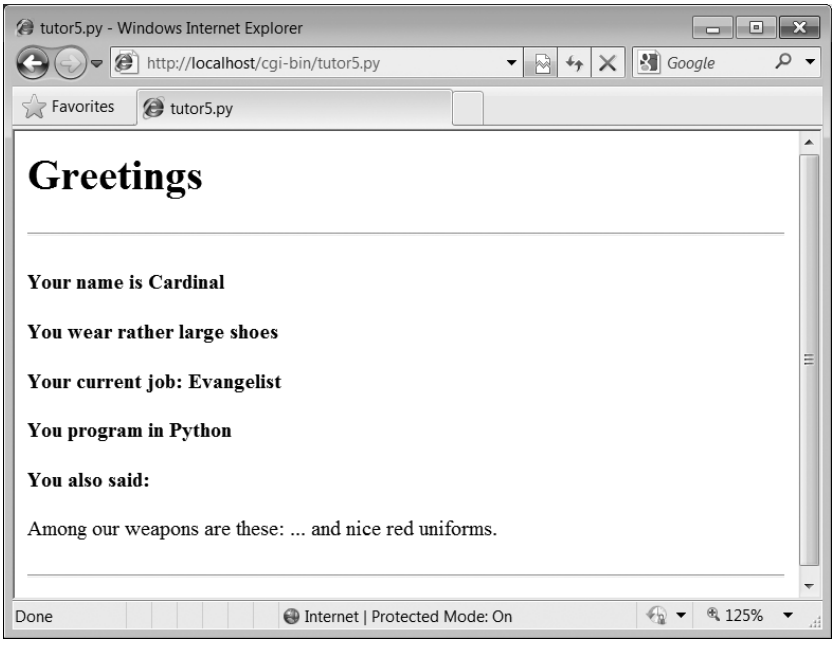


Рис. 15.16. Страница ответа, созданная сценарием tutor5.cgi (2)

Отделение отображения от логики обработки

В действительности, этот пример иллюстрирует очень важный принцип, используемый при создании крупных веб-сайтов: если приложить все силы, чтобы отделить разметку HTML от программного кода сценариев, в результате получается весьма удобное разделение на отображение и логику обработки – каждая часть может разрабатываться независимо, людьми с разной специализацией. Веб-дизайнеры, например, могут работать над интерфейсом, а программисты – над логикой выполнения.

Несмотря на небольшую величину примера, представленного в этом разделе, он уже выигрывает от такого отделения страницы с формой ввода. В некоторых случаях добиться подобного отделения бывает сложнее, потому что наши примеры сценариев содержат внутри разметку HTML страниц ответов. Однако стоит приложить совсем немного усилий, и мы сможем вынести разметку HTML страницы ответа в отдельный файл, которую также можно будет разрабатывать отдельно от логики сценария. Например, строку `html`, присутствующую в сценарии *tutor5.py* (пример 15.12), можно было бы сохранить в текстовом файле и загружать его во время выполнения сценария.

В крупных системах отделение логики от отображения может достигаться еще проще, с помощью таких инструментов, как механизмы шаблонов HTML, действующих на стороне сервера. Система Python Server Pages и фреймворки, такие как Zope и Django, например, способствуют отделению логики и отображения, предоставляя языки описания страниц ответов, которые позволяют включать фрагменты, сгенерированные логикой программ на языке Python. В некотором смысле языки шаблонов, применяемые на стороне сервера, встраивают программный код на языке Python в разметку HTML – в противоположность CGI-сценариям, которые встраивают разметку HTML в программный код Python, – и способны обеспечить более четкое разделение труда, позволяя оформлять программный код Python в виде отдельных компонентов. Подробности по этой теме ищите в Интернете. Аналогичные приемы можно использовать для отделения отображения и логики работы в графических интерфейсах, которые мы изучали ранее в этой книге, но для достижения поставленных целей они также требуют использования крупных фреймворков или моделей.

Передача параметров в жестко определенных адресах URL

Ранее мы передавали параметры CGI-сценариям, указывая их в конце адреса URL, вводимого в адресную строку броузера – в строке с параметрами запроса, следующей после символа ?. Но адресная строка броузера не накладывает каких-то особенных ограничений. В частности, тот же синтаксис URL можно использовать в гиперссылках, внутри веб-страниц.

Например, веб-страница, представленная в примере 15.14, определяет три гиперссылки (текст между тегами <A> и </A>), каждая из которых снова запускает наш первоначальный сценарий *tutor5.py*, но с тремя предопределенными наборами параметров.

Пример 15.14. PP4E\Internet\Web\tutor5c.html

```
<HTML><TITLE>CGI 101</TITLE>
<BODY>
<H1>Стандартные инструменты ввода: параметры в строке URL</H1>

<P>Эта страница вызывает серверный сценарий tutor5.py, подставляя
предопределенные данные в конец адреса URL сценария внутри простой
гиперссылки (вместо использования формы ввода). Выберите в своем броузере
пункт меню View Source (Исходный код страницы или Просмотр HTML-кода), чтобы
увидеть определение гиперссылок для каждого из элементов списка ниже.

<P>Этот пример в большей степени иллюстрирует особенности CGI, чем
языка Python. Обратите внимание, что модуль cgi из стандартной
библиотеки Python обрабатывает и этот способ ввода (который можно
сымитировать, определив метод GET в атрибуте action формы), и ввод методом
POST форм; для CGI-сценария на языке Python они выглядят одинаково. Иными
словами, пользователи модуля cgi не зависят от метода отправки данных.

<P>Следует также отметить, что адреса URL с добавленными в конец входными
значениями, как в данном примере, можно сгенерировать в процессе вывода
страницы другим CGI-сценарием, чтобы направить следующий щелчок
пользователя в нужное место; вместе со скрытыми полями ввода типа 'hidden'
они предоставляют один из способов сохранения состояния между щелчками.
</P><HR>

<UL>
<LI><A href="cgi-bin/tutor5.py?name=Bob&shoesize=small">Send Bob, small</A>
<LI><A href="cgi-bin/tutor5.py?name=Tom&language=Python">Send Tom, Python</A>

<LI><A href="http://localhost/cgi-bin/tutor5.py?job=Evangelist&comment=spam">
Send Evangelist, spam</A>
</UL>

<HR></BODY></HTML>
```

Этот статический файл HTML определяет три гиперссылки – первые две укороченные, а третья определена полностью, но все они действуют аналогичным образом (опять-таки, целевому сценарию это безразлично). При переходе по адресу URL этого файла отображается страница, представленная на рис. 15.17. В общем это просто страница для запуска готовых вызовов сценария CGI. (Я уменьшил размер шрифта в окне броузера, чтобы снимок экрана уместился на странице книги: выполните этот пример у себя, если не сможете разобрать детали на рисунке.)

При щелчке на второй ссылке в этой странице будет создана страница ответа, изображенная на рис. 15.18. Эта ссылка вызывает сценарий CGI

с параметром `name`, равным «Tom», и параметром `language`, равным «Python», поскольку эти параметры и их значения жестко определены в адресе URL, указанном в разметке HTML для второй гиперссылки. Гиперссылки с параметрами, подобные этим, иногда называют *ссылками, хранящими информацию о состоянии*, – они автоматически задают следующую операцию, которая будет выполнена сценарием. Эффект нажатия этой ссылки в точности такой, как если бы мы вручную ввели адрес URL в адресную строку броузера, как показано на рис. 15.18.

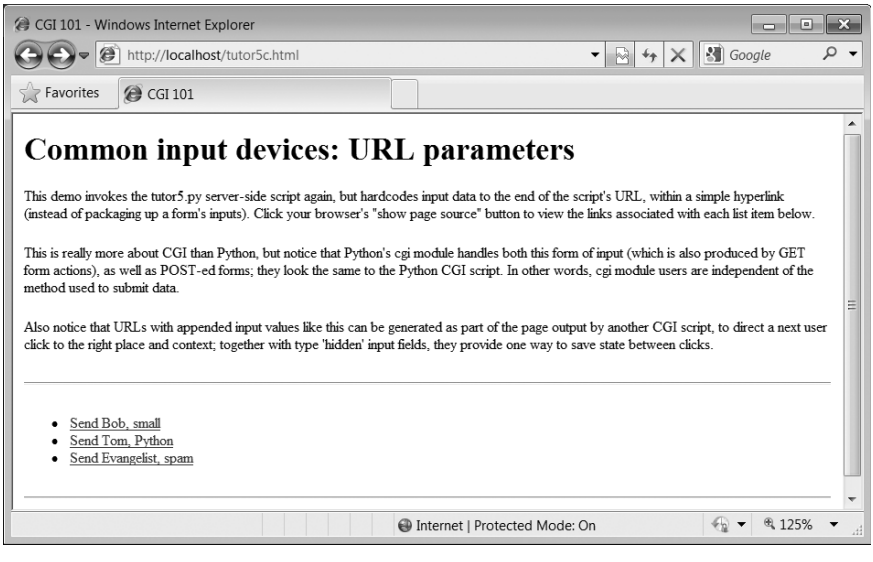


Рис. 15.17. *Страница с гиперссылками, созданная файлом tutor5c.html*

Обратите внимание, что здесь отсутствуют многие поля. Сценарий *tutor5.py* обнаружит отсутствующие поля, выдав для них сообщение `unknown` на странице ответа. Следует также подчеркнуть, что здесь мы опять повторно используем CGI-сценарий на языке Python. Сам сценарий совершенно не зависит от формата интерфейса пользователя в странице с формой и способа, которым он вызван, – при передаче формы или при переходе по жестко определенному адресу URL с параметрами запроса. В результате отделения интерфейса пользователя от логики обработки сценарии CGI становятся многократно используемыми программными компонентами – по крайней мере, в контексте окружения CGI.

Параметры запроса в адресах URL, встроенных в пример 15.14, были жестко определены в разметке HTML страницы. Но подобные адреса URL могут также генерироваться CGI-сценарием автоматически, как часть страницы ответа, – для передачи сценарию вводимых данных, что является следующим уровнем во взаимодействии с пользователем. Для веб-приложений они являются простейшим способом «хранить»

какие-то значения на протяжении сеанса. Тем же целям служат скрытые поля форм, которые рассматриваются далее.

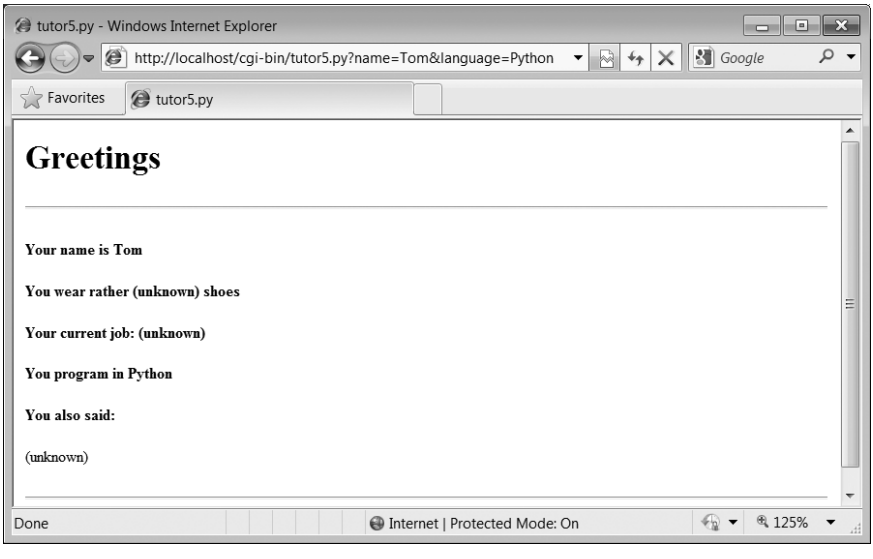


Рис. 15.18. Страница ответа, созданная сценарием tutor5.py (3)

Передача параметров в скрытых полях форм

Подобным же образом, в духе предыдущего раздела, входные данные для сценария можно жестко определять в разметке HTML страницы в виде скрытых полей ввода. Такие поля ввода не отображаются на странице, но передаются серверу вместе с формой. Так, страница в примере 15.15 позволяет вводить значение поля `job`, но параметры `name` и `language` заполняются автоматически, за счет использования скрытых полей ввода.

Пример 15.15. PP4E\Internet\Web\tutor5d.html

```
<HTML><TITLE>CGI 101</TITLE>
<BODY>
<H1> Стандартные инструменты ввода: скрытые поля ввода</H1>

<P>Эта страница также вызывает серверный сценарий tutor5.py, но она
определяет входные данные уже непосредственно в форме, в виде скрытых полей
ввода, а не в виде параметров в конце адреса URL в гиперссылке.
Как и прежде, определение этой формы, включая определение скрытых полей,
может быть сгенерировано при выводе страницы другим CGI-сценарием, чтобы
обеспечить передачу данных следующему сценарию; скрытые поля ввода форм
являются еще одним способом сохранения информации между вызовами страниц.
</P><HR><p>
```

```
<form method=post action="cgi-bin/tutor5.py">
    <input type=hidden name=name     value=Sue>
    <input type=hidden name=language value=Python>
    <input type=text   name=job      value="Enter job">
    <input type=submit value="Submit Sue">
</form>
</p><HR></BODY></HTML>
```

Если страницу, представленную в примере 15.15, открыть в броузере, мы получим страницу ввода, изображенную на рис. 15.19.

Отправка этой формы инициирует вызов оригинального сценария *tutor5.py*, как и прежде (пример 15.12), но при этом некоторые входные данные передаются в скрытых полях ввода. Сценарий возвращает страницу ответа, изображенную на рис. 15.20.

Как и в случае передачи данных в виде параметров в строке запроса, показанном в предыдущем разделе, здесь входные данные также жестко определены и встроены непосредственно в разметку HTML формы ввода. В отличие от параметров в строке запроса, значения скрытых полей ввода не отображаются в адресе следующей страницы. Подобно параметрам запроса определения таких полей ввода также можно генерировать динамически, при выводе ответа CGI-сценарием. Когда скрытые поля ввода служат целям передачи входных данных и получения следующей страницы, они являются своего рода средством сохранения информации между вызовами сценариев. Чтобы понять, где и когда их следует использовать, нам необходимо коротко познакомиться с другими альтернативными способами сохранения информации.

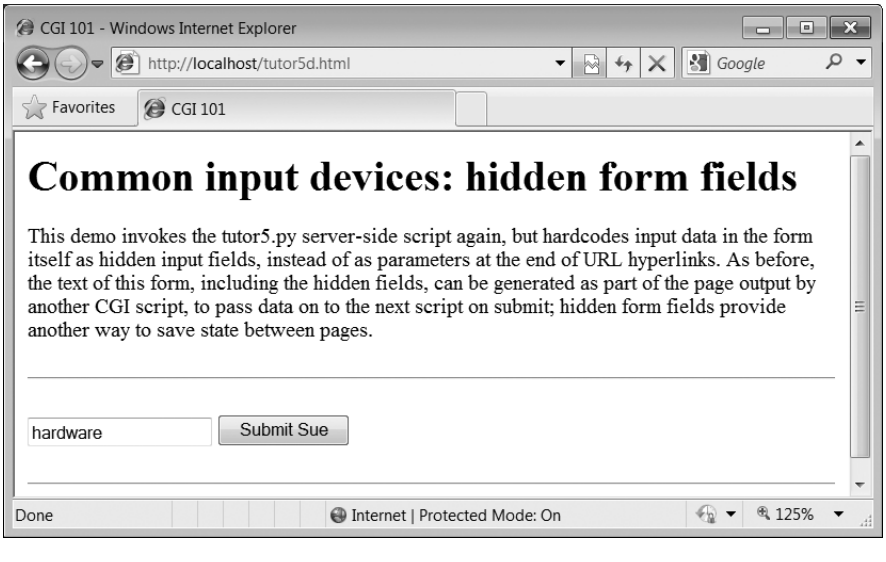


Рис. 15.19. *Страница с формой ввода tutor5d.html*

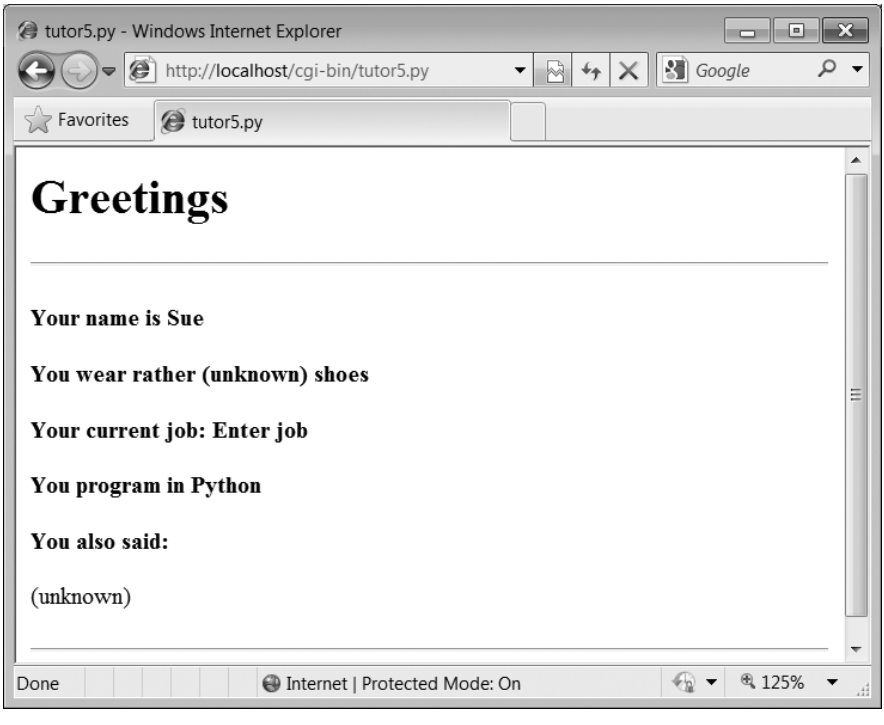


Рис. 15.20. *Страница ответа, созданная сценарием tutor5.py (4)*

Сохранение информации о состоянии в сценариях CGI

Один из самых необычных аспектов базовой модели CGI и одно из самых существенных отличий от модели программирования графических интерфейсов, которые мы изучали в предыдущей части книги, заключается в том, что сценарии CGI не имеют информации о состоянии – каждый из них является самостоятельной программой, обычно выполняемой автономно и не имеющей никакого представления о любых других сценариях, которые могли выполняться ранее или будут выполняться после. В модели CGI отсутствует такое понятие, как глобальные переменные или объекты, живущие дольше, чем выполняется единственный шаг взаимодействия, и сохраняющие контекст выполнения. Каждый сценарий начинает работу на пустом месте, не имея никакой информации о том, чем завершился предыдущий этап.

Тем самым обеспечивается простота и надежность веб-серверов – сценарий CGI, содержащий ошибку, не оказывает влияния на процесс сервера. Фактически ошибка в сценарии CGI затрагивает лишь единствен-

ную страницу, которую он реализует, и не влияет на все веб-приложение в целом. Это совсем иная модель программирования, чем модель функций-обработчиков обратного вызова, выполняющихся в едином процессе графического интерфейса; поэтому требуется приложить дополнительные усилия, чтобы обеспечить сохранение информации дольше, чем требуется на выполнение единственного сценария.

До сих пор отсутствие информации о состоянии никак не сказывалось на наших простых примерах, но крупные системы обычно предусматривают многоступенчатые взаимодействия с пользователем, реализуемые множеством сценариев, и им требуется иметь некоторый способ сохранения информации, собранной в ходе таких взаимодействий. Как было показано в двух предыдущих разделах, генерирование параметров запроса в адресах URL-ссылок и скрытых полей форм в страницах, отправляемых в ответ, являются для сценариев CGI двумя простейшими способами передачи данных следующему сценарию в приложении. В результате щелчка на ссылке или на кнопке отправки формы в виде таких параметров другому серверному сценарию отправляется запрограммированный выбор или информация о сеансе. В некотором смысле содержимое сгенерированной страницы ответа становится частью области памяти приложения.

Например, сайт, позволяющий читать вашу почту, может представлять список доступных для просмотра почтовых сообщений, реализованный в виде разметки HTML, как список гиперссылок, сгенерированный другим сценарием. Каждая гиперссылка может включать имя сценария просмотра сообщения, а также ряд параметров, идентифицирующих номер выбранного сообщения, имя почтового сервера и так далее – столько данных, сколько необходимо для получения сообщения, связанного с конкретной ссылкой. Сайт интернет-магазина, в свою очередь, мог бы возвращать сгенерированный список ссылок на товары, каждый элемент которого представляет собой жестко определенную ссылку с номером товара, его ценой и прочей информацией. Как вариант, страница оформления покупки на сайте интернет-магазина могла бы содержать перечень товаров, выбранных на предыдущей странице, в виде скрытых полей формы.

Одной из основных причин, по которым мы представляли вам приемы из двух предыдущих разделов, состоит в том, что мы собираемся широко использовать их в реализации крупного примера в следующей главе. Например, адреса URL с информацией о состоянии в виде параметров строки запроса мы будем использовать для реализации динамических списков, которые «знают», что предпринимать в случае щелчка мышью. Скрытые поля форм также будут использоваться для передачи информации аутентификации пользователя следующему сценарию. С более общей точки зрения оба приема являются способами сохранения информации между обращениями к страницам – они могут использоваться для прямого управления действиями следующего вызываемого сценария.

Приемы на основе адресов URL с параметрами и скрытых полей форм отлично подходят для передачи информации между страницами в течение единственного сеанса работы. Однако некоторые ситуации требуют большего. Например, что если нам потребуется сохранять имя пользователя между сеансами? Или следить за тем, какие страницы пользователь посещал в прошлом? Поскольку такая информация должна храниться дольше, чем длится один сеанс взаимодействия, параметров запроса и скрытых полей формы будет недостаточно. Кроме того, в некоторых случаях объем необходимой информации может оказаться слишком большим, чтобы встраивать его в страницу HTML ответа.

В целом существует множество различных способов передачи или сохранения информации между вызовами сценариев CGI и между сеансами взаимодействий:

Параметры запроса в строке URL

Информация сеанса встраивается в страницы ответа

Скрытые поля форм

Информация сеанса встраивается в страницы ответа

Cookies

Небольшие блоки данных, хранящиеся на стороне клиента, которые могут служить для сохранения информации между сеансами

Базы данных на стороне сервера

Могут сохранять большие объемы информации между сеансами

Расширения модели CGI

Постоянные процессы, механизмы управления сеансами и так далее

Большинство из них мы рассмотрим в последующих примерах, но, поскольку сохранение информации является одной из основных идей разработки серверных сценариев, коротко рассмотрим все эти приемы по порядку.

Параметры запроса в строке URL

С этим приемом мы познакомились выше в этой главе: жестко определенные параметры URL в динамически сгенерированных гиперссылках встраиваются в страницы ввода, возвращаемые в качестве ответа. Включение имени сценария и его входных параметров в такие ссылки позволяет непосредственно управлять действиями следующей страницы при их выборе. Параметры передаются от клиента на сервер автоматически, как часть запроса в стиле GET.

Передача параметров запроса является простой задачей – достаточно из CGI-сценария вывести в стандартный поток вывода корректно сформированный адрес URL в составе страницы ответа (соблюдая при этом некоторые соглашения по экранированию, с которыми мы познакомимся далее в этой главе). Ниже приводится фрагмент из реализации

клиента электронной почты с веб-интерфейсом, пример которого демонстрируется в следующей главе:

```
script = "onViewListLink.py"
user = 'bob'
mnum = 66
pswd = 'xxx'
site = ' pop.myisp.net'
print('<a href="%s?user=%s&pswd=%s&mnum=%d&site=%s">View %s</a>'
              % (script, user, pswd, mnum, site, mnum))
```

Получающийся в результате адрес URL содержит достаточно информации для управления следующим сценарием, который будет вызван щелчком по гиперссылке:

```
<a href="onViewListLink.py?user=bob&pswd=xxx&mnum=66&site=pop.myisp.
net">View 66</a>
```

Параметры запроса играют роль переменных в памяти и позволяют передавать информацию между страницами. Их удобно использовать для передачи информации между страницами в пределах одного и того же сеанса работы. Поскольку каждый сгенерированный адрес URL может включать различные параметры, такая схема позволяет обеспечить контекст в зависимости от действий пользователя. Каждая ссылка в списке доступных для выбора альтернатив, например, может подразумевать выполнение различных операций, определяемых различными значениями параметров. Кроме того, пользователи могут помещать ссылки с параметрами в закладки, чтобы вернуться к определенному этапу в последовательности взаимодействий.

Но поскольку сохраненная таким способом информация теряется при закрытии страницы, этот прием бесполезен, когда необходимо сохранять данные между сеансами. Кроме того, данные, добавленные в конец адреса URL, видны пользователю и могут сохраняться в файле журнала на сервере; в некоторых ситуациях может потребоваться зашифровать их вручную, чтобы исключить возможность подсмотреть их или подделать.

Скрытые поля форм

С этим приемом мы также познакомились в предыдущем разделе: скрытые поля форм включаются в данные формы, встроенные в веб-страницы ответа, но они не отображаются ни в веб-страницах, ни в адресах URL. Когда выполняется отправка формы, все скрытые поля передаются следующему сценарию вместе с фактически введенными данными, обеспечивая контекст выполнения. Таким способом обеспечивается контекст обработки всей формы ввода, а не какой-то отдельной гиперссылки. Уже введенное имя пользователя, пароль или результат выбора, например, могут передаваться в скрытых полях формы через целую последовательность динамически генерируемых страниц.

С программной точки зрения скрытые поля генерируются серверным сценарием как часть HTML-разметки страницы ответа и позднее возвращаются клиентом вместе с другими данными, введенными в форму. Приведем еще один пример из следующей главы:

```
print('<form method=post action="%s/onViewSubmit.py">' % urlroot)
print('<input type=hidden name=mnum value="%s">' % msgnum)
print('<input type=hidden name=user value="%s">' % user)
print('<input type=hidden name=site value="%s">' % site)
print('<input type=hidden name=pswd value="%s">' % pswd)
```

Подобно параметрам запроса скрытые поля форм могут также служить аналогами переменных в памяти, сохраняющими информацию о состоянии от страницы к странице. Кроме того, как и при использовании параметров запроса, такие своеобразные переменные встраиваются непосредственно в страницу, поэтому скрытые поля форм удобно использовать для сохранения информации между обращениями к страницам в пределах отдельного сеанса работы, но они непригодны для сохранения данных между сеансами.

И подобно параметрам запроса и cookies (рассматриваются далее) скрытые поля форм доступны пользователям – хотя они и не отображаются на страницах и в адресах URL, тем не менее их значения можно увидеть, отобразив исходную HTML-разметку страницы. Таким образом, скрытые поля форм не обеспечивают безопасность данных – в некоторых ситуациях также может потребоваться зашифровать встроенные данные, чтобы исключить возможность подсмотреть их на стороне клиента или подделать при отправке формы.

HTTP «Cookies»

Cookies, расширение лежащего в основе модели Веб протокола HTTP, предоставляют возможность северным веб-приложениям напрямую управлять сохранением информации на компьютере клиента. Поскольку эта информация не встраивается в разметку HTML веб-страниц, она сохраняется дольше, чем длится сеанс работы. Благодаря этому прием на основе cookies идеально подходит для сохранения информации в перерывах между сеансами.

Такие параметры, как имена пользователей или личные предпочтения, например, являются первыми кандидатами на включение в блоки данных cookies – они будут доступны при следующем посещении клиентом нашего сайта. Однако из-за того, что блоки данных cookies имеют ограниченный размер, иногда расцениваются как вторжение и могут быть отключены пользователем на стороне клиента, они не всегда могут справиться с задачей хранения данных. Часто их лучше использовать для сохранения между сеансами небольших фрагментов не самой важной информации, и веб-сайты, рассчитанные на широкое использование, должны сохранять работоспособность и в случае недоступности cookies.

С функциональной точки зрения HTTP cookies являются строками информации, хранящимися на компьютере клиента и передаваемыми между клиентом и сервером в заголовках сообщений HTTP. В процессе вывода страницы ответа серверные сценарии генерируют заголовки HTTP, предлагающие клиенту сохранить cookies. Позднее веб-броузер клиента генерирует заголовки HTTP, в которых обратно отправляются все cookies, соответствующие серверу и запрашиваемой странице. В результате информация, хранящаяся в cookies, встраивается в потоки данных подобно параметрам запроса и полям форм, только содержится она не в разметке HTML страницы, а в заголовках HTTP. Кроме того, данные в cookies могут постоянно храниться на стороне клиента и использоваться для сохранения информации между сеансами и обращениями к страницам.

Для разработчиков веб-приложений в состав стандартной библиотеки Python включены инструменты, упрощающие передачу и прием: модуль `http.cookiejar` выполняет обработку HTTP cookies на стороне клиента, взаимодействующего с веб-серверами, а модуль `http.cookies` упрощает задачи создания и приема cookies серверными сценариями. Кроме того, модуль `urllib.request`, с которым мы познакомились ранее, поддерживает операцию открытия адресов URL с автоматической обработкой cookies.

Создание cookies

Веб-броузеры, такие как Firefox и Internet Explorer, в целом обеспечивают корректную работу с этим протоколом на стороне клиента, сохраняя и отправляя данные в cookies. В этой главе нас в основном интересует обработка cookies на стороне сервера. Cookies создаются в результате передачи специальных заголовков HTTP в начале потока ответа:

```
Content-type: text/html
Set-Cookie: foo=bar;

<HTML>...
```

Полный формат заголовка «Set-Cookie» имеет следующий вид:

```
Set-Cookie: name=value; expires=date; path=pathname; domain=domainname;
secure
```

По умолчанию в параметре `domain` передается сетевое имя сервера, отправившего cookies, а в параметре `path` – путь к документу или сценарию, который устанавливает cookies. Эти параметры позднее используются клиентом, чтобы определить, когда отправлять cookies обратно на сервер. Создание cookies на языке Python реализуется очень просто – следующие инструкции в сценарии CGI сохраняют cookies с временем последнего посещения сайта:

```
import http.cookies, time
cook = http.cookies.SimpleCookie()
```

```
cook['visited'] = str(time.time()) # словарь
print(cook.output())           # выведет "Set-Cookie: visited=1276623053.89"
print('Content-type: text/html\n')
```

Вызов функции SimpleCookie здесь создает объект, напоминающий словарь, ключами которого являются строки (имена cookies), а значениями – объекты Morsel (описывающие значения cookies). Объекты Morsel в свою очередь также являются объектами, похожими на словари, содержащими по одному ключу для каждого свойства cookies: path и domain; expires, определяющее дату истечения срока хранения cookies (по умолчанию cookies считаются действительными только на протяжении сеанса работы броузера); и так далее. Кроме того, объекты Morsel имеют дополнительные атрибуты – например, key и value, определяющие имя и значение cookies соответственно. Операция присваивания ключу объекта cookies автоматически создает объект Morsel из строки, а метод output объектов cookies возвращает строку, которую можно использовать в качестве заголовка HTTP; тот же эффект дает непосредственная попытка вывода объекта с помощью функции print благодаря переопределению метода __str__ перегрузки оператора. Ниже приводится более полный пример использования этого интерфейса:

```
>>> import http.cookies, time
>>> cooks = http.cookies.SimpleCookie()
>>> cooks['visited'] = time.asctime()
>>> cooks['username'] = 'Bob'
>>> cooks['username']['path'] = '/myscript'

>>> cooks['visited'].value
'Tue Jun 15 13:35:20 2010'
>>> print(cooks['visited'])
Set-Cookie: visited="Tue Jun 15 13:35:20 2010"
>>> print(cooks)
Set-Cookie: username=Bob; Path=/myscript
Set-Cookie: visited="Tue Jun 15 13:35:20 2010"
```

Получение cookies

Если после этого клиент вновь посетит страницу, данные из cookies будут отправлены броузером на сервер в виде заголовка HTTP, имеющем вид «Cookie: name1=value1; name2=value2...». Например:

```
Cookie: visited=1276623053.89
```

Грубо говоря, броузер клиента возвращает все cookies, соответствующие доменному имени сервера и пути к сценарию. На стороне севера cookies будут доступны CGI-сценарию через переменную окружения HTTP_COOKIE, содержащую строки заголовков с данными cookies, выгруженными клиентом. Извлечь их в языке Python можно следующим образом:

```
import os, http.cookies
cooks = http.cookies.SimpleCookie(os.environ.get("HTTP_COOKIE"))
vcook = cooks.get("visited") # словарь типа Morsel
```

```
if vcook != None:
time = vcook.value
```

В данном случае конструктор SimpleCookie автоматически выполняет разбор строки с данными в cookies и преобразует ее в объект словаря типа Morsel – как обычно, метод get словаря по умолчанию возвращает None, если ключ отсутствует в словаре, а для получения строки со значением cookies используется атрибут value объекта Morsel.

Использование cookies в сценариях CGI

Чтобы помочь совместить все детали, описанные выше, в примере 15.16 приводится CGI-сценарий, который сохраняет cookies на стороне клиента при первом посещении и принимает и отображает его содержимое при последующих посещениях.

Пример 15.16. PP4E\Internet\Web\cgi-bin\cookies.py

```
"""
создает или использует имя пользователя, сохраненное в cookies на стороне
клиента; в этом примере отсутствуют данные, получаемые из формы ввода
"""


import http.cookies, os
cookstr = os.environ.get("HTTP_COOKIE")
cookies = http.cookies.SimpleCookie(cookstr)
usercook = cookies.get("user")               # извлечь, если был отправлен

if usercook == None:                         # создать при первом посещении
    cookies = http.cookies.SimpleCookie()    # вывести заголовок Set-cookie
    cookies['user'] = 'Brian'
    print(cookies)
    greeting = '<p>His name shall be... %s</p>' % cookies['user']
else:
    greeting = '<p>Welcome back, %s</p>' % usercook.value

print('Content-type: text/html\n')           # плюс пустая строка
print(greeting)                              # и фактическая разметка html
```

Если предположить, что вы используете локальный веб-сервер, представленный в примере 15.1, этот сценарий можно вызвать по адресу URL *http://localhost/cgi-bin/cookies.py* (введите его в адресной строке своего броузера или откройте его в интерактивном сеансе с помощью модуля urllib.request). При первом обращении к этому сценарию он установит cookies в заголовках ответа, и вы увидите в окне броузера такое сообщение:

```
His name shall be... Set-Cookie: user=Brian
```

Затем при повторном обращении по этому же адресу URL сценария в том же самом сеансе работы с броузером (воспользуйтесь кнопкой Обновить (Reload)) страница ответа будет содержать следующее сообщение:

```
Welcome back, Brian
```

Это объясняется тем, что при повторном обращении к сценарию клиент отправляет сохраненные ранее значения cookies обратно сценарию, по крайней мере, пока вы не закроете броузер и не перезапустите его вновь – по умолчанию срок действия cookies ограничивается продолжительностью сеанса работы с броузером. В реальных обстоятельствах такая схема может использоваться страницей входа в веб-приложение – пользователю достаточно будет пройти процедуру входа только один раз в течение одного сеанса работы с броузером.

Обработка cookies с помощью модуля urllib.request

Как уже упоминалось выше, модуль `urllib.request` предоставляет интерфейс для чтения страниц ответа по заданному адресу URL и использует модуль `http.cookiejar` для поддержки сохранения cookies на стороне клиента и отправки их на сервер. Однако эта поддержка не обеспечивается по умолчанию. Например, ниже выполняется тестирование сценария установки cookies из предыдущего раздела – при повторном обращении к сценарию cookies не возвращаются обратно на сервер:

```
>>> from urllib.request import urlopen
>>> reply = urlopen('http://localhost/cgi-bin/cookies.py').read()
>>> print(reply)
b'<p>His name shall be... Set-Cookie: user=Brian</p>\n'

>>> reply = urlopen('http://localhost/cgi-bin/cookies.py').read()
>>> print(reply)
b'<p>His name shall be... Set-Cookie: user=Brian</p>\n'
```

Для корректной поддержки cookies с помощью этого модуля нам нужно просто включить в работу класс обработки cookies – то же относится и к другим необязательным расширениям, реализованным в этом модуле. Попробуем вновь обратиться к сценарию из предыдущего раздела:

```
>>> import urllib.request as urllib
>>> opener = urllib.build_opener(urllib.HTTPCookieProcessor())
>>> urllib.install_opener(opener)
>>>
>>> reply = urllib.urlopen('http://localhost/cgi-bin/cookies.py').read()
>>> print(reply)
b'<p>His name shall be... Set-Cookie: user=Brian</p>\n'

>>> reply = urllib.urlopen('http://localhost/cgi-bin/cookies.py').read()
>>> print(reply)
b'<p>Welcome back, Brian</p>\n'

>>> reply = urllib.urlopen('http://localhost/cgi-bin/cookies.py').read()
>>> print(reply)
b'<p>Welcome back, Brian</p>\n'
```

Теперь все работает, потому что модуль `urllib.request` имитирует поведение веб-броузера в отношении cookies – он сохраняет cookies, когда эта операция запрашивается в заголовках страницы ответа, генерируемой сценарием, и добавляет их в заголовки при повторных обращениях к этому же сценарию. Кроме того, так же, как и броузер, этот модуль удалит cookies при завершении и повторном запуске интерактивного сеанса Python и повторном выполнении этого программного кода. Информацию об интерфейсах этого модуля смотрите в руководстве по стандартной библиотеке.

Несмотря на удобства в использовании, cookies имеют свои недостатки. Во-первых, они имеют ограниченный размер (типичные ограничения: 4 Кбайта на один блок данных cookies, не более 300 cookies всего и не более 20 cookies для каждого доменного имени). Во-вторых, в большинстве броузеров имеется возможность отключить cookies, что делает их непригодными для хранения критически важных данных. Некоторые даже рассматривают их, как вторжение в систему, потому что их можно использовать для слежения за действиями пользователя. (Многие сайты просто требуют, чтобы поддержка cookies была включена, устраняя тем самым эту проблему.) Наконец, поскольку cookies передаются между клиентом и сервером через сеть, они защищены в той же мере, в какой защищен сам поток данных – может быть далеко небезопасно передавать секретные данные, если для доступа к странице не используется защищенный протокол HTTP. Защищенные cookies и серверные концепции мы исследуем в следующей главе.

За дополнительной информацией о модулях, реализующих поддержку cookies, и о протоколах работы с cookies в целом обращайтесь к руководству по стандартной библиотеке Python и к ресурсам в Сети. Вполне возможно, что в будущем похожие механизмы хранения данных будут предоставлять реализации HTML.

Базы данных на стороне сервера

Сценарии на языке Python способны поддерживать полноценные базы данных на стороне сервера, обеспечивающие более надежное хранение данных. Более подробно мы будем рассматривать эту возможность в главе 17. Сценарии на языке Python могут использовать самые разные механизмы хранения данных на стороне сервера, включая плоские файлы, хранилища объектов, объектно-ориентированные базы данных, такие как ZODB, и реляционные базы данных SQL, такие как MySQL, PostgreSQL, Oracle и SQLite. Помимо хранения данных такие системы могут предоставлять дополнительные инструменты, такие как транзакции и отмена операций, синхронизация параллельных изменений и многие другие.

Полноценные базы данных являются самым мощным решением проблемы хранения данных. Они могут использоваться для сохранения состояния между обращениями к страницам в пределах одного сеанса

(помечая данные ключами, генерируемыми для каждого сеанса) и между сеансами (помечая данные ключами, генерируемыми для каждого пользователя).

Для каждого конкретного пользователя, например, CGI-сценарии могут извлекать из базы данных на сервере все данные об этом пользователе, собранные в прошлом. Серверные базы данных идеально подходят для сохранения более сложной информации о сеансах – приложение покупательской корзины, например, может сохранять элементы, добавленные в прошлом в базе данных на сервере.

Базы данных сохраняют данные между обращениями к страницам и сеансами. Поскольку данные сохраняются явно, нет никакой необходимости встраивать их в параметры запроса или в скрытые поля форм страниц ответа. Поскольку данные хранятся на сервере, нет никакой необходимости сохранять их на стороне клиента в виде cookies. И поскольку в таких схемах хранения используются многоцелевые базы данных, на них не действуют ограничения на объем данных и другие количественные показатели, характерные для cookies.

В обмен на дополнительные возможности, полноценные базы данных требуют приложения дополнительных усилий по установке, обслуживанию и использованию в программах. К счастью, как будет показано в главе 17, реализация доступа к базам данных из программ на языке Python отличается удивительной простотой. Кроме того, интерфейсы к базам данных в языке Python могут использоваться в любых приложениях, как с веб-интерфейсом, так и в других.

Расширения модели CGI

Наконец, существует множество дополнительных протоколов и фреймворков, обеспечивающих сохранение информации о состоянии на стороне сервера, которые мы не будем рассматривать в этой книге. Например, фреймворк Zope для построения веб-приложений, кратко обсуждавшийся в главе 12, предоставляет программный интерфейс, позволяющий конструировать объекты для использования в веб-приложениях, которые сохраняются автоматически.

Другие схемы, такие как FastCGI, а также расширения для определенных серверов, такие как `mod_python` для Apache, могут обеспечивать некоторые способы обойти одноразовую природу CGI-сценариев или как-то иначе расширять базовую модель CGI средствами долговременного хранения данных. Например:

- *FastCGI* позволяет веб-приложениям выполняться как постоянно действующим процессам, которые принимают входные данные и отправляют ответы веб-серверу HTTP посредством механизмов взаимодействий между процессами (Inter-Process Communication, IPC), таких как сокеты. Этим данная модель отличается от обычной модели CGI, которая предусматривает обмен данными через переменные

окружения, стандартные потоки ввода-вывода и аргументы командной строки и предусматривает запуск сценариев для завершения каждого запроса. Поскольку процесс FastCGI может существовать дольше, чем одна страница, он способен сохранять информацию между обращениями к страницам и избежать потерь производительности на запуск сценариев.

- `mod_python` расширяет открытый веб-сервер Apache, встраивая интерпретатор Python внутрь Apache. Программный код на языке Python выполняется непосредственно внутри процесса сервера Apache, устраняя необходимость запускать внешние процессы. Этот пакет также поддерживает понятие сеансов, которое может использоваться для сохранения данных между обращениями к страницам. Параллельные операции с сеансовыми данными выполняются под защитой блокировок, а сами данные могут храниться в файлах или в памяти, в зависимости от того, в каком режиме выполняется Apache – как несколько процессов или как один процесс с несколькими потоками выполнения. Расширение `mod_python` также включает инструменты веб-разработки, такие как серверный механизм Python Server Pages (PSP) шаблонов HTML, упоминавшийся в главе 12 и в этой главе.

Однако такие модели не поддерживаются повсеместно и могут нести в себе дополнительные сложности, например синхронизация доступа к хранимым данным с помощью блокировок. Кроме того, ошибки в веб-приложениях, использующих модель FastCGI, оказывают влияние на все приложение, а не только на единственную страницу, а такие проблемы, как утечки памяти, обходятся намного дороже. Дополнительную информацию о хранении данных в модели CGI и о поддержке в Python таких моделей, как FastCGI, ищите в Веб или в специализированных ресурсах.

Комбинирование приемов

Естественно, эти приемы могут комбинироваться для реализации различных стратегий хранения данных в рамках сеансов или на более продолжительное время. Например:

- Веб-приложение может использовать cookies для сохранения пользовательских или сеансовых ключей на стороне клиента и позднее использовать их для получения полной информации о пользователе или сеансе из базы данных на стороне сервера.
- Даже информация, живущая в течение одного сеанса, – параметры запросов в адресах URL или скрытые поля форм – может использоваться для передачи от страницы к странице ключей, идентифицирующих сеанс, что позволяет использовать их в следующем сценарии для обращения к базе данных на стороне сервера.
- Кроме того, для сохранения временной информации, объединяющей страницы, между обращениями к страницам могут использоваться

параметры запроса в адресе URL и скрытые поля, а для сохранения информации, объединяющей сеансы, – cookies и базы данных.

Выбор того или иного приема сохранения данных во многом зависит от потребностей приложения. Несмотря на то, что описанные механизмы не так просты в обращении, как переменные и объекты в памяти общего процесса графического интерфейса, выполняющегося на компьютере клиента, при творческом подходе проблема сохранения данных в сценариях CGI полностью разрешима.

Переключатель «Hello World»

Вернемся к программному коду. Настало время заняться чем-нибудь более полезным, чем примеры, что мы видели до сих пор (или хотя бы более увлекательным). В этом разделе представлена программа, демонстрирующая базовый синтаксис различных языков программирования, необходимый для вывода строки «Hello World», классической начальной программы.

Для простоты предполагается, что вывод строки будет производиться в стандартный поток вывода, а не в графический интерфейс или веб-страницу. Кроме того, будет показана лишь сама команда вывода, а не полная программа. Версия на языке Python оказывается при этом законченной программой, но мы здесь не ставим себе цели противопоставлять ее реализациям на конкурирующих языках.

Структурно этот пример в первом приближении состоит из файла HTML главной страницы и CGI-сценария на языке Python, запускаемого из формы на главной странице HTML. Поскольку здесь не требуется сохранять какую-либо информацию между переходами от страницы к странице, данный пример все еще достаточно прост. На самом деле главная страница HTML, представленная в примере 15.17, является просто одним большим раскрывающимся списком выбора, находящимся внутри формы.

Пример 15.17. PP4E\Internet\Web\languages.html

```
<html><title>Languages</title>
<body>
<h1>Hello World selector</h1>

<P>Эта демонстрационная страница показывает, как выглядит вывод сообщения
"hello world" на различных языках программирования. Для простоты этот пример
показывает лишь команду вывода (чтобы создать законченную программу,
на некоторых языках программирования требуется написать дополнительный
программный код) и только простейшее решение, обеспечивающее вывод текста
в консоль (без графического интерфейса или логики конструирования HTML).
Эта страница представляет собой простой файл HTML; следующая страница,
которая появится после щелчка на кнопке ниже, генерируется CGI-сценарием
на языке Python, выполняющимся на сервере. Дополнительные указания:
```

```
<UL>
<LI>Чтобы увидеть разметку HTML страницы, выберите пункт меню 'View Source'
    ('Исходный код страницы' или 'Просмотр HTML-кода') в броузере.
<LI>Чтобы увидеть исходный программный код Python сценария CGI,
    выполняющегося на сервере,
    <A HREF="cgi-bin/languages-src.py">щелкните здесь</A> или
    <A HREF="cgi-bin/getfile.py?filename=cgi-bin\languages.py">здесь</A>.
<LI>Чтобы увидеть альтернативную версию этой страницы,
    сгенерированную динамически,
    <A HREF="cgi-bin/languages2.py">щелкните здесь</A>.
</UL></P>
<hr>
<form method=POST action="cgi-bin/languages.py">
    <P><B>Select a programming language:</B>
    <P><select name=language>
        <option>All
        <option>Python
        <option>Python2
        <option>Perl
        <option>Tcl
        <option>Scheme
        <option>SmallTalk
        <option>Java
        <option>C
        <option>C++
        <option>Basic
        <option>Fortran
        <option>Pascal
        <option>Other
    </select>
    <P><input type=Submit>
</form>
</body></html>
```

Не будем пока обращать внимание на некоторые гиперссылки в середине этого файла – они ведут к более фундаментальным понятиям, таким как передача файлов и удобство сопровождения, которыми мы займемся в следующих двух разделах. При открытии этого файла HTML в броузере клиента выводится страница, представленная на рис. 15.21.

Виджет над кнопкой Submit Query (Отправить запрос) является раскрывающимся списком, позволяющим выбрать одно из значений тега `<option>` в файле HTML. Как обычно, выбрав название одного из языков и нажав кнопку Submit Query (Отправить запрос) внизу (или клавишу Enter), мы отправим название выбранного языка экземпляру программы сценария CGI на сервере, имя которой указано в параметре `action` формы. В примере 15.18 представлен сценарий Python, выполняемый на сервере в результате передачи.

Рис. 15.21. *Главная страница «Hello World»*

Пример 15.18. PP4E\Internet\Web\cgi-bin\languages.py

```
#!/usr/bin/python
"""
демонстрирует синтаксис вывода сообщения 'hello world' на выбранном языке
программирования; обратите внимание, что в сценарии используются "сырые"
строки вида r'...', чтобы исключить интерпретацию последовательностей
символов '\n' в таблице, и к строкам применяется функция cgi.escape(), чтобы
такие строки, как '<<', корректно интерпретировались броузером - они будут
преобразованы в допустимый код разметки HTML; сценарию может быть передано
название любого языка программирования, так как в броузере можно явно ввести
полную строку URL вида "http://servername/cgi-bin/languages.py?language=
Cobol" или передать ее из сценария (с помощью urllib.request.urlopen).
предупреждение: список языков отображается в обеих версиях страницы,
CGI и HTML, - его можно было бы импортировать из общего файла,
если список выбора также генерируется сценарием CGI;
"""

debugme  = False      # True=параметр в виде аргумента командной строки
inputkey = 'language' # имя входного параметра

hellos = {
    'Python':    r" print('Hello World')               ",
    'Python2':   r" print 'Hello World'                ",
    'Perl':      r' print "Hello World\n";             ',
    'Tcl':       r' puts "Hello World"                 ',
    'Scheme':    r' (display "Hello World") (newline)  ',
    'SmallTalk': r" 'Hello World' print.               ",
```

```
    'Java':      r' System.out.println("Hello World"); ',
    'C':         r' printf("Hello World\n");           ',
    'C++':       r' cout << "Hello World" << endl;     ',
    'Basic':     r' 10 PRINT "Hello World"             ',
    'Fortran':   r" print *, 'Hello World'             ",
    'Pascal':    r" WriteLn('Hello World');            "
}

class dummy:                                # имитация входного объекта
    def __init__(self, str): self.value = str
import cgi, sys
if debugme:
    form = {inputkey: dummy(sys.argv[1])}  # имя в командной строке
else:
    form = cgi.FieldStorage()               # разбор действительного ввода

print('Content-type: text/html\n')          # добавить пустую строку
print('<TITLE>Languages</TITLE>')
print('<H1>Syntax</H1><HR>')

def showHello(form):                        # разметка HTML для одного языка
    choice = form[inputkey].value
    print('<H3>%s</H3><P><PRE>' % choice)
    try:
        print(cgi.escape(hellos[choice]))
    except KeyError:
        print("Sorry--I don't know that language")
    print('</PRE></P><BR>')

if not inputkey in form or form[inputkey].value == 'All':
    for lang in hellos.keys():
        mock = {inputkey: dummy(lang)}
        showHello(mock)
else:
    showHello(form)
print('<HR>')
```

Как обычно, этот сценарий выводит разметку HTML в стандартный поток вывода, чтобы создать страницу ответа в броузере клиента. В этом сценарии немного нового, о чем стоило бы рассказать, но в нем использованы некоторые приемы, заслуживающие особого внимания:

Необрабатываемые «сырые» строки и кавычки

Обратите внимание на использование *необрабатываемых строк* (raw strings, константы строк, которым предшествует символ «r») в определении словаря языков. Напомню, что необрабатываемые строки сохраняют в строке символы обратного слэша \ как литералы, не интерпретируя их как начало экранированных последовательностей. Без этого последовательность символов перевода строки \n в фрагментах программного кода на некоторых языках воспринималась

бы интерпретатором Python как перевод строки и не выводилась бы в разметке HTML ответа как \n. Здесь также используются двойные кавычки для строк, включающих неэкранированные символы апострофов, – обычное правило оформления литералов строк в языке Python.

Экранирование текста, встраиваемого в HTML и URL

Этот сценарий следит за форматированием текста фрагментов программного кода на каждом языке с помощью вспомогательной функции cgi.escape. Эта стандартная утилита Python автоматически преобразует специальные символы HTML в экранированные последовательности HTML, чтобы броузеры не интерпретировали их как операторы HTML. Технически функция cgi.escape преобразует символы в экранированные последовательности в соответствии со стандартными соглашениями HTML: <, > и & превращаются в <, > и &. Если передать ей значение True во втором аргументе, то и символ двойной кавычки (") будет преобразован в ".

Например, оператор << сдвига влево в программном коде на языке C++ транслируется в последовательность << – пару экранированных последовательностей языка разметки HTML. Поскольку каждый выводимый фрагмент кода в итоге встраивается в поток вывода разметки HTML ответа, необходимо преобразовывать все содержащиеся в нем специальные символы HTML. Инструменты анализа HTML (в том числе стандартный модуль Python html.parser, который будет рассматриваться в главе 19) транслируют экранированные последовательности обратно в исходные символы при отображении страницы.

В более широком смысле, поскольку модель CGI основывается на концепции передачи форматированных строк через Сеть, операция экранирования специальных символов встречается повсеместно. Сценариям CGI почти всегда для надежности приходится экранировать текст, генерируемый для включения в ответ. Например, если возвращается произвольный текст, введенный пользователем или полученный из источника данных на сервере, обычно нельзя гарантировать, что в нем не будет специальных символов HTML, поэтому на всякий случай нужно его экранировать.

В последующих примерах мы также увидим, что символы, вставляемые в строки адресов URL, генерируемых нашими сценариями, тоже могут потребовать экранирования. Например, литеральный символ & в адресе URL является специальным и должен быть экранирован, если появляется в тексте, вставляемом в URL. Однако синтаксис URL резервирует другие специальные символы в сравнении с синтаксисом HTML, а потому должны использоваться другие инструменты и соглашения по экранированию. Как мы увидим далее в этой главе, функция cgi.escape реализует экранирование символов в разметке HTML, а urllib.parse.quote (и родственные ей) экранируют символы в строках URL.

Имитация данных, вводимых через форму

Здесь снова ввод данных из формы имитируется (моделируется) как для целей отладки, так и для ответа на запрос всех языков в списке. Если глобальная переменная debugme в сценарии получит значение True, к примеру, то сценарий создаст словарь, взаимозаменяемый с результатом вызова cgi.FieldStorage – его ключ «языка» будет ссылаться на экземпляр подставного класса dummy. Этот класс, в свою очередь, создаст объект с таким же интерфейсом, как в содержимом результата вызова функции cgi.FieldStorage, – он сконструирует объект, атрибут value которого будет установлен равным переданной строке.

В итоге мы получаем возможность тестировать этот сценарий, запуская его из командной строки: сгенерированный словарь заставит сценарий считать, что он был вызван броузером через Сеть. Аналогично, если названием запрашиваемого языка является «All», сценарий обойдет все записи в таблице языков, создавая из них искусственный словарь (как если бы пользователь поочередно запросил все языки).

Это позволяет повторно использовать имеющуюся логику функции showHello и отобразить программный код на всех языках на одной странице. Как всегда в Python, мы программируем интерфейсы объектов и протоколы, а не конкретные типы данных. Функция showHello успешно обработает любой объект, откликающийся на обращение form['language'].value.[1] Теперь снова вернемся к взаимодействию с этой программой. Обратите внимание, что того же результата можно было бы добиться, использовав в функции showHello аргумент со значением по умолчанию, правда при этом пришлось бы предусматривать обработку специального случая.

Теперь снова вернемся к взаимодействию с этой программой. Если выбрать конкретный язык, наш сценарий CGI сгенерирует в ответ разметку HTML, как показано ниже (а также необходимые заголовок «Content-type» и пустую строку). Воспользуйтесь пунктом меню View Source (Исходный код страницы или Просмотр HTML-кода) в своем броузере, чтобы убедиться в этом:

```
<TITLE>Languages</TITLE>
<H1>Syntax</H1><HR>
<H3>Scheme</H3><P><PRE>
```

[1] Наиболее внимательные читатели могли заметить, что мы уже второй раз использовали прием моделирования в этой главе (смотрите пример *tutor4.py* выше). Если вы считаете этот прием в целом полезным, может иметь смысл поместить класс dummy и функцию заполнения словаря формы по требованию в отдельный модуль, который можно будет многократно использовать. Мы и сделаем это в следующем разделе. Даже для таких классов-двустрочников ввод одного и того же программного кода третий раз подряд вполне может убедить в мощи повторного использования кода.

```
 (display "Hello World") (newline)
</PRE></P><BR>
<HR>
```

Программный код заключен в тег <PRE>, который определяет предварительно отформатированный текст (броузер не станет форматировать его как абзац обычного текста). Этот ответ показывает, что мы получаем при выборе языка программирования Scheme. На рис. 15.22 показана страница, отправленная сценарием после выбора в раскрывающемся списке языка «Python» (который в этом издании и на ближайшее будущее, конечно же, соответствует версии Python 3.X).

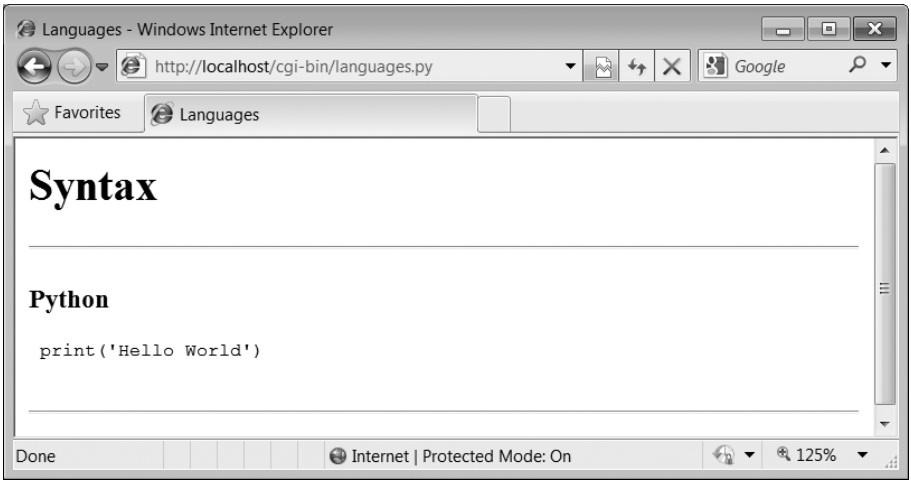


Рис. 15.22. *Страница ответа, созданная сценарием languages.py*

Наш сценарий принимает также значение «All» в качестве имени языка и интерпретирует его как запрос на вывод синтаксических конструкций для всех известных ему языков. Например, ниже приводится разметка HTML, которая генерируется, если установить значение True в глобальной переменной debugme и запустить сценарий из командной строки с одним аргументом All. Это та же разметка, которую получит броузер клиента в ответ на выбор пункта «All»[1]:

```
C:\...\PP4E\Internet\Web\cgi-bin> python languages.py All
Content-type: text/html
```

[1] Интересно, что мы также получим ответ «All», если установить значение False в переменной debugme при запуске сценария из командной строки. При вызове не из окружения CGI конструктор cgi.FieldStorage вернет пустой словарь, не возбуждая исключительной ситуации, поэтому требуется выполнить проверку отсутствия ключа. Однако вряд ли следует полагаться на такое поведение как на гарантированное.

```
<TITLE>Languages</TITLE>
<H1>Syntax</H1><HR>
<H3>C</H3><P><PRE>
 printf("Hello World\n");
</PRE></P><BR>
<H3>Java</H3><P><PRE>
 System.out.println("Hello World");
</PRE></P><BR>
<H3>C++</H3><P><PRE>
 cout &lt;&lt; "Hello World" &lt;&lt; endl;
</PRE></P><BR>
<H3>Perl</H3><P><PRE>
 print "Hello World\n";
</PRE></P><BR>
<H3>Fortran</H3><P><PRE>
 print *, 'Hello World'
</PRE></P><BR>
<H3>Basic</H3><P><PRE>
 10 PRINT "Hello World"
</PRE></P><BR>
<H3>Scheme</H3><P><PRE>
 (display "Hello World") (newline)
</PRE></P><BR>
<H3>SmallTalk</H3><P><PRE>
 'Hello World' print.
</PRE></P><BR>
<H3>Python</H3><P><PRE>
 print('Hello World')
</PRE></P><BR>
<H3>Pascal</H3><P><PRE>
 WriteLn('Hello World');
</PRE></P><BR>
<H3>Tcl</H3><P><PRE>
 puts "Hello World"
</PRE></P><BR>
<H3>Python2</H3><P><PRE>
 print 'Hello World'
</PRE></P><BR>
<HR>
```

Все языки представлены здесь по одной и той же схеме – функция `showHello` вызывается для каждой записи в таблице вместе с моделируемым объектом формы. Обратите внимание на то, как экранируется программный код на языке C++ для встраивания в поток HTML – это результат вызова функции `cgi.escape`. При отображении страницы веб-броузер преобразует экранированные последовательности `<` в символы `<`. На рис. 15.23 показано, как выглядит в броузере страница ответа на выбор пункта «All» – порядок следования языков программирования псевдослучайный, потому что для их хранения используется словарь, а не последовательность.

Рис. 15.23. *Страница, возвращаемая в ответ на выбор пункта «All»*

Проверка отсутствующих или недопустимых данных

До сих пор мы запускали этот сценарий CGI, выбирая название языка из раскрывающегося списка на главной странице HTML. В данном контексте можно быть вполне уверенным, что сценарий получит допустимые входные данные. Однако обратите внимание, что ничто не мешает пользователю передать сценарию CGI название языка в конце адреса URL, в виде явно заданного параметра, не используя форму страницы HTML. Например, следующий адрес URL, введенный в адресной строке броузера или переданный с помощью модуля `urllib.request`:

```
http://localhost/cgi-bin/languages.py?language=Python
```

дает ту же страницу ответа «Python», которая изображена на рис. 15.22. Однако, поскольку пользователь всегда может обойти файл HTML и использовать явный адрес URL, существует возможность вызвать сценарий с названием неизвестного ему языка программирования, которого нет в раскрывающемся списке HTML (и, соответственно, нет в таблице нашего сценария). На самом деле сценарий можно запустить вообще без параметра `language` (или без значения в параметре), если явно ввести адрес URL в адресную строку броузера или отправить его из другого сценария с помощью `urllib.request`, как было показано ранее в этой главе. Корректные запросы будут обработаны нормально:

```
>>> from urllib.request import urlopen
>>> request = 'http://localhost/cgi-bin/languages.py?language=Python'
>>> reply = urlopen(request).read()
>>> print(reply.decode())
<TITLE>Languages</TITLE>
<H1>Syntax</H1><HR>
<H3>Python</H3><P><PRE>
 print('Hello World')
</PRE></P><BR>
<HR>
```

При этом для надежности сценарий явно проверяет оба случая, ведущие к ошибке, как должны в целом делать все сценарии CGI. Например, ниже приводится разметка HTML, сгенерированная в ответ на запрос фиктивного языка GuiDO (и снова вы можете увидеть код разметки, выбрав в броузере пункт меню View Source (Исходный код страницы или Просмотр HTML-кода) после того, как введете адрес URL вручную):

```
>>> request = 'http://localhost/cgi-bin/languages.py?language=GuiDO'
>>> reply = urlopen(request).read()
>>> print(reply.decode())
<TITLE>Languages</TITLE>
<H1>Syntax</H1><HR>
<H3>GuiDO</H3><P><PRE>
Sorry--I don't know that language
</PRE></P><BR>
<HR>
```

Если сценарий не получает во входных данных названия языка, по умолчанию обрабатывается вариант «All» (то же самое относится к случаю, когда сценарий получает в URL параметр `?language=` без названия языка):

```
>>> reply = urlopen('http://localhost/cgi-bin/languages.py').read()
>>> print(reply.decode())
<TITLE>Languages</TITLE>
<H1>Syntax</H1><HR>
<H3>C</H3><P><PRE>
 printf("Hello World\n");
</PRE></P><BR>
<H3>Java</H3><P><PRE>
 System.out.println("Hello World");
</PRE></P><BR>
<H3>C++</H3><P><PRE>
 cout &lt;&lt; "Hello World" &lt;&lt; endl;
</PRE></P><BR>
```

...часть строк опущена...

Если не отслеживать такие случаи, то весьма возможно, что сценарий завершился бы в результате исключения Python, а пользователь получил бы по большей части бесполезную и незавершенную страницу или страницу по умолчанию с сообщением об ошибке (здесь мы не связали

потоки вывода stderr и stdout, поэтому сообщение Python об ошибке не выводилось бы). Наглядно это показано на рис. 15.24, где представлена страница, генерируемая при вызове такого явно заданного URL:

```
http://localhost/cgi-bin/languages.py?language=COBOL
```

Для проверки такого варианта ошибки в раскрывающийся список включено название «Other», при выборе которого создается аналогичная ответная страница ошибки. Добавление в таблицу сценария программного кода программы «Hello World» на языке COBOL (и на других языках, которые вы могли бы вспомнить из опыта своей работы) оставляется читателю в качестве упражнения.

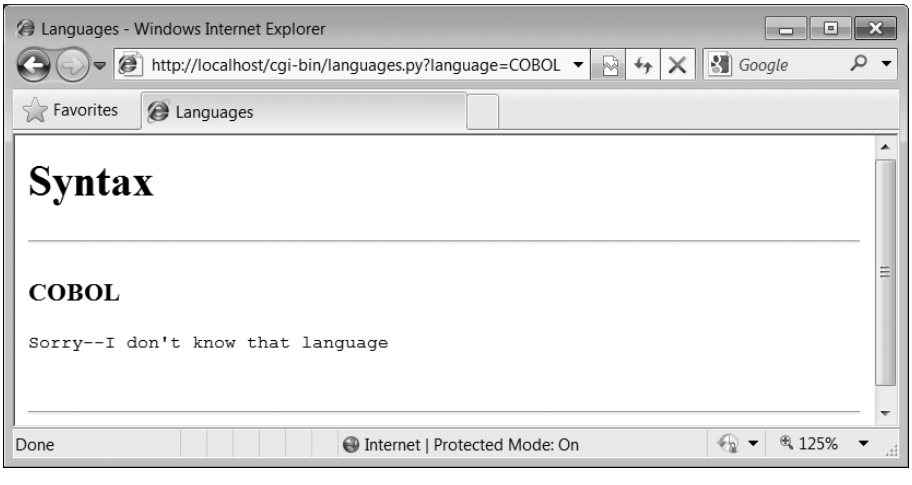


Рис. 15.24. *Страница ответа для неизвестного языка*

Дополнительные примеры запуска сценария *languages.py* вы найдете в конце главы 13. Там мы использовали его для тестирования вызова сценария из клиентских сценариев, использующих низкоуровневый интерфейс к протоколу HTTP и модуль urllib, но теперь у вас должно быть еще и представление о том, как вызываются сценарии на сервере.

Рефакторинг программного кода с целью облегчения его сопровождения

Отойдем на некоторое время от деталей написания программного кода, чтобы взглянуть на него с точки зрения архитектуры. Как мы видели, программный код Python в целом автоматически образует системы, которые легко читать и сопровождать; он имеет простой синтаксис, в значительной мере освобождающий от нагромождений, создаваемых другими инструментами. С другой стороны, стиль программирования

и архитектура программы часто могут не меньше, чем синтаксис, влиять на простоту сопровождения. Например, страницы переключателя «Hello World», представленного выше в этой главе, выполняют требуемую функцию, и «слепить» их удалось быстро и легко. Однако в существующем виде этот переключатель языков имеет существенные недостатки в отношении простоты сопровождения.

Представим, например, что вы действительно возьметесь за задачу, которую я предложил в конце предыдущего раздела, и попробуете добавить еще одну запись для языка COBOL. Если добавить язык COBOL в таблицу сценария CGI, это будет лишь половина решения: список поддерживаемых языков избыточно размещен в двух местах – в разметке HTML для главной страницы и в словаре синтаксиса сценария. При изменении одного из них другой не меняется. Фактически так и произошло, когда я добавлял пункт «Python2» в версию для этого издания (я забыл изменить файл HTML). В более широком случае эта программа не прошла бы придирчивого контроля качества по следующим параметрам:

Список выбора

Как только что было отмечено, список языков, которые поддерживает эта программа, находится в двух местах – в файле HTML и в таблице сценария CGI, а избыточность всегда усложняет сопровождение.

Имя поля

Имя поля входного параметра, `language`, жестко определено в обоих файлах. Если вы измените его в одном файле, то можете не вспомнить, что нужно изменить его также и в другом.

Имитация формы

В этой главе мы уже второй раз написали класс для имитации ввода из полей формы, поэтому класс «dummy» явно нужно сделать повторно используемым механизмом.

Код разметки HTML

Встроенная в сценарий и генерируемая им разметка HTML разбросана по всей программе в вызовах функции `print`, из-за чего трудно осуществлять значительные изменения структуры веб-страницы или передать разработку веб-страницы непрограммистам.

Конечно, это небольшой пример, но проблемы избыточности и повторного использования по мере роста размеров сценариев встают более остро. Следует придерживаться практического правила, что если для изменения какой-то одной особенности поведения приходится редактировать несколько исходных файлов или при написании программ используется прием копирования фрагментов уже существующего программного кода, то надо подумать о рационализации структуры программы. Чтобы проиллюстрировать стиль и приемы программирования, упрощающие сопровождение, перепишем наш пример (или выполним *рефакторинг*) и исправим сразу все отмеченные слабости.

Шаг 1: совместное использование объектов разными страницами – новая форма ввода

Первые две из перечисленных выше проблем, осложняющие сопровождение, можно снять с помощью простого преобразования. Решение заключается в динамической генерации главной страницы сценарием вместо использования готового файла HTML. Сценарий может импортировать имя поля ввода и значения для списка выбора из общего файла модуля на языке Python, совместно используемого сценариями генерации главной страницы и страницы ответа. При изменении списка выбора или имени поля в общем модуле автоматически изменяются оба клиента. Сначала переместим совместно используемые объекты в файл общего модуля, как показано в примере 15.19.

Пример 15.19. PP4E\Internet\Web\cgi-bin\languages2common.py

```
"""
общие объекты, совместно используемые сценариями главной страницы
и страницы ответа; при добавлении нового языка программирования
достаточно будет изменить только этот файл.
"""


inputkey = 'language'                   # имя входного параметра

hellos = {
    'Python':    r" print('Hello World')               ",
    'Python2':   r" print 'Hello World'                ",
    'Perl':      r' print "Hello World\n";             ',
    'Tcl':       r' puts "Hello World"                 ',
    'Scheme':    r' (display "Hello World") (newline)  ',
    'SmallTalk': r" 'Hello World' print.               ",
    'Java':      r' System.out.println("Hello World"); ',
    'C':         r' printf("Hello World\n");           ',
    'C++':       r' cout << "Hello World" << endl;     ',
    'Basic':     r' 10 PRINT "Hello World"             ',
    'Fortran':   r" print *, 'Hello World'             ",
    'Pascal':    r" WriteLn('Hello World');            "
}
```

В модуле `languages2common` содержатся все данные, которые должны быть согласованы на страницах: имя поля ввода и словарь синтаксиса. Словарь синтаксиса `hellos` содержит не совсем тот код разметки HTML, который нужен, но список его ключей можно использовать для динамического создания разметки HTML списка выбора на главной странице.

Обратите внимание, что этот модуль находится в том же каталоге *cgi-bin*, что и сценарии CGI, которые его используют. Это упрощает процедуру импортирования – модуль будет найден в текущем рабочем каталоге сценария без дополнительной настройки пути поиска модулей. В целом внешние ссылки в сценариях CGI разрешаются следующим образом:

- *Импортирование* модулей выполняется относительно текущего рабочего каталога сценария CGI (*cgi-bin*) с учетом всех настроек пути поиска, выполненных при запуске сценария.
- При использовании *минимальных* адресов URL поиск страниц и сценариев, указанных в ссылках и атрибутах `action` форм HTML, выполняется как обычно относительно местоположения предшествующей страницы. Для сценариев CGI такие минимальные адреса URL откладываются относительно местоположения самого сценария.
- Поиск *файлов*, имена которых указываются в параметрах запросов и передаются сценариям, обычно выполняется относительно каталога, содержащего сценарий CGI (*cgi-bin*). Однако на некоторых платформах и серверах пути могут откладываться относительно каталога веб-сервера. В нашем локальном веб-сервере действует последний вариант.

Чтобы убедиться в том, что это действительно так, посмотрите и запустите сценарий CGI, находящийся в пакете примеров и доступный по адресу URL *http://localhost/cgi-bin/test-context.py*: при выполнении в Windows под управлением нашего локального веб-сервера он способен импортировать модули, находящиеся в его собственном каталоге, но поиск файлов выполняется относительно родительского каталога, откуда запущен веб-сервер (вновь создаваемые файлы появляются здесь). Ниже приводится программный код этого сценария на случай, если вам будет интересно узнать, как отображаются пути в вашем веб-сервере и на вашей платформе. Такая зависимость интерпретации относительных путей к файлам от типа веб-сервера, возможно, и не способствует переносимости, но это всего лишь одна из множества особенностей, отличающихся для разных серверов:

```
import languages2common                  # из моего каталоге
f = open('test-context-output.txt', 'w') # в каталоге сервера ..
f.write(languages2common.inputkey)
f.close()
print('context-type: text/html\n\nDone.\n')
```

Далее в примере 15.20 мы перепишем главную страницу в виде выполняемого сценария, помещающего в разметку HTML ответа значения, импортированные из файла общего модуля, представленного в предыдущем примере.

Пример 15.20. PP4E\Internet\Web\cgi-bin\languages2.py

```
#!/usr/bin/python
"""
разметка HTML главной страницы генерируется сценарием Python, а не готовым
файлом HTML; это позволяет импортировать ожидаемое имя поля ввода и значения
таблицы выбора языков из общего файла модуля Python; теперь изменения нужно
выполнять только в одном месте, в файле модуля Python;
"""
```

```
REPLY = """Content-type: text/html

<html><title>Languages2</title>
<body>
<h1>Hello World selector</h1>
<P>Эта страница похожа на страницу в файле <a href="../languages.html">
languages.html</a>, но генерируется динамически с помощью сценария CGI
на языке Python, используемый здесь список выбора и имена полей ввода
импортируются из общего модуля Python на сервере. При добавлении новых
языков достаточно будет изменить только общий модуль, потому что он
совместно используется сценариями, производящими страницы ответа.

Чтобы увидеть программный код, генерирующий эту страницу и ответ, щелкните
<a href="getfile.py?filename=cgi-bin\languages2.py">здесь</a>,
<a href="getfile.py?filename=cgi-bin\languages2reply.py">здесь</a>,
<a href="getfile.py?filename=cgi-bin\languages2common.py">здесь</a> и
<a href="getfile.py?filename=cgi-bin\\formMockup.py">здесь</a>.</P>
<hr>
<form method=POST action="languages2reply.py">
    <P><B>Select a programming language:</B>
    <P><select name=%s>
    <option>All
    %s
    <option>Other
    </select>
    <P><input type=Submit>
</form>
</body></html>
"""

from languages2common import hellos, inputkey

options = []
for lang in hellos:                     # можно было бы отсортировать
                                        # по ключам
    options.append('<option>' + lang)   # обернуть таблицу ключей
                                        # разметкой HTML
options = '\n\t'.join(options)
print(REPLY % (inputkey, options))      # имя поля и значения из модуля
```

Пока не обращайте внимания на гиперссылки `getfile` в этом файле, их значение будет объяснено в следующем разделе. Обратите, однако, внимание, что определение страницы HTML превратилось в строку Python (с именем `REPLY`), включающую спецификаторы форматирования `%s`, которые будут замещены значениями, импортированными из общего модуля. В остальных отношениях оно аналогично содержимому оригинального файла HTML. При переходе по адресу URL этого сценария выводится сходная страница, изображенная на рис. 15.25. Но на этот раз страница создается в процессе выполнения сценария на стороне сервера, который заполняет раскрывающийся список выбора значениями

из списка ключей в общей таблице синтаксиса. Используйте пункт меню броузера View Source (Исходный код страницы или Просмотр HTML-кода), чтобы увидеть сгенерированную разметку HTML – она практически идентична содержимому файла HTML в примере 15.17, хотя порядок следования языков в списке может отличаться, что обусловлено особенностями поведения словарей.

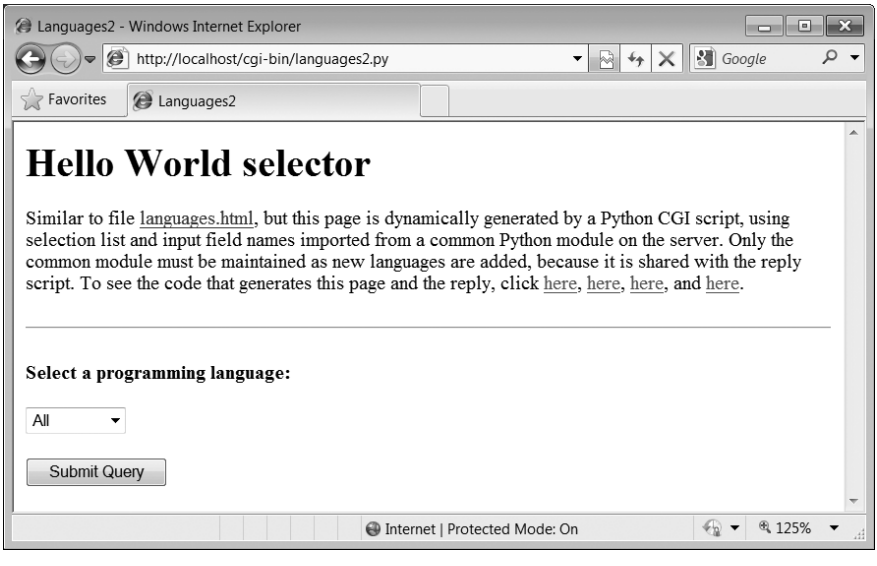


Рис. 15.25. *Альтернативная главная страница, созданная сценарием languages2.py*

Одно замечание, касающееся сопровождения: содержимое строки REPLY с шаблоном разметки HTML в примере 15.20 можно было бы загружать из внешнего текстового файла, чтобы над ним можно было работать независимо от сценария на языке Python. Однако в целом работать над внешними текстовыми файлами ничуть не проще, чем над сценариями Python. Фактически сценарии Python являются текстовыми файлами, и это важная особенность языка – сценарии Python легко можно изменять непосредственно в установленной системе без необходимости выполнять этапы компиляции и связывания. Однако внешние файлы HTML могут храниться в системе управления версиями отдельно, что, возможно, придется учитывать в вашей организации.

Шаг 2: многократно используемая утилита имитации формы

Перемещение таблицы со списком языков и имени поля ввода в файл модуля решает первые две отмеченные проблемы сопровождения. Но

если нам потребуется избавиться от необходимости писать фиктивный класс, имитирующий поля в каждом создаваемом сценарии CGI, следует сделать еще кое-что. И снова для этого нужно просто воспользоваться модулями Python, обеспечивающими возможность повторного использования программного кода: переместим фиктивный класс во вспомогательный модуль, представленный в примере 15.21.

Пример 15.21. PP4E\Internet\Web\cgi-bin\formMockup.py

```
"""
Инструменты имитации результатов, возвращаемых конструктором cgi.
FieldStorage(); удобно для тестирования сценариев CGI из командной строки
"""


class FieldMockup:              # имитируемый объект с входными данными
    def __init__(self, str):
        self.value = str

def formMockup(**kwargs):       # принимает аргументы в виде поле=значение
    mockup = {}                 # множественный выбор: [value,...]
    for (key, value) in kwargs.items():
        if type(value) != list: # простые поля имеют атрибут .value
            mockup[key] = FieldMockup(str(value))
        else:                   # поля множественного выбора
                                # являются списками
        mockup[key] = []        # что сделать: добавить поля выгрузки файлов
        for pick in value:
            mockup[key].append(FieldMockup(pick))
    return mockup

def selftest():
    # использовать эту форму, если поля можно определить жестко
    form = formMockup(name='Bob', job='hacker',
                      food=['Spam', 'eggs', 'ham'])
    print(form['name'].value)
    print(form['job'].value)
    for item in form['food']:
        print(item.value, end=' ')
    # использовать настоящий словарь, если значения ключей находятся
    # в переменных или вычисляются
    print()
    form = {'name': FieldMockup('Brian'),
            'age': FieldMockup(38)}             # или dict()
    for key in form.keys():
        print(form[key].value)

if __name__ == '__main__': selftest()
```

Благодаря переносу имитирующего класса в модуль `formMockup.py` он автоматически становится многократно используемым инструментом и может импортироваться любым сценарием, который мы соберемся на-

писать.[1] Для удобочитаемости класс dummy, моделирующий отдельное поле, переименован в FieldMockup. Для удобства введена также вспомогательная функция formMockup, которая конструирует полный словарь формы в соответствии с переданными ей именованными аргументами. В случае когда имена фиктивной формы могут быть жестко определены, модель создается за один вызов. В модуле есть также функция самотестирования, вызываемая при запуске этого файла из командной строки и демонстрирующая использование экспортируемых им данных. Ниже приводится контрольный вывод, образуемый путем создания и опроса двух объектов, имитирующих формы:

```
C:\...\PP4E\Internet\Web\cgi-bin> python formMockup.py
Bob
hacker
Spam eggs ham
38
Brian
```

Так как теперь реализация имитирующего класса находится в модуле, ею можно будет воспользоваться всякий раз, когда потребуется протестировать сценарий CGI в автономном режиме. Для иллюстрации в примере 15.22 приводится переработанная версия сценария *tutor5.py*, представленного ранее, в котором использована утилита имитации формы для моделирования полей ввода. Если бы мы спланировали это заранее, то таким способом могли бы проверить этот сценарий безо всякого подключения к Сети.

Пример 15.22. PP4E\Internet\Web\cgi-bin\tutor5_mockup.py

```
#!/usr/bin/python
"""
выполняет логику сценария tutor5 с применением formMockup
вместо cgi.FieldStorage()
для проверки: python tutor5_mockup.py > temp.html и открыть temp.html
"""


from formMockup import formMockup
form = formMockup(name='Bob',
                  shoesize='Small',
                  language=['Python', 'C++', 'HTML'],
                  comment='ni, Ni, NI')

# остальная часть сценария, как в оригинале, кроме операции присваивания form
```

[1] При этом, конечно, предполагается, что этот модуль находится в пути поиска модулей Python. Так как по умолчанию интерпретатор ведет поиск импортируемых модулей в текущем каталоге, это условие всегда выполняется без изменения sys.path, если все файлы находятся в главном веб-каталоге. Для других приложений может потребоваться добавить этот каталог в переменную окружения PYTHONPATH или использовать синтаксис импортирования пакетов (с указанием пути к каталогу).

Если запустить этот сценарий из командной строки, можно увидеть, как будет выглядеть разметка HTML ответа:

```
C:\...\PP4E\Internet\Web\cgi-bin> python tutor5_mockup.py
Content-type: text/html

<TITLE>tutor5.py</TITLE>
<H1>Greetings</H1>
<HR>
<H4>Your name is Bob</H4>
<H4>You wear rather Small shoes</H4>
<H4>Your current job: (unknown)</H4>
<H4>You program in Python and C++ and HTML</H4>
<H4>You also said:</H4>
<P>ni, Ni, NI</P>
<HR>
```

При запуске из броузера будет выведена страница, показанная на рис. 15.26. Значения полей ввода здесь жестко определены, что похоже по духу на расширение *tutor5*, в котором входные параметры встраивались в конец адреса URL внутри гиперссылки. Здесь они поступают от объектов имитации формы, создаваемых в сценарии ответа, которые можно изменить только редактированием сценария. Но поскольку программный код на языке Python выполняется немедленно, модификация сценария Python во время цикла отладки происходит со скоростью ввода с клавиатуры.

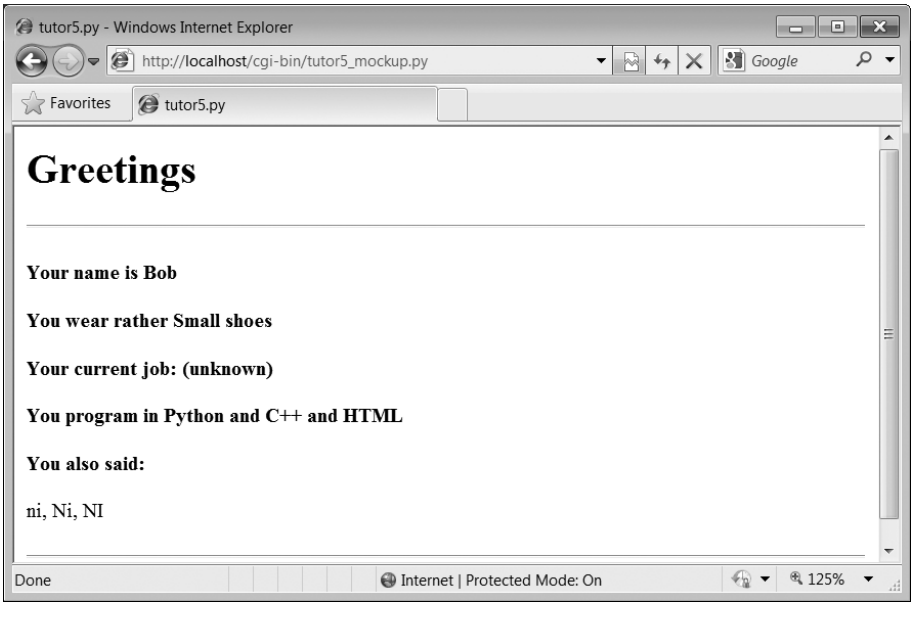


Рис. 15.26. *Страница ответа с имитируемыми входными данными*

Шаг 3: объединим все вместе – новый сценарий ответа

Остался последний шаг на пути к нирване сопровождения программного обеспечения: мы еще должны переписать сценарий, возвращающий страницу ответа, добавив в него импортирование данных, выделенных в общий модуль, и инструменты из многократно используемого модуля имитации форм. Раз уж мы об этом заговорили, поместим программный код в функции (на случай, если мы когда-нибудь поместим в этот файл то, что захочется импортировать в другой сценарий), а всю разметку HTML – в блоки строк, заключенные в тройные кавычки. Получившийся результат представлен в примере 15.23. Изменять разметку HTML обычно проще, когда она выделена в такие отдельные строки, а не разбросана по всей программе.

Пример 15.23. PP4E\Internet\Web\cgi-bin\languages2reply.py

```
#!/usr/bin/python
"""
То же самое, но проще для сопровождения, использует строки шаблонов
разметки HTML, получает таблицу со списком языков и имя входного параметра
из общего модуля и импортирует многократно используемый модуль
имитации полей форм для нужд тестирования.
"""

import cgi, sys
from formMockup import FieldMockup            # имитация полей ввода
from languages2common import hellos, inputkey # общая таблица, имя параметра
debugme = False

hdrhtml = """Content-type: text/html\n
<TITLE>Languages</TITLE>
<H1>Syntax</H1><HR>"""

langhtml = """
<H3>%s</H3><P><PRE>
%s
</PRE></P><BR>"""

def showHello(form):                # разметка HTML для одного языка
    choice = form[inputkey].value   # экранировать имя языка тоже
    try:
        print(langhtml % (cgi.escape(choice),
                          cgi.escape(hellos[choice])))
    except KeyError:
        print(langhtml % (cgi.escape(choice),
                          "Sorry--I don't know that language"))

def main():
    if debugme:
        form = {inputkey: FieldMockup(sys.argv[1])} # имя в командной строке
    else:
        form = cgi.FieldStorage()   # разбор действительных входных данных
```

```
    print(hdrhtml)
    if not inputkey in form or form[inputkey].value == 'All':
        for lang in hellos.keys():
            mock = {inputkey: FieldMockup(lang)}    # здесь не dict(n=v)!
            showHello(mock)
    else:
        showHello(form)
    print('<HR>')

if __name__ == '__main__': main()
```

Если в глобальной переменной `debugme` установить значение `True`, сценарий можно будет тестировать из командной строки:

```
C:\...\PP4E\Internet\Web\cgi-bin> python languages2reply.py Python
Content-type: text/html

<TITLE>Languages</TITLE>
<H1>Syntax</H1><HR>

<H3>Python</H3><P><PRE>
 print('Hello World')
</PRE></P><BR>
<HR>
```

При вызове этого сценария из страницы на рис. 15.25 или путем ввода адреса URL с параметрами запроса вручную возвращаются те же страницы, которые мы видели при использовании первоначальной версии этого примера (не станем повторять их снова). Данное преобразование изменило архитектуру программы, но не ее интерфейс пользователя. Однако теперь обе страницы, ввода данных и ответа, создаются сценариями CGI, а не с помощью статических файлов HTML.

Большая часть изменений, внесенных в эту версию сценария ответа, проста. При практическом испытании этих страниц единственными отличиями, которые можно обнаружить, будут адреса URL в строке ввода адреса броузера (все-таки, это разные файлы), дополнительные пустые строки в генерируемой разметке HTML (игнорируются броузером) и, возможно, иной порядок следования названий языков в раскрывающемся списке главной страницы.

Это различие в порядке следования элементов списка обусловлено тем, что данная версия зависит от порядка в списке ключей словаря Python, а не от фиксированного списка в файле HTML. Словари, если вы помните, упорядочивают свои записи так, чтобы скорейшим образом осуществлять их выборку. Если вы хотите, чтобы список выбора был более предсказуем, просто отсортируйте список ключей с помощью функции `sort`, появившейся в версии Python 2.4:

```
for lang in sorted(hellos): # итератор словаря вместо метода .keys()
    mock = {inputkey: FieldMockup(lang)}
```

Имитация ввода с помощью переменных оболочки

Если вы легко работаете с командной оболочкой, то на некоторых платформах можно также протестировать сценарии CGI из командной строки, установив переменные окружения так же, как это делают серверы HTTP, и затем запуская свой сценарий. Например, можно прикинуться веб-сервером, записав входные параметры в переменную окружения QUERY_STRING, применив тот же синтаксис, который используется в конце строки URL после ?:

```
$ setenv QUERY_STRING "name=Mel&job=trainer,+writer"
$ python tutor5.py
Content       type: text/html

<TITLE>tutor5.py<?TITLE>
<H1>Greetings</H1>
<HR>
<H4>Your name is Mel</H4>
<H4>You wear rather (unknown) shoes</H4>
<H4>Your current job: trainer, writer</H4>
<H4>You program in (unknown)</H4>
<H4>You also said:</H4>
<P>(unknown)</P>
<HR>
```

Здесь мы имитируем отправку формы в стиле GET или явную передачу параметров в адресе URL. Серверы HTTP помещают строку запроса (параметры) в переменную оболочки QUERY_STRING. Модуль Python cgi обнаруживает их там, как если бы они были переданы броузером. Передачу данных в стиле POST тоже можно имитировать с помощью переменных оболочки, но это сложнее, и лучше было бы не пытаться узнать, как это сделать. В действительности имитация ввода с помощью объектов Python (как в *formMockup.py*) должна работать надежнее. Но для некоторых сценариев CGI могут существовать дополнительные ограничения окружения или тестирования, требующие особой обработки.

Подробнее об экранировании HTML и URL

Возможно, самое тонкое изменение в версии из предыдущего раздела состоит в том, что для надежности в этой редакции сценария ответа (пример 15.23) функция cgi.escape вызывается также для обработки *названий* языков, а не только для фрагмента программного кода на этих языках. Этого не требовалось в сценарии *languages2.py* (пример 15.20), потому что названия языков в таблице выбора были нам известны. Хоть

и маловероятно, но все же возможно, что некто передаст сценарию название языка, содержащее символ HTML. Например, следующий URL:

```
http://localhost/cgi-bin/languages2reply.py?language=a<b
```

вставляет символ < в параметр с названием языка (названием является a<b). При передаче такого параметра эта версия с помощью функции cgi.escape правильно преобразует < для использования в HTML-разметке ответа согласно обсуждавшимся выше стандартным соглашениям HTML. Ниже приводится текст сгенерированного ответа:

```
<TITLE>Languages</TITLE>
<H1>Syntax</H1><HR>

<H3>a&lt;b</H3><P><PRE>
Sorry--I don't know that language
</PRE></P><BR>
<HR>
```

В исходной версии, в примере 15.18, название языка не экранируется, поэтому последовательность символов <b будет интерпретироваться как тег HTML (в результате чего оставшаяся часть страницы может быть выведена полужирным шрифтом!). Как вы уже можете теперь судить, экранирование текста повсеместно используется в сценариях CGI – даже текст, который кажется вам безопасным, обычно должен преобразовываться в экранированные последовательности перед вставкой в разметку HTML ответа.

Фактически, так как Веб является преимущественно текстовой средой, объединяющей множество языков программирования, к текстам могут применяться множество различных правил форматирования: одни для адресов URL и другие для разметки HTML. Выше в этой главе мы уже сталкивались с экранированием разметки HTML, теперь необходимо сказать несколько дополнительных слов об экранировании адресов URL и комбинации HTML и URL.

Соглашения по экранированию адресов URL

Обратите, однако, внимание, что хотя неэкранированные < нельзя вставлять в разметку HTML ответа, тем не менее, их вполне допустимо включать в адреса URL, используемые для получения ответа. На самом деле для HTML и URL определены совершенно различные специальные символы. Например, символ & в разметке HTML необходимо преобразовывать в &amp, но для представления символа & в адресе URL должна использоваться совсем другая схема кодирования (там этот символ обычно разделяет параметры). Для передачи сценарию названия языка, такого как a&b, нужно ввести такой адрес URL:

```
http://localhost/cgi-bin/languages2reply.py?language=a%26b
```

Здесь %26 представляет символ &, то есть символ & заменяется комбинацией символа % с шестнадцатеричным значением (0x26) & в кодировке

ASCII (38). Аналогично, как уже отмечалось в конце главы 13, чтобы передать название языка C++ в виде параметра запроса, символ + необходимо заменить экранированной последовательностью %2b:

```
http://localhost/cgi-bin/languages2reply.py?language=C%2b%2b
```

Передача названия C++ в неэкранированном виде не даст желаемого результата, потому что символ + в строках URL имеет специальное значение – он представляет пробел. По стандартам URL большинство не алфавитно-цифровых символов должно преобразовываться в такие экранированные последовательности, а пробелы – замещаться символами +. Это соглашение известно как формат строки запроса *application/x-www-form-urlencoded*, и оно стоит за теми странными адресами URL, которые вы можете часто видеть в адресной строке броузера, путешествуя в Веб.

Инструменты Python для экранирования HTML и URL

Вероятно, в вашей памяти, как и в моей, не запечатлелось шестнадцатеричное значение кода ASCII для символа & (впрочем, в этом вам может помочь вызов hex(ord(c))). К счастью, в языке Python имеются инструменты автоматического экранирования адресов URL, подобные функции cgi.escape, используемой для экранирования разметки HTML. Главное, о чем нужно помнить, – это то, что код разметки HTML и строки URL имеют совершенно разный синтаксис и потому для них используются различные соглашения по экранированию. Пользователям Веб это обычно безразлично, если только им не потребуется ввести сложный адрес URL в явном виде – внутренняя реализация броузеров обычно предусматривает все, что необходимо для экранирования. Но если вы пишете сценарии, которые должны генерировать разметку HTML или адреса URL, нужно следить за экранированием символов, имеющих специальное значение.

Поскольку для HTML и URL используется разный синтаксис, в языке Python имеются два разных набора инструментов их экранирования. В стандартной библиотеке Python:

- cgi.escape экранирует текст, который должен быть вставлен в разметку HTML
- urllib.parse.quote и quote_plus экранируют текст, который должен быть вставлен в адреса URL

В модуле urllib.parse есть также инструменты для обратного преобразования экранированных адресов URL (unquote, unquote_plus), но экранированная разметка HTML обычно не преобразуется обратно во время синтаксического анализа HTML (например, с помощью модуля html.parser). Чтобы проиллюстрировать действия этих двух соглашений по экранированию и инструментам, применим каждый из этих инструментов к нескольким простым примерам.

По совершенно необъяснимым причинам разработчики Python решили в версии 3.2 переместить функцию `cgi.escape`, широко используемую в этой книге, в другой модуль и переименовать ее в `html.escape`, объявить прежнюю функцию нерекомендуемой к использованию и несколько изменить ее поведение. И это несмотря на тот факт, что эта функция используется уже целую вечность, практически во всех CGI-сценариях на языке Python: яркий пример, когда понятия об эстетике, сложившиеся в маленькой группе лиц, наносят удар по распространенной практике в 3.X и нарушают работоспособность уже имеющегося программного кода. В будущих версиях Python вам может потребоваться использовать новую функцию `html.escape` – в том случае если пользователи Python не выскажут достаточно громких претензий (да, это намек!).

Экранирование разметки HTML

Как было показано ранее, `cgi.escape` преобразует текст, который должен быть включен в разметку HTML. Обычно эта утилита вызывается в сценариях CGI, но несложно исследовать ее действие и интерактивно:

```
>>> import cgi
>>> cgi.escape('a < b > c & d "spam"', 1)
'a &lt; b &gt; c &amp; d &quot;spam&quot;'

>>> s = cgi.escape("1<2 <b>hello</b>")
>>> s
'1&lt;2 &lt;b&gt;hello&lt;/b&gt;'
```

Модуль `cgi` автоматически преобразует специальные символы HTML в соответствии с соглашениями HTML. Он преобразует символы `<`, `>`, `&`, а при передаче дополнительного аргумента `True` также символ `"`, в экранированные последовательности вида `&X;`, где `X` – мнемоника, обозначающая исходный символ. Например, `<` обозначает оператор «меньше» (`<`), а `&` обозначает литерал амперсанда (`&`).

В модуле `cgi` отсутствуют инструменты *обратного* преобразования экранированных последовательностей, потому что экранированные последовательности HTML распознаются в контексте анализатора HTML, подобно тому, как это делают веб-броузеры при загрузке страницы. В Python также есть полный анализатор HTML в виде стандартного модуля `html.parser`. Мы не станем здесь вникать в детали инструментов синтаксического анализа HTML (они описываются в главе 19, вместе с инструментами обработки текста), но чтобы показать, как экранированные последовательности в конечном итоге преобразуются обратно в неэкранированный вид, приведем пример работы модуля анализатора HTML, воспроизводящий последнюю строку в примере выше:

```
>>> import cgi, html.parser
>>> s = cgi.escape("1<2 <b>hello</b>")
```

```
>>> s
'1&lt;2 &lt;b&gt;hello&lt;/b&gt;'
>>>
>>> html.parser.HTMLParser().unescape(s)
'1<2 <b>hello</b>'
```

Здесь для обратного преобразования используется вспомогательный метод класса, реализующего синтаксический анализ HTML. В главе 19 мы увидим, как использовать этот класс для решения более реальных задач, с созданием подклассов, переопределяющих методы, которые вызываются в процессе анализа при обнаружении тегов, данных, мнемоник и других элементов разметки. Продолжение этой истории и примеры более полноценного синтаксического анализа HTML вы найдете в главе 19.

Экранирование адресов URL

В адресах URL используются иные специальные символы и иные соглашения по их экранированию. По этой причине для экранирования адресов URL применяются другие библиотечные инструменты Python. Модуль Python `urllib.parse` предоставляет два инструмента, осуществляющие преобразование: функцию `quote`, возвращающую стандартные шестнадцатеричные экранированные последовательности `%XX` URL для большинства не алфавитно-цифровых символов, и функцию `quote_plus`, для преобразования пробелов в символы `+`. Кроме того, модуль `urllib.parse` предоставляет функции для обратного преобразования экранированных последовательностей в адресах URL: функция `unquote` преобразует последовательности `%XX`, а функция `unquote_plus` также преобразует символы `+` в пробелы. Ниже приводится пример использования модуля в интерактивной оболочке:

```
>>> import urllib.parse
>>> urllib.parse.quote("a & b #! c")
'a%20%26%20b%20%23%21%20c'

>>> urllib.parse.quote_plus("C:\stuff\spam.txt")
'C%3A%5Cstuff%5Cspam.txt'

>>> x = urllib.parse.quote_plus("a & b #! c")
>>> x
'a+%26+b+%23%21+c'

>>> urllib.parse.unquote_plus(x)
'a & b #! c'
```

Экранированные последовательности URL помещают шестнадцатеричные значения небезопасных символов вслед за знаком `%` (обычно это их ASCII-коды). В `urllib.parse` небезопасными символами обычно считаются все кроме букв, цифр и некоторых безопасных специальных символов (символов `'_.-'`), но эти два инструмента по-разному интерпретируют

символы слэша, а кроме того, имеется возможность расширить набор безопасных символов, передав функции quote дополнительный строковый аргумент:

```
>>> urllib.parse.quote_plus("uploads/index.txt")
'uploads%2Findex.txt'
>>> urllib.parse.quote("uploads/index.txt")
'uploads/index.txt'
>>>
>>> urllib.parse.quote_plus("uploads/index.txt", '/')
'uploads/index.txt'
>>> urllib.parse.quote("uploads/index.txt", '/')
'uploads/index.txt'
>>> urllib.parse.quote("uploads/index.txt", '')
'uploads%2Findex.txt'
>>>
>>> urllib.parse.quote_plus("uploads\index.txt")
'uploads%5Cindex.txt'
>>> urllib.parse.quote("uploads\index.txt")
'uploads%5Cindex.txt'
>>> urllib.parse.quote_plus("uploads\index.txt", '\\')
'uploads\\index.txt'
```

Обратите внимание, что модуль Python cgi в процессе извлечения входных данных также преобразует экранированные последовательности URL обратно в их исходные символы и меняет знаки + на пробелы. Внутренне конструктор cgi.FieldStorage при необходимости автоматически вызывает инструменты urllib.parse для обратного преобразования параметров, передаваемых в конце адреса URL. В итоге сценарии CGI получают исходные, неэкранированные строки URL, и им не требуется самостоятельно производить обратное преобразование значений. Как мы видели, сценариям CGI вообще не требуется знать, что входные данные поступили из URL.

Экранирование адресов URL, встроенных в разметку HTML

Теперь мы знаем, как экранировать текст, вставляемый в разметку HTML и в адреса URL. Но что делать с адресами URL внутри HTML? То есть как выполнять экранирование, когда генерируется текст, вставляемый в адреса URL, в свою очередь встраиваемые в разметку HTML? В некоторых предыдущих примерах внутри тегов гиперссылок <A HREF>, используются жестко определенные адреса URL, дополненные входными параметрами. Например, файл *languages2.py* выводит разметку HTML, содержащую адрес URL:

```
<a href="getfile.py?filename=cgi-bin\languages2.py">
```

Поскольку строка URL здесь встроена в разметку в HTML, она должна быть экранирована, по меньшей мере, в соответствии с соглашениями

HTML (например, все символы < должны быть превращены в последовательности <), а все пробелы должны быть преобразованы в знаки +. Этого можно достичь с помощью вызова cgi.escape(url) с последующим вызовом строкового метода url.replace(" ", "+"), чего в большинстве случаев должно быть достаточно.

Однако в общем случае этого мало, потому что соглашения по экранированию HTML отличны от соглашений для URL. Надежное экранирование адресов URL, встраиваемых в разметку HTML, следует осуществлять путем применения функции urllib.parse.quote_plus к строке URL или хотя бы к большинству ее компонентов, перед тем как поместить ее в текст HTML. Результат такого экранирования будет также удовлетворять соглашениям по экранированию HTML, потому что модуль urllib.parse транслирует больше символов, чем функция cgi.escape, а символ % в экранированных последовательностях внутри URL не является специальным символом HTML.

Конфликты HTML и URL: &

Но здесь есть еще одна удивительно тонкая (и, к счастью, редкая) особенность: нужно быть осторожными с символами & в строках URL, которые встраиваются в разметку HTML (например, в теги гиперссылок <A>). В адресах URL символ & используется как разделитель параметров в строке запроса (?a=1&b=2), а в HTML он является первым символом в экранированных последовательностях (<). Следовательно, есть вероятность конфликтов, если определение параметра запроса совпадет с экранированной последовательностью HTML. Например, параметр запроса с именем amp, который записывается как &amp=1 и имеет порядковый номер 2 или выше в адресе URL, может интерпретироваться некоторыми анализаторами HTML как экранированная последовательность HTML и преобразовываться в &=1.

Даже если части строки URL экранированы в соответствии с соглашениями для URL, при наличии нескольких параметров они разделяются символом &, а разделитель & также может оказаться преобразованным в & в соответствии с соглашениями для HTML. Чтобы понять причину, взгляните на следующий тег гиперссылки HTML:

```
<A HREF="file.py?name=a&job=b&amp=c&sect=d&lt=e">hello</a>
```

При отображении в большинстве проверенных мной броузеров, включая Internet Explorer в Windows 7, этот адрес URL-ссылки выглядит неправильно, как показано ниже (символ S в первом из них интерпретирован как маркер раздела, код которого выходит за пределы диапазона ASCII):

```
file.py?name=a&job=b&=cS=d<=e          результат в IE
file.py?name=a&job=b&=c%A7=d%3C=e      результат в Chrome (0x3C – это <)
```

Первые два параметра будут сохранены, как и можно было бы ожидать (name=a, job=b), потому что перед name нет символа &, а &job не распознает-

ся как допустимая экранированная последовательность HTML. Однако последовательности символов &amp, &sect и &lt интерпретируются как специальные символы, потому что совпадают с именами допустимых в HTML экранированных последовательностей, даже несмотря на отсутствие завершающего символа точки с запятой.

Чтобы убедиться в этом, откройте файл *test-escapes.html*, входящий в состав пакета с примерами, в своем броузере и выделите или выберите эту ссылку – имена параметров в строке запроса могут быть интерпретированы как экранированные последовательности HTML. Похоже, что этот текст правильно интерпретируется модулем синтаксического анализа разметки HTML, описанным выше (если только фрагменты строки запроса не завершаются точкой с запятой), – это может помочь при получении ответов вручную, с помощью модуля urllib.request, но не при отображении в броузерах:

```
>>> from html.parser import HTMLParser
>>> html = open('test-escapes.html').read()
>>> HTMLParser().unescape(html)
'<HTML>\n<A HREF="file.py?name=a&job=b&amp=c&sect=d&lt=e">hello</a>\n</HTML>'
```

Исключение конфликтов

Так что же делать? Чтобы все действовало так, как нужно, необходимо экранировать разделители &, если имена параметров могут конфликтовать с экранированными последовательностями HTML:

```
<A HREF="file.py?name=a&amp;job=b&amp;amp=c&amp;sect=d&amp;lt=e">hello</a>
```

Такую полностью экранированную ссылку броузеры Python выводят как надо (откройте файл *test-escapes2.html*, чтобы убедиться), и синтаксический анализатор HTML из стандартной библиотеки также корректно интерпретирует ее:

```
file.py?name=a&job=b&amp=c&sect=d&lt=e        результат в IE и Chrome

>>> h = '<A HREF="file.py?name=a&amp;job=b&amp;amp=c&amp;
                                    sect=d&amp;lt=e">hello</a>'
>>> HTMLParser().unescape(h)
'<A HREF="file.py?name=a&job=b&amp=c&sect=d&lt=e">hello</a>'
```

Из-за этого конфликта между синтаксисом HTML и URL большинство серверных инструментов (включая инструменты разбора параметров запроса в модуле Python urlib.parse, используемые модулем cgi) позволяют также использовать символ точки с запятой вместо & в качестве разделителя параметров. Следующая ссылка, например, действует точно так же, как и полностью экранированный адрес URL, но для нее не требуется выполнять дополнительное экранирование HTML (по крайней мере для ;):

```
file.py?name=a;job=b;amp=c;sect=d;lt=e
```

Инструменты обратного преобразования экранированных последовательностей в модуле Python `html.parser` позволяют передавать символы точки с запятой без изменений, просто потому, что они не являются специальными символами HTML. Чтобы полностью протестировать все три эти ссылки одновременно, поместите их в файл HTML, откройте его в броузере, используя адрес URL *http://localhost/badlink.html*, и посмотрите, как выглядят эти ссылки. Файла HTML, представленного в примере 15.24, будет вполне достаточно.

Пример 15.24. PP4E\Internet\Web\badlink.html

```
<HTML><BODY>

<p><A HREF=
"cgi-bin/badlink.py?name=a&job=b&amp=c&sect=d&lt=e">unescaped</a>

<p><A HREF=
"cgi-bin/badlink.py?name=a&amp;job=b&amp;amp=c&amp;sect=d&amp;lt=e">
escaped</a>

<p><A HREF=
"cgi-bin/badlink.py?name=a;job=b;amp=c;sect=d;lt=e">alternative</a>

</BODY></HTML>
```

Щелчок на любой из этих ссылок будет вызывать запуск простого сценария CGI, представленного в примере 15.25. Этот сценарий выводит входные параметры, полученные им от клиента, в стандартный поток ошибок, чтобы избежать необходимости дополнительных преобразований (при использовании нашего локального веб-сервера из примера 15.1 вывод будет осуществляться в окно консоли сервера).

Пример 15.25. PP4E\Internet\Web\cgi-bin\badlink.py

```
import cgi, sys
form = cgi.FieldStorage()   # выводит все входные параметры
                            # в stderr; stodout=reply page
for name in form.keys():
    print('[%s:%s]' % (name, form[name].value), end=' ', file=sys.stderr)
```

Ниже приводится (отредактированный, чтобы уместился по ширине страницы) вывод, полученный в окне консоли нашего локального веб-сервера для каждой из трех ссылок в странице HTML при использовании Internet Explorer. Второй и третий результаты содержат корректные наборы параметров благодаря экранированию в соответствии с соглашениями для HTML или URL, но случайные совпадения имен параметров с экранированными последовательностями HTML в первой неэкранированной ссылке вызывают серьезные проблемы – синтаксический анализатор HTML на стороне клиента интерпретирует их не так, как пред-

полагалось (в Chrome получаются похожие результаты, но в первой ссылке отображается символ раздела, не входящий в набор ASCII, соответствующий другой экранированной последовательности):

```
mark-VAIO - [16/Jun/2010 10:43:24] b'[:c\xa7=d<=e] [job:b] [name:a] '
mark-VAIO - [16/Jun/2010 10:43:24] CGI script exited OK

mark-VAIO - [16/Jun/2010 10:43:27] b'[amp:c] [job:b] [lt:e] [name:a] [sect:d]'
mark-VAIO - [16/Jun/2010 10:43:27] CGI script exited OK

mark-VAIO - [16/Jun/2010 10:43:30] b'[amp:c] [job:b] [lt:e] [name:a] [sect:d]'
mark-VAIO - [16/Jun/2010 10:43:30] CGI script exited OK
```

Из всей этой истории должен быть сделан следующий вывод: если нет уверенности, что все имена параметров в строке запроса URL, встраиваемой в разметку HTML, кроме самого левого, отличны от экранированных последовательностей HTML, таких как `amp`, вы должны или использовать в качестве разделителя точку с запятой (если это поддерживается инструментами, с которыми вы имеете дело), или пропустить весь URL через функцию `cgi.escape` после того, как имена параметров и их значения будут экранированы с помощью `urllib.parse.quote_plus`:

```
>>> link = 'file.py?name=a&job=b&amp=c&sect=d&lt=e'

# экранирование в соответствии с соглашениями для HTML
>>> import cgi
>>> cgi.escape(link)
'file.py?name=a&amp;job=b&amp;amp=c&amp;sect=d&amp;lt=e'

# экранирование в соответствии с соглашениями для URL
>>> import urllib.parse
>>> elink = urllib.parse.quote_plus(link)
>>> elink
'file.py%3Fname%3Da%26job%3Db%26amp%3Dc%26sect%3Dd%26lt%3De'

# экранирование URL также удовлетворяет соглашениям для HTML:
тот же результат
>>> cgi.escape(elink)
'file.py%3Fname%3Da%26job%3Db%26amp%3Dc%26sect%3Dd%26lt%3De'
```

При этом я должен добавить, что в некоторых примерах этой книги разделители `&` в адресах URL, встраиваемых в HTML, не были экранированы, поскольку известно, что имена параметров в этих URL не конфликтуют с экранированными последовательностями HTML. В действительности, эта проблема редко возникает на практике, поскольку имена параметров, ожидаемых программой, полностью подконтрольны программистам. Однако нельзя считать, что этой проблемы не существует, особенно если имена параметров могут извлекаться динамически из базы данных – при наличии сомнений экранируйте много и часто.

«Взгляните на жизнь с ее приятной стороны»

Какими бы корявыми (и заставляющими с криком просыпаться ночью!) не казались вам все эти правила форматирования HTML и URL, обратите внимание, что эти соглашения по экранированию HTML и URL диктуются самим Интернетом, а не Python. (Мы уже знаем, что в Python существует другой механизм экранирования специальных символов в строковых константах – с помощью символов обратного слэша.) Эти правила проистекают из того обстоятельства, что Сеть основана на идее пересылки по всему свету форматированных текстовых строк, а правила вырабатывались в ответ на то, что разные группы разработчиков использовали различные системы обозначений.

Можете, однако, утешиться тем, что часто не требуется заботиться о подобных таинственных вещах, а если требуется, то Python автоматизирует процесс с помощью библиотечных средств. Просто имейте в виду, что всякому сценарию, генерирующему разметку HTML или адреса URL динамически, возможно, требуется для устойчивости воспользоваться средствами Python преобразования в экранированные последовательности. В дальнейших примерах этой и следующей глав часто будут использоваться средства экранирования HTML и URL. Кроме того, фреймворки и инструменты для разработки веб-приложений, такие как Zope и другие, стремятся избавить нас от некоторых низкоуровневых сложностей, с которыми сталкиваются разработчики сценариев CGI. И, как обычно в программировании, ничто не может заменить интеллект – ценой доступа к поразительным технологиям вроде Интернета является их сложность.

Передача файлов между клиентами и серверами

Пришла пора объяснить ту часть разметки HTML, которая до этого скрывалась в тени. Обратили внимание на гиперссылки на главной странице примера выбора языка, вызывающие отображение исходного программного кода сценария CGI (на которые я предлагал не обращать внимание)? Обычно мы не видим исходный программный код таких сценариев, потому что обращение к сценарию CGI заставляет его выполняться – мы можем видеть только разметку HTML, генерируемую им для создания новой страницы. Сценарий в примере 15.26, на который ссылается гиперссылка на главной странице `language.html`, обходит это правило, открывая исходный файл и пересылая его содержимое как часть ответа HTML. Текст заключен в тег `<PRE>` как предварительно форматированный текст и преобразован с помощью `cgi.escape` для передачи в виде HTML.

Пример 15.26. PP4E\Internet\Web\cgi-bin\languages-src.py

```
#!/usr/bin/python
"Отображает содержимое сценария languages.py не выполняя его."

import cgi
filename = 'cgi-bin/languages.py'

print('Content-type: text/html\n')       # обернуть в разметку HTML
print('<TITLE>Languages</TITLE>')
print("<H1>Source code: '%s'</H1>" % filename)
print('<HR><PRE>')
print(cgi.escape(open(filename).read())) # декодирование выполняется
print('</PRE><HR>')                      # с применением кодировки по умолчанию
```

Для нашего веб-сервера, выполняющегося в Windows, путь к файлу `filename` откладывается относительно каталога сервера (смотрите предыдущее обсуждение этой темы и удалите часть `cgi-bin` пути при опробовании примера на других платформах). Если запустить этот сценарий из Веб, щелкнув в примере 15.17 на первой гиперссылке из числа ведущих к исходным программным кодам или введя адрес URL вручную, то сценарий отправит клиенту ответ, содержащий текст файла с исходным программным кодом сценария CGI. Он показан на рис. 15.27.

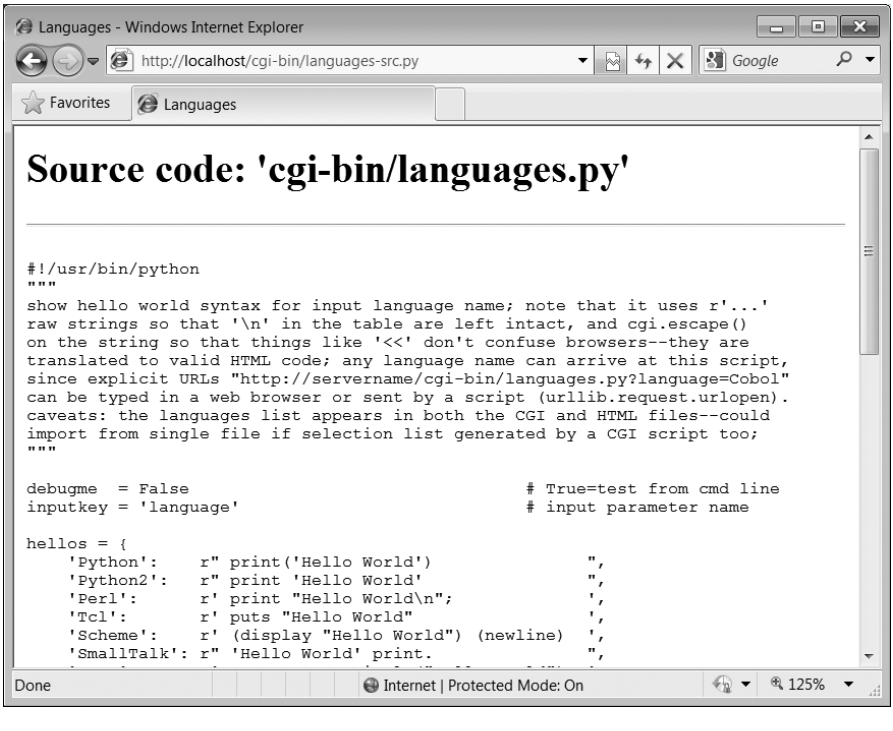


Рис. 15.27. *Страница просмотра исходного программного кода*

Обратите внимание, что здесь также решающим обстоятельством является экранирование текста файла с помощью функции `cgi.escape`, потому что он встраивается в разметку HTML ответа. Если этого не сделать, все символы текста, имеющие какое-либо специальное значение в HTML, будут интерпретированы как теги HTML. Например, символ оператора C++ < в тексте этого файла может привести к странным результатам, если не экранировать его надлежащим образом. Функция `cgi.escape` превратит его в стандартную последовательность `<`, которая может быть безопасно встроена в разметку HTML.

Отображение произвольных файлов сервера на стороне клиента

Почти сразу после того, как я написал сценарий просмотра исходного программного кода, представленный в предыдущем примере, мне пришло в голову, что, приложив совсем немного усилий, я смогу написать более обобщенную версию, которая принесет больше пользы, – такую, которая по переданному ей имени файла отобразит *любой* файл на сайте. На стороне сервера это простое видоизменение: необходимо лишь разрешить передачу имени файла в качестве входных данных. Сценарий Python *getfile.py*, представленный в примере 15.27, реализует это обобщение. Он предполагает, что имя файла введено в форму на веб-странице или дописано в конец URL в качестве параметра. Напомню, что модуль Python `cgi` прозрачным образом обрабатывает оба эти случая, поэтому данный сценарий не содержит программного кода, который как-то разделял бы эти варианты.

Пример 15.27. PP4E\Internet\Web\cgi-bin\getfile.py

```
#!/usr/bin/python
"""
###############################################################################
Отображает содержимое любого сценария CGI (или другого файла), имеющегося
на стороне сервера, не выполняя его. Имя файла можно передать в параметре
строки URL или с помощью поля формы (используйте имя сервера "localhost",
если используется локальный сервер):

    http://servername/cgi-bin/getfile.py?filename=somefile.html
    http://servername/cgi-bin/getfile.py?filename=cgi-bin\somefile.py
    http://servername/cgi-bin/getfile.py?filename=cgi-bin%2Fsomefile.py

Пользователи могут сохранить файл у себя, скопировав текст через буфер
обмена или воспользовавшись пунктом меню "View Source" ("Исходный код
страницы" или "Просмотр HTML-кода"). При обращении к этому сценарию из IE
для получения версии text/plain (formatted=False) может запускаться
программа Блокнот (Notepad), при этом не всегда используются символы конца
строки в стиле DOS; Netscape, напротив, корректно отображает текст на
странице броузера. Отправка файла в версии text/HTML действует в обоих типах
броузеров - текст правильно отображается на странице ответа в броузере.
```

```
Мы также проверяем имя файла, чтобы избежать отображения закрытых файлов;
в целом это может не предотвратить доступ к таким файлам: не устанавливайте
этот сценарий, если исходные тексты закрытых сценариев у вас не защищены
каким-то иным способом!
##############################################################################
"""

import cgi, os, sys
formatted = True                            # True=обернуть текст в HTML
privates = ['PyMailCgi/cgi-bin/secret.py']  # эти файлы не показывать

try:
    samefile = os.path.samefile         # проверка устройства, номера inode
except:
    def samefile(path1, path2):                 # не доступна в Windows
        apath1 = os.path.abspath(path1).lower() # близкая аппроксимация
        apath2 = os.path.abspath(path2).lower() # нормализовать пути,
        return apath1 == apath2                 # привести к одному регистру

html = """
<html><title>Getfile response</title>
<h1>Source code for: '%s'</h1>
<hr>
<pre>%s</pre>
<hr></html>"""

def restricted(filename):
    for path in privates:
        if samefile(path, filename):    # пути унифицированы вызовом os.stat
            return True                 # иначе вернет None=false

try:
    form = cgi.FieldStorage()
    filename = form['filename'].value   # параметр URL или поле формы
except:
    filename = 'cgi-bin\getfile.py'     # иначе имя файла по умолчанию

try:
    assert not restricted(filename)     # загрузить, если не закрытый файл
    filetext = open(filename).read()    # кодировка Юникода для платформы
except AssertionError:
    filetext = '(File access denied)'
except:
    filetext = '(Error opening file: %s)' % sys.exc_info()[1]

if not formatted:
    print('Content-type: text/plain\n') # отправить простой текст
    print(filetext)                     # действует в NS, но не в IE?
else:
print('Content-type: text/html\n')      # обернуть в HTML
print(html % (filename, cgi.escape(filetext)))
```

Этот сценарий на языке Python, выполняемый на сервере, просто извлекает имя файла из объекта с входными данными, читает и выводит текст файла для отправки броузеру клиента. В зависимости от значения глобальной переменной `formatted`, файл отправляется либо в режиме простого текста (с указанием типа `text/plain` в заголовке ответа), либо обертывает его разметкой HTML страницы (`text/html`).

Оба режима (а также другие) в целом корректно интерпретируются большинством броузеров, но Internet Explorer обрабатывает режим простого текста не так элегантно, как это делает Netscape, – при проверке он открыл загруженный текст в редакторе Блокнот (Notepad), но из-за символов конца строки в стиле Unix файл был показан как одна длинная строка. (Netscape правильно выводит текст в теле самой веб-страницы ответа.) Режим отображения HTML в современных броузерах работает более переносимым образом. О логике этого сценария в работе с файлами, доступ к которым ограничен, будет рассказываться несколько ниже.

Запустим сценарий, набрав его URL в адресной строке броузера вместе с именем нужного файла в конце. На рис. 15.28 показана страница, которая была получена при посещении этого адреса URL (вторая ссылка на файл с исходным программным кодом в странице выбора языка, представленной в примере 15.17, имеет аналогичный эффект, но возвращает другой файл):

```
http://localhost/cgi-bin/getfile.py?filename=cgi-bin\languages-src.py
```

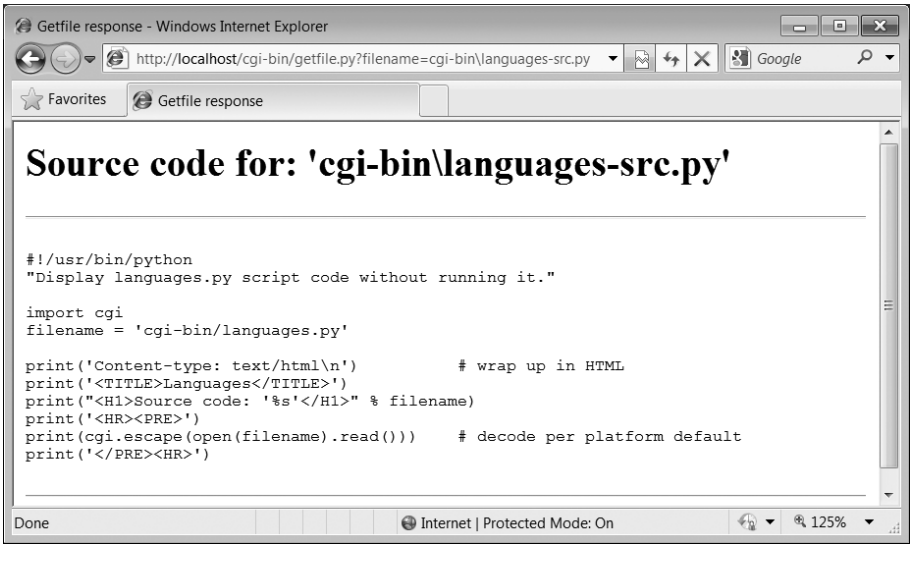


Рис. 15.28. Универсальная страница просмотра исходного программного кода

В теле этой страницы показан текст находящегося на сервере файла, имя которого было передано в конце URL. После получения файла можно просматривать его текст, выполнять операции удаления и вставки для сохранения в файле на стороне клиента и так далее. В действительности теперь, имея такую обобщенную программу просмотра содержимого файлов, можно заменить гиперссылку на сценарий *languages-src.py* в *language.html* на адрес URL следующего вида (я включил обе для иллюстрации):

```
http://localhost/cgi-bin/getfile.py?filename=cgi-bin\languages.py
```

Тонкое замечание: обратите внимание, что в параметре запроса в этой и в других строках URL в данной книге в качестве разделителя каталогов в путях используется символ обратного слэша в стиле Windows. В Windows при использовании локального веб-сервера из примера 15.1 и Internet Explorer мы можем использовать любую из первых двух экранированных форм URL, которые показаны ниже, но попытка использовать литерал символа слэша, как в последней строке, приведет к неудаче (в URL экранированная последовательность `%5C` соответствует символу `\`, а `%2F` – `/`):

```
http://localhost/cgi-bin/getfile.py?filename=cgi-bin%5Clanguages.py
                                                                  работает
http://localhost/cgi-bin/getfile.py?filename=cgi-bin%2Flanguages.py
                                                                  работает
http://localhost/cgi-bin/getfile.py?filename=cgi-bin/languages.py   неудача
```

Этим данная версия сценария отличается от версии в предыдущем издании этой книги (где для переносимости использовалась последняя форма URL), и, возможно, данное решение не является идеальным (хотя, как и контекст рабочего каталога, это одна из многих особенностей, отличающих серверы и платформы, с которыми вы наверняка столкнетесь в Веб). Проблема, похоже, заключается в том, что функция `quote` из модуля `urllib.parse` считает символ `/` безопасным, а `quote_plus` – нет. Если вы заинтересованы в обеспечении переносимости строк URL в этом контексте, то вторая форма из трех приведенных является, вероятно, самой лучшей, однако она сложна для запоминания, если придется вводить ее вручную (инструменты экранирования могут автоматизировать преобразование в эту форму). В противном случае вы можете вводить двойные символы обратного слэша, чтобы избежать конфликтов с другими экранированными последовательностями в строках, обусловленными особенностями обработки параметров URL; смотрите ссылки на этот сценарий в примере 15.20, где предпринята попытка исключить конфликт со специальным символом `\f`.

С более общей точки зрения, такие адреса URL в действительности являются непосредственными вызовами (хотя и через веб) нашего сценария Python с явным параметром, содержащим имя файла. Мы исполь-

зуем сценарий как своеобразную функцию, находящуюся где-то в киберпространстве, которая возвращает текстовое содержимое файла. Как мы видели, параметры, которые передаются в адресе URL, обрабатываются так же, как данные, введенные в поля формы. Давайте для удобства напишем простую веб-страницу, позволяющую ввести имя нужного файла прямо в форме, как показано в примере 15.28.

Пример 15.28. PP4E\Internet\Web\getfile.html

```
<html><title>Getfile: download page</title>
<body>
<form method=get action="cgi-bin/getfile.py">
    <h1>Type name of server file to be viewed</h1>
    <p><input type=text size=50 name=filename>
    <p><input type=submit value=Download>
</form>
<hr><a href="cgi-bin/getfile.py?filename=cgi-bin\getfile.py">View script
code</a>
</body></html>
```

На рис. 15.29 показана страница, полученная при посещении URL этого сценария. На этой странице требуется ввести только имя файла, а не полный адрес сценария CGI. Обратите внимание, что здесь есть возможность использовать символы слэша, потому что броузер будет экранировать их при передаче запроса, а функция Python `open` принимает любые типы слэшей в Windows (при создании параметров запроса вручную ответственность за правильность выбора ложится на плечи пользователя или сценария, генерирующего эти параметры).

Если для отправки формы нажать кнопку `Download` (Загрузить) на этой странице, имя файла передается на сервер, и мы получаем ту же страницу, что и раньше, когда имя файла добавлялось в адрес URL (как на рис. 15.28, хотя и с другим символом-разделителем каталогов). На самом деле имя файла здесь тоже будет добавлено в адрес URL: метод `get` в разметке HTML формы указывает броузеру, что он должен добавить имя файла в конец URL, как мы делали это вручную. Оно появляется в конце адреса URL, в строке адреса страницы ответа, хотя в действительности мы ввели его в форму. Щелчок на ссылке внизу страницы на рис. 15.29 открывает исходный программный код самого сценария, возвращающего файлы, правда при этом адрес URL определяется явно.[1]

[1] Можно заметить одно отличие в страницах ответа, получаемых через форму и через явно введенные адреса URL: при использовании формы значение параметра «filename» в конце адреса URL страницы ответа может содержать экранированные последовательности вместо некоторых символов в пути к файлу, имя которого мы ввели. Броузеры автоматически преобразуют некоторые отсутствующие в ASCII символы в экранированные последовательности URL (как `urllib.parse.quote`). Экранирование адресов URL обсуждалось выше в этой главе. Пример такого автоматического преобразования, выполненного броузером, мы увидим чуть ниже.

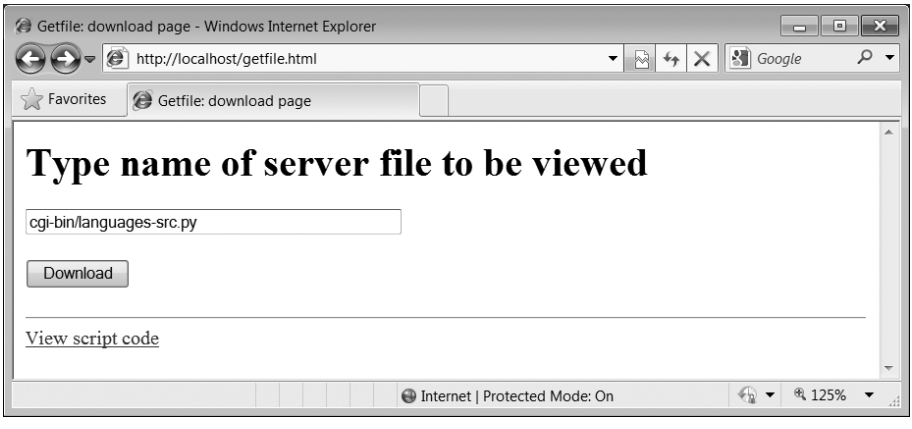


Рис. 15.29. *Страница выбора файла для просмотра его содержимого*

Обработка закрытых файлов и ошибок

Если сценарий CGI обладает правами, достаточными для открытия нужного файла на сервере, этот сценарий можно использовать для просмотра *любого* файла на сервере и сохранения его на локальном компьютере.[1] Например, на рис. 15.30 изображена страница, полученная при запросе файла *PyMailCgi/pymailcgi.html* – текстового файла HTML в подкаталоге другого приложения, находящегося в каталоге, родительском для данного сценария (приложение PyMailCGI мы будем исследовать в следующей главе). Пользователи могут указывать как относительные, так и абсолютные пути к файлам – годится любой синтаксис пути, понятный серверу.

Вообще говоря, этот сценарий отобразит файл с любым путем, доступный для чтения пользователю, с привилегиями которого выполняется сценарий CGI. На некоторых серверах для этого обычно используется предопределенная учетная запись «nobody» с ограниченными привилегиями. Это касается почти любого файла на сервере, используемого в веб-приложениях, иначе они оказались бы недоступными для броузеров. При использовании нашего локального веб-сервера можно получить содержимое любого файла, имеющегося на компьютере: напри-

[1] Стоит заметить, что помещая подобный сценарий на свой веб-сервер, вы можете существенно ослабить защиту этого сервера. Вам будет трудно поместить все файлы, которые вы хотите защитить, в список privates. Полагаю, автор поместил этот пример в книгу исключительно в учебных целях, без намерения использовать его на реально работающем сервере. Для того чтобы сделать вышеприведенный сценарий более безопасным, следует, по крайней мере, ограничить область, в которой лежат файлы, доступные для просмотра, одним каталогом и при запросе проверять, находится ли запрашиваемый файл в этом каталоге. – *Прим. науч. ред.*

мер, я смог получить файл *C:\Users\mark\Stuff\Websites\public_html\index.html*, находящийся на моем ноутбуке, введя путь к нему в форму на рис. 15.29.

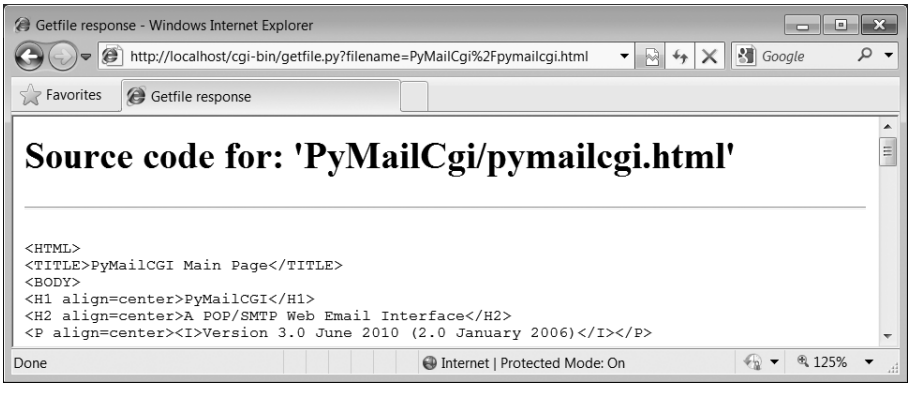


Рис. 15.30. *Просмотр файлов по относительному пути*

Усиливая гибкость инструмента, это также представляет потенциальную опасность, если сервер выполняется на удаленном компьютере. Как быть, если потребуется закрыть для пользователей доступ к некоторым файлам на сервере? Например, в следующей главе будет реализован модуль для шифрования паролей учетных записей электронной почты. На сервере он будет находиться по адресу *PyMailCgi/cgi-bin/secret.py*. Если пользователям будет доступен просмотр реализации этого модуля, это существенно облегчит им возможность взлома зашифрованных паролей, передаваемых через Сеть.

Чтобы уменьшить эту опасность, сценарий `getfile` хранит имена файлов с ограниченным доступом в списке `privates` и с помощью встроенной функции `os.path.samefile` проверяет, не совпадает ли имя запрашиваемого файла с одним из имен в списке `privates`. Функция `samefile` сравнивает идентификационную информацию обоих файлов, полученную с помощью встроенной функции `os.stat` (устройство и номера индексных узлов). Поэтому пути, выглядящие синтаксически различными, но ссылающиеся на один и тот же файл, считаются идентичными. Например, на сервере, использовавшемся в процессе работы над вторым изданием этой книги, следующие пути к модулю шифрования являются разными строками, но вызов `os.path.samefile` для них возвращает «истину»:

```
../PyMailCgi/secret.py
/home/crew/lutz/public_html/PyMailCgi/secret.py
```

К сожалению, функция `os.path.samefile` поддерживается в Unix, Linux и Macs, но не поддерживается в Windows. Чтобы имитировать ее действие в Windows, мы разворачиваем пути к файлам до абсолютных путей, приводим их к общему регистру символов и сравниваем (в следующем

примере я сократил пути, заменив их части троеточием ..., чтобы уместить по ширине страницы):

```
>>> import os
>>> os.path.samefile
AttributeError: 'module' object has no attribute 'samefile'
>>> os.getcwd()
'C:\\...\\PP4E\\dev\\Examples\\PP4E\\Internet\\Web'
>>>
>>> x = os.path.abspath('../Web/PYMailCgi/cgi-bin/secret.py').lower()
>>> y = os.path.abspath('PyMailCgi/cgi-bin/secret.py').lower()
>>> z = os.path.abspath('./PYMailCGI/cgi-bin/../cgi-bin/SECRET.py').lower()
>>> x
'c:\\...\\dev\\examples\\pp4e\\internet\\web\\pymailcgi\\cgi-bin\\secret.py'
>>> y
'c:\\...\\dev\\examples\\pp4e\\internet\\web\\pymailcgi\\cgi-bin\\secret.py'
>>> z
'c:\\...\\dev\\examples\\pp4e\\internet\\web\\pymailcgi\\cgi-bin\\secret.py'
>>>
>>> x == y, y == z
(True, True)
```

Попытка обращения по любому из этих путей влечет вывод страницы с сообщением об ошибке, как показано на рис. 15.31. Обратите внимание, что имена закрытых файлов хранятся в этом модуле как глобальные данные и предполагается, что они принадлежат к числу файлов, доступных для всего сайта. Мы могли бы хранить этот список в файле с настройками сайта, однако внесение изменений в сценарии в глобальные переменные для каждого сайта ничем особенно не отличается от внесения изменений в файлы с настройками сайта.

Рис. 15.31. *Результат попытки доступа к закрытому файлу*

Заметьте также, что настоящие ошибки доступа к файлам обрабатываются иначе. Проблемы нехватки прав доступа и попытки доступа к несуществующим файлам, например, перехватываются различными обработчиками исключений, которые выводят сообщения (извлекаемые

с помощью функции sys.exc_info), передавая дополнительный контекст. На рис. 15.32 показана такая страница с сообщением об ошибке.

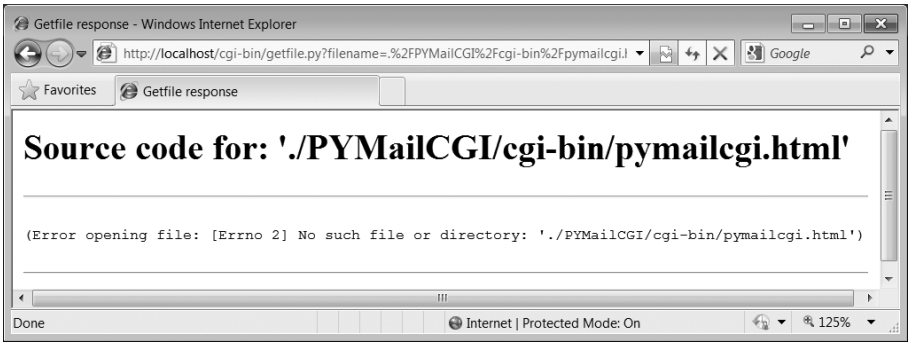


Рис. 15.32. *Вывод ошибки доступа к файлу*

Общее практическое правило требует подробно сообщать об исключительных ситуациях, возникших в процессе обработки файлов, особенно во время отладки сценариев. При перехвате таких исключений в сценарии программист должен позаботиться о выводе подробностей (перенаправление потока вывода sys.stderr в sys.stdout здесь не поможет, так как интерпретатор в данном случае не выводит сообщение об ошибке). Объекты с типом текущего исключения, данными и трассировочной информацией всегда доступны в модуле sys и могут быть выведены вручную.

Не устанавливайте сценарий *getfile.py*, если вы действительно хотите сохранить в тайне содержимое своих файлов! Проверка по списку закрытых файлов помешает непосредственно просмотреть модуль шифрования с помощью этого сценария, но он может не учитывать некоторые дополнительные возможности, особенно в Windows. Безопасность не является темой этой книги, поэтому я не стану углубляться в ее обсуждение, отмечу только, что некоторая степень паранойи в Интернете не помешает. Вы всегда должны исходить из того, что в конечном счете осуществится худший из возможных сценариев, особенно при установке систем в общедоступном Интернете.

Выгрузка файлов клиента на сервер

Сценарий getfile позволяет клиенту просматривать файлы, находящиеся на сервере, но в некотором смысле он может считаться инструментом общего назначения для загрузки файлов с сервера. Хотя и не так прямо, как при получении файлов по FTP или непосредственно через сокеты, но он служит аналогичным целям. Пользователи сценария могут копировать отображаемый код с веб-страницы или через пункт меню броузера View Source (Исходный код страницы или Просмотр HTML-кода).

Как уже описывалось выше, сценарии, контактируя со сценарием getfile с помощью модуля urllib, также способны извлекать содержимое файлов с помощью модуля Python, реализующего синтаксический анализ HTML.

Но как быть с движением в обратном направлении – выгрузкой файла с компьютера клиента на сервер? Например, предположим, что вы пишете систему электронной почты с веб-интерфейсом и вам необходимо реализовать способ, с помощью которого пользователи могли бы выгружать файлы вложений. В целом это вполне возможная ситуация – в следующей главе, где мы будем разрабатывать веб-приложение электронной почты PyMailCGI, будет представлена фактическая реализация этой идеи.

В главе 13 мы видели, что выгрузка достаточно просто осуществляется сценарием, выполняемым на стороне клиента и использующим модуль Python поддержки FTP. Однако такое решение не применимо в контексте веб-броузера – вряд ли вы станете предлагать всем клиентам программы запустить сценарий Python для FTP в другом окне, чтобы отправить файл. Кроме того, нет простого способа явно запросить отправку файла в сценарии на стороне сервера, если только на компьютере клиента не работает сервер FTP (что случается отнюдь не часто). Пользователи могут посылать файлы по почте отдельно, но это может быть неудобно, особенно в случае вложений в электронные письма.

Так есть ли способ написать веб-приложение, которое разрешает своим пользователям отправлять файлы на общий сервер? На самом деле есть, хотя он имеет большее отношение к HTML, чем собственно к Python. Теги HTML <input> поддерживают тип type=file, с помощью которого создается поле ввода вместе с кнопкой, щелчок на которой выводит диалог выбора файла. Имя файла на компьютере клиента, который нужно выгрузить на сервер, можно ввести в это поле или выбрать с помощью диалога. Чтобы продемонстрировать это, в примере 15.29 приводится файл страницы HTML, которая позволяет выбрать любой файл на стороне клиента и выгрузить его с помощью серверного сценария, имя которого указано в параметре action формы.

Пример 15.29. PP4E\Internet\Web\putfile.html

```
<html><title>Putfile: upload page</title>
<body>
<form enctype="multipart/form-data"
      method=post
      action="cgi-bin/putfile.py">
    <h1>Select client file to be uploaded</h1>
    <p><input type=file size=50 name=clientfile>
    <p><input type=submit value=Upload>
</form>
<hr><a href="cgi-bin/getfile.py?filename=cgi-bin\putfile.py">View script
code</a>
</body></html>
```

Обратите внимание на одно ограничение: формы, использующие поля ввода с типом `file`, должны указывать тип кодировки `multipart/form-data` и метод передачи `post`, как и сделано в этом файле. Адреса URL в стиле `get` не годятся для выгрузки файлов на сервер. При обращении к этому файлу HTML отображается страница, изображенная на рис. 15.33. Нажатие кнопки Browse (Обзор) открывает диалог выбора файла, а нажатие кнопки Upload (Выгрузить) – отсылает выбранный файл.

Рис. 15.33. *Страница выбора файла для отправки*

Когда на стороне клиента на этой странице нажимается кнопка Upload (Выгрузить), броузер открывает выбранный файл, читает его и упаковывает его содержимое вместе с остальными полями ввода формы (если они есть). Когда эти данные попадают на сервер, выполняется сценарий Python, указанный в параметре `action` формы и представленный в примере 15.30.

Пример 15.30. PP4E\Internet\Web\cgi-bin\putfile.py

```
#!/usr/bin/python
"""
###############################################################################
извлекает файл, выгруженный веб-броузером по протоколу HTTP; пользователи
открывают страницу putfile.html, чтобы получить страницу с формой выгрузки,
которая затем запускает этот сценарий на сервере; способ очень мощный
и очень опасный: обычно желательно проверять имя файла и так далее; выгрузка
возможна, только если файл или каталог доступен для записи: команды Unix
'chmod 777 uploads' может оказаться достаточно; путь к файлу поступает
в формате пути на стороне клиента: его требуется обработать здесь;

предупреждение: выходной файл можно было бы открыть в текстовом режиме,
чтобы подставить символы конца строки, используемые на принимающей
```

```
платформе, поскольку содержимое файла всегда возвращается модулем cgi
в виде строки str, но это временное решение - модуль cgi в 3.1 вообще
не поддерживает выгрузку двоичных файлов;
##############################################################################
"""

import cgi, os, sys
import posixpath, ntpath, macpath   # для обработки клиентских путей
debugmode = False                   # True=вывод данных формы
loadtextauto = False                # True=читать файл целиком
uploaddir = './uploads'             # каталог для сохранения файлов

sys.stderr = sys.stdout             # для вывода ошибок
form = cgi.FieldStorage()           # выполнить анализ данных формы
print("Content-type: text/html\n")  # с пустой строкой
if debugmode: cgi.print_form(form)  # вывести поля формы

# шаблоны html

html = """
<html><title>Putfile response page</title>
<body>
<h1>Putfile response page</h1>
%s
</body></html>"""

goodhtml = html % """
<p>Your file, '%s', has been saved on the server as '%s'.
<p>An echo of the file's contents received and saved appears below.
</p><hr>
<p><pre>%s</pre>
</p><hr>
"""

# обработать данные формы

def splitpath(origpath):            # получить имя файла без пути
    for pathmodule in [posixpath, ntpath, macpath]: # проверить все типы
        basename = pathmodule.split(origpath)[1]   # сервер может быть любой
        if basename != origpath:
            return basename                     # пробелы допустимы
    return origpath                             # неудача или нет каталога

def saveonserver(fileinfo):                     # поле с именем файла data
    basename = splitpath(fileinfo.filename)     # имя без пути
    srvrname = os.path.join(uploaddir, basename)  # в каталог, если указан
    srvrfile = open(srvrname, 'wb')             # всегда в двоичном режиме
    if loadtextauto:
        filetext = fileinfo.value               # прочитать текст в строку
        if isinstance(filetext, str):           # прием для Python 3.1
            filedata = filetext.encode()
```

```
        srvrfile.write(filedata)                # сохранить на сервере
    else:                                       # иначе читать построчно
        numlines, filetext = 0, ''              # напр. для больших файлов
        while True:                             # содержимое всегда str
            line = fileinfo.file.readline()     # или цикл for и итератор
            if not line: break
            if isinstance(line, str):           # прием для Python 3.1
                line = line.encode()
            srvrfile.write(line)
            filetext += line.decode()           # то же самое
            numlines += 1
        filetext = ('[Lines=%d]\n' % numlines) + filetext
    srvrfile.close()
    os.chmod(srvrname, 0o666)      # разрешить запись: владелец 'nobody'
    return filetext, srvrname

def main():
    if not 'clientfile' in form:
        print(html % 'Error: no file was received')
    elif not form['clientfile'].filename:
        print(html % 'Error: filename is missing')
    else:
        fileinfo = form['clientfile']
        try:
            filetext, srvrname = saveonserver(fileinfo)
        except:
            errmsg = '<h2>Error</h2><p>%s<p>%s' % tuple(sys.exc_info()[:2])
            print(html % errmsg)
        else:
            print(goodhtml % (cgi.escape(fileinfo.filename),
                              cgi.escape(srvrname),
                              cgi.escape(filetext)))
main()
```

В этом сценарии для обработки выгружаемых на сервер файлов применены интерфейсы, свойственные Python. На самом деле в этом нет ничего нового: файл поступает сценарию в виде записи в объекте, возвращаемом конструктором `cgi.FieldStorage` в результате анализа формы, как обычно; ключом является `clientfile`, определяемый значением атрибута `name` поля ввода в странице HTML.

Однако на этот раз запись имеет дополнительные атрибуты в виде имени этого файла на стороне клиента. Кроме того, при обращении к атрибуту `value` входного объекта выгруженного файла содержимое файла автоматически целиком считывается в строку на сервере. Если файл очень большой, его можно читать построчно (или блоками байтов заданного размера). Внутренняя реализация модуля Python `cgi` автоматически сохраняет выгруженный файл во временном файле – наш сценарий просто выполняет чтение этого временного файла. Однако, если файл слишком велик, он может не уместиться целиком в памяти.

Для иллюстрации сценарий реализует обе схемы: в зависимости от значения глобальной переменной `loadtextauto` он либо запрашивает содержимое файла как строку, либо читает его построчно. Вообще говоря, для полей выгружаемых файлов модуль CGI возвращает объекты со следующими атрибутами:

`filename`

Имя файла, как оно указано на стороне клиента

`file`

Объект файла, из которого может быть прочитано содержимое файла, выгруженного на сервер

`value`

Содержимое выгруженного файла (читаемое из атрибута `file`)

Существуют дополнительные атрибуты, не используемые в нашем сценарии. Файлы представляют собой третий тип объектов полей ввода. Мы уже видели, что атрибут `value` является *строкой* для простых полей ввода, а для элементов с множественным выбором может быть получен *список* объектов.

Для сохранения выгруженных файлов на сервере сценарии CGI (выполняемые с привилегиями пользователя «nobody») должны иметь право записи в каталог, если файл еще не существует, или в сам файл, если он уже имеется. Чтобы изолировать выгружаемые файлы, сценарий записывает их в тот каталог, который указан в глобальной переменной `uploaddir`. На сервере Linux моего сайта я должен был с помощью команды `chmod` установить для этого каталога биты разрешений 777 (все права чтения/записи/выполнения), чтобы заставить загрузку работать вообще. Для локального веб-сервера, используемого в этой главе, это не является проблемой, но у вас может быть иная ситуация, поэтому, если данный сценарий откажется работать, проверьте права доступа.

В этом сценарии также вызывается функция `os.chmod`, чтобы установить для файла на сервере права доступа, в соответствии с которыми его чтение и запись могут осуществлять все пользователи. Если при выгрузке файл создается заново, его владельцем станет пользователь «nobody», что означает возможность просмотра файла и загрузки на сервер для всех пользователей в киберпространстве. Однако на моем сервере Linux этот файл будет также по умолчанию доступен для записи только пользователю «nobody», что может оказаться неудобным, когда потребуется изменить этот файл непосредственно на сервере (степень неудобства может различаться в зависимости от операций).

Изоляция файлов, выгружаемых клиентами, путем помещения их в отдельный каталог на сервере способствует уменьшению угрозы безопасности: исключается возможность произвольного затирания существующих файлов. Но при этом может потребоваться копирование файлов на сервере после их выгрузки и не устраняются все

угрозы безопасности – злонамеренные клиенты сохраняют возможность выгрузки очень больших файлов; для отлавливания этого нужна дополнительная логика, отсутствующая в данном сценарии. В целом, такие ловушки могут понадобиться только в сценариях, открытых для Интернета.

Если клиент и сервер оба выполнят свои роли, то сценарий CGI, после того как сохранит содержимое файла клиента в новом или существующем уже файле на сервере, вернет страницу, изображенную на рис. 15.34. Для контроля в странице ответа указываются пути к файлу на стороне клиента и сервера, а также повторяется содержимое выгруженного файла вместе со счетчиком строк, если чтение производилось в построчном режиме.

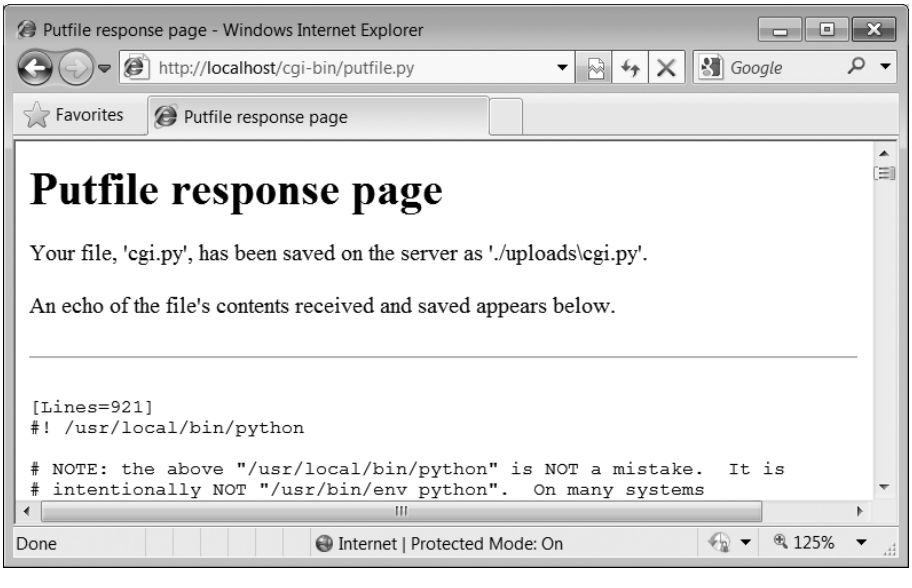


Рис. 15.34. *Ответная страница сценария putfile*

Обратите внимание, что вывод файла в странице ответа предполагает, что его содержимым является текст. Как оказывается, это вполне безопасное предположение, потому что модуль `cgi` всегда возвращает содержимое файла в виде строки `str`, а не `bytes`. Это проистекает из несколько неприятного факта, что в версии 3.1 модуль `cgi` вообще не поддерживает выгрузку двоичных файлов (подробнее об этом ограничении рассказывается во врезке ниже).

Этот файл, выгруженный и сохраненный в отдельном каталоге, идентичен оригинальному файлу (выполните команду `fc` в Windows, чтобы убедиться в этом). Кстати, проверить выгруженный файл можно также с помощью программы `getfile`, написанной в предыдущем разделе.

Просто зайдите на страницу выбора файла и введите путь и имя файла на сервере, как показано на рис. 15.35.

Рис. 15.35. *Проверка результатов работы putfile с помощью getfile – выбор файла*

Если выгрузка файла на сервер прошла успешно, то будет получена результирующая страница, изображенная на рис. 15.36. Поскольку пользователь «nobody» (сценарии CGI) смог записать файл, «nobody» должен быть в состоянии и прочесть его.

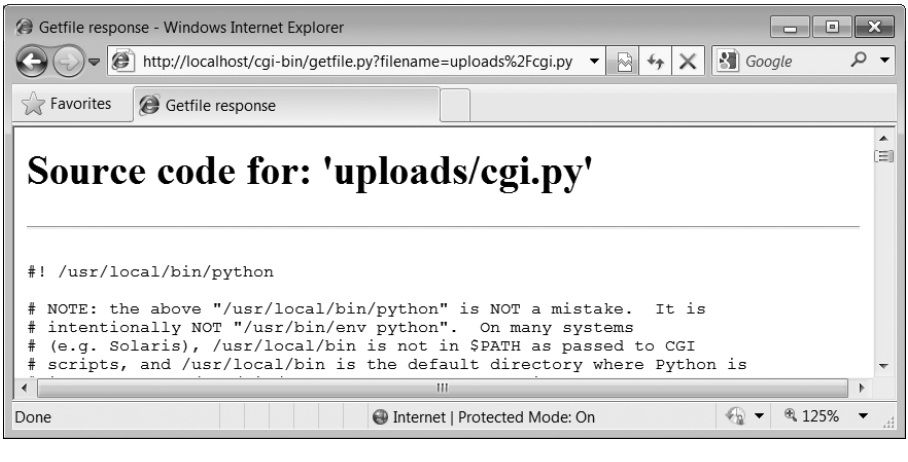


Рис. 15.36. *Проверка результатов работы putfile с помощью getfile – ответ*

Обратите внимание на адрес URL в адресной строке броузера для этой страницы: броузер преобразовал символ /, введенный на странице выбора файла, в шестнадцатеричную экранированную последовательность %2F перед тем, как поместить его в конец URL в качестве параметра. С экранированными последовательностями URL мы познакомились выше в этой главе. В данном случае преобразование осуществил броузер, но конечный результат получился таким же, как если бы мы вручную применили к строке пути одну из функций экранирования из модуля urllib.parse.

Технически экранированная последовательность %2F представляет здесь стандартное преобразование URL для символов, не входящих в диапазон ASCII, в стандартной схеме кодирования, используемой броузерами. Пробелы обычно также транслируются в символы +. Часто можно обойтись без преобразования вручную большинства символов, не входящих в диапазон ASCII, при явной отправке путей (во вводимых адресах URL). Но как мы видели раньше, иногда нужно следить за тем, чтобы экранировать символы, имеющие в строках URL особое значение (например, &), с помощью инструментов в модуле urllib.parse.

Обработка формата пути, используемого клиентом

В конце своей работы сценарий *putfile.py* сохраняет выгруженный файл на сервер в жестко определенном каталоге *uploaddir* под тем именем, которое задано в конце пути к файлу на стороне клиента (то есть без пути каталога на стороне клиента). Обратите, однако, внимание, что функции splitpath в этом сценарии приходится делать дополнительную работу для извлечения базового имени файла в правой части. Броузеры могут посылать имя файла в том формате пути каталога, который используется на компьютере *клиента*. Этот формат пути может не совпадать с тем, который используется на сервере, где выполняется сценарий CGI. Разные броузеры могут действовать по-разному, но об этой проблеме следует помнить, чтобы обеспечить переносимость.

Стандартный инструмент разделения путей на составные части, функция os.path.split, умеет извлекать базовое имя файла, но распознает только символы-разделители пути, которые используются на платформе, где она выполняется. Это означает, что если выполнить этот сценарий CGI в системе Unix, функция os.path.split отрубит путь к каталогу от базового имени по разделителю /. Однако если пользователь выгружает файл из DOS или Windows, то разделителем в переданном имени файла будет \, а не /. Броузеры, работающие в некоторых версиях Macintosh, могут отправлять еще более отличающиеся пути.

Чтобы обеспечить универсальную обработку путей в формате клиентов, этот сценарий импортирует из библиотеки Python специфические для платформ модули обработки путей для каждого клиента, который должен поддерживаться, и пытается отделить путь поочередно с помощью каждого из них, пока не будет обнаружено имя файла в правой

части. Например, posixpath обрабатывает пути, используемые в системах Unix, а ntpath распознает пути клиентов DOS и Windows. Обычно мы не импортируем эти модули напрямую, поскольку os.path.split автоматически загружает тот из них, который соответствует платформе, где действует сервер; однако в данном случае нам необходимо использовать специализированные версии, поскольку путь поступает с другого компьютера. Обратите внимание, что можно было бы запрограммировать логику разделения пути иным образом, чтобы избежать нескольких вызовов split:

```
def splitpath(origpath):                          # получить имя в конце
    basename = os.path.split(origpath)[1]         # формат платформы сервера
    if basename == origpath:                      # без изменений?
        if '\\' in origpath:
            basename = origpath.split('\\')[-1] # формат DOS
        elif '/' in origpath:
            basename = origpath.split('/')[-1]  # формат Unix
    return basename
```

Но эта альтернативная версия может терпеть неудачу для некоторых форматов путей (например, для пути DOS с именем диска, но без символов обратного слэша). В настоящем виде в обоих вариантах попусту тратится время, если имя файла уже является базовым (то есть не содержит слева пути к каталогу), но в целом мы должны учитывать более сложные случаи.

Этот сценарий выгрузки файлов на сервер работает так, как задумано, но следует сделать некоторые замечания, прежде чем перевернуть страницу после этого примера:

- Во-первых, сценарий putfile никак не учитывает несовместимости, существующие между разными платформами в самих именах файлов. Например, пробелы в имени файла, передаваемом клиентом DOS, не преобразуются в другие символы – они остаются пробелами в имени файла на стороне сервера, что может быть допустимо, но в некоторых ситуациях затрудняет обработку.
- Во-вторых, предусмотренная процедура построчного чтения означает, что этот сценарий ориентирован на выгрузку текстовых файлов. Он использует режим wb, чтобы сохранить содержимое выгруженного файла в двоичном режиме, но в других местах предполагается, что данные являются текстом, включая создание страницы ответа. Подробнее о двоичных файлах рассказывается в главе 4. Это самый спорный пункт в Python 3.1, так как выгрузка двоичных файлов вообще не поддерживается (смотрите врезку «Ограничения на выгрузку файлов модели CGI в версии 3.1»). Однако в будущих версиях эта проблема, возможно, будет решена.

Если вы столкнетесь с каким-то из этих ограничений, их преодоление я предлагаю рассматривать, как упражнение для самостоятельного решения.

Ограничения на выгрузку файлов модели CGI в версии 3.1

К сожалению, я должен особо отметить тот факт, что поддержка выгрузки файлов в модели CGI, реализованной в стандартной библиотеке, была частично нарушена в Python 3.1. Если говорить кратко, в настоящее время внутренняя реализация модуля `cgi` возбуждает исключение, если попытаться выгрузить на сервер двоичные данные или текст в несовместимой кодировке. Это исключение возникает еще до того, как сценарий получит возможность вмешаться, что делает невозможной реализацию обходных решений. Выгрузка файлов в модели CGI в версии Python 2.X действовала потому, что строки могли обрабатывать двоичные данные, но в современной версии 3.X возникает ошибка.

Отчасти данная потеря функциональности проистекает из того факта, что модуль `cgi` использует механизм анализа, реализованный в пакете `email`, для извлечения входных данных, состоящих из нескольких частей, и потому испытывает те же самые проблемы, свойственные пакету `email`, о которых подробно рассказывалось в главе 13. Механизм анализа в пакете `email` ожидает получить полный текст сообщения в виде строки `str`, что может приводить к ошибкам в реализации поддержки модели CGI, при выгрузке некоторых типов данных. Как отмечалось в главе 13, при выгрузке файлов через CGI они могут содержать смесь текстовых и двоичных данных – включая исключительно двоичные данные, *не* преобразованные в формат MIME, текст в любой возможной кодировке и произвольные их комбинации. Требование текущей версии пакета `email` декодировать эти данные в строку `str` для анализа является совершенно несовместимым, хотя, похоже, что и сам модуль `cgi` в некоторых случаях действует не совсем корректно.

Если у вас появится желание своими глазами увидеть, в каком виде данные выгружаются броузерами, откройте и попробуйте выполнить файлы HTML и Python с именами *test-cgiu-uploads-bug**, включенные в пакет примеров, и с их помощью выгрузить текст, двоичные и смешанные данные:

- *test-cgi-uploads-bug.html/py* пытается выполнить анализ обычным способом, который дает положительные результаты для некоторых типов текстовых файлов, но всегда терпит неудачу при попытке декодировать двоичные файлы
- *test-cgi-uploads-bug0.html/py* пытается использовать двоичный режим при работе с входным потоком, но всегда терпит неудачу с ошибкой, сообщающей о недопустимом типе, для обоих типов файлов, текстовых и двоичных, из-за того, что пакет `email` ожидает получить данные в виде строки `str`

- *test-cgi-uploads-bug1.html/py* сохраняет входной поток с единственным файлом
- *test-cgi-uploads-bug.html/py* сохраняет входной поток с несколькими файлами

Последние два сценария в этом списке просто читают и сохраняют данные в двоичном режиме в файл для дальнейшего просмотра и отображают два заголовка, переданных в переменных окружения, которые используются при анализе (тип содержимого «multipart/form-data» и объем данных). Попытка проанализировать сохраняемые входные данные с помощью модуля `cgi` будет терпеть неудачу, если только это не будут текстовые данные в совместимой кодировке. В действительности, из-за того, что данные могут представлять собой произвольную смесь текста и двоичных данных, корректный механизм анализа должен будет читать их в двоичном режиме и свободно переключаться между текстовым и двоичным режимами в процессе анализа.

Весьма вероятно, что такое положение дел улучшится в будущем, но, скорее всего, не ранее, чем в версии Python 3.3 или выше. Даже спустя почти два года после выхода версии 3.0 эта книга играет роль бета-тестера гораздо чаще, чем могла бы. В первую очередь это обусловлено тем, что разделение типов `str/bytes` в Python 3.X не получило полной реализации в стандартной библиотеке Python к моменту появления этой версии. Конечно, это не означает, что нужно относиться с пренебрежением к тем, кто тратил свое время и силы на создание версии 3.X. Тем не менее, для меня, принадлежащего к числу тех, кто помнит еще версию 0.X, эта ситуация выглядит далекой от идеала.

Разработка программного кода взамен модуля `cgi` и используемого им пакета `email` – единственное настоящее решение всех проблем – практически невозможная задача для этой книги. Поэтому пока сценарии CGI, выполняющие выгрузку файлов и реализованные в этой книге, будут работать только с текстовыми файлами и только с текстом в совместимой кодировке. Это распространяется и на вложения электронной почты, выгружаемые с помощью веб-приложения электронной почты PyMailCGI, рассматриваемого в следующей главе, – еще одна причина, почему в этом издании этот пример не был дополнен новыми функциональными возможностями, как это произошло с приложением PyMailGUI из предыдущей главы. Отсутствие возможности прикреплять таким способом изображения к электронным письмам – это серьезное ограничение, которое сужает область применения в целом.

Обновления, которые, возможно, в будущем справятся с этой проблемой, ищите на веб-сайте книги (описывается в предисловии). Исправления, вероятно, окажутся несовместимыми с текущим прикладным интерфейсом библиотечных модулей, но такова действительность в мире разработки программного обеспечения, за исключением случаев, когда новая система пишется с чистого листа. (А «бегство» – это не выход...)

Как же все-таки протолкнуть биты через Сеть

В заключение обсудим некоторый контекст. К этому моменту мы уже увидели три сценария `getfile`. Тот, который приведен в этой главе, отличается от двух других, написанных в более ранних главах, но решает сходную задачу:

- Версия сценария `getfile` в этой главе является сценарием CGI, который выполняется на стороне сервера и выводит файлы по протоколу HTTP (через порт 80).
- В главе 12 мы создали клиентскую и серверную части `getfile` для передачи файлов через простые сокеты (через порт 50001).
- В главе 13 мы реализовали `getfile`, выполняющийся на стороне клиента, для передачи по протоколу FTP (через порт 21).

В действительности CGI-сценарий `getfile`, реализованный в этой главе, лишь отображает файлы, но может считаться инструментом загрузки, если дополнить его операциями копирования и вставки в веб-броузере. Кроме того, сценарий `putfile` из этой главы, основанный на CGI и HTTP, также отличается от сценария `putfile` из главы 13, использующего протокол FTP, но может рассматриваться как альтернатива обоим вариантам отправки, через сокеты и FTP.

Следует подчеркнуть, что есть много способов передачи файлов через Интернет – сокеты, FTP и HTTP (веб-страницы); все они могут перемещать файлы между компьютерами. С технической точки зрения, файлы можно пересылать и с помощью других технологий и протоколов – электронной почты POP, телеконференций NNTP, Telnet и так далее.

У каждой технологии есть свои уникальные особенности, но в итоге они делают одно дело: перемещают биты через Сеть. Все они, в конечном счете, используют сокеты и определенные порты, но такие протоколы, как FTP и HTTP, создают дополнительную структуру над слоем сокетов, а прикладные модели, такие как CGI, создают как структуру, так и возможность программирования.

В следующей главе мы используем полученные здесь знания при создании более существенного приложения, целиком выполняющегося

в Веб, – PyMailCGI, веб-приложения электронной почты, позволяющего отправлять и просматривать электронные письма в броузере, обрабатывать вложения и многое другое. Однако, в конечном счете, все сводится к передаче байтов через сокеты с помощью пользовательского интерфейса.

Загрузка через CGI: принудительное решение

В примере 15.27 мы написали сценарий с именем *getfile.py* – CGI-программу на языке Python, предназначенную для отображения в веб-броузере (или в других программах) любых общедоступных файлов, находящихся на сервере, по запросу с компьютера клиента. Для корректного отображения текстовых файлов в окне броузера он использует заголовок «Content-type» со значением `text/plain` или `text/html`. В описании мы сравнили *getfile.py* с обобщенным CGI-инструментом загрузки, дополненным возможностью копирования или сохранения.

Хотя это и верно, тем не менее, сценарий *getfile.py* в первую очередь задумывался как инструмент отображения содержимого файлов, а не как демонстрация возможности загрузки файлов через CGI. Если вам действительно потребуется реализовать непосредственную загрузку файлов через CGI (вместо того, чтобы отображать его содержимое в броузере или открывать в другом приложении), вы можете заставить броузер на стороне клиента вывести диалог сохранения файла, добавив соответствующий заголовок «Content-type» в ответе сценария CGI.

Что делать с файлом, броузеры решают исходя из расширения в имени файла (например, файл с именем *xxx.jpg* интерпретируется как изображение) или из значения заголовка «Content-type» (например, значение `text/html` означает файл с разметкой HTML). Используя различные значения MIME в заголовке, можно указать неизвестный броузеру тип данных и тем самым заставить его отказаться от обработки данных. Например, значение `application/octet-stream` в заголовке «Content-type» ответа вынудит броузер вывести диалог сохранения файла.

Тем не менее такой подход иногда осуждается, потому что он прячет истинную природу файла, – обычно лучше позволить пользователю/клиенту самому решать, как обрабатывать загружаемые данные, чем принудительно выводить диалог сохранения. Кроме того, этот прием практически не имеет отношения к языку Python – за дополнительной информацией обращайтесь к книгам о CGI или попробуйте выполнить поиск в Веб по фразе «CGI download».

16

Сервер PyMailCGI

«Список дел на поездку в Чикаго»

Эта глава – пятая в нашем обзоре интернет-программирования на языке Python. В ней продолжается обсуждение, начатое в главе 15. Там мы исследовали основы создания серверных CGI-сценариев на языке Python. Вооружившись полученными знаниями, в этой главе мы перейдем к большому конкретному примеру, в котором делается акцент на более сложные темы CGI.

Эта глава представляет PyMailCGI – веб-приложение для чтения и отправки электронной почты, иллюстрирующее концепции системы безопасности, скрытые поля форм, генерацию адресов URL и многое другое. Поскольку эта система близка по духу программе PyMailGUI, представленной в главе 14, этот пример позволяет также сравнить веб-приложения с обычными. Этот пример основан на использовании сценариев CGI и реализует законченный веб-сайт, имеющий более высокую практическую ценность, чем примеры в главе 15.

Как обычно, в этой главе внимание фокусируется на деталях прикладного уровня, а также на принципах программирования на языке Python. Поскольку это полновесный практический пример, он демонстрирует концепции проектирования систем, важные для реальных проектов. Он также более полно представляет сценарии CGI в целом: PyMailCGI знакомит с понятиями сохранения информации о состоянии, а также с соображениями безопасности и шифрованием.

Система, представленная здесь, не отличается особенной броскостью или богатством функций, встречающимися на сайтах (по правде говоря, первоначальный набросок PyMailCGI был сделан во время задержки в аэропорту Чикаго). Увы, вы не найдете на сайте ни танцующих

медвежат, ни мерцающих огней. С другой стороны, система была написана для практического использования, более широкого освещения сценариев CGI и чтобы показать, чего можно достичь с помощью программ Python, выполняемых на сервере. Как отмечалось в начале этой части книги, существует множество высокоуровневых фреймворков, систем и библиотек инструментов, построенных на идеях, которые мы применяем здесь. Поэтому посмотрим, что нам может предложить Python, примененный в Веб.

Веб-сайт PyMailCGI

В главе 14 мы создали программу под названием PyMailGUI, в которой с помощью Python+tkinter был реализован законченный почтовый клиент с графическим интерфейсом (если вы не читали соответствующую главу, вы можете бегло просмотреть ее сейчас). Здесь мы собираемся сделать нечто в том же роде, но в Веб: представленная в данном разделе система PyMailCGI является совокупностью сценариев CGI, реализующих простой веб-интерфейс для отправки и чтения электронной почты в любом броузере. В результате мы получим веб-приложение электронной почты – пусть оно и не такое мощное, как можно встретить у провайдеров услуг Интернета (ISP), но возможность вносить в него изменения дает вам полный контроль над его возможностями и дальнейшим развитием.

При изучении этой системы наша задача отчасти состоит в том, чтобы научиться еще нескольким приемам CGI, отчасти – в том, чтобы получить некоторые сведения о проектировании крупных систем Python в целом, и частично – в том, чтобы проанализировать соотношение преимуществ и недостатков между веб-приложениями (PyMailCGI) и обычными приложениями, разработанными для выполнения на локальном компьютере (PyMailGUI). Мы будем проводить сравнение приложений по мере продвижения по главе, а после знакомства со всей системой вернемся к нему для более глубокого рассмотрения.

Обзор реализации

На самом верхнем уровне PyMailCGI разрешает пользователям просматривать входящую почту с помощью интерфейса POP и отправлять новую почту по протоколу SMTP. Пользователям также предоставлена возможность составить ответ, переслать или удалить входящую почту во время ее просмотра. В данной реализации отправить электронное письмо с сайта PyMailCGI может всякий, но для просмотра почты, как правило, нужно установить PyMailCGI на своем собственном локальном компьютере или на сайте вместе с информацией о своем собственном почтовом сервере (из-за соображений безопасности, излагаемых ниже).

Просмотр и отправка почты выглядят достаточно просто, и мы уже несколько раз реализовывали эти возможности. Но в процессе взаимодей-

ствия участвует несколько отдельных страниц, для каждой из которых требуется отдельный сценарий CGI или файл HTML. На самом деле PyMailCGI представляет собой довольно *линейную* систему – в самом сложном варианте схемы взаимодействия с пользователем взаимодействие проходит от начала до конца шесть состояний (и потому создано шесть веб-страниц). Так как каждая страница в мире CGI обычно генерируется отдельным файлом, это предполагает также наличие шести файлов с исходным программным кодом.

Технически, PyMailCGI можно было бы описать как *конечный автомат*, хотя при переходе от одного состояния к другому передается очень немного информации о состоянии. Сценарии передают друг другу имя пользователя и информацию о сообщении в скрытых полях форм и в параметрах запросов; в текущей версии не используются cookies, сохраняемые на стороне клиента, и базы данных на стороне сервера. В этой разработке мы столкнемся с ситуациями, когда более совершенные механизмы сохранения информации о состоянии могли бы дать дополнительные преимущества.

Чтобы помочь уследить за тем, как все файлы PyMailCGI вписываются в общую систему, перед началом практического программирования я написал файл, представленный в примере 16.1. В нем сделан неформальный набросок прохождения пользователя через систему и вызываемых при этом файлов. Разумеется, для описания передачи управления и информации через состояния можно использовать более формальные обозначения, такие как веб-страницы (например, диаграммы потоков данных), но для данного простого примера этого файла достаточно.

Пример 16.1. PP4E\Internet\Web\PyMailCgi\pageflow.txt

```
файл или сценарий                          создает
-----------------                          -------

[pymailcgi.html]                           Главная страница
 => [onRootViewLink.py]                    Ввод пароля
     => [onViewPswdSubmit.py]              Список (загружает всю почту)
         => [onViewListLink.py]            Просмотр+выбор=del|reply|fwd
                                           (загрузка)
             => [onViewPageAction.py]      Редактор или удалить+
                                           подтвердить (del)
                 => [onEditPageSend.py]    Подтверждение (отправка почты)
                     => обратно к главной странице
 => [onRootSendLink.py]                    Редактор
     => [onEditPageSend.py]                Подтверждение (отправка почты)
         => обратно к главной странице
```

В этом файле просто перечислены все файлы с исходными текстами системы, где запускаемые ими сценарии отмечены символами => и отступами.

Например, ссылки на главной странице `pymailcgi.html` вызывают сценарии *onRootViewLink.py* и *onRootSendLink.py*. Сценарий *onRootViewLink.py* создает страницу для ввода пароля, нажатие на кнопку Submit (Отправить) запускает *onViewPswdSubmit.py* и так далее. Обратите внимание, что оба действия, просмотр и отправка почты, могут заканчиваться запуском сценария *onSendSubmit.py* для отправки новой почты – операции просмотра попадают в этот сценарий, если пользователь решит ответить на входящую почту или переслать ее.

В такой системе отдельные сценарии CGI несут немного смысла, поэтому полезно помнить общую взаимосвязь страниц. Обращайтесь к этому файлу, если потеряете нить изложения. Дополнительно обстановку проясняет рис. 16.1, на котором представлено общее содержание этого сайта при просмотре структуры каталогов в окне оболочки Cygwin в Windows.

```
/cygdrive/c/Users/mark/Stuff/Books/4E/PP4E/dev/Examples/PP4E/Internet/Web/PyMailCgi
mark@mark-VAIO /cygdrive/c/Users/mark/Stuff/Books/4E/PP4E/dev/Examples/PP4E/Inte
rnet/Web/PyMailCgi
$ ls --file-type
PythonPoweredSmall.gif  pageflow.txt    ppsmall.gif     sentmail.txt
README-imports.txt      partsdownload/  pymailcgi.html
cgi-bin/                partsupload/    sendurl.py

mark@mark-VAIO /cygdrive/c/Users/mark/Stuff/Books/4E/PP4E/dev/Examples/PP4E/Inte
rnet/Web/PyMailCgi
$ ls --file-type cgi-bin
commonhtml.py  mailconfig.py      onRootViewLink.py    onViewPswdSubmit.py
externs.py     onEditPageSend.py  onViewListLink.py    secret.py
loadmail.py    onRootSendLink.py  onViewPageAction.py

mark@mark-VAIO /cygdrive/c/Users/mark/Stuff/Books/4E/PP4E/dev/Examples/PP4E/Inte
rnet/Web/PyMailCgi
$ _
```

Рис. 16.1. *Содержание PyMailCgi*

Все файлы, которые вы здесь видите, были выгружены в ходе установки сайта в подкаталог *PyMailCgi*, находящийся в корневом веб-каталоге на сервере. Помимо файлов страниц HTML и сценариев CGI, вызываемых в ответ на действия пользователя, в PyMailCGI используется также несколько вспомогательных модулей:

`commonhtml.py`

Служит библиотекой инструментов HTML.

`externs.py`

Объединяет в себе доступ к модулям, импортируемым из других систем.

`loadmail.py`

Инкапсулирует загрузку из почтового ящика для расширения в будущем.

`secret.py`

Реализует шифрование пароля с возможностью настройки.

PyMailCGI также повторно использует части пакета `mailtools` и модуля `mailconfig.py`, созданных в главе 13. Первый из них доступен для импорта из корневого каталога *PP4E* пакета примеров, а второй является локальной версией, скопированной в каталог *PyMailCgi*, вследствие чего реализация этого модуля для PyMailGUI и PyMailCGI может отличаться. Модуль `externs.py` предназначен для сокрытия особенностей размещения модулей, на случай, если будет выбрана иная структура каталогов установки.

Эта система вновь демонстрирует практические выгоды повторного использования программного кода. Версия системы в этом издании импортирует значительный объем логики из нового пакета `mailtools`, представленного в главе 13, – загрузку сообщений, отправку, удаление, анализ, составление, кодирование и декодирование и добавление вложений – несмотря на то, что первоначально модули этого пакета разрабатывались для программы PyMailGUI. Когда позднее пришло время обновить PyMailCGI, инструменты для выполнения сложных операций, таких как добавление вложений и поиск текста в сообщениях, уже были готовы. Исходные тексты пакета `mailtools` вы найдете в главе 13.

Как обычно, PyMailCGI использует также ряд модулей из стандартной библиотеки Python: `smtplib`, `poplib`, `email.*`, `cgi`, `urllib.*` и другие. Благодаря повторному использованию нашего собственного и стандартного программного кода эта система позволяет сделать много минимальным объемом программного кода. В общей сложности PyMailCGI содержит всего 846 строк нового программного кода, включая пробельные строки, комментарии и главный файл HTML (более подробная информация приводится в файле *linecounts.xls*, в исходном каталоге системы; версия в предыдущем издании содержала около 835 новых строк).

Сравнение размеров говорит не в пользу «настольного» клиентского приложения PyMailGUI из главы 14, но различия в значительной степени объясняются более ограниченными функциональными возможностями PyMailCGI – здесь не используются локальные файлы для сохранения почты, не предусматривается перекрытие во времени потоков выполнения операций передачи, не предусматривается кэширование сообщений, не проверяется и не восстанавливается синхронизация с ящиком входящих сообщений, не предусматривается возможность выбора нескольких сообщений, отсутствует возможность просмотра исходного текста сообщений и так далее. Кроме того, как рассказывается в сле-

дующем разделе, в этой версии PyMailCGI используется более ограниченная политика работы с Юникодом, и, хотя просматривать можно любые вложения, отправка двоичных и некоторых текстовых вложений не поддерживается в текущей версии из-за проблем в Python 3.1.

Иными словами, PyMailCGI в действительности является чем-то вроде *прототипа*, предназначенного для иллюстрации концепций веб-программирования и проектирования систем в этой книге, который должен служить отправной точкой для дальнейшего развития. Система PyMailGUI продвинулась по эволюционной шкале программного обеспечения гораздо дальше. Однако мы увидим, что правильная организация программного кода PyMailCGI и повторное использование существующих модулей позволяют реализовать массу возможностей в небольшом объеме программного кода.

Новое в версии для четвертого издания (версия 3.0)

В четвертом издании веб-приложение PyMailCGI было адаптировано для работы под управлением Python 3.X. Кроме того, эта версия наследует и использует различные новые возможности, реализованные в пакете `mailtools`, включая декодирование и кодирование почтовых заголовков, кодирование текста сообщения, возможность ограничивать количество извлекаемых заголовков и многие другие. Особенно примечательной является новая реализация поддержки Юникода и национальных наборов символов:

- Перед отображением основной текст сообщения и его заголовки декодируются в соответствии со стандартами электронной почты, MIME и Юникода – текст декодируется с учетом информации в заголовках, а заголовки декодируются в соответствии с их содержимым.
- При необходимости перед отправкой выполняется кодирование основного текста сообщения, текстовых вложений и заголовков в соответствии с теми же стандартами, с использованием по умолчанию кодировки UTF-8.
- При создании ответов или при пересылке писем заголовки, копируемые в цитируемый текст сообщения, также декодируются для отображения.

Обратите внимание, что данная версия опирается на возможности веб-броузеров отображать текст Юникода в произвольных кодировках. Она не добавляет теги «meta» с объявлением кодировки в страницах HTML, генерируемых для просмотра и составления сообщений. Например, корректно сформированный документ HTML обычно содержит объявление кодировки, как показано ниже:

```
<HTML><HEAD>
<META http-equiv=Content-Type content="text/html; charset=windows-1251">
</HEAD>
```

Подобные заголовки здесь отсутствуют. Отчасти это обусловлено тем, что почтовые сообщения могут включать текст с произвольной кодировкой и даже фрагменты с отличающимися кодировками в самих сообщениях и в заголовках, которые могут быть несовместимы с кодировкой, указанной в странице ответа HTML. Представьте себе страницу, на которой отображается список заголовков нескольких сообщений – поскольку заголовки «Subject» и «From» могут содержать различные наборы символов (один – кириллические символы, другой – китайские иероглифы и так далее), объявления единой кодировки будет недостаточно (хотя часто ситуацию может спасти универсальная кодировка UTF-8). Разрешение конфликтов при отображении таких страниц, содержащих смешанные наборы символов, остается за броузером, который, в крайнем случае, может прибегнуть к помощи пользователя в выборе кодировки. В PyMailGUI отображение разнородных наборов символов в одном окне было возможно благодаря тому, что мы передавали уже декодированный текст Юникода текстовому виджету `Text` из библиотеки tkinter, который прекрасно справляется с произвольными кодовыми пунктами Юникода. В PyMailCGI мы в значительной степени будем обходить эту проблему, чтобы сохранить пример как можно более простым.

Кроме того, вложения обоих типов, текстовые и двоичные, просто сохраняются в двоичном виде и открываются броузером по имени файла, если щелкнуть на ссылках, ведущих к ним; при этом в выборе правильного решения мы опять же полагаемся на броузер. На текстовые вложения в отправляемые письма также накладываются ограничения, касающиеся выгрузки файлов через CGI, описанные во врезке ниже. Помимо всего перечисленного, похоже, что в Python 3.1 имеется проблема с выводом некоторых типов текста Юникода в стандартный поток вывода в контексте CGI, для преодоления которой необходимо использовать обходное решение в главном вспомогательном модуле – открывать `stdout` в двоичном режиме и выводить текст в виде кодированной строки `bytes` (дополнительные подробности вы найдете в исходном программном коде).

Поддержка Юникода/i18n здесь не такая обширная, как в PyMailGUI. Однако, учитывая, что здесь мы не можем запросить кодировку, и принимая во внимание нехватку времени и места в книге в целом, дальнейшее усовершенствование этой поддержки для различных случаев и броузеров мы оставляем в качестве самостоятельного упражнения.

Более конкретное описание изменений, внесенных в версию 3.0, в четвертом издании вы найдете в программном коде, в комментариях, начинающихся с номера версии «3.0». Кроме того, сохранились все возможности, добавленные в предыдущем издании. Они описываются в следующем разделе.

Ограничения на отправку вложений (в этом издании)

Если вы этого еще не сделали, прочитайте текст врезки «Ограничения на выгрузку файлов модели CGI в версии 3.1» в конце предыдущей главы. Если коротко, модуль `cgi` в Python 3.1, а также используемый им механизм синтаксического анализа в пакете `email`, возбуждает исключение при попытке выгрузить на сервер файл с двоичными данными или с текстом в несовместимой кодировке. К сожалению, поскольку система PyMailCGI, рассматриваемая в этой главе, использует для выгрузки вложений модель CGI, это ограничение означает, что данная система в настоящее время не поддерживает отправку почты с двоичными вложениями, такими как изображения и аудиофайлы. Такая поддержка была в версии для предыдущего издания, выполнявшейся под управлением Python 2.X.

Отправка таких вложений поддерживается обычным настольным приложением PyMailGUI, представленным в главе 14, по той простой причине, что оно имеет возможность напрямую читать данные из локальных файлов (используя двоичный режим, если это необходимо, и преобразовывая данные в формат MIME перед включением их в сообщение электронной почты). Однако, так как веб-приложение PyMailCGI, рассматриваемое здесь, опирается на получение файлов вложений, выгружаемых на сервере через CGI, оно полностью находится во власти ограничений, обусловленных нарушенной поддержкой выгрузки файлов в модуле `cgi`. Разработка другого модуля взамен `cgi` – слишком грандиозная задача для этой книги.

Исправление этой проблемы ожидается в будущем, и оно может оказаться выполненным к моменту, когда вы будете читать эти строки. Однако версия PyMailCGI в этом издании, опирающаяся на версию Python 3.1, просто не в состоянии обеспечить поддержку отправки таких вложений, хотя во входящих сообщениях они свободно могут просматриваться. Фактически, несмотря на то, что версия PyMailCGI в этом издании наследует некоторые новые возможности из пакета `mailtools`, такие как декодирование и кодирование интернационализированных заголовков, данное ограничение на отправку вложений оказывается серьезным препятствием для дальнейшего расширения возможностей этой системы до уровня версии PyMailGUI в этом же издании. Например, политика работы с Юникодом здесь чрезвычайно проста, если не примитивна.

За счет применения некоторых технологий на стороне клиента, таких как AJAX, вполне возможно реализовать отправку двоичных файлов вообще без использования модели CGI. Однако для этого потребовалось бы развертывать фреймворки и использовать технологии, не рассматривающиеся в этой книге, что подразумевало бы совершенно иную, более сложную организацию программы, которая бы избавляла от любой зависимости от ограничений в Python 3.X. Но прежде чем писать новую программу (PyMailRIA?), подождем окончательного вердикта относительно поддержки CGI в Python 3.X.

Новое в версии для предыдущего издания (версия 2.0)

В третьем издании в веб-приложении PyMailCGI был задействован новый пакет `mailtools`, представленный в главе 13; применялся для шифрования паролей пакет PyCrypto – в случае если он установлен; добавлена поддержка просмотра и отправки вложений и была повышена эффективность выполнения. Все эти улучшения были унаследованы версией 3.0.

Со всеми этими улучшениями мы еще столкнемся в процессе обсуждения, однако последние два заслуживают, чтобы сказать о них несколько слов перед тем, как двинуться дальше. Поддержка вложений реализована в достаточно упрощенном виде, но она вполне пригодна для использования и достаточно широко использует программный код из пакета `mailtools`:

- Для *просмотра* вложений соответствующие части сообщений извлекаются из электронных писем и сохраняются в локальных файлах на сервере. После этого в страницы просмотра добавляются гиперссылки, ссылающиеся на временные файлы. После щелчка на этих ссылках вложения открываются веб-броузером, как предусмотрено его настройками для типов файлов выбранной части.
- Для *отправки* вложений используется прием выгрузки файлов, представленный в конце главы 15. Страницы редактирования сообщений теперь снабжены элементами управления выгрузкой файлов, что позволяет присоединять к письму до трех вложений. Выбранные файлы, как обычно, выгружаются броузером на сервер с оставшейся частью страницы, сохраняются во временных файлах на сервере и добавляются в исходящую почту из локальных файлов на сервере с помощью инструментов из пакета `mailtools`. Как описывалось во врезке в предыдущем разделе, версия 3.0 позволяет отправлять только текстовые вложения в совместимой кодировке, включая и файлы HTML, и не поддерживает двоичные вложения.

Обе функции будут терпеть неудачу при одновременном обращении нескольких пользователей, но, поскольку файл с настройками PyMailCGI (описывается ниже в этой главе) и так может хранить только одно имя пользователя, это ограничение является вполне оправданным. Ссылки на временные файлы, генерируемые для просмотра вложений, действуют только для последнего выбранного сообщения и только при соблюдении обычного порядка следования через страницы. Добавление поддержки одновременной работы нескольких пользователей, а также дополнительных особенностей, таких как параметры сохранения и открытия локальных файлов, оставляется в качестве самостоятельного упражнения.

Для эффективности эта версия PyMailCGI также сводит к минимуму число повторных загрузок почты. В предыдущей версии полные тексты всех входящих сообщений загружались при каждом посещении страницы со списком почты и каждый раз при выборе сообщения для просмотра. В этой версии при посещении страницы со списком загружаются только заголовки сообщений, а полный текст сообщения загружается только при выборе его для просмотра. Кроме того, автоматически применяется ограничение на количество загружаемых заголовков, определяемое настройками, добавленными в пакет `mailtools` в четвертом издании этой книги, которое призвано сократить время загрузки (более ранние сообщения, не входящие в указанное число, игнорируются).

Но даже с этими улучшениями загрузка страницы со списком заголовков может выполняться довольно долго, при большом количестве входящих сообщений в почтовом ящике (как я уже признавался в главе 14, в одном из моих почтовых ящиков скопилось несколько тысяч писем). Более удачное решение заключалось бы в том, чтобы каким-то образом кэшировать почту, – для ограничения объема загружаемых данных хотя бы в течение сеанса работы с броузером. Например, можно было бы загружать заголовки только вновь прибывших сообщений и помещать в кэш заголовки, прочитанные ранее, как это сделано в клиенте PyMailGUI в главе 14.

Однако из-за отсутствия встроенного механизма сохранения информации о состоянии в сценариях CGI для этого пришлось бы задействовать на стороне сервера базу данных того или иного вида. Мы могли бы, например, сохранять уже полученные заголовки под ключом, идентифицирующим сеанс (состоящим, например, из номера процесса и времени), и передавать эти ключи между страницами через cookies, скрытые поля форм или параметры запроса URL. Каждая страница могла бы использовать этот ключ для получения кэшированных данных непосредственно с веб-сервера, вместо того чтобы снова загружать их с почтового сервера. Загрузка данных из локального файла кэша наверняка выполнялась бы быстрее, чем загрузка с почтового сервера через сетевое соединение.

Дальнейшее расширение системы было бы для вас интересным упражнением, но обсуждение этого сильно увеличило бы объем этой главы (честно признаться, я уже исчерпал время и объем, отведенные на этот проект и на эту главу задолго до того, как стал задумываться о потенциальных расширениях).

Обзорное представление программы

Большая часть «действий», происходящих в PyMailCGI, сосредоточена в совместно используемых вспомогательных модулях (особенно в одном, под именем `commonhtml.py`). Как будет показано чуть ниже, сценарии CGI, реализующие взаимодействие с пользователем, сами по себе выполняют немного. Такая архитектура была выбрана намеренно, чтобы сделать сценарии максимально простыми, избежать избыточности программного кода и придать им одинаковый внешний вид. Но это означает, что приходится прыгать между файлами, чтобы понять, как работает система.

Чтобы проще было разобраться в этом примере, мы будем рассматривать его программный код по частям: сначала сценарии страниц, а затем – вспомогательные модули. Во-первых, мы изучим снимки экранов основных веб-страниц, создаваемых системой, а также файлы HTML и CGI-сценарии Python верхнего уровня, которые их генерируют. Мы начнем с того, что проследим, как выполняется отправка почты, а затем – как выполняется чтение и последующая обработка имеющейся почты. В этих разделах мы познакомимся с большинством деталей реализации, но для понимания того, что в действительности делают сценарии, будьте готовы смотреть вперед, в описание вспомогательных модулей.

Я должен также подчеркнуть, что это довольно сложная система, и я не стану описывать ее исчерпывающим образом. Как и при изучении PyMailGUI в главе 14, читайте попутно исходный программный код, чтобы разобраться в деталях, о которых явно не сказано в тексте. Все исходные тексты входят в состав доступного для загрузки пакета с примерами к книге, и мы изучим здесь ключевые концепции этой системы. Как обычно в отношении конкретных примеров, приводимых в книге, предполагается, что читатель на данной стадии изучения в состоянии читать программный код на языке Python и за дополнительными деталями будет обращаться к программному коду примера. Синтаксис Python настолько близок к исполняемому псевдокоду, что иногда системы лучше описываются на Python, чем на обычном языке.

Опробование примеров из этой главы

Страницы HTML и CGI-сценарии, входящие в состав PyMailCGI, могут быть установлены на любой веб-сервер, доступный вам. Однако для простоты мы будем следовать той же политике, что и в главе 15, – мы будем использовать сценарий *webserver.py* локального веб-сервера, пред-

ставленный в примере 15.1, выполняющийся на одном компьютере с веб-броузером. Как уже говорилось в предыдущей главе, это означает, что для доступа к страницам системы из броузера или с помощью модуля `urllib.request` мы будем использовать доменное имя сервера «localhost» (или эквивалентный ему IP-адрес «127.0.0.1).

Запустите этот сервер на своем компьютере, чтобы опробовать программу в действии. В конечном счете, эта система должна иметь возможность взаимодействовать с почтовым сервером через Интернет, чтобы получать или отправлять сообщения, но в данном случае веб-сервер будет выполняться локально, на вашем компьютере.

Одно небольшое замечание: программный код PyMailCGI находится в отдельном каталоге, на один уровень ниже каталога со сценарием *webserver.py*. По этой причине при запуске веб-сервера необходимо явно указывать каталог приложения и номер порта в командной строке, как показано ниже:

```
C:\...\PP4E\Internet\Web> webserver.py PyMailCgi 8000
```

Введите эту команду в окне консоли Windows или в командной оболочке на Unix-подобной платформе. При запуске таким способом веб-сервер будет принимать запросы на соединение на компьютере «localhost», на сокете с портом 8000. Он будет обслуживать страницы из подкаталога *PyMailCgi*, находящегося на один уровень ниже каталога со сценарием, и выполнять сценарии CGI, расположенные в каталоге *PyMailCgi\cgi-bin*. Это возможно благодаря тому, что сценарий веб-сервера делает текущим рабочим каталогом тот, который был передан в командной строке при запуске.

Тонкое замечание: поскольку при запуске таким способом мы указываем в командной строке уникальный номер порта, веб-сервер не будет конфликтовать с другим его экземпляром, обслуживающим примеры из предыдущей главы, находящиеся в дереве каталогов на один уровень выше, – тот экземпляр веб-сервера будет принимать соединения на сокете с портом 80, а наш новый экземпляр будет обрабатывать запросы, поступающие на порт 8000. Фактически, имеется возможность взаимодействовать с любым сервером из одного и того же броузера, указывая тот или иной номер порта. Если вы запустите два экземпляра сервера, обслуживающих каталоги с примерами для различных глав, то для доступа к страницам и сценариям из предыдущей главы используйте адреса URL следующего вида:

```
http://localhost/languages.html
http://localhost/cgi-bin/languages.py?language=All
```

А для доступа к страницам и сценариям из этой главы используйте адреса URL такого вида:

```
http://localhost:8000/pymailcgi.html
http://localhost:8000/cgi-bin/onRootSendLink.py
```

В процессе работы с примерами вы увидите, как будут появляться трассировочные сообщения в окне того сервера, с которым вы взаимодействуете. Чтобы понять, почему происходит именно так, как происходит, смотрите введение в адреса сетевых сокетов в главе 12 и обсуждение адресов URL в главе 15.

Если вы решите установить программный код этого примера на другой сервер, просто замените часть «localhost: 8000/cgi-bin» в адресах URL, используемых здесь, именем своего сервера, номером порта и путем к каталогу. На практике такие системы, как PyMailCGI, более полезны, когда они устанавливаются на удаленном сервере, что позволяет обрабатывать почту с помощью любого веб-клиента.[1]

Как и в примере приложения PyMailGUI, вам необходимо будет отредактировать настройки в модуле `mailconfig.py`, чтобы использовать эту систему для чтения своей почты. Информация о почтовом сервере, указанная в настройках в настоящий момент, не подойдет для чтения вашей почты – подробнее об этом чуть ниже.

Программное обеспечение типа «ручная кладь»

Система PyMailCGI действует согласно замыслу и иллюстрирует новые принципы CGI и электронной почты, но я хочу заранее сделать несколько предостережений. Приложение было первоначально написано во время двухчасовой задержки в чикагском аэропорту O'Hare (правда, отладка потребовала еще нескольких часов). Я написал его в связи со специфической потребностью – иметь возможность читать и отправлять почту с помощью любого веб-броузера во время путешествий по свету для преподавания Python. Я не стремился к тому, чтобы она доставляла кому-то эстетическое удовольствие, и не очень усердствовал над повышением ее эффективности.

[1] Один из недостатков локального веб-сервера *webserver.py*, который я отмечал в ходе разработки, состоит в том, что на платформах, где сценарии CGI выполняются в том же процессе, что и сам сервер, вам придется останавливать и перезапускать сервер всякий раз при изменении импортируемого модуля. В противном случае последующие операции импорта в сценариях CGI не будут оказывать желаемого эффекта: модуль уже будет импортирован в процесс. Эта проблема отсутствует в Windows и на платформах, где сценарии CGI выполняются в отдельных, новых процессах. Реализация серверных классов изменяется со временем, но если внесение изменений в сценарии CGI не вступает в силу, причина для вашей платформы может оказаться как раз такой; попробуйте остановить и перезапустить локальный веб-сервер.

Я также умышленно сохранил простоту примера, чтобы включить его в эту книгу. Например, PyMailCGI предоставляет не все функции из имеющихся в программе PyMailGUI из главы 14 и перегружает почту чаще, чем, вероятно, следовало бы. Вследствие этого производительность системы оказывается не очень высокой при большом количестве входящих сообщений в почтовом ящике.

Фактически эта система почти требует использования дополнительных механизмов сохранения информации о состоянии. В настоящий момент информация о пользователе и о сообщении передается генерируемым страницам через скрытые поля форм и параметры запроса, но мы могли бы также избежать необходимости повторной загрузки почты, развернув на сервере базу данных и используя приемы работы с ней, о которых рассказывается в главе 17. С помощью такого расширения можно было бы в конечном счете довести функциональность PyMailCGI до уровня PyMailGUI, хотя и ценой некоторого усложнения программного кода. Но даже в этом случае данная система будет страдать от ограничений на выгрузку вложений, уходящих корнями в Python 3.1 и описанных выше, которые также необходимо решать.

Иными словами, эту систему следует рассматривать, как прототип, работа над которым не завершена. Это пока не та программа, которую можно продать, и в целом не тот программный продукт, который хотелось бы использовать для обработки важной почты. С другой стороны, она вполне справляется со своей работой, и ее можно индивидуально настраивать, внося изменения в ее исходный программный код на языке Python, что можно сказать далеко не о всякой продаваемой программе.

Корневая страница

Начнем с реализации главной страницы этого примера. Файл, представленный в примере 16.2, используется преимущественно для вывода ссылок на страницы отправки и просмотра. Она реализована в виде статического файла HTML, поскольку не содержит динамической информации.

Пример 16.2. PP4E\Internet\Web\PyMailCgi\pymailcgi.html

```
<HTML>
<TITLE>PyMailCGI Main Page</TITLE>
<BODY>
<H1 align=center>PyMailCGI</H1>
<H2 align=center>A POP/SMTP Web Email Interface</H2>
```

```
<P align=center><I>Version 3.0 June 2010 (2.0 January 2006)</I></P>

<table>

<tr><td><hr>
<h2>Actions</h2>
<P>
<UL>
<LI><a href="cgi-bin/onRootViewLink.py">Просмотреть, Ответить, Переслать,
Удалить входящее сообщение</a>
<LI><a href="cgi-bin/onRootSendLink.py">Отправить новое сообщение по SMTP</a>
</UL></P>

<tr><td><hr>
<h2>Обзор</h2>
<P>
<A href="http://rmi.net/~lutz/about-pp.html">
<IMG src="ppsmall.gif" align=left
alt="[Book Cover]" border=1 hspace=10></A>
Этот сайт реализует простой веб-интерфейс к учетной записи электронной
почты по протоколам POP/SMTP. С помощью этого интерфейса любой желающий
сможет отправить письмо, но из-за ограничений безопасности вы не сможете
просматривать электронную почту, не определив параметры своей учетной
записи на почтовом сервере. Веб-приложение PyMailCgi реализовано как набор
CGI-сценариев на языке Python, выполняющихся на сервере (не на вашем
локальном компьютере) и генерирующих разметку HTML при взаимодействии
с броузером. Подробности смотрите в книге <I>Программирование на Python,
4-е издание</I>.</P>

<tr><td><hr>
<h2>Примечания</h2>
<P>Внимание: Версия PyMailCgi 1.0 первоначально была создана во время
2-часового ожидания вылета из чикагского аэропорта O'Hare. Эта версия
не такая быстрая и полнофункциональная, как PyMailGUI (например, каждый
щелчок запускает выполнение операции через Интернет, здесь отсутствует
операция сохранения электронной почты и не поддерживается многопоточный
режим выполнения, кроме того, здесь не предусматривается кэширование
заголовков или уже просмотренных сообщений). С другой стороны, PyMailCgi
может взаимодействовать с любым веб-броузером и не требует устанавливать
Python (и Tk) на ваш компьютер.

<P> Если вы решите использовать эти сценарии для чтения своей почты,
то следует учесть, что PyMailCgi не гарантирует безопасность пароля вашей
учетной записи. Смотрите примечания в странице операции просмотра сообщения,
а также дополнительную информацию в книге относительно безопасности.

<p><I><U>Новое в версии 2</U></I>: программа PyMailCGI теперь поддерживает
просмотр и отправку вложений для одного пользователя, и в ней удалось
избежать необходимости излишней повторной загрузки почты в некоторых случаях.
При отображении страницы со списком она загружает только заголовки
```

```
сообщений, а загрузка полного текста сообщения выполняется только
при выборе его для просмотра.

<p><I><U>Новое в версии 3</U></I>: программа PyMailCGI теперь выполняется
под управлением Python 3.X (только) и использует множество новых
особенностей из пакета mailtools: декодирование и кодирование
интернационализированных заголовков, декодирование основного текста
почтового сообщения и так далее. Из-за снижения функциональных возможностей
модуля cgi и пакета email в Python 3.1 версия 3.0 не поддерживает отправку
двоичных вложений или текстовых вложений с несовместимой кодировкой, однако
сохраняется возможность просмотра вложений во входящих сообщениях
(смотрите главы 15 и 16).

<p>Смотрите также:
<UL>
<li>Программу <I>PyMailGUI</I> в каталоге Internet, которая
    реализует более полноценный графический интерфейс к электронной почте
    на основе Python+Tk
<li>Программу <I>pymail.py</I> в каталоге Email, которая реализует простой
    интерфейс командной строки к электронной почте
<li>Модуль Python imaplib, реализующий поддержку протокола IMAP электронной
    почты вместо POP
</UL></P>
</table><hr>

<A href="http://www.python.org">
<IMG SRC="PythonPoweredSmall.gif" ALIGN=left
ALT="[Python Logo]" border=0 hspace=15></A>
<A href="http://rmi.net/~lutz/about-pp.html">[Book]</a>
<A href="http://www.oreilly.com">[O'Reilly]</a>
</BODY></HTML>
```

Файл *pymailcgi.html* описывает корневую страницу системы и располагается в подкаталоге *PyMailCgi*, выделенном специально для этого приложения и позволяющем хранить его отдельно от других примеров. Для доступа к этой системе запустите локальный веб-сервер, как было описано в предыдущем разделе, и введите в адресной строке броузера следующий адрес URL (или другой, в зависимости от используемого вами веб-сервера):

```
http://localhost:8000/pymailcgi.html
```

После этого сервер вернет страницу, подобную той, что показана на рис. 16.2 в веб-броузере Google Chrome в Windows 7. На протяжении всей этой главы вместо Internet Explorer я буду использовать Chrome – для разнообразия, а также потому, что он позволяет отображать страницы с большим числом подробностей. Откройте ее в своем броузере, чтобы убедиться в работоспособности системы – она переносима так же, как Веб, HTML и CGI-сценарии на языке Python.

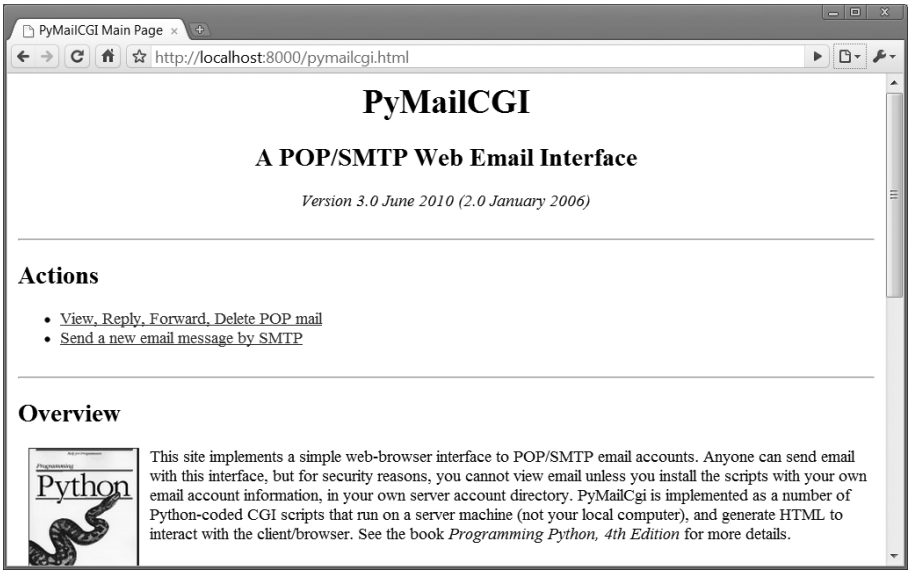


Рис. 16.2. *Главная страница PyMailCGI*

Настройка PyMailCGI

Теперь, прежде чем вы щелкните на ссылке View, ... (Просмотреть, ...), которую можно заметить на рис. 16.2, рассчитывая прочесть свою почту, я должен отметить, что по умолчанию система PyMailCGI позволяет любому желающему отправить почту с этой страницы, воспользовавшись ссылкой Send... (Отправить...) (как мы уже знаем, в SMTP нет паролей). Однако она разрешает произвольным пользователям прочесть свою почту только ценой ввода явного и небезопасного URL или после выполнения некоторых действий по установке и настройке.

Это сделано нарочно и связано, как будет показано далее, с ограничениями безопасности. Как мы увидим позднее, система написана так, что никогда не связывает вместе имя пользователя и пароль электронной почты без применения шифрования. Конечно, в этом нет никакой проблемы, если используется локальный веб-сервер, но это правило приобретает особую значимость, когда система выполняется на удаленном сервере.

По умолчанию эта страница устроена так, чтобы читать почтовую учетную запись, созданную специально для этой книги, – с адресом *PP4E@learning-python.com* – и требует использовать пароль этой учетной записи POP. Так как вы вряд ли угадаете этот пароль (и даже если сделаете это, вряд ли найдете в этом почтовом ящике что-то интересное!), то PyMailCGI не слишком полезна в том виде, в котором она распространяется. Чтобы использовать ее для чтения собственной почты, вы должны

отредактировать файл с настройками *mailconfig.py*, указав в нем параметры своей учетной записи. Этот файл мы увидим чуть позже. А пока продолжим исследование примера, используя учетную запись электронной почты для этой книги, – она работает одинаково независимо от того, к чьей учетной записи обращается.

Отправка почты по SMTP

PyMailCGI поддерживает две основные функции в виде ссылок на корневой странице: составление и отправку новой почты и просмотр входящей почты. Функция просмотра (ссылка View...) ведет на страницы, которые позволяют пользователям читать, отвечать, пересылать и удалять имеющуюся почту. Поскольку самой простой является функция отправки, начнем с ее страниц и сценариев.

Страница составления сообщений

Функция отправки (ссылка Send...) на корневой странице проводит пользователей через две страницы: одну для редактирования сообщения и другую для подтверждения отправки. Щелчок на ссылке Send... на главной странице, которую можно заметить на рис. 16.2, заставляет веб-сервер выполнить сценарий CGI, представленный в примере 16.3.

Пример 16.3. PP4E\Internet\Web\PyMailCgi\cgi-bin\onRootSendLink.py

```
#!/usr/bin/python
"""
##############################################################################
Вызывается щелчком на ссылке 'send' в главной странице: отображает страницу
составления нового сообщения
##############################################################################
"""


import commonhtml
from externs import mailconfig

commonhtml.editpage(kind='Write', headers={'From': mailconfig.myaddress})
```

Нет, этот файл не был урезан. Смотреть в этом сценарии особенно не на что, поскольку все действия инкапсулированы в модулях commonhtml и externs. Здесь можно лишь сказать, что для создания ответа этот сценарий вызывает некоторую функцию с именем editpage, передавая ей нечто с именем myaddress для заголовка «From».

Так было предусмотрено изначально – скрыв детали во вспомогательных модулях, мы значительно облегчим чтение и создание таких сценариев верхнего уровня, избавляясь от избыточности программного кода и обеспечивая единство внешнего вида всех наших страниц. Кроме того, в этом сценарии нет входных данных. При выполнении он создает страницу для составления нового сообщения, изображенную на рис. 16.3.

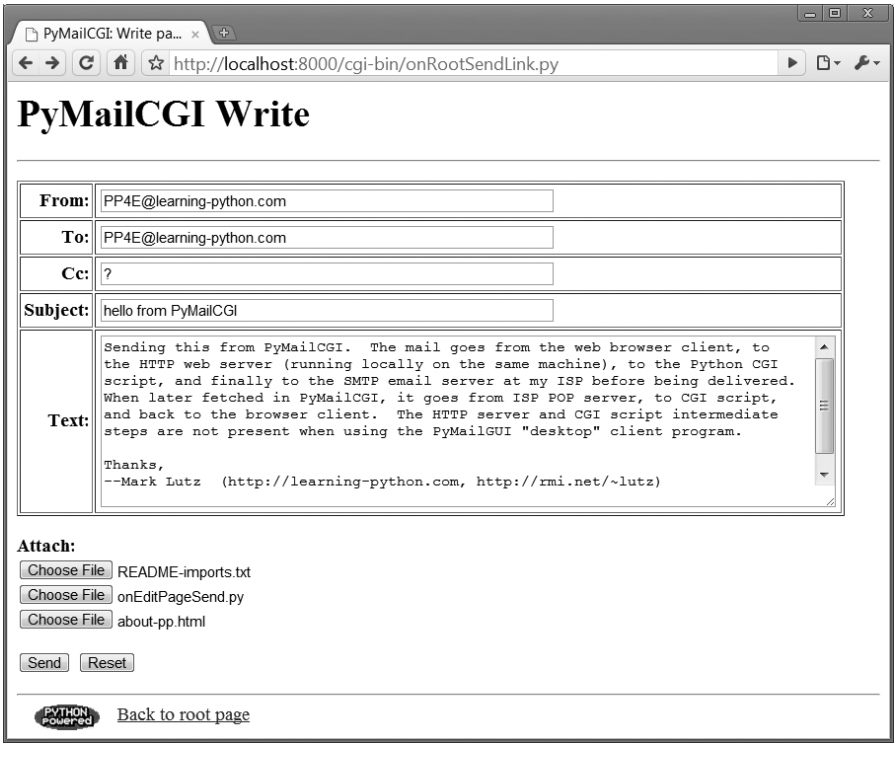


Рис. 16.3. *Страница PyMailCGI отправки (составления) сообщения*

В значительной степени страница составления нового сообщения понятна без лишних пояснений – вам нужно лишь заполнить поля заголовков и основного текста сообщения (содержимое заголовка «From» и текст подписи берутся из настроек в модуле `mailconfig`, который обсуждается ниже). Щелчок на любой из кнопок Choose File (Выбрать файл) открывает диалог выбора файла, который будет отправлен в виде вложения. Интерфейс этой страницы выглядит совершенно отличным от интерфейса клиентской программы PyMailGUI из главы 14, но по своей функциональности они очень похожи. Обратите также внимание на верхнюю и нижнюю части страницы – по причинам, описанным в следующем разделе, эти области будут выглядеть одинаково на всех страницах нашей системы.

Сценарий отправки почты

Как обычно, страница HTML редактирования сообщения, изображенная на рис. 16.3, ссылается на свой сценарий-обработчик. Если щелкнуть на кнопке Send (Отправить), на сервере будет запущен сценарий, представленный в примере 16.4, который обработает введенные нами данные и отправит почтовое сообщение.

Пример 16.4. PP4E\Internet\Web\PyMailCgi\cgi-bin\onEditPageSend.py

```
#!/usr/bin/python
"""
###############################################################################
Вызывается при отправке формы в окне редактирования: завершает составление
нового сообщения, ответа или пересылаемого сообщения;

в 2.0+: мы повторно используем инструменты из пакета mailtools
для конструирования и отправки сообщения вместо устаревшей схемы, основанной
на строковых методах; из этого модуля мы также наследуем возможность
добавлять вложения и преобразовывать отправляемые сообщения в формат MIME;

3.0: выгрузка через CGI двоичных вложений и текстовых вложений
в несовместимой кодировке не допускается из-за ограничений модуля cgi
в py3.1, поэтому мы просто используем здесь кодировку по умолчанию
для данной платформы (механизм синтаксического анализа, используемый
модулем cgi, не может предложить ничего лучше);
3.0: кроме того, для основного текста и для вложений используются
простые правила кодирования Юникода;
###############################################################################
"""


import cgi, sys, commonhtml, os
from externs import mailtools

savedir = 'partsupload'
if not os.path.exists(savedir):
    os.mkdir(savedir)

def saveAttachments(form, maxattach=3, savedir=savedir):
    """
    сохраняет выгруженные файлы вложений в локальных файлах на сервере,
    откуда mailtools будет добавлять их в сообщение; класс FieldStorage
    в 3.1 и другие компоненты модуля cgi могут вызывать появление ошибки
    для многих типов выгружаемых файлов, поэтому мы не будем прилагать
    особых усилий, чтобы попытаться определить корректную
    кодировку символов;
    """
    partnames = []
    for i in range(1, maxattach+1):
        fieldname = 'attach%d' % i
        if fieldname in form and form[fieldname].filename:
            fileinfo = form[fieldname]           # передана и заполнена?
            filedata = fileinfo.value            # прочитать в строку
            filename = fileinfo.filename         # путь на стороне клиента
            if '\\' in filename:
                basename = filename.split('\\')[-1]  # для клиентов DOS
            elif '/' in filename:
                basename = filename.split('/')[-1]   # для клиентов Unix
            else:
                basename = filename              # видимо, путь отсутствует
```

```
            pathname = os.path.join(savedir, basename)
            if isinstance(filedata, str):         # 3.0: rb требует bytes
                filedata = filedata.encode()      # 3.0: использ. кодировку?
            savefile = open(pathname, 'wb')
            savefile.write(filedata)              # или с инструментом with
            savefile.close()                      # но EIBTI[1]
            os.chmod(pathname, 0o666)             # треб. некот. серверами
            partnames.append(pathname)            # список локальных путей
    return partnames                              # определить тип по имени

#commonhtml.dumpstatepage(0)
form = cgi.FieldStorage()         # извлечь данные из формы
attaches = saveAttachments(form)  # cgi.print_form(form), чтобы посмотреть

# имя сервера из модуля или из URL, полученного методом GET
smtpservername = commonhtml.getstandardsmtpfields(form)

# здесь предполагается, что параметры получены из формы или из URL
from commonhtml import getfield   # для получения значений атрибутов
From = getfield(form, 'From')     # пустые поля не должны отправляться
To   = getfield(form, 'To')
Cc   = getfield(form, 'Cc')
Subj = getfield(form, 'Subject')
text = getfield(form, 'text')
if Cc == '?': Cc = ''

# 3.0: не-ascii заголовки кодируются в utf8 в пакете mailtools
parser = mailtools.MailParser()
Tos = parser.splitAddresses(To)   # списки получателей: разделитель ','
Ccs = (Cc and parser.splitAddresses(Cc)) or ''
extraHdrs = [('Cc', Ccs), ('X-Mailer', 'PyMailCGI 3.0')]

# 3.0: определить кодировку для основного текста и текстовых вложений;
# по умолчанию=ascii в mailtools
bodyencoding = 'ascii'
try:
    text.encode(bodyencoding)     # сначала попробовать ascii (или latin-1?)
except (UnicodeError, LookupError): # иначе использ. utf8 (или из настроек?)
    bodyencoding = 'utf-8'          # что сделать: это более ограниченное
                                    # решение, чем в PyMailGUI

# 3.0: использовать utf8 для всех вложений;
# здесь мы не можем спросить у пользователя
attachencodings = ['utf-8'] * len(attaches)       # игнорировать
                                                  # нетекстовые части

# кодировать и отправить
sender = mailtools.SilentMailSender(smtpservername)
```

[1] Аббревиатура от «Explicit Is Better Than Implicit» («явное лучше неявного») – один из девизов Python. – *Прим. перев.*

```
try:
    sender.sendMessage(From, Tos, Subj, extraHdrs, text, attaches,
                       bodytextEncoding=bodyencoding,
                       attachesEncodings=attachencodings)
except:
    commonhtml.errorpage('Send mail error')
else:
    commonhtml.confirmationpage('Send mail')
```

Этот сценарий получает входную информацию из заголовков почты и текст из формы страницы редактирования (или из параметров запроса в адресе URL) и отправляет сообщение с помощью стандартного модуля Python `smtplib` посредством пакета `mailtools`. Мы подробно изучили пакет `mailtools` в главе 13, поэтому я не стану подробно говорить о нем сейчас. Однако обратите внимание, что благодаря повторному использованию его функции отправки отправленное сообщение автоматически сохраняется в файле *sentmail.txt* на сервере – в PyMailCGI отсутствуют инструменты для просмотра содержимого этого файла, но он может служить своеобразным журналом.

Появившаяся в версии 2.0 функция `saveAttachments` получает все файлы, отправленные броузером, и сохраняет их в локальных файлах во временном каталоге на сервере. Позднее, при отправке, эти файлы будут вложены в почтовое сообщение. Мы подробно рассматривали механизм выгрузки файлов на сервер через CGI в конце главы 15 – там вы найдете объяснение, как действует программный код здесь (а также ограничения в Python 3.1 и в этом издании – из-за этих ограничений мы можем прикреплять к сообщениям только простой текст). Непосредственное прикрепление файлов к электронному письму осуществляется пакетом `mailtools` автоматически.

Утилита `commonhtml` в конечном счете получает имя сервера SMTP, который примет сообщение, из модуля `mailconfig` или входных данных сценария (в поле формы или в параметре запроса URL). Если все пройдет успешно, мы получим сгенерированную страницу подтверждения, как на рис. 16.4.

Откройте файл *sentmail.txt* в исходном каталоге программы PyMailCGI, если у вас появится желание увидеть, как выглядит полный исходный текст отправленного сообщения (или получите сообщение в клиенте электронной почты с возможностью просмотра исходного текста письма, в таком как PyMailGUI). В этой версии каждое вложение кодируется в кодировке UTF-8 и преобразуется в формат MIME Base64, но основная текстовая часть отправляется как простой текст ASCII, если это возможно.

Как будет показано далее, этот же сценарий используется для отправки *ответа* и *пересылки* входящего письма по другому адресу. Пользовательский интерфейс этих операций немного отличается от интерфейса составления нового письма. Но, как и в программе PyMailGUI, логика

обработчика, выполняющего отправку, была оформлена в виде общего совместно используемого программного кода – отправка ответа и пересылка по другому адресу, по сути, такие же операции отправки писем, включающих цитированный текст и предустановленные заголовки.

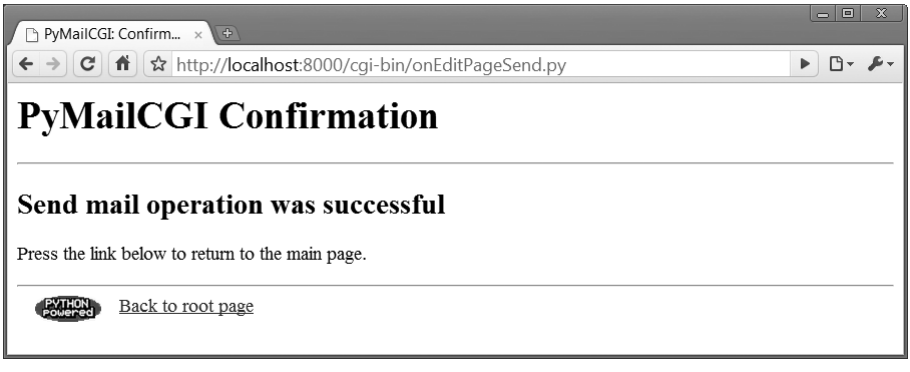


Рис. 16.4. *Страница PyMailCGI подтверждения отправки*

Обратите внимание, что здесь не видно ни имени пользователя, ни пароля: как отмечалось в главе 13, для SMTP обычно нужен только сервер, который прослушивает порт SMTP, а не учетная запись пользователя или пароль. В той же главе мы видели, что неудачные операции отправки по протоколу SMTP либо возбуждают исключительную ситуацию Python (например, если нельзя установить соединение с хостом сервера), либо возвращают словарь, содержащий получателей, почту которым не удалось отправить. Пакет `mailtools` избавляет нас от необходимости помнить эти тонкости, всегда возбуждая исключение.

Страницы с сообщениями об ошибках

Если во время доставки почты возникнут проблемы, будет получена страница с сообщением об ошибке, как показано на рис. 16.5. На этой странице приводится адрес получателя, которому почта не была отправлена, и трассировочная информация, полученная с помощью модуля `traceback` из стандартной библиотеки. В случае появления ошибки выводится само сообщение об ошибке, возвращаемое интерпретатором, и дополнительные данные.

Стоит также подчеркнуть, что в модуле `commonhtml` инкапсулирована генерация как страниц подтверждения, так и страниц с сообщениями об ошибках, чтобы все эти страницы выглядели в PyMailCGI одинаково, независимо от того, где и когда они созданы. Логика, генерирующая страницу редактирования почты в `commonhtml`, повторно используется также в операциях создания ответа и пересылаемого письма (но с другими почтовыми заголовками).

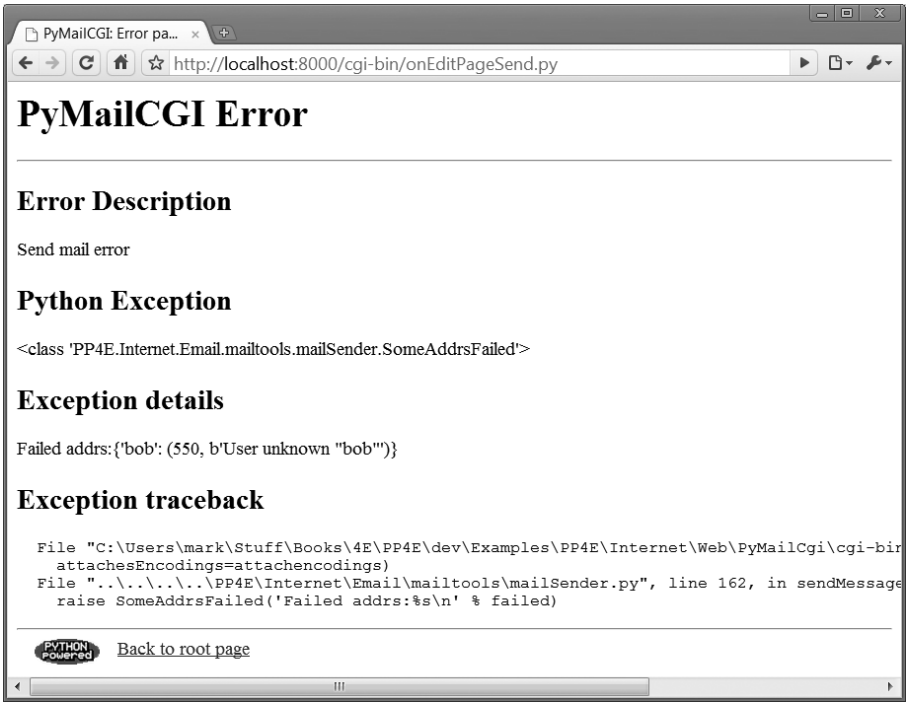


Рис. 16.5. *Страница PyMailCGI с сообщением об ошибке отправки*

Единство внешнего вида

Модуль `commonhtml` обеспечивает одинаковый внешний вид всех страниц – он также содержит функции для создания заголовков и нижних колонтитулов, всюду используемых в системе. Вы могли уже заметить, что все страницы построены по одинаковой схеме: они начинаются с заголовка и горизонтальной линии, содержат что-то свое в середине и заканчиваются внизу другой горизонтальной линией, за которой следуют значок Python и ссылка. Этот стандартный внешний вид обеспечивается модулем `commonhtml`, в котором генерируется все необходимое каждой странице в системе, кроме ее средней части (за исключением корневой страницы – статического файла HTML).

Что особенно важно, если когда-нибудь нам придется изменить функции формирования заголовка и нижнего колонтитула в модуле `commonhtml`, заголовки и нижние колонтитулы на всех страницах обновятся автоматически. Если вы прямо сейчас хотите узнать, как действует эта инкапсулированная логика, пролистайте вперед до примера 16.14. Его мы рассмотрим после того, как изучим остальные страницы этого почтового сайта.

Использование сценария отправки почты без броузера

Вначале я написал сценарий отправки для применения только в PyMailCGI с использованием значений, введенных в форме редактирования почты. Но, как мы видели, входные данные могут передаваться сценариям в полях формы или в параметрах запроса в адресе URL; так как сценарий отправки почты ищет входные данные во входных данных CGI, прежде чем импортировать их из модуля `mailconfig`, этот сценарий можно вызывать для отправки почты не только со страницы редактирования. Например, явно введя в броузере адрес URL такого вида (в одну строку без пробелов):

```
http://localhost:8000/cgi-bin/
               onEditPageSend.py?site=smtp.rmi.net&
                                 From=lutz@rmi.net&
                                 To=lutz@rmi.net&
                                 Subject=test+url&
                                 text=Hello+Mark;this+is+Mark
```

мы отправим почтовое сообщение в соответствии с входными параметрами, указанными в конце URL. Конечно, такую строку URL долго вводить в поле адреса броузера, но ее можно автоматически генерировать в другом сценарии. Как мы видели в главах 13 и 15, с помощью модуля `urllib.request` можно передать такую строку URL на сервер из программы на языке Python. В примере 16.5 показан один из способов автоматизировать эту операцию.

Пример 16.5. PP4E\Internet\Web\PyMailCgi\sendurl.py

```
"""
##################################################################
Отправляет почтовое сообщение, конструируя из входных данных
адрес URL следующего вида:
http://servername/pathname/
          onEditPageSend.py?site=smtp.rmi.net&
                            From=lutz@rmi.net&
                            To=lutz@rmi.net&
                            Subject=test+url&
                            text=Hello+Mark;this+is+Mark
##################################################################
"""

from urllib.request import urlopen
from urllib.parse import quote_plus

url = 'http://localhost:8000/cgi-bin/onEditPageSend.py'
url += '?site=%s' % quote_plus(input('Site>'))
url += '&From=%s' % quote_plus(input('From>'))
url += '&To=%s' % quote_plus(input('To >'))
url += '&Subject=%s' % quote_plus(input('Subj>'))
url += '&text=%s' % quote_plus(input('text>'))   # или цикл ввода
```

```
print('Reply html:')
print(urlopen(url).read().decode())          # страница подтверждения
                                             # или с сообщением об ошибке
```

Запуская этот сценарий из командной строки, мы получаем еще один способ отправить электронное письмо – на этот раз путем обращения к нашему сценарию CGI на удаленном сервере, который должен выполнить всю работу. Сценарий *sendurl.py* выполняется на любой машине, где есть Python и сокеты, позволяет вводить параметры почты интерактивно и вызывает другой сценарий Python, находящийся на удаленной машине. Он выводит разметку HTML, возвращаемую нашим сценарием CGI:

```
C:\...\PP4E\Internet\Web\PyMailCgi> sendurl.py
Site>smtpout.secureserver.net
From>PP4E@learning-python.com
To  >lutz@learning-python.com
Subj>testing sendurl.py
text>But sir, it's only wafer-thin...
Reply html:
<html><head><title>PyMailCGI: Confirmation page (PP4E)</title></head>
<body bgcolor="#FFFFFF"><h1>PyMailCGI Confirmation</h1><hr>
<h2>Send mail operation was successful</h2>
<p>Press the link below to return to the main page.</p>
</p><hr><a href="http://www.python.org">
<img src="../PythonPoweredSmall.gif"
align=left alt="[Python Logo]" border=0 hspace=15></a>
<a href="../pymailcgi.html">Back to root page</a>
</body></html>
```

Ответ в формате HTML, выведенный этим сценарием, предназначен для отображения в виде новой веб-страницы, когда его получает броузер. Такой замысловатый вывод нельзя назвать идеальным, но можно легко найти в нем строку ответа, чтобы определить результат (например, искать строку «successful» с помощью строкового метода `find` или оператора `in` проверки на вхождение), проанализировать состав с помощью стандартного модуля Python `html.parse` или `re` (описывается в главе 19) и так далее. Получившееся почтовое сообщение, которое мы для разнообразия просмотрим с помощью программы PyMailGUI из главы 14, появляется в моем почтовом ящике и показано на рис. 16.6 (это сообщение состоит из единственной текстовой части).

Конечно, есть другие, менее окольные способы отправки почты с компьютера клиента. Например, сам модуль Python `smtplib` (используется пакетом `mailtools`) зависит только от работоспособности клиента и возможности установить соединение с сервером SMTP, в то время как этот сценарий зависит также от работоспособности веб-сервера и сценария CGI (запросы проходят от клиента на веб-сервер, передаются сценарию CGI и далее серверу SMTP). Но поскольку наш сценарий CGI поддерживает работу с адресами URL, он способен на большее, чем тег HTML

mailto:, и может вызываться с помощью модуля urllib.request вне контекста работающего веб-броузера. Например, с помощью сценариев типа *sendurl.py* можно вызывать и *тестировать* программы, выполняемые на сервере.

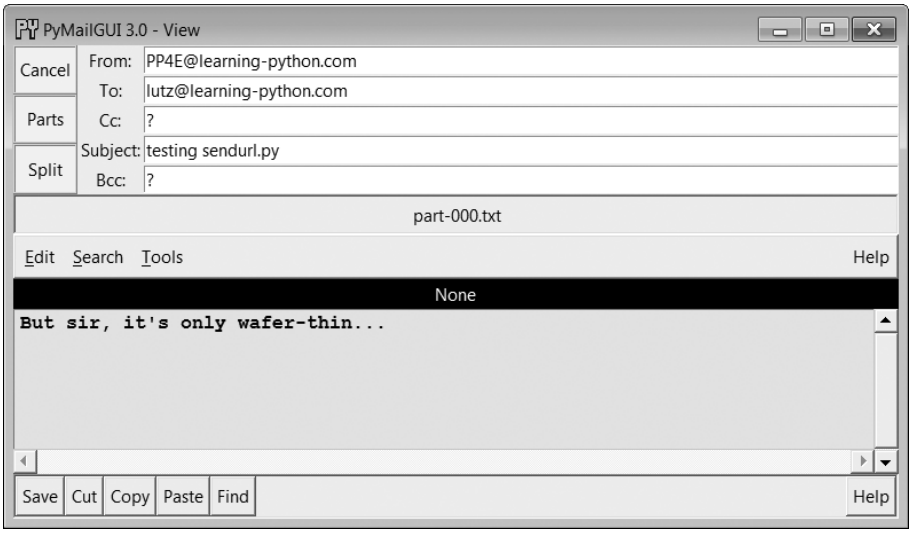


Рис. 16.6. *Результат выполнения сценария sendurl.py*

Чтение электронной почты по протоколу POP

К настоящему моменту мы проследили путь, по которому система *отправляет* новую почту. Теперь посмотрим, что происходит при *просмотре* входящей почты POP.

Страница ввода пароля POP

Если вернуться к главной странице, изображенной на рис. 16.2, можно увидеть ссылку View..., щелчок на которой запускает сценарий на сервере, представленный в примере 16.6:

Пример 16.6. PP4E\Internet\Web\PyMailCgi\cgi-bin\onRootViewLink.py

```
#!/usr/bin/python
"""

###############################################################################
Вызывается щелчком на ссылке View... в главной странице HTML:
создает страницу ввода пароля;

этот сценарий можно было бы заменить статической страницей HTML,
потому что этот сценарий не имеет никаких входных параметров,
```

```
но мне требовалось задействовать стандартные функции создания заголовка
и нижнего колонтитула и отобразить имя сайта и имя пользователя, которые
необходимо получить из настроек; при отправке этой формы пароль передается
без имени пользователя, пароль и имя пользователя передаются вместе в виде
параметров URL или в скрытых полях формы только после того, как пароль будет
зашифрован с помощью модуля шифрования, выгружаемого пользователем;
##########################################################################
"""


# шаблон страницы
pswdhtml = """
<form method=post action=%sonViewPswdSubmit.py>
<p>
Введите пароль учетной записи пользователя "%s" на сервере POP "%s".
<p><input name=pswd type=password>
<input type=submit value="Submit"></form></p>
<hr><p><i>Примечание, касающееся безопасности</i>: Пароль, введенный
в поле выше, будет отправлен на сервер через Интернет, но он нигде
не отображается, никогда не передается в паре с именем пользователя
в незашифрованном виде и нигде не сохраняется: ни на сервере (он только
передается последующим страницам в скрытых полях форм), ни на стороне
клиента (система не генерирует никаких cookies). Тем не менее, полная
безопасность не гарантируется; при работе с PyMailCGI вы можете
использовать кнопку "Назад" ("Back") своего броузера в любой
момент времени.</p>
"""


# создание страницы ввода пароля
import commonhtml                       # обычный прием работы с параметрами:
user, pswd, site = commonhtml.getstandardpopfields({})  # сначала из модуля,
commonhtml.pageheader(kind='POP password input')        # затем из html|url
print(pswdhtml % (commonhtml.urlroot, user, site))
commonhtml.pagefooter()
```

Этот сценарий почти целиком состоит из разметки HTML: заключенная в тройные кавычки строка `pswdhtml` выводится с применением операции форматирования строки, с целью вставить значения, за один шаг. Но поскольку нужно получить имя пользователя и сервера, чтобы показать их на странице, использован выполняемый сценарий, а не статический файл HTML. Модуль `commonhtml` загружает имена пользователя и сервера из входных данных сценария (например, если они добавлены в конец адреса URL сценария) или импортирует их из файла `mailconfig` – ни в том, ни в другом случае нежелательно жестко прописывать их в сценарии или в его разметке HTML, поэтому файл HTML не подойдет. В мире сценариев CGI программный код на языке Python включает разметку HTML и выполняет подстановку в нее необходимых значений, как показано выше (серверные механизмы шаблонов, такие как PSP, действуют похожим образом, только при их использовании наоборот, разметка HTML включает программный код Python, воспроизводящий необходимые значения).

Поскольку это сценарий, можно воспользоваться функциями создания заголовка и нижнего колонтитула страницы из модуля `commonhtml`, чтобы страница ответа выглядела стандартным образом, как показано на рис. 16.7.

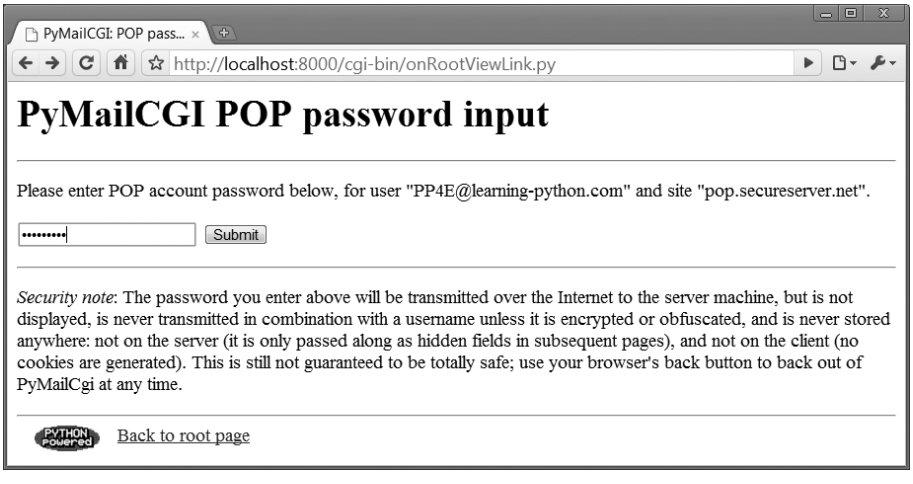


Рис. 16.7. *Страница PyMailCGI ввода пароля для просмотра почты*

Предполагается, что на этой странице пользователь должен ввести пароль к учетной записи для указанного имени пользователя и сервера POP. Обратите внимание, что фактический пароль не отображается; поле ввода в разметке HTML определено как `type=password`, благодаря чему оно действует как обычное текстовое поле, но введенные символы отображаются в виде звездочек. (Смотрите также пример программы pymail в главе 13, как то же сделать в консоли, и пример программы PyMailGUI в главе 14, как то же сделать в графическом интерфейсе на базе tkinter.)

Страница выбора почты из списка

После заполнения поля пароля в предыдущей странице и нажатия кнопки Submit (Отправить) пароль передается сценарию, представленному в примере 16.7.

Пример 16.7. PP4E\Internet\Web\PyMailCgi\cgi-bin\onViewPswdSubmit.py

```
#!/usr/bin/python
"""

###############################################################################
Вызывается при отправке пароля POP из окна ввода: создает страницу
со списком входящей почты;

в 2.0+ загружаются только заголовки писем, позднее, по запросу, загружается
полный текст только одного выбранного сообщения; однако нам по-прежнему
```

```
необходимо загружать все заголовки всякий раз, когда создается страница
с оглавлением: для реализации кэширования сообщений на стороне сервера
может потребоваться использовать базу данных на стороне сервера (?)
и ключ сеанса или использовать какие-то другие механизмы;
3.0: выполняет декодирование заголовков перед отображением списка,
однако принтер и броузер должны уметь обрабатывать их;
##############################################################################
"""

import cgi
import loadmail, commonhtml
from   externs import mailtools
from   secret import encode # модуль шифрования, определяемый пользователем
MaxHdr = 35                 # максимальная длина заголовков в списке

# с предыдущей страницы поступает только пароль, остальное обычно в модуле
formdata = cgi.FieldStorage()
mailuser, mailpswd, mailsite = commonhtml.getstandardpopfields(formdata)
parser = mailtools.MailParser()

try:
    newmails = loadmail.loadmailhdrs(mailsite, mailuser, mailpswd)
    mailnum = 1
    maillist = []                               # или использ. enumerate()
    for mail in newmails:                       # список заголовков
        msginfo = []
        hdrs = parser.parseHeaders(mail)        # email.message.Message
        addrhdrs = ('From', 'To', 'Cc', 'Bcc')  # декодировать только имена
        for key in ('Subject', 'From', 'Date'):
            rawhdr = hdrs.get(key, '?')
            if key not in addrhdrs:
                dechdr = parser.decodeHeader(rawhdr) # 3.0: декодир.
                                                     # для отображения
            else:                               # закодиров. при отправке
                dechdr = parser.decodeAddrHeader(rawhdr) # только имена
            msginfo.append(dechdr[:MaxHdr])
        msginfo = ' | '.join(msginfo)
        maillist.append((msginfo, commonhtml.urlroot + 'onViewListLink.py',
                                 {'mnum': mailnum,
                                  'user': mailuser,         # параметры
                                  'pswd': encode(mailpswd), # передаются
                                  'site': mailsite}))       # в URL,
                                                            # а не в полях
        mailnum += 1
    commonhtml.listpage(maillist, 'mail selection list')
except:
    commonhtml.errorpage('Error loading mail index')
```

Основным назначением этого сценария является создание страницы со списком сообщений в почтовом ящике пользователя с использованием пароля, введенного на предыдущей странице (или переданного в адресе URL). Как обычно, благодаря инкапсуляции большинство деталей скрыто в других файлах:

`loadmail.loadmailhdrs`

Повторно использует пакет `mailtools` из главы 13 для загрузки сообщений по протоколу POP. Для вывода списка требуется счетчик сообщений и почтовые заголовки. В этой версии, чтобы сэкономить время, сценарий загружает только текст заголовков, а не сообщения целиком (обеспечивается командой `TOP` интерфейса POP, которую поддерживают большинство серверов, – если это не так, смотрите модуль `mailconfig`, который позволяет отключить эту возможность).

`commonhtml.listpage`

Генерирует HTML для вывода переданного ему списка кортежей (*текст*, *адрес_URL*, *словарь_параметров*) в виде списка гиперссылок на странице ответа. Значения параметров отображаются в конце URL в странице ответа.

Создаваемый здесь список `maillist` используется при построении тела следующей страницы – списка для выбора сообщений электронной почты. Каждая гиперссылка в списке включает адрес URL, содержащий достаточно информации, чтобы следующий сценарий мог загрузить и вывести конкретное сообщение. Как мы узнали в предыдущей главе, это простейший прием сохранения информации о состоянии для передачи между страницами и сценариями.

Если все в порядке, этот сценарий сгенерирует разметку HTML страницы со списком для выбора почтовых сообщений, изображенную на рис. 16.8. Если вы, как и я, получаете много почты, то, чтобы увидеть конец списка, может понадобиться прокрутить страницу вниз. Благодаря модулю `commonhtml` эта страница получает оформление, общее для всех страниц PyMailCGI.

Если сценарий не сможет получить доступ к вашей учетной записи почты (например, из-за ввода неправильного пароля), то обработчик в инструкции `try` создаст стандартную страницу с сообщением об ошибке.

На рис. 16.9 показана такая страница, сообщающая о возникшем исключении Python и его деталях после перехвата действительной исключительной ситуации. Как обычно, информация об исключении извлекается с помощью функции `sys.exc_info`, а для получения трассировочной информации используется модуль Python `traceback`.

PyMailCGI: mail sele...

http://localhost:8000/cgi-bin/onViewPswdSubmit.py

PyMailCGI mail selection list

View	I'm a Lumberjack, and I'm Okay | lutz@rmi.net | Wed, 5 May 2010 11:29:36 -0400 (EDT
View	A B C D E F G | Eric.the.Half.a.Bee@yahoo.com | Thu, 06 May 2010 14:11:07 -0000
View	a b c d e f g | Eric.the.Half.a.Bee@aol.com | Thu, 06 May 2010 14:32:32 -0000
View	test interactive smtplib | PP4E@learning-python.com | Thu, 06 May 2010 10:58:40 -0400
View	Among our weapons are these | Cardinal@hotmail.com | Fri, 07 May 2010 20:32:38 -0000
View	test pymail2 send | lutz@rmi.net | Sat, 08 May 2010 19:44:25 -0000
View	testing mailtools package | PP4E@learning-python.com | Thu, 13 May 2010 20:43:37 -0000
View	Re: Conheça mais uma novidade LC2! | PP4E@learning-python.com | Tue, 01 Jun 2010 23:54:08 -0000
View	testing pymailgui send | lutz@learning-python.com | Thu, 03 Jun 2010 21:05:57 -0000
View	Live Organ Transplants! | PP4E@learning-python.com | Fri, 04 Jun 2010 00:44:32 -0000
View	Re: Live Organ Transplants! | PP4E@learning-python.com | Fri, 04 Jun 2010 02:05:30 -0000
View	Fwd: Live Organ Transplants! | PP4E@learning-python.com | Fri, 04 Jun 2010 02:26:35 -0000
View	Already got one... | PP4E@learning-python.com | Fri, 04 Jun 2010 06:30:26 -0000
View	test reply all | PP4E@learning-python.com | Sat, 12 Jun 2010 15:47:01 -0000
View	source code line count sheets, june | PP4E@learning-python.com | Sun, 13 Jun 2010 18:25:22 -0000
View	hello from PyMailCGI | PP4E@learning-python.com | Sat, 19 Jun 2010 21:09:00 -0000

PYTHON powered Back to root page

Рис. 16.8. *Страница PyMailCGI списка выбора сообщений, начало*

PyMailCGI: Error pa...

http://localhost:8000/cgi-bin/onViewPswdSubmit.py

PyMailCGI Error

Error Description

Error loading mail index

Python Exception

<class 'socket.gaierror'>

Exception details

[Errno 11004] getaddrinfo failed

Exception traceback

```
File "C:\Users\mark\Stuff\Books\4E\PP4E\dev\Examples\PP4E\Internet\Web\PyMailCgi\cg
  newmails = loadmail.loadmailhdrs(mailsite, mailuser, mailpswd)
File "C:\Users\mark\Stuff\Books\4E\PP4E\dev\Examples\PP4E\Internet\Web\PyMailCgi\cg
  hdrs, sizes, full = fetcher.downloadAllHeaders()     # get list of hdr text
File "..\..\..\..\PP4E\Internet\Email\mailtools\mailFetcher.py", line 126, in downl
  server = self.connect()                # mbox now locked until quit
File "..\..\..\..\PP4E\Internet\Email\mailtools\mailFetcher.py", line 34, in connec
  server = poplib.POP3(self.popServer)
```

Рис. 16.9. *Страница PyMailCGI с сообщением об ошибке регистрации*

Передача информации о состоянии в параметрах URL-ссылки

Центральным механизмом, действующим в примере 16.7, является генерация адресов URL, содержащих номера сообщений и информацию об учетной записи почты. Щелчок на любой из ссылок View (Просмотреть) в списке запускает другой сценарий, который использует информацию в параметрах URL-ссылки для загрузки и вывода выбранного сообщения. Как говорилось в главе 15, поскольку ссылки в списке запрограммированы так, чтобы «уметь» загружать конкретное сообщение, они фактически несут в себе информацию о том, какое действие должно быть следующим. На рис. 16.10 показана часть разметки HTML, генерируемой этим сценарием (чтобы увидеть ее у себя, воспользуйтесь пунктом меню View Source (Исходный код страницы или Просмотр HTML-кода) в своем броузере, – я выбрал пункт меню Save As (Сохранить как) и затем открыл полученный результат в окне просмотра исходного кода в Internet Explorer).

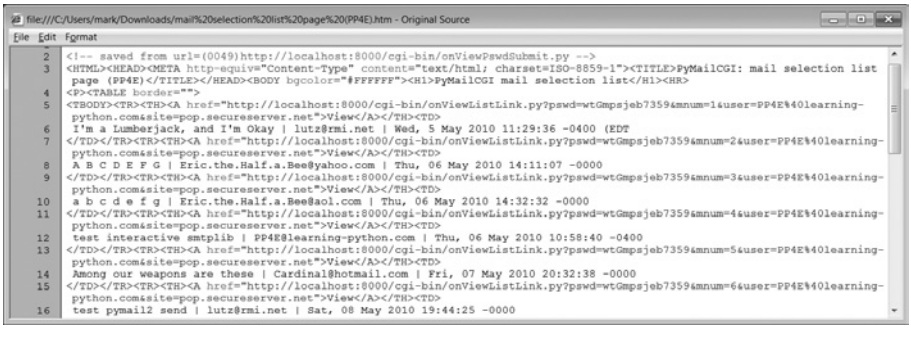


Рис. 16.10. *Код разметки HTML списка просмотра, сгенерированный PyMailCGI*

Вам все понятно на рис. 16.10? Если вы не сможете прочесть такой код разметки HTML, то сможет ваш броузер. Для читателей с ограниченными возможностями синтаксического анализа ниже приводится отдельно одна из ссылок, в которую добавлено форматирование в виде переносов строк и пробелов, облегчающее понимание:

```
<tr><th><a href="onViewListLink.py?
                    pswd=wtGmpsjeb7359&
                    mnum=5&
                    user=PP4E%40learning-python.com&
                    site=pop.secureserver.net">View</a>
<td>Among our weapons are these | Cardinal@hotmail.com
                                | Fri, 07 May 2010 20:32...
```

Программа PyMailCGI генерирует относительные минимальные адреса URL (имя сервера и путь определяются по предыдущей странице, если они не были установлены модулем commonhtml). Щелчок на слове «View» в гиперссылке, отображаемой этой разметкой HTML, запускает сценарий onViewListLink, которому передаются все параметры, добавленные в конец URL: имя пользователя POP, номер сообщения POP для письма, ассоциированного со ссылкой, а также пароль POP и данные о сайте. В сценарии, который выполняется следующим, эти значения можно будет получить из объекта, возвращаемого конструктором cgi.FieldStorage. Обратите внимание, что параметр номера сообщения POP mnum отличается для каждой ссылки, потому что щелчок на каждой из них открывает определенное сообщение, и что текст после тега <td> извлекается из заголовков сообщений с помощью пакета mailtools, использующего пакет email.

Модуль commonhtml выполняет экранирование параметров ссылки с помощью модуля urllib.parse, а не cgi.escape, потому что они являются частью URL. Это очевидно только в параметре пароля pswd – его значение зашифровано и может содержать произвольные байты, но urllib.parse дополнительно экранирует небезопасные символы в зашифрованной строке в соответствии с соглашениями об адресах URL (вот откуда берутся все эти последовательности символов %xx). Ничего страшного, если при шифровании получаются странные и даже непечатаемые символы – операция экранирования URL сделает их доступными для передачи. Когда пароль попадет в следующий сценарий, конструктор cgi.FieldStorage выполнит обратное преобразование экранированных последовательностей в адресе URL, заменив в строке с зашифрованным паролем экранированные последовательности %.

Полезно посмотреть, как commonhtml формирует параметры ссылок с информацией о состоянии. Ранее мы узнали, как с помощью функции urllib.parse.quote_plus выполнить экранирование строки перед включением ее в адрес URL:

```
>>> import urllib.parse
>>> urllib.parse.quote_plus("There's bugger all down here on Earth")
'There%27s+bugger+all+down+here+on+Earth'
```

Однако модуль commonhtml вызывает функцию urllib.parse.urlencode более высокого уровня, транслирующую словарь, состоящий из пар *имя:значение*, в законченную строку запроса с параметрами, которую можно добавить в URL после маркера ?. Ниже приводится пример использования функции urlencode в интерактивной оболочке:

```
>>> parmdict = {'user': 'Brian',
...             'pswd': '#!/spam',
...             'text': 'Say no more, squire!'}
>>> urllib.parse.urlencode(parmdict)
'text=Say+no+more%2C+squire%21&pswd=%23%21%2Fspam&user=Brian'
```

```
>>> "%s?%s" % ("http://scriptname.py", urllib.parse.urlencode(parmdict))
'http://scriptname.py?text=Say+no+more%2C+squire%21&pswd=%23%21%2Fspam&
user=Brian'
```

Внутри функция urlencode передает каждое имя и значение из словаря встроенной функции str (чтобы создать из них строки), а затем, при добавлении к результату, пропускает их все через функцию urllib.parse.quote_plus. Сценарий CGI строит список аналогичных словарей и передает его модулю commonhtml для формирования страницы со списком выбора.[1]

В общем случае такая генерация URL с параметрами является одним из способов передачи информации о состоянии следующему сценарию (наряду со скрытыми полями форм и базами данных и, обсуждавшимися в главе 15). Без такой информации о состоянии пользователю пришлось бы заново вводить имя, пароль и имя сайта на каждой посещаемой странице.

Между прочим, генерируемый этим сценарием список не сильно отличается по своим возможностям от того, который мы конструировали в программе PyMailGUI в главе 14, а два имеющихся отличия носят исключительно косметический характер. На рис. 16.11 показано, как выглядит тот же список, показанный ранее на рис. 16.8, в графическом интерфейсе почтового клиента.

Важно заметить, что программа PyMailGUI не посылает разметку HTML броузеру, а использует для создания интерфейса пользователя библиотеку tkinter. Она также целиком выполняется на стороне клиента и взаимодействует с почтовыми серверами напрямую, загружая почту с сервера POP на компьютер клиента через сокеты по требованию. Поскольку на протяжении всего сеанса она хранит всю необходимую информацию в памяти, программа PyMailGUI легко может минимизировать количество обращений к почтовому серверу. После начальной загрузки заголовков при следующем запуске ей достаточно будет загрузить заголовки только вновь поступивших сообщений. Кроме того, при удалении писем она может обновлять список сообщений в памяти, не загружая его заново с сервера, и обладает достаточным объемом информации, чтобы проверить соответствие с почтовым ящиком на сервере и выполнить удаление максимально безопасно. Программа PyMailGUI также сохраняет письма, которые уже просматривались, – в течение сеанса работы программы эти письма не требуется загружать повторно.

---

[1] Формально кроме этого обычно желательно экранировать разделители & в созданных ссылках URL с помощью cgi.escape, на случай если имя какого-либо параметра (такое, как &amp=high) может совпасть с именем экранированной последовательности HTML. Смотрите подробности в главе 15; здесь такое экранирование не производится, потому что нет конфликтов между URL и HTML.

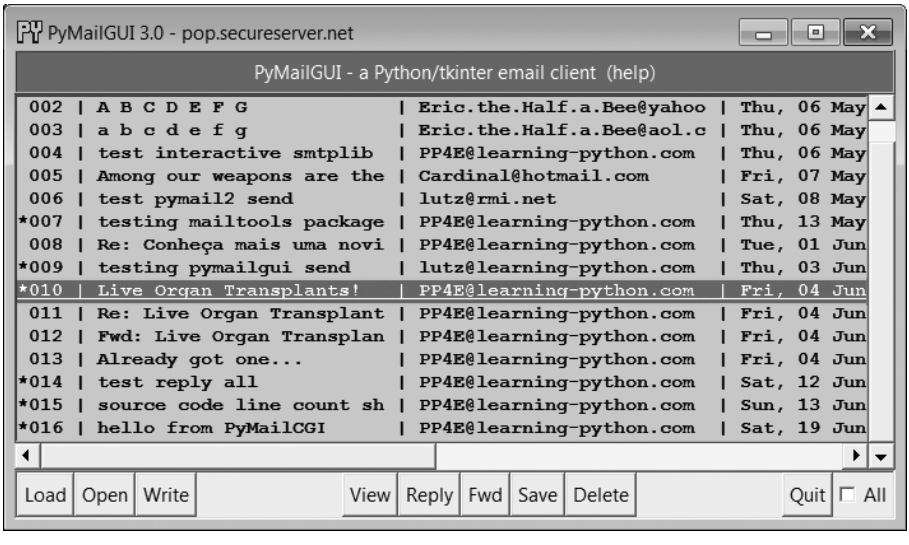


Рис. 16.11. *Тот же самый список сообщений в PyMailGUI*

Напротив, PyMailCGI выполняется на веб-сервере и просто отображает текст почты в броузере клиента – почта загружается с почтового сервера POP на веб-сервер, где выполняются сценарии CGI. Из-за автономной природы сценариев CGI приложение PyMailCGI не имеет автоматической памяти для сохранения информации между вызовами страниц, и на протяжении одного и того же сеанса ей может потребоваться повторно загружать заголовки и ранее просматривавшиеся сообщения. У этих архитектурных различий есть некоторые важные следствия, которые мы обсудим несколько позже.

Протоколы защиты данных

Обратите внимание, что в исходном программном коде сценария `onViewPswdSubmit` (пример 16.7) введенный пароль при добавлении в словарь параметров передается функции `encode`, благодаря чему появляется в URL гиперссылок в зашифрованном виде. Пароли также кодируются для передачи (с использованием последовательностей `%`, если это необходимо) и позднее декодируются и расшифровываются в других сценариях при необходимости обратиться к учетной записи POP. Шифрование пароля с помощью `encode` лежит в основе политики безопасности PyMailCGI.

В современных версиях Python имеется стандартный модуль `ssl`, поддерживающий защищенные сокеты (Secure Sockets Layer, SSL) с помощью своих функций-оберток, если необходимая библиотека встроена в интерпретатор. Эта поддержка автоматически шифрует передаваемые данные, обеспечивая их безопасность в Сети. К сожалению, по причи-

нам, которые мы будем обсуждать далее в этой главе, когда дойдем до модуля secret.py (пример 16.13), это решение не является универсальным для передачи паролей в PyMailCGI. Коротко скажу лишь, что веб-сервер, реализованный на языке Python и используемый здесь, не поддерживает HTTPS – защищенную версию протокола HTTP. Поэтому была разработана альтернативная схема, уменьшающая вероятность кражи информации об учетной записи почты при передаче ее по Сети.

Вот как она действует. Когда сценарий вызывается из формы страницы ввода пароля, он получает только один входной параметр: пароль, введенный в форму. Имя пользователя импортируется из модуля mailconfig, установленного на сервере, – оно не передается вместе с незашифрованным паролем, потому что перехват такой комбинации злоумышленником значительно облегчил бы ему жизнь.

Чтобы передать имя пользователя и пароль следующей странице как информацию о состоянии, этот сценарий помещает их в конец URL страницы списка почтовых сообщений, но только после шифрования пароля с помощью secret.encode – функции из модуля, расположенного на сервере, и который может быть различным для каждого места, где установлена программа PyMailCGI. В действительности PyMailCGI написана так, что она не обязана ничего знать о функции шифрования – поскольку шифрование обеспечивается отдельным модулем, можно выбрать тот, который вам больше нравится. Если вы не станете публиковать реализацию своего модуля шифрования, то зашифрованный пароль, передаваемый вместе с именем пользователя, не принесет пользы злоумышленнику.

В результате PyMailCGI обычно не передает и не получает одновременно имя пользователя и пароль в одной операции, если только пароль не зашифрован с помощью выбранной вами функции. Это несколько ограничивает полезность программы (поскольку на сервере можно определить имя пользователя только для одной учетной записи почты), но альтернатива в виде двух страниц – одной для ввода пароля и другой для ввода имени пользователя – была бы еще менее дружелюбной. В целом при желании читать свою почту с помощью этой системы в том виде, как она реализована, необходимо установить ее файлы на своем сервере, поправить *mailconfig.py*, чтобы он отражал параметры конкретной учетной записи, и изменить модуль шифрования secret.py по желанию.

Чтение почты с использованием прямых адресов URL

Одно исключение: поскольку любой сценарий CGI можно вызвать с явными параметрами в URL вместо значений полей формы, и так как модуль commonhtml пытается получить входные данные из объекта формы, прежде чем импортировать их из mailconfig, любой пользователь может проверить свою почту с помощью PyMailCGI, не устанавливая и не настраивая ее. Например, следующий адрес URL (без переноса строки, добавленного, чтобы уместить его на странице), если ввести его в адрес-

ной строке броузера или отправить его с помощью такого инструмента, как urllib.request:

```
http://localhost:8000/cgi-bin/
  onViewPswdSubmit.py?user=lutz&pswd=guess&site=pop.earthlink.net
```

действительно загрузит почту в список выбора, подобный тому, как показано на рис. 16.8, какие бы значения имени пользователя, пароля и сайта ни были в него добавлены. Из списка выбора можно затем осуществлять просмотр почты, отвечать на нее, переадресовывать и удалять.

Обратите внимание, что при такой организации взаимодействий пароль, отправляемый в URL такого вида, не шифруется. В последующих сценариях, однако, предполагается отправка введенного пароля в зашифрованном виде, что затрудняет их использование с явными URL (потребуется обеспечить соответствие с зашифрованной формой, создаваемой на сервере модулем secret). Пароли шифруются при добавлении их в ссылки из списка выбора страницы ответа и остаются зашифрованными в последующих URL и скрытых полях форм.

Не используйте URL такого типа, если только вам не безразлично, что ваш почтовый пароль окажется открытым. Одновременная пересылка незашифрованного имени пользователя и пароля через Сеть в таком URL крайне небезопасна и делает их открытыми для посторонних. Это все равно что выдать всю информацию о своей электронной почте – всякий, кто перехватит этот URL или подсмотрит его в файлах журналов на сервере, получит полный доступ к вашей почтовой учетной записи. Еще более ненадежными такие адреса URL делает тот факт, что формат URL опубликован в книге, которая будет широко распространяться по всему свету.

Если вас заботит безопасность, и вы хотите пользоваться программой PyMailCGI на удаленном сервере, установите ее на собственном сервере и настройте модули mailconfig и secret. Это должно, по крайней мере, гарантировать, что информация с вашим именем и паролем никогда не будет передаваться незашифрованной в одной операции. Эта схема все же не является абсолютно надежной, поэтому будьте осторожны. Без защищенных сокетов и HTTPS Интернет станет средой типа «используйте на свой страх и риск».

Страница просмотра сообщений

Вернемся к нашей последовательности страниц. В настоящий момент мы все еще рассматриваем список выбора сообщений на рис. 16.8. Если щелкнуть на одной из этих сгенерированных гиперссылок, то URL, хранящий информацию о состоянии, запустит на сервере сценарий, представленный в примере 16.8, передавая ему номер выбранного сообщения и информацию о почтовой учетной записи (пользователь, пароль и сайт) в виде параметров в конце адреса URL сценария.

Пример 16.8. PP4E\Internet\Web\PyMailCgi\cgi-bin\onViewListLink.py

```
#!/usr/bin/python
"""
##############################################################################
Вызывается щелчком мыши на ссылке, указывающей на сообщение в главном
списке: создает страницу просмотра;

конструктор cgi.FieldStorage преобразует экранированные последовательности
в ссылках с помощью urllib.parse (%xx и '+', замещающие пробелы, уже
преобразованы обратно); в 2.0+ здесь загружается только одно сообщение,
а не весь список; в 2.0+ мы также отыскиваем основную текстовую часть
сообщения, вместо того чтобы вслепую отображать полный текст (со всеми
вложениями), и генерируем ссылки на файлы вложений, сохраненные на сервере;
сохранение файлов вложений возможно только для 1 пользователя и 1 сообщения;
большая часть улучшений в 2.0 обусловлена использованием пакета mailtools;

3.0: перед анализом с помощью пакета email пакет mailtools декодирует байты
полного текста сообщения;
3.0: для отображения пакет mailtools декодирует основной текст,
commonhtml декодирует заголовки сообщения;
##############################################################################
"""


import cgi
import commonhtml, secret
from externs import mailtools
#commonhtml.dumpstatepage(0)

def saveAttachments(message, parser, savedir='partsdownload'):
    """
    сохраняет части полученного сообщения в файлы на
    сервере для дальнейшего просмотра в веб-броузере пользователя
    """
    import os
    if not os.path.exists(savedir):      # CWD CGI-сценария на сервере
        os.mkdir(savedir)                # будет открываться в броузере
    for filename in os.listdir(savedir): # удалить прежние файлы: временные!
        dirpath = os.path.join(savedir, filename)
        os.remove(dirpath)
    typesAndNames = parser.saveParts(savedir, message)
    filenames = [fname for (ctype, fname) in typesAndNames]
    for filename in filenames:
        os.chmod(filename, 0o666)         # некоторые серверы требуют права
                                          # на чтение/запись
    return filenames

form = cgi.FieldStorage()
user, pswd, site = commonhtml.getstandardpopfields(form)
pswd = secret.decode(pswd)
```

```
try:
    msgnum   = form['mnum'].value                    # из URL-ссылки
    parser   = mailtools.MailParser()
    fetcher  = mailtools.SilentMailFetcher(site, user, pswd)
    fulltext = fetcher.downloadMessage(int(msgnum)) # не используйте eval!
    message  = parser.parseMessage(fulltext)        # Message в пакете email
    parts    = saveAttachments(message, parser)     # для URL-ссылок
    mtype, content = parser.findMainText(message)   # первая текстовая часть
    commonhtml.viewpage(msgnum, message, content, form, parts)  # зашифр.
                                                                # пароль
except:
    commonhtml.errorpage('Error loading message')
```

И снова большая часть работы здесь выполняется в модуле `commonhtml`, который приведен ниже в этом разделе (пример 16.14). Этот сценарий вводит дополнительную логику для расшифровывания переданного пароля (с помощью настраиваемого модуля шифрования `secret`) и извлечения заголовков и текста выбранного сообщения с помощью пакета `mailtools` из главы 13. В конечном счете сценарий загружает полный текст выбранного сообщения, анализирует его и декодирует с помощью пакета `mailtools`, использующего стандартный модуль `poplib` и пакет `email`. И хотя при повторной попытке просмотреть сообщение его снова придется загрузить, тем не менее, начиная с версии 2.0, PyMailCGI больше не загружает все сообщения целиком, чтобы получить выбранное.[1]

Еще одна новая, появившаяся в версии 2.0 функция `saveAttachments` используется в этом сценарии для выделения частей полученного сообщения и сохранения их в каталоге на веб-сервере. Как уже говорилось выше в этой главе, после этого в страницу просмотра добавляются ссылки с адресами URL, указывающими на сохраненные в файлах части. Ваш веб-броузер сможет открывать их исходя из их имен и содержимого. Все операции по извлечению, декодированию и именованию унаследованы из пакета `mailtools`. Файлы вложений хранятся временно – они будут удалены при выборе следующего сообщения. Кроме того, в данной реализации все файлы сохраняются в единственном каталоге, и по этой причине система может использоваться только одним пользователем.

Если сообщение было успешно загружено и проанализировано, в результате будет создана страница, изображенная на рис. 16.12, позволяющая просматривать текст сообщения, но не редактировать его. Функция `commonhtml.viewpage` генерирует параметр HTML «read-only» («только для чтения») для всех текстовых элементов на этой странице. При внимательном рассмотрении можно заметить, что это то сообщение, которое

[1] Обратите внимание, что номер сообщения передается в виде строки и использовать его для загрузки сообщения можно только после преобразования в целое число. Но здесь нельзя выполнять преобразование с помощью функции `eval`, так как эта строка была передана через Сеть и могла быть включена в конец произвольного URL (помните старое предупреждение по этому поводу?).

было отправлено из страницы на рис. 16.3 и которое видно в конце списка на рис. 16.8.

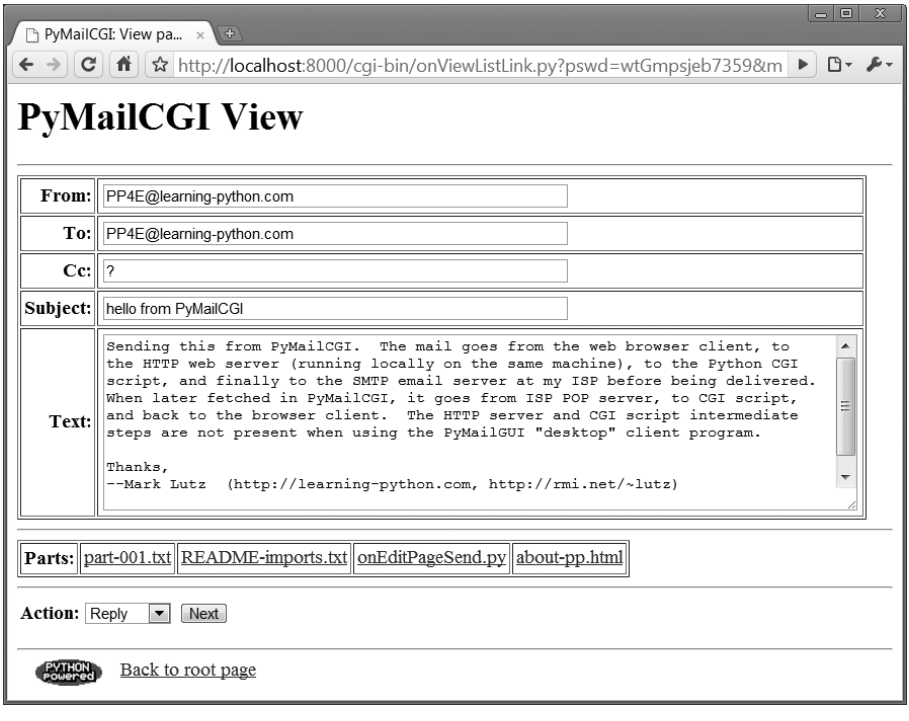


Рис. 16.12. *Страница PyMailCGI просмотра сообщения*

В нижней части страницы просмотра имеется раскрывающийся список выбора операции – если нужно что-то сделать с письмом, выберите в этом списке требуемое действие (Reply (Ответить), Forward (Переслать) или Delete (Удалить)) и щелкните на кнопке Next (Далее), чтобы перейти к следующей странице. Если вы собираетесь только просматривать сообщения, щелкните внизу на ссылке Back to root page (Назад на главную страницу) или на кнопке Назад (Back) броузера, чтобы вернуться на главную страницу со списком.

Как уже говорилось, на рис. 16.12 отображается сообщение, отправленное нами ранее в этой главе, которое мы решили просмотреть после получения. Обратите внимание на ссылки «Parts:» – щелчок на любой из них открывает адрес URL, указывающий на временный файл, находящийся на сервере, в соответствии с правилами, действующими в броузере для данного типа файлов. Например, щелчок на файле с расширением «.txt» вероятнее всего приведет к открытию этого файла в броузере или в текстовом редакторе. В других сообщениях щелчок на файле с расширением «.jpg» может привести к его открытию в программе про-

смотра графических изображений, на файле с расширением «.pdf» может запустить программу Adobe Reader и так далее. На рис. 16.13 показан результат щелчка на файле вложения с расширением «.py» в странице на рис. 16.12, отображенной в броузере Chrome.

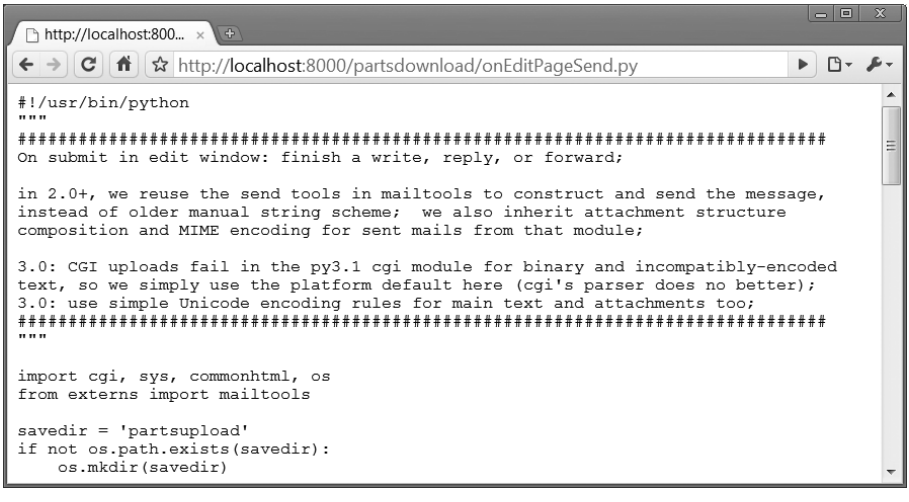


Рис. 16.13. *Результат попытки просмотра файла вложения*

Передача информации о состоянии в скрытых полях форм HTML

То, чего вы не видите на странице просмотра, изображенной на рис. 16.12, так же важно, как то, что вы видите. За деталями реализации обращайтесь к примеру 16.14, но замечу, что здесь происходит нечто новое. Номер исходного сообщения, а также имя пользователя и пароль (по-прежнему зашифрованный), переданные этому сценарию в составе URL-ссылки с информацией о состоянии, оказываются скопированными в разметку HTML этой страницы в виде значений скрытых полей ввода формы. Ниже приводится программный код в модуле `commonhtml`, создающий скрытые поля:

```
print('<form method=post action="%s/onViewPageAction.py">' % urlroot)
print('<input type=hidden name=mnum value="%s">' % msgnum)
print('<input type=hidden name=user value="%s">' % user)  # из страницы|url
print('<input type=hidden name=site value="%s">' % site)  # для удаления
print('<input type=hidden name=pswd value="%s">' % pswd)  # зашифр. пароль
```

Как мы уже знаем, скрытые поля, подобно параметрам URL в генерируемых гиперссылках, позволяют встраивать информацию о состоянии внутрь самой веб-страницы. Увидеть эту скрытую информацию можно только в исходном коде разметки страницы, потому что скры-

тые поля не отображаются. Но при щелчке на кнопке отправки этой формы скрытые значения автоматически передаются очередному сценарию вместе с видимыми полями формы.

На рис. 16.14 показан исходный код разметки, сгенерированной для страницы просмотра другого сообщения, – скрытые поля ввода, используемые для передачи информации о состоянии выбранного сообщения, находятся в начале файла.

http://localhost:8000/cgi-bin/onViewListLink.py?pswd=wtGmpsjeb7359&mnum=16&user=PP4E%40learning-python.c...

File Edit Format

```
1  <html><head><title>PyMailCGI: View page (PP4E)</title></head>
2  <body bgcolor="#FFFFFF"><h1>PyMailCGI View</h1><hr>
3  <form method=post action="onViewPageAction.py">
4  <input type=hidden name=mnum value="16">
5  <input type=hidden name=user value="PP4E@learning-python.com">
6  <input type=hidden name=site value="pop.secureserver.net">
7  <input type=hidden name=pswd value="wtGmpsjeb7359">
8  <table border cellpadding=3>
9  <tr><th align=right>From:
10     <td><input type=text
11     name=From value="PP4E@learning-python.com" readonly size=60>
12 <tr><th align=right>To:
13     <td><input type=text
14     name=To value="PP4E@learning-python.com" readonly size=60>
15 <tr><th align=right>Cc:
16     <td><input type=text
17     name=Cc value="?" readonly size=60>
18 <tr><th align=right>Subject:
19     <td><input type=text
20     name=Subject value="hello from PyMailCGI" readonly size=60>
21 <tr><th align=right>Text:
22 <td><textarea name=text cols=80 rows=10 readonly>
23 Sending this from PyMailCGI.  The mail goes from the web browser client, to
24 the HTTP web server (running locally on the same machine), to the Python CGI
25 script, and finally to the SMTP email server at my ISP before being delivered.
26 When later fetched in PyMailCGI, it goes from ISP POP server, to CGI script,
27 and back to the browser client.  The HTTP server and CGI script intermediate
28 steps are not present when using the PyMailGUI "desktop" client program.
29
30 Thanks,
31 --Mark Lutz  (http://learning-python.com, http://rmi.net/~lutz)
32
33 </textarea></table>
34 <hr><table border cellpadding=3><tr><th>Parts:
35 <td><a href=../partsdownload/part-001.txt>part-001.txt</a>
36 <td><a href=../partsdownload/README-imports.txt>README-imports.txt</a>
37 <td><a href=../partsdownload/onEditPageSend.py>onEditPageSend.py</a>
38 <td><a href=../partsdownload/about-pp.html>about-pp.html</a>
```

Рис. 16.14. *Разметка HTML, сгенерированная для страницы просмотра PyMailCGI*

В результате скрытые поля ввода в разметке HTML, так же как параметры в конце генерируемых URL, действуют как временное хранилище и сохраняют информацию о состоянии, передавая ее между страницами и этапами взаимодействия с пользователем. Оба механизма представляют собой веб-эквивалент переменных в языках программирования. Их удобно использовать, когда приложению требуется что-то передать из одной страницы в другую.

Особенно полезными скрытые поля оказываются, когда невозможно вызвать следующий сценарий с помощью гиперссылки со сгенерированным адресом URL, содержащим параметры. Например, следующее дей-

ствие в нашем сценарии запускается щелчком на кнопке передачи формы (Next (Далее)), а не на гиперссылке, поэтому для передачи состояния используются скрытые поля. Как и прежде, без этих скрытых полей пользователи вынуждены были бы снова вводить сведения об учетной записи на сервере POP в странице просмотра, если они нужны очередному сценарию (в нашем случае они нужны, если следующим действием является удаление сообщения).

Экранирование текста сообщения и паролей в HTML

Обратите внимание, что все видимое на странице просмотра сообщения, изображенной на рис. 16.14, подверглось экранированию с помощью `cgi.escape`. Поля заголовков и сам текст сообщения могут содержать специальные символы HTML и должны транслироваться как обычно. Например, поскольку некоторые почтовые программы позволяют отправлять сообщения в формате HTML, текст сообщения может содержать тег `</textarea>`, который без экранирования безнадежно испортит страницу ответа.

Здесь есть одна тонкость: экранированные последовательности HTML важны, только когда текст посылается броузеру сценарием CGI. Если этот текст затем передается другому сценарию (например, при отправке ответа), текст вернется в исходный, непреобразованный формат, когда будет снова получен на сервере. Броузер анализирует экранированные последовательности и не возвращает их обратно при отправке данных формы, поэтому в дальнейшем не требуется делать обратное преобразование. Например, ниже приводится часть текста после экранирования, посылаемого броузеру во время выполнения операции Reply (Ответить) (воспользуйтесь пунктом меню броузера View Source (Исходный код страницы или Просмотр HTML-кода), чтобы увидеть его в реальности):

```
<tr><th align=right>Text:
<td><textarea name=text cols=80 rows=10 readonly>
more stuff

--Mark Lutz (http://rmi.net/~lutz) [PyMailCgi 2.0]


&gt; -----Original Message-----
&gt; From: lutz@rmi.net
&gt; To: lutz@rmi.net
&gt; Date: Tue May 2 18:28:41 2000
&gt;
&gt; &lt;table&gt;&lt;textarea&gt;
&gt; &lt;/textarea&gt;&lt;/table&gt;
&gt; --Mark Lutz (http://rmi.net/~lutz) [PyMailCgi 2.0]
&gt;
&gt;
&gt; &gt; -----Original Message-----
```

После отправки этого ответа его текст выглядит так же, как перед экранированием (и в точности каким он был у пользователя на странице редактирования сообщения):

```
more stuff

--Mark Lutz (http://rmi.net/~lutz) [PyMailCgi 2.0]


> -----Original Message-----
> From: lutz@rmi.net
> To: lutz@rmi.net
> Date: Tue May 2 18:28:41 2000
>
> <table><textarea>
> </textarea></table>
> --Mark Lutz (http://rmi.net/~lutz) [PyMailCgi 2.0]
>
>
> > -----Original Message-----
```

Помимо обычного текста экранированию в соответствии с правилами HTML подвергается также и пароль. Хотя этого не видно в наших примерах, тем не менее, скрытое поле пароля в сгенерированной разметке HTML (как видно на рис. 16.14) после шифрования может иметь совершенно причудливый вид. Оказывается, пароль POP остается зашифрованным при помещении в скрытые поля форм HTML. Так и должно быть по соображениям безопасности. Значения скрытых полей страницы можно увидеть при просмотре исходного кода разметки в броузере, воспользовавшись пунктом меню броузера View Source (Исходный код страницы или Просмотр HTML-кода), и нельзя исключить возможность перехвата текста этой страницы при передаче в Сети.

Однако когда пароль помещается в скрытое поле, он уже не подвергается преобразованию в соответствии с правилами оформления адресов URL, как при добавлении в конец URL-ссылки с информацией о состоянии. В зависимости от алгоритма шифрования пароль, сгенерированный здесь как значение скрытого поля, может содержать непечатаемые символы, но броузеру это безразлично, поскольку поле пропускается через функцию `cgi.escape`, как и все другое, помещаемое в поток HTML ответа. Модуль `commonhtml` заботится о том, чтобы при создании страницы просмотра весь текст и заголовки были обработаны функцией `cgi.escape`.

Для сравнения на рис. 16.15 показано, как выглядит сообщение, приведенное на рис. 16.12, при просмотре в PyMailGUI – почтовом клиенте с графическим интерфейсом на основе tkinter из главы 14. В этой программе список частей сообщения можно получить с помощью кнопки Parts (Части) и извлечь, сохранить и открыть с помощью кнопки Split (Разбить); кроме того, в нашем распоряжении имеются кнопки быстрого

доступа к частям и вложениям, расположенные ниже заголовков сообщения. С точки зрения конечного пользователя интерфейс выглядит похожим.

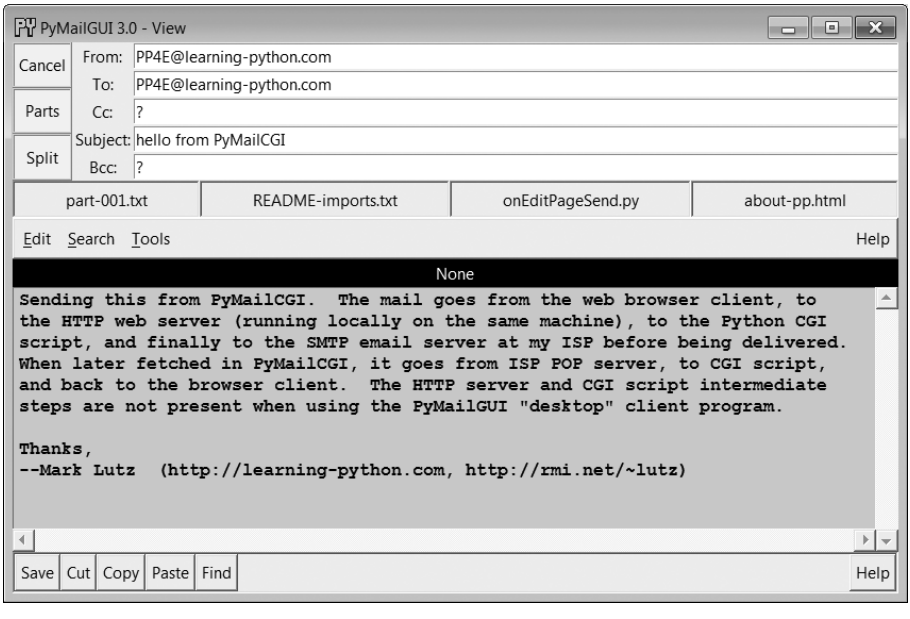


Рис. 16.15. *Так выглядит в PyMailGUI то же сообщение, что и на рис. 16.12*

Однако, с точки зрения реализации, модели отличаются существенно. Программе PyMailGUI не требуется беспокоиться о передаче информации о состоянии в адресах URL или скрытых полях (она хранит состояние в переменных Python) или экранировать HTML и строки URL (броузеры не используются, а после загрузки почты не требуется выполнять передачу через сеть). Ей также не приходится создавать ссылки на временные файлы, чтобы обеспечить доступ к частям сообщения – сообщение хранится в памяти, присоединенное к объекту окна, и продолжает существовать между операциями. С другой стороны, программе PyMailGUI требуется, чтобы на компьютере клиента был установлен Python, но к этому мы еще вернемся через несколько страниц.

Обработка загруженной почты

В данный момент нашего гипотетического взаимодействия с веб-приложением PyMailCGI мы просматриваем почтовое сообщение (рис. 16.12), выбранное на странице со списком. Если, находясь на странице просмотра сообщения, выбрать в раскрывающемся списке некоторое действие и щелкнуть на кнопке Next (Далее), на сервере будет вызван сценарий,

представленный в примере 16.9, который выполнит операцию создания ответа, пересылки или удаления для просматриваемого сообщения.

Пример 16.9. PP4E\Internet\Web\PyMailCgi\cgi-bin\onViewPageAction.py

```
#!/usr/bin/python
"""
###############################################################################
Вызывается при отправке формы в окне просмотра сообщения: выполняет
выбранное действие=(fwd, reply, delete);
в 2.0+ повторно используется логика удаления в пакете mailtools,
первоначально реализованная для PyMailGUI;
###############################################################################
"""


import cgi, commonhtml, secret
from externs import mailtools, mailconfig
from commonhtml import getfield

def quotetext(form):
    """
    обратите внимание, что заголовки поступают из формы предыдущей страницы,
    а не получаются в результате повторного анализа почтового сообщения;
    это означает, что функция commonhtml.viewpage должна передавать дату
    в скрытом поле
    """
    parser = mailtools.MailParser()
    addrhdrs = ('From', 'To', 'Cc', 'Bcc')        # декодируется только имя
    quoted = '\n-----Original Message-----\n'
    for hdr in ('From', 'To', 'Date'):
        rawhdr = getfield(form, hdr)
        if hdr not in addrhdrs:
            dechdr = parser.decodeHeader(rawhdr)  # 3.0: декод. для отображ.
        else:                                     # закодиров. при отправке
            dechdr = parser.decodeAddrHeader(rawhdr) # только имена в адресах
        quoted += '%s: %s\n' % (hdr, dechdr)
    quoted += '\n' + getfield(form, 'text')
    quoted = '\n' + quoted.replace('\n', '\n> ')
    return quoted

form = cgi.FieldStorage()      # извлечь данные из формы или из URL
user, pswd, site = commonhtml.getstandardpopfields(form)
pswd = secret.decode(pswd)

try:
    if form['action'].value == 'Reply':
        headers = {'From': mailconfig.myaddress,  # 3.0: декодирование
                   'To': getfield(form, 'From'),  # выполняет commonhtml
                   'Cc': mailconfig.myaddress,
                   'Subject': 'Re: ' + getfield(form, 'Subject')}
        commonhtml.editpage('Reply', headers, quotetext(form))
```

```
        elif form['action'].value == 'Forward':
            headers = {'From': mailconfig.myaddress, # 3.0: декодирование
                       'To': '',                     # выполняет commonhtml
                       'Cc': mailconfig.myaddress,
                       'Subject': 'Fwd: ' + getfield(form, 'Subject')}
            commonhtml.editpage('Forward', headers, quotetext(form))

        elif form['action'].value == 'Delete': # поле mnum необходимо здесь
            msgnum  = int(form['mnum'].value)  # но не eval(): может быть код
            fetcher = mailtools.SilentMailFetcher(site, user, pswd)
            fetcher.deleteMessages([msgnum])
            commonhtml.confirmationpage('Delete')

        else:
            assert False, 'Invalid view action requested'
    except:
        commonhtml.errorpage('Cannot process view action')
```

Этот сценарий получает всю информацию о выбранном сообщении в виде полей формы (некоторые из них могут быть скрытыми и зашифрованными) вместе с именем выбранного действия. Следующий шаг зависит от выбранного действия:

Действия Reply (Ответить) и Forward (Переслать)

Генерируют страницу редактирования сообщения, в которой строки исходного сообщения автоматически цитируются с указанием символа > перед каждой из них.

Действие Delete (Удалить)

Вызывает немедленное удаление просматриваемого сообщения с помощью инструмента, импортированного из пакета `mailtools` из главы 13.

Во всех этих действиях используются данные, передаваемые из формы предыдущей страницы, но только для действия Delete (Удалить) требуются имя пользователя и пароль и декодирование полученного пароля (они поступают из скрытых полей формы в разметке HTML, сгенерированной предыдущей страницей).

Ответ и пересылка

Если в качестве очередного действия выбрать операцию Reply (Ответить), сценарий сгенерирует страницу редактирования сообщения, изображенную на рис. 16.16. Текст сообщения на этой странице можно редактировать, а при нажатии кнопки Send (Отправить) запускается сценарий отправки почты, который мы видели в примере 16.4. Если все пойдет удачно, будет получена та же страница подтверждения, которую мы получали раньше при создании нового сообщения (рис. 16.4).

Операция Forward (Переслать) осуществляется практически так же, за исключением некоторых отличий в заголовках сообщений. Всю эту рабо-

ту мы получаем «бесплатно», так как страницы Reply (Ответить) и Forward (Переслать) генерируются вызовом `commonhtml.editpage` — той же утилиты, с помощью которой создается страница составления нового сообщения. Мы просто передаем этой утилите готовые строки заголовков (например, при создании ответа к тексту темы добавляется приставка «Re:»). Такого же рода прием с повторным использованием применялся в PyMailGUI, но в ином контексте. В PyMailCGI один сценарий обрабатывает три страницы; в PyMailGUI один суперкласс и одна функция обратного вызова обрабатывает три кнопки, но архитектура аналогична.

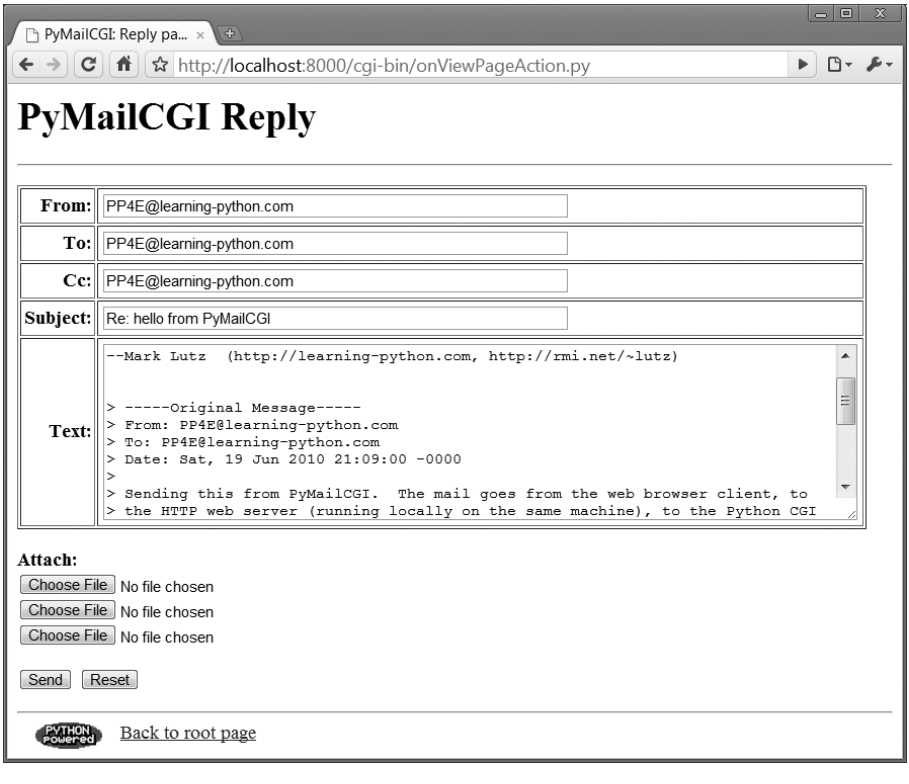


Рис. 16.16. *Страница PyMailCGI создания ответа*

Удаление

При выборе операции Delete (Удалить) в странице просмотра сообщения после щелчка на кнопке Next (Далее) сценарий `onViewPageAction` немедленно удалит просматриваемое сообщение. Удаление осуществляется путем вызова повторно используемой вспомогательной функции удаления, реализованной в пакете `mailtools` (см. главу 13). В предыдущей версии вызов этой утилиты был заключен в вызов `commonhtml.runsilent`,

который предотвращает вывод данных функциями `print` в поток HTML ответа (они являются лишь сообщениями о состоянии, а не кодом разметки HTML). В этой версии тот же эффект достигается за счет использования классов `Silent` из пакета `mailtools`. На рис. 16.17 показана операция удаления в действии.

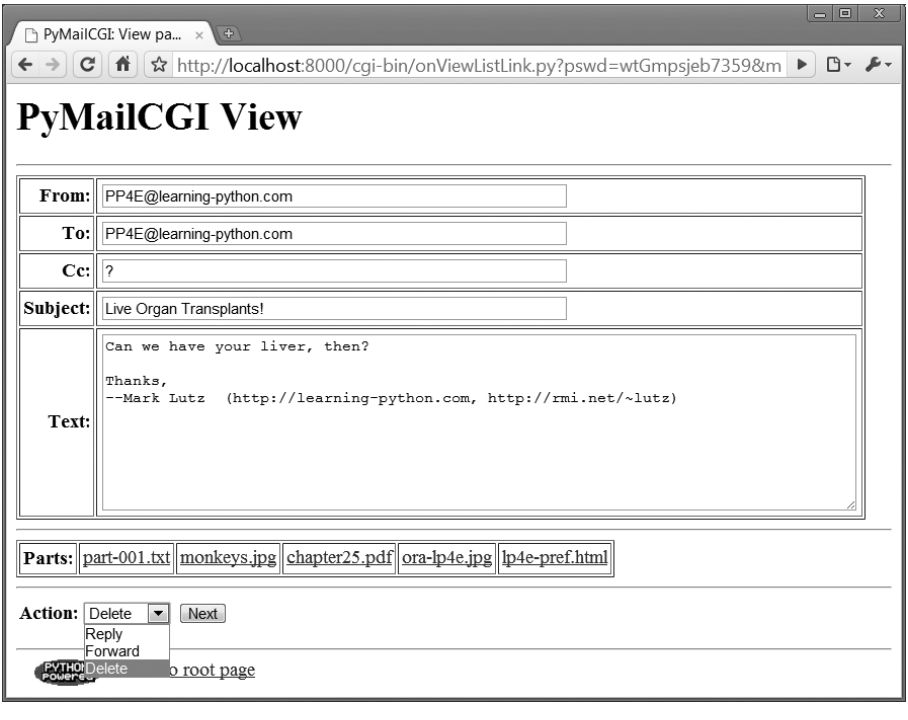


Рис. 16.17. *Страница PyMailCGI просмотра сообщений, выбрана операция удаления*

Между прочим, обратите внимание на присутствие вложений различных типов на рис. 16.17. В версии 3.0 можно отправлять только текстовые вложения из-за ухудшенной поддержки выгрузки файлов через CGI в Python 3.1, описывавшейся выше, но мы по-прежнему можем просматривать произвольные вложения во входящих сообщениях, получаемых от других отправителей. В этом сообщении присутствуют изображения и документ PDF. Такие вложения открываются в соответствии с настройками броузера – на рис. 16.18 показано, как Chrome обрабатывает щелчок на ссылке *monkeys.jpg* в нижней части страницы PyMailCGI, изображенной на рис. 16.17. Это то же самое изображение, что мы отправляли по протоколу FTP в главе 13 и с помощью PyMailGUI в главе 14, но здесь оно извлекается с помощью сценария CGI приложения PyMailCGI и возвращается локальным веб-сервером.

Рис. 16.18. *В PyMailCGI было выбрано вложение с изображением*

Вернемся к операции удаления. Как отмечалось, операция удаления является единственной, использующей данные учетной записи POP (пользователь, пароль и сайт), переданные в скрытых полях из предыдущей страницы просмотра сообщения. Напротив, операции создания ответа и пересылки формируют страницу редактирования, которая в конечном счете отошлет сообщение серверу SMTP – никакие данные POP им не требуются и не передаются.

Но к этому моменту пароль POP накрутил в своем перемещении не одну милю. В действительности он мог пройти по телефонным линиям, спутниковым каналам связи и пересечь целые континенты, путешествуя с компьютера на компьютер. Его маршрут описывается ниже:

1. Ввод (клиент): пароль начинает свою жизнь с ввода на странице регистрации у клиента (или встраивания в явный адрес URL) в незашифрованном виде. При вводе в форму в веб-броузере каждый символ отображается в виде звездочки (*).
2. Загрузка списка сообщений (от клиента через CGI-сервер на POP-сервер): затем он передается от клиента сценарию CGI на сервере, который пересылает его вашему POP-серверу для загрузки списка почтовых сообщений. Клиент посылает пароль в незашифрованном виде.
3. Страница со списком URL (CGI-сервер клиенту): Для управления поведением следующего сценария пароль встраивается в саму веб-страницу со списком для выбора почтовых сообщений в виде параметров

запроса URL гиперссылок, зашифрованный и экранированный в соответствии с правилами URL.

4. Загрузка сообщения (от клиента через CGI-сервер на POP-сервер): Когда сообщение выбирается в списке, пароль посылается следующему сценарию, указанному в адресе URL, – сценарий CGI расшифровывает его и посылает POP-серверу для загрузки выбранного сообщения.
5. Поля на странице просмотра (CGI-сервер клиенту): Для управления поведением следующего сценария пароль встраивается в саму страницу просмотра в виде скрытых полей ввода, в зашифрованном виде и экранированный в соответствии с правилами HTML.
6. Удаление (от клиента через CGI-сервер на POP-сервер): Наконец, пароль снова передается от клиента на сервер CGI, на этот раз в виде значений скрытых полей формы – сценарий CGI расшифровывает его и посылает серверу POP для удаления сообщения.

Попутно сценарии передавали пароль между страницами как параметр запроса в адресе URL или как скрытое поле ввода HTML – в том и другом случае всегда передается зашифрованная строка, и никогда в одной операции не передаются одновременно незашифрованный пароль и имя пользователя. При запросе операции удаления перед передачей серверу POP пароль должен быть расшифрован с помощью модуля `secret`. Если сценарий смог снова обратиться к серверу POP и удалить выбранное сообщение, появляется еще одна страница подтверждения, которая показана на рис. 16.19 (в настоящее время операция удаления не спрашивает подтверждения, так что будьте внимательны).

Рис. 16.19. *Подтверждение удаления в PyMailCGI*

Одна тонкость: при выполнении операций создания ответа и пересылки сценарий действий с почтой `onViewPageAction` цитирует исходное сообщение, добавляя символ > в начало каждой строки и вставляя исходные строки заголовков «From:», «To:» и «Date:» перед исходным текстом сообщения. Обратите, однако, внимание, что заголовки исходных сооб-

щений берутся из формы ввода, а не за счет выполнения синтаксического анализа исходного сообщения (в этот момент сообщение недоступно непосредственно). Иными словами, сценарий берет значения заголовков из полей ввода формы на странице просмотра. Так как поля «Date:» на странице просмотра нет, дата исходного сообщения также передается сценарию вместе с сообщением в виде скрытого поля ввода, чтобы не загружать сообщение заново. Попробуйте проследить по программному коду в листингах этой главы, как даты перемещаются с одной страницы на другую.

Операция удаления и номера POP-сообщений

Обратите внимание, что после успешного удаления вам действительно *придется* щелкнуть на ссылке Back to root page (Назад на главную страницу) – не пользуйтесь в этот момент кнопкой броузера Назад (Back), чтобы вернуться в список выбора сообщений, потому что в результате удаления относительные номера некоторых сообщений в этом списке изменились. В программе PyMailGUI эта проблема решалась за счет автоматического обновления кэша сообщений в памяти и списка на экране, но в PyMailCGI в настоящее время отсутствует возможность пометить прежние страницы как устаревшие.

Если в результате щелчка на кнопке Назад (Back) броузер повторно запустит серверный сценарий, он обновит страницу и вы получите актуальный список. Однако если броузер отобразит страницу из кэша, вы увидите, что удаленное сообщение по-прежнему присутствует в списке. Хуже того, если щелкнуть на ссылке View (Просмотреть) в старой странице со списком, может быть вызвано не то сообщение, которое вы предполагаете, если оно следует после сообщения, которое удалено.

Это характерная особенность POP в целом, которую мы обсуждали ранее в этой книге: поступающая почта добавляется в список с более высокими номерами сообщений, но при удалении почта изымается из произвольного места в списке, и потому изменяются номера всех сообщений, следующих за удаленными.

Потенциальная ошибка рассинхронизации с почтовым ящиком

Как мы видели в главе 14, даже в PyMailGUI номера некоторых сообщений могут оказаться неверными, если удалить почту другой программой в то время, когда открыт графический интерфейс, – например во втором экземпляре программы PyMailGUI или с помощью PyMailCGI. Такое также возможно, если сервер автоматически удаляет сообщения после загрузки списка, например в случае ошибки перемещает сообщения из папки входящих сообщений в папку недоставленных.

Именно по этой причине в PyMailGUI предусмотрен свой способ определения рассинхронизации с почтовым сервером при выполнении опера-

ций загрузки и удаления сообщений – с помощью пакета mailtools. Например, при выполнении операции удаления для оценки соответствия она сопоставляет хранящиеся в программе заголовки сообщений с заголовками на сервере. Аналогичная проверка выполняется при загрузке сообщений. При несовпадении автоматически выполняется повторная загрузка списка заголовков. К сожалению, без дополнительной информации PyMailCGI не может определять такие ошибки: в ней нет списка сообщений, который можно было бы сравнить со списком на сервере при выполнении операции просмотра или удаления, – только номер сообщения в ссылке или в скрытом поле формы.

В худшем случае PyMailCGI не сможет гарантировать, что операция удаления удалит именно выбранное сообщение, – маловероятно, но вполне возможно, что сообщение, присутствовавшее в списке, будет удалено в промежутке времени между моментом получения номеров сообщений и моментом вызова операции удаления сообщения на сервере. Без дополнительной информации, хранящейся на сервере, программа PyMailCGI не может использовать инструменты безопасного удаления или проверки синхронизации, имеющиеся в пакете mailtools, чтобы убедиться в корректности номеров сообщений.

Чтобы гарантировать невозможность ошибки при удалении, программе PyMailCGI необходим механизм сохранения информации о состоянии, который позволил бы отображать номера сообщений, передаваемые страницам, на хранящиеся заголовки сообщений, полученные, когда номера были определены в последний раз, или какое-то более широкое решение, позволяющее решить эту проблему полностью. В следующих трех разделах кратко описываются возможные улучшения, которые можно рассматривать как упражнения для самостоятельного решения.

Альтернатива: передача текста заголовка в скрытых полях (PyMailCGI_2.1)

Пожалуй, самый простой способ гарантировать точность операции удаления заключается во встраивании полного текста заголовков отображаемого сообщения в страницу просмотра сообщения в виде скрытого поля формы с применением следующей схемы:

onViewListLink.py

Встраивает текст заголовков в скрытое поле формы, экранируя его в соответствии с соглашениями HTML вызовом функции cgi.escape (с аргументом quote, установленным в значение True, чтобы обеспечить экранирование всех вложенных кавычек в тексте заголовка).

onViewPageAction.py

Извлекает встроенный текст заголовков из поля формы и передает его функции безопасного удаления в пакете mailtools для сопоставления с заголовками на сервере.

Для реализации этого решения не придется существенно изменять программный код, но может потребоваться загружать дополнительные заголовки в первом из этих сценариев (в настоящее время он загружает полный текст сообщения) и конструировать список всех заголовков сообщения (здесь удаляется одно сообщение и нужны заголовки только для одного сообщения). Как вариант текст заголовков можно было бы извлекать из полного текста сообщения, разбивая его по пустой строке, отделяющей заголовки от тела сообщения.

Кроме того, это решение может привести к увеличению объема данных, передаваемых между клиентом и сервером, – объем текста заголовков сообщения обычно превышает 1 Кбайт, а может быть еще больше. Это небольшой объем данных по современным меркам, но есть вероятность, что он может превысить ограничения, действующие на некоторых клиентах или серверах.

И действительно, данная схема недостаточно полна. Она решает только проблему ошибочного удаления не того сообщения и никак не затрагивает проблему рассинхронизации в целом. Например, система по-прежнему может пытаться загружать и отображать не те сообщения, что были выбраны в списке, после удаления на сервере более ранних сообщений, выполненного из другого места. Фактически, этот прием лишь гарантирует, что будет удалено именно то сообщение, которое отображалось в окне просмотра в момент выбора операции удаления. Он не гарантирует, что будет удалено или открыто в окне просмотра сообщение, выбранное в списке.

В частности, так как эта схема предусматривает встраивание заголовков в разметку HTML окна просмотра, сопоставление заголовков при удалении будет полезно, только если удаление более ранних входящих сообщений из другого места произойдет лишь после того, как сообщение будет открыто для просмотра. Если содержимое почтового ящика будет изменено из другого места, прежде чем сообщение будет открыто для просмотра, из страницы со списком может быть получен номер не того сообщения. В этом случае данная схема не позволит удалить сообщение, отличное от того, что отображается в окне просмотра, но предполагает, что пользователь сам заметит ошибку и не удалит сообщение, по ошибке загруженное из главной страницы. Хотя эти ситуации крайне редки, тем не менее, такое поведение системы нельзя назвать дружественным по отношению к пользователю.

Несмотря на свою неполноту, это решение все же позволяет хотя бы избежать удаления не тех сообщений, если содержимое почтового ящика на сервере изменится во время просмотра сообщения, – удалено может быть только то сообщение, которое отображается в окне просмотра. Экспериментальная, но работающая реализация этой схемы находится в следующем каталоге в пакете с примерами для книги:

```
PP4E\Internet\Web\dev\PyMailCGI_2.1
```

Во время разработки эта версия опробовалась в веб-броузере Firefox, и для внедрения описанной схемы потребовалось изменить чуть больше 10 строк программного кода в трех файлах, перечисленных ниже (ищите изменения по строкам, помеченным комментарием «#EXPERIMENTAL»):

```
# onViewListLink.py
    ...
    hdrstext = fulltext.split('\n\n')[0]  # использовать пустую строку
    commonhtml.viewpage(                  # шифрует пароль
                msgnum, message, content, form, hdrstext, parts)

# commonhtml.py
    ...
    def viewpage(msgnum, headers, text, form, hdrstext, parts=[]):
        ...
        # операции удаления необходим текст заголовков, чтобы проверить
        # синхронизацию с почтовым ящиком: может иметь объем
        # в несколько килобайтов
        hdrstext = cgi.escape(hdrstext, quote=True) # экранировать '"' тоже
        print('<input type=hidden name=Hdrstext value="%s">' % hdrstext)

# onViewPageAction.py
    ...
     fetcher = mailtools.SilentMailFetcher(site, user, pswd)
     #fetcher.deleteMessages([msgnum])
     hdrstext = getfield(form, 'Hdrstext') + '\n'
     hdrstext = hdrstext.replace('\r\n', '\n')        # получ. \n от top
     dummyhdrslist = [None] * msgnum                  # только 1 заголов.
     dummyhdrslist[msgnum-1] = hdrstext               # в скрытом поле
     fetcher.deleteMessagesSafely([msgnum], dummyhdrslist)  # исключение
     commonhtml.confirmationpage('Delete')            # при рассинхрониз.
```

Чтобы опробовать эту версию локально, запустите сценарий веб-сервера, представленный в примере 15.1 (в главе 15), с именем подкаталога dev и уникальным номером порта, если собираетесь одновременно отработать обе версии, оригинальную и экспериментальную. Например:

```
C:\...\PP4E\Internet\Web> webserver.py dev\PyMailCGI_2.1 9000   команда
http://localhost:9000/pymailcgi.html                            URL для веб-броузера
```

Хотя эта версия работает в проверенных броузерах, тем не менее, она считается экспериментальной (она не использовалась в этой главе и не была перенесена на Python 3.X), потому что реализует неполное решение. В тех редких случаях, когда изменения содержимого почтового ящика на сервере могут сделать номера сообщений недействительными после получения заголовков с сервера, эта версия позволяет избежать удаления не тех сообщений по ошибке, но список почты по-прежнему может оставаться рассинхронизированным с почтовым ящиком. При выполнении операции просмотра все еще могут загружаться не те сооб-

щения, и для решения этой проблемы наверняка потребуются более сложные механизмы сохранения информации о состоянии.

Обратите внимание, что в большинстве случаев заголовка `message-id` вполне достаточно для идентификации сообщения в почтовом ящике при выполнении операции удаления, и между страницами можно было бы передавать только его. Однако поскольку этот заголовок является необязательным и может быть подделан, он не является надежным способом идентификации сообщений – для большей надежности необходимо сопоставление всех имеющихся заголовков. За дополнительной информацией обращайтесь к обсуждению пакета `mailtools` в главе 13.

Альтернатива: сохранение заголовков в файлах на стороне сервера

Основной недостаток приема, описанного в предыдущем разделе, заключается в том, что он решает только проблему удаления уже полученного сообщения. Чтобы решить остальные проблемы, связанные с рассинхронизацией списка и содержимого почтового ящика, необходимо также было бы сохранять заголовки, полученные при конструировании страницы со списком.

Для сохранения информации о состоянии в главной странице используются параметры запроса в адресах URL, однако добавление длинных текстов заголовков в адреса URL в виде дополнительных параметров является практически невозможным. В принципе, тексты заголовков всех сообщений в списке можно было бы встраивать в главную страницу в виде скрытого поля, но это может привести к необходимости передавать чрезмерно большие объемы информации.

В качестве более полного решения каждый раз при создании главной страницы со списком сообщений в сценарии `onViewPswdSubmit.py` полученные заголовки всех сообщений можно было бы сохранять на сервере, в плоском файле с уникальным именем, генерируемым динамически (например, из текущего времени, идентификатора процесса и имени пользователя). Имя этого файла можно было бы передавать вместе с номерами сообщений между страницами в виде дополнительного скрытого поля или параметра запроса.

При выполнении операции удаления сценарий `onViewPageAction.py` мог бы загружать заголовки сообщений из файла и передавать их инструментам безопасного удаления в пакете `mailtools`. При загрузке полного текста сообщения этот файл также можно было бы использовать для проверки синхронизации, чтобы избежать ошибочной загрузки и отображения не тех сообщений. При этом необходимо предусмотреть какое-то решение для удаления файлов с заголовками (удалять старые файлы мог бы сценарий создания главной страницы), а кроме того, возможно, придется рассмотреть проблемы обслуживания нескольких пользователей одновременно.

По сути эта схема использует файлы на стороне сервера для имитации оперативной памяти в программе PyMailGUI, однако ситуация осложняется тем, что пользователи могут использовать возвраты в своих броузерах. Это обстоятельство может приводить к тому, что пользователи будут пытаться удалять сообщения из страниц просмотра, полученных из страниц с устаревшими списками, или повторно загружать сообщения, пользуясь устаревшими страницами со списками, и так далее. В целом может потребоваться проанализировать все возможные способы перемещения между страницами (это, по сути, конечный автомат). Файлы с сохраненными заголовками можно также использовать для проверки синхронизации в операциях загрузки сообщений и удалять в операциях удаления сообщений, чтобы эффективно отключить в предыдущих страницах возможность выполнения операций на основе устаревшей информации, хотя сопоставления заголовков может быть вполне достаточно, чтобы обеспечить точность операции удаления.

Альтернатива: удаление при загрузке

В качестве последнего варианта почтовые клиенты могли бы удалять все сообщения с сервера сразу после их загрузки, чтобы операция удаления не воздействовала на идентификаторы POP (эту схему по умолчанию использует Microsoft Outlook, например). Однако это решение требует реализации дополнительных механизмов сохранения удаляемых сообщений, чтобы обеспечить возможность обратиться к ним позднее, и это также означает, что вы сможете просматривать загруженную почту только на компьютере, куда она была загружена. Поскольку программы PyMailGUI и PyMailCGI предназначены для использования на различных компьютерах, почта по умолчанию хранится на сервере POP.

Из-за отсутствия проверки синхронизации в текущей версии PyMailCGI вы не должны с ее помощью удалять почту из почтовых ящиков, где могут храниться важные сообщения, без применения одной из описанных схем. В этом случае используйте другие инструменты для сохранения сообщений перед удалением. Добавление поддержки сохранения информации о состоянии для обеспечения синхронизации с почтовым ящиком могло бы превратиться в интересное упражнение, но для этого пришлось бы писать дополнительный программный код, для которого у нас здесь недостаточно места, особенно, если поставить себе задачу обобщить его для обслуживания нескольких пользователей.

Вспомогательные модули

В этом разделе представлен исходный программный код вспомогательных модулей, импортируемых и используемых сценариями страниц, рассмотренными выше. В данной установке все эти модули находятся в том же каталоге, что и сценарии CGI, чтобы упростить их импортирование, – интерпретатор обнаруживает их в текущем рабочем каталоге.

Здесь вы не увидите новых снимков экранов, потому что это вспомогательные модули, а не сценарии CGI. Кроме того, изучать эти модули отдельно не слишком полезно, и они помещены здесь, в основном, чтобы обращаться к ним для справки при изучении программного кода сценариев CGI. Дополнительные сведения, которые приводились в данной главе ранее, здесь повторно не приводятся.

Внешние компоненты и настройки

Когда программа PyMailCGI запускается из своего собственного каталога в дереве примеров к книге, она полагается на ряд внешних модулей, которые теоретически могут находиться в любом месте. Поскольку все модули доступны из корневого каталога *PP4E* пакета примеров, они могут импортироваться как обычно, с применением точечной нотации, относительно корневого каталога. На случай, если данная структура каталогов когда-либо изменится, я заключил все внешние зависимости во вспомогательный модуль, представленный в примере 16.10, – при изменении структуры каталогов достаточно будет исправить только этот модуль.

Пример 16.10. PP4E\Internet\Web\PyMailCgi\cgi-bin\externs.py

```
"""
Изолирует операции импортирования модулей, находящихся за пределами
каталога PyMailCgi, благодаря чему при изменении их местоположения
достаточно будет изменить только этот модуль; мы повторно используем
настройки в модуле mailconfig, использовавшиеся в pymailgui2 в главе 13;
вмещающий каталог PP4E/ должен находиться в списке sys.path, чтобы
обеспечить возможность выполнения последней инструкции import здесь;
"""


import sys
#sys.path.insert(0, r'C:\Users\mark\Stuff\Books\4E\PP4E\dev\Examples')
sys.path.insert(0, r'..\..\..\..')         # относительно каталога сценария

import mailconfig                          # локальная версия
from PP4E.Internet.Email import mailtools  # пакет mailtools
```

Этот модуль просто импортирует все внешние имена, необходимые программе PyMailCGI, в свое собственное пространство имен. Дополнительную информацию о модулях из пакета `mailtools`, импортируемого и повторно используемого здесь, смотрите в главе 13 – как и в PyMailGUI, основные операции, которые выполняются за спиной PyMailCGI, реализованы в пакете `mailtools`.

Данная версия PyMailCGI имеет собственную локальную копию модуля `mailconfig`, реализованного нами в главе 13 и расширенного в главе 14, при этом он просто импортирует версию из главы 13, чтобы избежать избыточности, и настраивает необходимые параметры – локальная версия представлена в примере 16.11.

Пример 16.11. PP4E\Internet\Email\PyMailCgi\cgi-bin\mailconfig.py

```
"""
Пользовательские настройки для различных почтовых программ
(версия для PyMailCGI);
Сценарии для работы с электронной почтой получают имена серверов и другие
параметры из этого модуля: измените модуль так, чтобы он отражал имена
ваших серверов, вашу подпись и предпочтения;
"""

from PP4E.Internet.Email.mailconfig import *  # использовать настройки
                                              # из главы 13
fetchlimit = 50  # 4E: максимальное число загружаемых заголовков/сообщений
                 # (по умолчанию = 25)
```

Интерфейс к протоколу POP

Следующий вспомогательный модуль, файл `loadmail`, представленный в примере 16.12, зависит от внешних файлов и инкапсулирует операции доступа к почте на удаленном POP-сервере. В настоящее время он экспортирует одну функцию, `loadmailhdrs`, которая возвращает список заголовков (только) всех почтовых сообщений для указанной учетной записи POP. Вызывающей программе неизвестно, загружается ли эта почта из Сети, находится ли в памяти или загружается из постоянного хранилища на сервере CGI. Так сделано намеренно – изменения в `loadmail` не оказывают влияния на его клиентов, что оставляет возможность для расширения модуля в будущем.

Пример 16.12. PP4E\Internet\Web\PyMailCgi\cgi-bin\loadmail.py

```
"""
загружает список заголовков сообщений; на будущее - добавить сохранение
списка в промежутках между вызовами сценария CGI, чтобы избежать
необходимости всякий раз повторно загружать весь список; если все сделать
правильно, это никак не отразится на клиентах; пока, для простоты,
при каждом выводе страницы перезагружается весь список;
2.0+: теперь загружаются только заголовки сообщений (с использованием
команды TOP), а не все сообщения целиком, но по-прежнему при каждом
обращении к странице со списком загружаются все заголовки - кэширование
списка в памяти сценариев CGI невозможно, т.к. они не сохраняют информацию
о состоянии, и для этого необходимо использовать настоящую базу данных
(скорее всего, на стороне сервера);
"""

from commonhtml import runsilent  # подавляет вывод (без флага verbose)
from externs    import mailtools  # используется совместно с PyMailGUI

# загрузить все письма начиная с номера 1
# может возбудить исключение
```

```
import sys
def progress(*args): # не используется
    sys.stderr.write(str(args) + '\n')

def loadmailhdrs(mailserver, mailuser, mailpswd):
    fetcher = mailtools.SilentMailFetcher(mailserver, mailuser, mailpswd)
    hdrs, sizes, full = fetcher.downloadAllHeaders()  # получить список
                                                      # заголовков
    return hdrs
```

Особо интересного здесь ничего нет – просто интерфейс и обращения к другим модулям. Класс `mailtools.SilentMailFetcher` (повторно используется реализация из главы 13) с помощью модуля Python `poplib` загружает почту через сокеты. Класс `SilentMailFetcher` подавляет вывод функций `print` в пакете `mailtools`, чтобы он не попадал в ответный поток HTML (хотя все исключения распространяются дальше).

Данная версия `loadmail` загружает только заголовки всех входящих сообщений, чтобы создать страницу со списком выбора. Однако при каждом обращении к странице со списком приходится заново загружать все заголовки. Как говорилось выше, такая реализация более эффективна, чем предыдущая версия, но она все равно оказывается слишком медленной, когда на сервере хранится достаточно много сообщений. Применение базы данных в комбинации с алгоритмом проверки списка при удалении и получении новых сообщений могло бы помочь ликвидировать это узкое место. Поскольку интерфейс, экспортируемый модулем `loadmail`, не должен измениться при реализации механизма кэширования, клиенты этого модуля смогут использовать его, как и прежде, без необходимости вносить в них какие-либо изменения.

Шифрование паролей

Выше мы вкратце обсуждали подход к защите паролей, принятый в PyMailCGI. Здесь мы рассмотрим его конкретную реализацию. Программа PyMailCGI передает имя пользователя и пароль между страницами с помощью скрытых полей форм и параметров запроса в адресах URL, встраиваемых в разметку HTML страниц. Мы рассмотрели эти приемы в предыдущей главе. Такие данные передаются через сетевые сокеты в виде простого текста – внутри разметки HTML ответа сервера – и в виде параметров в запросах, посылаемых клиентом. При таком подходе возникает проблема защиты секретных данных.

Эта особенность не является проблемой при использовании локального веб-сервера, как и во всех примерах до сих пор. Данные в этом случае передаются между двумя программами, выполняющимися на одном и том же компьютере, и недоступны внешнему миру. Однако если вам потребуется установить PyMailCGI на удаленный веб-сервер, это может стать проблемой. Поскольку эти данные желательно хранить втайне от

посторонних, в идеале хотелось бы иметь способ скрывать их при передаче и предотвращать возможность подсмотреть их в файлах журналов на сервере. С появлением новых и уходом старых возможностей приемы решения этой проблемы неоднократно изменялись на протяжении жизни этой книги:

- Во втором издании этой книги был разработан собственный модуль шифрования, использующий модуль шифрования `rotor` из стандартной библиотеки. Этот модуль использовался для шифрования данных, вставляемых в поток ответа сервера, и затем для расшифровывания данных, возвращаемых клиентом в параметрах. К сожалению, модуль `rotor` был исключен из стандартной библиотеки в версии Python 2.4 из-за проблем, связанных с безопасностью. Возможно, это было слишком радикальное решение (модуль `rotor` вполне пригоден для использования в простых приложениях), тем не менее, в последних версиях Python модуль `rotor` более недоступен.
- В третьем издании модель второго издания была расширена за счет добавления поддержки шифрования паролей с помощью сторонних модулей и открытой системы PyCrypto. К сожалению, эта система доступна только для Python 2.X и к моменту написания этих строк для четвертого издания в середине 2010 года версия для 3.X еще не вышла (хотя некоторый прогресс в этом направлении имеется). Кроме того, классы Python с реализацией веб-сервера, выполняемого локально и используемого в этом издании для опробования примеров, в Python 3.1 все еще не поддерживают защищенный протокол HTTPS – законченное решение, обеспечивающее безопасность в Веб, о котором я расскажу чуть ниже.
- Вследствие всего вышеперечисленного в этом четвертом издании сохранена унаследованная поддержка модуля `rotor` и системы PyCrypto, если они будут установлены, а на крайний случай реализовано упрощенное шифрование пароля, алгоритм которого можно изменять для каждой установки PyMailCGI. Поскольку эту версию в целом можно считать лишь прототипом, дальнейшее улучшение этой модели, включая поддержку HTTPS при выполнении под управлением более надежных веб-серверов, я оставляю в качестве самостоятельного упражнения.

В целом существуют различные подходы к шифрованию информации, передаваемой между клиентом и сервером. Но, к сожалению, ни один из них нельзя реализовать настолько просто, чтобы использовать в качестве примера в этой книге, ни один из них не является универсальным и большинство из них требуют использования инструментов и приемов, знакомство с которыми выходит далеко за рамки этой книги. Тем не менее, для общего знакомства в следующем разделе приводится краткое описание некоторых из наиболее распространенных методик.

Шифрование данных вручную: rotor (более не существующий)

В принципе, сценарии CGI могут вручную шифровать любые данные, добавляемые в поток ответа, как это было реализовано в версии PyMailCGI для второго издания этой книги. Однако с исключением модуля rotor из версии Python 2.4 в стандартной библиотеке не осталось инструментов шифрования для решения этой задачи. Кроме того, использование программного кода из оригинального модуля rotor нежелательно, с точки зрения сопровождения; к тому же задействовать его не так просто, потому что он был написан на языке C (недостаточно будет просто скопировать файл *.py* из более ранней версии Python). Если только вы не используете старую версию Python, модуль rotor практически недоступен.

Главным образом из исторического интереса и для сравнения с современными приемами ниже демонстрируется, как использовался этот модуль. Он был основан на алгоритме шифрования Enigma: создавался новый объект rotor с ключом (и, при необходимости, со счетчиком циклов) и вызывались его методы encrypt и decrypt:

```
>>> import rotor
>>> r = rotor.newrotor('pymailcgi') # (ключ, [,счетчик])
>>> r.encrypt('abc123')             # может возвращать непечатаемые символы
' \323an\021\224'

>>> x = r.encrypt('spam123')        # результат имеет ту же длину, что
>>> x                               # и исходная строка
'* _\344\011pY'
>>> len(x)
7
>>> r.decrypt(x)
'spam123'
```

Обратите внимание, что один и тот же объект rotor может зашифровывать несколько строк, результат может содержать непечатаемые символы (выводимые как экранированные последовательности \ascii) и длина результата всегда совпадает с длиной исходной строки. Самое главное, что строка, зашифрованная с помощью объекта rotor, может быть расшифрована в другом процессе (например, позднее другим сценарием CGI), если объект rotor создать заново:

```
>>> import rotor
>>> r = rotor.newrotor('pymailcgi') # может быть расшифрована в др. процессе
>>> r.decrypt('* _\344\011pY')      # 2 символа представлены экранированными
'spam123'                           # последовательностями "\ascii"
```

Наш модуль secret по умолчанию использует для шифрования только объект rotor и не реализует никаких собственных алгоритмов шифрования. Он полагается на экранирование адресов URL при встраивании пароля в параметр URL и экранирование HTML при встраивании пароля в скрытые поля форм. Для URL производятся следующие вызовы:

```
>>> from secret import encode, decode
>>> x = encode('abc$#<>&+')            # это делают сценарии CGI
>>> x
' \323a\016\317\326\023\0163'

>>> import urllib.parse                # это делает urlencode
>>> y = urllib.parse.quote_plus(x)
>>> y
'+%d3a%0e%cf%d6%13%0e3'

>>> a = urllib.parse.unquote_plus(y)   # это делает cgi.FieldStorage
>>> a
' \323a\016\317\326\023\0163'

>>> decode(a)                          # это делают сценарии CGI
'abc$#<>&+'
```

Несмотря на то, что в настоящее время модуль `rotor` используется уже не так широко, те же приемы можно использовать для реализации других алгоритмов шифрования.

Шифрование данных вручную: PyCrypto

Среди сторонних разработок можно найти множество свободно доступных инструментов шифрования, включая популярный набор инструментов Python Cryptography Toolkit, известный также как PyCrypto. Этот пакет содержит модули, реализующие алгоритмы шифрования с открытым и закрытым ключом, такие как AES, DES, IDEA и RSA, предоставляет модуль Python для чтения и расшифровывания файлов PGP и многие другие. Ниже приводится пример использования шифрования по алгоритму AES после установки PyCrypto в Windows на моем компьютере:

```
>>> from Crypto.Cipher import AES
>>> AES.block_size = 16
>>> mykey = 'pymailcgi'.ljust(16, '-') # ключ должен быть 16, 24 или 32 байта
>>> mykey
'pymailcgi-------'
>>>
>>> password = 'Already got one.'       # длина должна быть кратна 16
>>> aesobj1 = AES.new(mykey, AES.MODE_ECB)
>>> cyphertext = aesobj1.encrypt(password)
>>> cyphertext
'\xfez\x95\xb7\x07_¨\xd4\xb6\xe3r\x07g~X]'
>>>
>>> aesobj2 = AES.new(mykey, AES.MODE_ECB)
>>> aesobj2.decrypt(cyphertext)
'Already got one.'
```

Этот интерфейс напоминает интерфейс модуля `rotor`, но в нем используются более совершенные алгоритмы шифрования. Алгоритм AES – это

популярный алгоритм шифрования с закрытым ключом. Он требует наличия ключа фиксированной длины и строку данных с длиной кратной 16 байтам.

К сожалению, этот пакет не является частью стандартной библиотеки Python, его распространение в двоичной форме может регулироваться законом США (и других стран) о контроле за экспортом, а описание его – это слишком большая и сложная тема для этой книги. Все это делает его не таким универсальным, как хотелось бы. По крайней мере, распространение двоичной программы установки вместе с примерами этой книги может оказаться не совсем законным. А поскольку шифрование данных является одной из базовых потребностей PyMailCGI, зависимость от внешнего инструмента может оказаться слишком сильной.

Однако самый главный фактор, препятствующий использованию PyCrypto в этой книге, заключается в том, что этот пакет поддерживает только Python 2.X и пока еще не вышла версия для Python 3.X. Это обстоятельство делает невозможным использование пакета в примерах для данной книги. Однако, если у вас появится возможность установить пакет PyCrypto и разобраться в нем, это может стать мощным решением проблемы. Дополнительную информацию о PyCrypto ищите в Интернете.

HTTPS: защищенная передача данных по протоколу HTTP

Если вы пользуетесь веб-сервером, поддерживающим защищенный протокол HTTP, вы можете просто создавать страницы HTML и переложить шифрование данных на веб-сервер и броузер. Если этот протокол поддерживается с обоих концов соединения, он обеспечит законченное решение и обеспечит безопасность в Сети. В настоящее время он используется большинством сайтов электронной коммерции.

Защищенный протокол HTTP (Secure HTTP, HTTPS) определяется в адресах URL за счет указания имени протокола `https://` вместо `http://`. При использовании протокола HTTPS данные по-прежнему передаются с использованием обычного протокола HTTP, но весь трафик при этом шифруется уровнем защищенных сокетов (Secure Sockets Layer, SSL). Протокол HTTPS поддерживается большинством веб-броузеров и может быть настроен на большинстве веб-серверов, включая Apache и сценарий *webserver.py*, который мы запускали локально в этой главе. Если интерпретатор Python скомпилирован с поддержкой SSL, сокеты Python будут поддерживать этот механизм посредством модуля `ssl`, реализующего обертки вокруг сокетов, и клиентский модуль `urllib.request`, с которым мы познакомились в главе 13, также будет поддерживать протокол HTTPS.

К сожалению, включение поддержки защищенного протокола HTTP на веб-сервере требует дополнительных знаний и настройки, для описания которых у нас здесь нет места, и может потребовать установки инструментов, не входящих в стандартную версию Python. Если в будущем

у вас появится желание глубже исследовать эту проблему, поищите в Интернете ресурсы, описывающие настройку HTTPS-сервера, написанного на языке Python, поддерживающего защищенные взаимодействия с применением SSL. Например, обратите внимание на сторонний пакет M2Crypto с поддержкой OpenSSL, который можно использовать для шифрования паролей, поддержку HTTPS в пакете `urllib` и другие инструменты – это достаточно привлекательные альтернативы для реализации шифрования вручную, хотя к моменту написания этих строк они все еще не были доступны в Python 3.X.

В целом дополнительные подробности о протоколе HTTPS ищите в Интернете. Вполне возможно, что некоторая поддержка HTTPS появится со временем в классах реализации веб-серверов, присутствующих в стандартной библиотеке Python. В последние годы такая поддержка отсутствовала, что, возможно, обусловлено предназначением этих классов – они обеспечивают ограниченную функциональность для использования в локальных серверах, в первую очередь ориентированных на тестирование, а не на развертывание действующих серверов.

Защищенные cookies

В настоящее время скрытые поля форм и параметры запросов в PyMailCGI вполне можно заменить клиентскими cookies, помеченными как защищенные. Такие cookies автоматически шифруются при передаче. К сожалению, cookies, помеченные как защищенные, могут передаваться только через защищенное соединение с удаленным хостом. Для них не предусматривается дополнительное шифрование. Вследствие этого данная возможность сама по себе не является сколько-нибудь ценной – для ее поддержки все еще необходим сервер с поддержкой HTTPS.

Модуль secret.py

Как вы уже могли понять, безопасность в Веб является слишком объемной темой, чтобы ее можно было рассмотреть здесь. Вследствие этого модуль `secret.py`, представленный в примере 16.13, старается обойти эту проблему, обеспечивая различные варианты ее решения:

- Если у вас есть возможность получить и установить стороннюю систему PyCrypto, описанную выше, модуль будет использовать инструменты шифрования по алгоритму AES из этого пакета для шифрования пароля при передаче вместе с именем пользователя.
- В противном случае следующей будет попытка использовать пакет `rotor`, если вам удастся отыскать и установить оригинальный модуль `rotor` для используемой вами версии Python.
- И, наконец, в крайнем случае он будет использовать реализацию простейшего алгоритма шифрования, основанного на искажении символов, которую вы сможете заменить своей реализацией при установке программы в Интернете.

Дополнительные детали смотрите в примере 16.13 – там используются определения функций, вложенные в условные инструкции `if`, которые позволяют сгенерировать функции для выбранной схемы шифрования во время выполнения.

Пример 16.13. PP4E\Internet\Web\PyMailCgi\cgi-bin\secret.py

```
"""
###############################################################################
PyMailCGI шифрует пароль, когда он пересылается клиенту или от него через
сеть вместе с именем пользователя в скрытых полях форм или в параметрах
запроса URL; использует функции encode/decode в этом модуле для шифрования
пароля - выгрузите на сервер собственную версию этого модуля, чтобы
использовать другой механизм шифрования; PyMailCGI не сохраняет пароли
на сервере и не отображает его при вводе пользователем в форме ввода,
но это не дает 100% защиты - файл этого модуля сам может оказаться уязвим;
использование протокола HTTPS может оказаться более удачным и более простым
решением, но классы реализации веб-сервера в стандартной библиотеке Python
не поддерживают его;
###############################################################################
"""

import sys, time
dayofweek = time.localtime(time.time())[6] # для реализации собственных схем
forceReadablePassword = False

###############################################################################
# схемы преобразования строк
###############################################################################

if not forceReadablePassword:
    ###########################################################################
    # по умолчанию не делать ничего: вызовы urllib.parse.quote
    # или cgi.escape в commonhtml.py выполнят необходимое экранирование
    # пароля для встраивания его в URL или HTML; модуль cgi
    # автоматически выполнит обратное преобразование;
    ###########################################################################
    def stringify(old): return old
    def unstringify(old): return old

else:
    ###########################################################################
    # преобразование кодированной строки в/из строки цифр,
    # чтобы избежать проблем с некоторыми специальными/непечатаемыми
    # символами, но сохранить возможность чтения результата
    # (хотя и зашифрованного); в некоторых броузерах есть проблемы
    # с преобразованными амперсандами и т.д.;
    ###########################################################################

    separator = '-'
```

```
    def stringify(old):
        new = ''
        for char in old:
            ascii = str(ord(char))
            new = new + separator + ascii # '-ascii-ascii-ascii'
        return new

    def unstringify(old):
        new = ''
        for ascii in old.split(separator)[1:]:
            new = new + chr(int(ascii))
        return new

##################################################
# схемы шифрования: пробует PyCrypto, затем rotor,
# затем простейший/нестандартный алгоритм
##################################################

useCrypto = useRotor = True
try:
    import Crypto
except:
    useCrypto = False
    try:
        import rotor
    except:
        useRotor = False

if useCrypto:
    ############################################################
    # использовать алгоритм AES из стороннего пакета pycrypto
    # предполагается, что в конце строки пароля отсутствует
    # символ '\0': используется для дополнения справа
    # измените закрытый ключ здесь, если используете этот метод
    ############################################################

    sys.stderr.write('using PyCrypto\n')
    from Crypto.Cipher import AES
    mykey = 'pymailcgi3'.ljust(16, '-')   # ключ должен иметь длину 16, 24
                                          # или 32 байта

    def do_encode(pswd):
        over = len(pswd) % 16
        if over: pswd += '\0' * (16-over)    # дополнение: длина должна быть
        aesobj = AES.new(mykey, AES.MODE_ECB)  # кратна 16
        return aesobj.encrypt(pswd)

    def do_decode(pswd):
        aesobj = AES.new(mykey, AES.MODE_ECB)
        pswd = aesobj.decrypt(pswd)
        return pswd.rstrip('\0')
```

```
elif useRotor:
    ##############################################################
    # использовать для шифрования стандартный модуль rotor
    # он лучше подходит для шифрования, чем программный код выше
    # к сожалению, он больше недоступен в Py 2.4+
    ##############################################################

    sys.stderr.write('using rotor\n')
    import rotor
    mykey = 'pymailcgi3'

    def do_encode(pswd):
        robj = rotor.newrotor(mykey) # использовать алгоритм enigma
        return robj.encrypt(pswd)

    def do_decode(pswd):
        robj = rotor.newrotor(mykey)
        return robj.decrypt(pswd)

else:
    ##############################################################
    # в крайнем случае использовать собственную схему, основанную
    # на искажении символов некоторым обратимым способом
    # предупреждение: слишком простой алгоритм - замените своим
    ##############################################################

    sys.stderr.write('using simple\n')
    adder = 1

    def do_encode(pswd):
        pswd = 'vs' + pswd + '48'
        res = ''
        for char in pswd:
            res += chr(ord(char) + adder) # увеличить каждый код ASCII
        return str(res)

    def do_decode(pswd):
        pswd = pswd[2:-2]
        res = ''
        for char in pswd:
            res += chr(ord(char) - adder)
        return res

##############################################################################
# функции верхнего уровня
##############################################################################

def encode(pswd):
    return stringify(do_encode(pswd))  # шифрование плюс
                                       # преобразование строки
```

```
def decode(pswd):
    return do_decode(unstringify(pswd))
```

В дополнение к шифрованию в этом модуле реализованы также методы преобразования уже зашифрованных строк, которые преобразуют их содержимое в печатаемые символы и обратно. По умолчанию функции преобразования ничего не делают, и система полностью полагается на то, что зашифрованные пароли будут корректно обработаны функциями экранирования URL и HTML. Дополнительная схема преобразования транслирует зашифрованные строки в строки, содержащие цифровые коды ASCII символов, разделенные дефисами. Такой метод позволяет преобразовать непечатаемые символы в зашифрованной строке в печатаемые.

Для иллюстрации проверим инструменты из этого модуля в интерактивном режиме. В этих тестах переменной `forceReadablePassword` было присвоено значение `True`. Функции верхнего уровня `encode` и `decode` будут воспроизводить печатаемые символы (для иллюстрации этот тест проводился в Python 2.X с установленным пакетом PyCrypto):

```
>>> from secret import *
using PyCrypto
>>> data = encode('spam@123+')
>>> data
'-47-248-2-170-107-242-175-18-227-249-53-130-14-140-163-107'
>>> decode(data)
'spam@123+'
```

А ниже шифрование выполняется в два этапа – собственно шифрование и преобразование в печатаемые символы:

```
>>> raw = do_encode('spam@123+')
>>> raw
'/\xf8\x02\xaak\xf2\xaf\x12\xe3\xf95\x82\x0e\x8c\xa3k'
>>> text = stringify(raw)
>>> text
'-47-248-2-170-107-242-175-18-227-249-53-130-14-140-163-107'
>>> len(raw), len(text)
(16, 58)
```

Ниже показано, как выглядит шифрование без дополнительного преобразования в печатаемые символы:

```
>>> raw = do_encode('spam@123+')
>>> raw
'/\xf8\x02\xaak\xf2\xaf\x12\xe3\xf95\x82\x0e\x8c\xa3k'
>>> do_decode(raw)
'spam@123+'
```

Реализация собственного алгоритма шифрования

В текущей реализации PyMailCGI никогда не передает имя пользователя и пароль учетной записи POP через Сеть в одной операции, если па-

роль не был зашифрован с помощью модуля secret.py на сервере. Этот модуль может быть разным в разных установках PyMailCGI, и в будущем на сервер может быть выгружен новый модуль – зашифрованные пароли нигде не сохраняются и существуют только на протяжении одного сеанса обработки почты. Если вы не опубликуете реализацию своего алгоритма шифрования или свои закрытые ключи, ваши данные будут защищены в той же степени, в какой будет защищен ваш собственный модуль шифрования на сервере.

Если вы соберетесь использовать эту систему в Интернете, вам необходимо будет модифицировать этот программный код. В идеале было бы желательно установить PyCrypto и изменить строку закрытого ключа. Если это невозможно, замените сценарий в примере 16.13 собственной реализацией шифрования или используйте один из подходов, описанных выше, такой как поддержка HTTPS на веб-сервере. В любом случае, это программное обеспечение не дает надежных гарантий – защита вашего пароля целиком лежит на ваших плечах.

Дополнительную информацию об инструментах и приемах защиты ищите в Интернете и в книгах, посвященных исключительно приемам веб-программирования. Так как в целом эта система не более, чем прототип, обеспечение безопасности – это лишь одна из множества проблем, которые необходимо решить в более устойчивой, готовой для эксплуатации версии.

Поскольку в PyMailCGI используются симметричные алгоритмы шифрования, вы смогли бы реконструировать мой пароль, подсмотрев его зашифрованную форму на снимках экрана, – при условии, что закрытый ключ, приведенный в secret.py, не отличается от того, что использовался в показанных тестах. Чтобы обойти эту проблему, в этой книге была использована временная учетная запись, которая будет удалена к тому моменту, когда вы будете читать эти строки. Пожалуйста, используйте свою учетную запись электронной почты для тестирования системы.

Общий вспомогательный модуль

Наконец, файл *commonhtml.py*, представленный в примере 16.14, служит «центральным вокзалом» этого приложения – его программный код используется почти всеми остальными файлами системы. По большей части он самодокументирован, а все основные идеи, заложенные в нем, мы уже изучали ранее, когда знакомились со сценариями CGI, использующими его.

Однако я ничего не сказал о поддержке *отладки* в нем. Обратите внимание, что этот модуль связывает поток вывода sys.stderr с потоком sys.stdout, чтобы заставить текст сообщений Python об ошибках выводиться в броузере клиента (напомню, что информация о необработанных исключениях выводится в sys.stderr). Иногда это действует в PyMailCGI,

но не всегда – текст ошибки появляется на веб-странице, только если функция page_header уже вывела преамбулу ответа. Если нужно увидеть все сообщения об ошибках, вызывайте page_header (или выводите строки Content-type: вручную) перед тем, как выполнять какую-либо обработку.

В этом модуле также определены функции, сбрасывающие в броузер массу информации окружения CGI (dumpstatepage) и служащие обертками для функций, выводящих сообщения о состоянии, чтобы их вывод не попадал в поток HTML (runsilent). Дополнительно версия 3.0 пытается также учитывать тот факт, что вызов встроенной функции print может терпеть неудачу в Python 3.1 при выводе некоторых типов текста Юникода (например, при выводе заголовков, содержащих национальные символы, не входящие в диапазон набора ASCII), принудительно устанавливая двоичный режим и отправляя байты в поток вывода (print).

Я предоставляю читателю самому исследовать магию, оставшуюся в программном коде, представленном в примере 16.14. А мы пойдем дальше; нам предстоит читать, обращаться за справками и повторно использовать.

Пример 16.14. PP4E\Internet\Web\PyMailCgi\cgi-bin\commonhtml.py

```
#!/usr/bin/python
"""
###############################################################################
Генерирует стандартный заголовок страницы, список и нижний колонтитул;
в этом файле скрыты детали, относящиеся к созданию разметки HTML;
выводимый здесь текст поступает клиенту через сокеты, создавая части
новой веб-страницы в веб-броузере; каждая строка выводится отдельным
вызовом print, использует urllib для экранирования параметров
в ссылках URL, создаваемых из словаря, а для вставки их в скрытые
поля форм HTML - cgi.escape; некоторые инструменты отсюда могут
использоваться вне pymailcgi; можно было бы возвращать генерируемую
разметку HTML вместо вывода в поток, для включения в другие страницы;
можно было бы выстроить структуру в виде единого сценария CGI, который
получает и проверяет имя следующей операции в скрытом поле формы;
предупреждение: система действует, но была написана в основном во время
2-часовой задержки в чикагском аэропорту O'Hare: некоторые компоненты
стоило бы улучшить и оптимизировать;
###############################################################################
"""

import cgi, urllib.parse, sys, os

# 3.0: в Python 3.1 наблюдаются проблемы при выводе некоторых
# декодированных строк str в поток вывода stdout
import builtins
bstdout = open(sys.stdout.fileno(), 'wb')
def print(*args, end='\n'):
    try:
        builtins.print(*args, end=end)
```

```
        sys.stdout.flush()
    except:
        for arg in args:
            bstdout.write(str(arg).encode('utf-8'))
        if end: bstdout.write(end.encode('utf-8'))
        bstdout.flush()

sys.stderr = sys.stdout         # выводить сообщения об ошибках в броузер
from externs import mailconfig  # из пакета, находящегося на сервере
from externs import mailtools   # для анализа и декодирования заголовков
parser = mailtools.MailParser() # один парсер на процесс в этом модуле

# корневой каталог cgi-сценариев
#urlroot = 'http://starship.python.net/~lutz/PyMailCgi/'
#urlroot = 'http://localhost:8000/cgi-bin/'

urlroot = ''         # использовать минимальные, относительные пути

def pageheader(app='PyMailCGI', color='#FFFFFF', kind='main', info=''):
    print('Content-type: text/html\n')
    print('<html><head><title>%s: %s page (PP4E)</title></head>'
          % (app, kind))
    print('<body bgcolor="%s"><h1>%s %s</h1><hr>'
          % (color, app, (info or kind)))

def pagefooter(root='pymailcgi.html'):
    print('</p><hr><a href="http://www.python.org">')
    print('<img src="../PythonPoweredSmall.gif" ')
    print('align=left alt="[Python Logo]" border=0 hspace=15></a>')
    print('<a href="../%s">Back to root page</a>' % root)
    print('</body></html>')

def formatlink(cgiurl, parmdict):
    """
    создает ссылку запроса "%url?key=val&key=val" из словаря;
    экранирует str() всех ключей и значений,
    подставляя %xx, замещает ' ' на +
    обратите внимание, что адреса URL экранируются иначе, чем HTML
    (cgi.escape)
    """
    parmtext = urllib.parse.urlencode(parmdict) # вызовет parse.quote_plus
    return '%s?%s' % (cgiurl, parmtext)         # всю работу делает urllib

def pagelistsimple(linklist):   # выводит простой нумерованный список
    print('<ol>')
    for (text, cgiurl, parmdict) in linklist:
        link = formatlink(cgiurl, parmdict)
        text = cgi.escape(text)
        print('<li><a href="%s">\n %s</a>' % (link, text))
    print('</ol>')
```

```
def pagelisttable(linklist):    # выводит список в виде таблицы
    print('<p><table border>')  # для верности выполняет экранирование
    for (text, cgiurl, parmdict) in linklist:
        link = formatlink(cgiurl, parmdict)
        text = cgi.escape(text)
        print('<tr><th><a href="%s">View</a><td>\n %s' % (link, text))
    print('</table>')

def listpage(linkslist, kind='selection list'):
    pageheader(kind=kind)
    pagelisttable(linkslist)   # [('text', 'cgiurl', {'parm':'value'})]
    pagefooter()

def messagearea(headers, text, extra=''):  # extra - для readonly
    addrhdrs = ('From', 'To', 'Cc', 'Bcc') # декодировать только имена
    print('<table border cellpadding=3>')
    for hdr in ('From', 'To', 'Cc', 'Subject'):
        rawhdr = headers.get(hdr, '?')
        if hdr not in addrhdrs:
            dechdr = parser.decodeHeader(rawhdr) # 3.0: декодировать
                                                 # для отобр.
        else:                                    # закодированы при отправке
            dechdr = parser.decodeAddrHeader(rawhdr)  # только имена
                                                      # в адресах
        val = cgi.escape(dechdr, quote=1)
        print('<tr><th align=right>%s:' % hdr)
        print('    <td><input type=text ')
        print('    name=%s value="%s" %s size=60>' % (hdr, val, extra))
        print('<tr><th align=right>Text:')
        print('<td><textarea name=text cols=80 rows=10 %s>' % extra)
        print('%s\n</textarea></table>' % (cgi.escape(text) or '?')) # если
                                                                  # имеются </>s

def viewattachmentlinks(partnames):
    """
    создает гиперссылки для сохраняемых локально файлов частей/вложений,
    открытие файлов будет выполнять веб-броузер
    предполагается наличие единственного пользователя, файлы сохраняются
    только для одного сообщения
    """
    print('<hr><table border cellpadding=3><tr><th>Parts:')
    for filename in partnames:
        basename = os.path.basename(filename)
        filename = filename.replace('\\', '/') # грубый прием для Windows
        print('<td><a href=../%s>%s</a>' % (filename, basename))
    print('</table><hr>')

def viewpage(msgnum, headers, text, form, parts=[]):
    """
    при выборе сообщения в списке и выполнении операции просмотра
    (вызывается щелчком на созданной ссылке)
```

```
    очень тонкое место: к этому моменту пароль был закодирован в ссылке,
    в формате URL, и затем декодировался при анализе входных данных;
    здесь он встроен в разметку HTML, поэтому мы применяем cgi.escape;
    обычно в скрытых полях появляются непечатные символы,
    но в IE и NS как-то работает:
    в url:  ?user=lutz&mnum=3&pswd=%8cg%c2P%1e%f0%5b%c5J%1c%f3&...
    в html: <input type=hidden name=pswd value="...непечатные..">
    можно бы пропустить поле HTML через urllib.parse.quote,
    но это потребовало бы вызывать urllib.parse.unquote в следующем
    сценарии (что не мешает передавать данные в URL, а не форме);
    можно вернуться к цифровому формату строк из secret.py
    """
    pageheader(kind='View')
    user, pswd, site = list(map(cgi.escape, getstandardpopfields(form)))
    print('<form method=post action="%sonViewPageAction.py">' % urlroot)
    print('<input type=hidden name=mnum value="%s">' % msgnum)
    print('<input type=hidden name=user value="%s">' % user) # из стран.|url
    print('<input type=hidden name=site value="%s">' % site) # для удаления
    print('<input type=hidden name=pswd value="%s">' % pswd) # кодиров. пароль
    messagearea(headers, text, 'readonly')
    if parts: viewattachmentlinks(parts)

    # onViewPageAction.quotetext требует передачи даты в странице
    print('<input type=hidden name=Date value="%s">'
          % headers.get('Date','?'))
    print('<table><tr><th align=right>Action:')
    print('<td><select name=action>')
    print('    <option>Reply<option>Forward<option>Delete</select>')
    print('<input type=submit value="Next">')
    print('</table></form>')   # 'сброс' здесь не требуется
    pagefooter()

def sendattachmentwidgets(maxattach=3):
    print('<p><b>Attach:</b><br>')
    for i in range(1, maxattach+1):
        print('<input size=80 type=file name=attach%d><br>' % i)
    print('</p>')

def editpage(kind, headers={}, text=''):
    # вызывается при отправке, View+выбор+Reply, View+выбор+Fwd
    pageheader(kind=kind)
    print('<p><form enctype="multipart/form-data" method=post', end=' ')
    print('action="%sonEditPageSend.py">' % urlroot)
    if mailconfig.mysignature:
        text = '\n%s\n%s' % (mailconfig.mysignature, text)
    messagearea(headers, text)
    sendattachmentwidgets()
    print('<input type=submit value="Send">')
    print('<input type=reset value="Reset">')
    print('</form>')
    pagefooter()
```

```
def errorpage(message, stacktrace=True):
    pageheader(kind='Error')                    # было sys.exc_type/exc_value
    exc_type, exc_value, exc_tb = sys.exc_info()
    print('<h2>Error Description</h2><p>', message)
    print('<h2>Python Exception</h2><p>',  cgi.escape(str(exc_type)))
    print('<h2>Exception details</h2><p>', cgi.escape(str(exc_value)))
    if stacktrace:
        print('<h2>Exception traceback</h2><p><pre>')
        import traceback
        traceback.print_tb(exc_tb, None, sys.stdout)
        print('</pre>')
    pagefooter()

def confirmationpage(kind):
    pageheader(kind='Confirmation')
    print('<h2>%s operation was successful</h2>' % kind)
    print('<p>Press the link below to return to the main page.</p>')
    pagefooter()

def getfield(form, field, default=''):
    # имитация метода get словаря
    return (field in form and form[field].value) or default

def getstandardpopfields(form):
    """
    поля могут отсутствовать, быть пустыми или содержать значение, жестко
    определены в URL; по умолчанию используются настройки из mailconfig
    """
    return (getfield(form, 'user', mailconfig.popusername),
            getfield(form, 'pswd', '?'),
            getfield(form, 'site', mailconfig.popservername))

def getstandardsmtpfields(form):
    return getfield(form, 'site', mailconfig.smtpservername)

def runsilent(func, args):
    """
    выполняет функцию, подавляя вывод в stdout
    например: подавляет вывод из импортируемых инструментов, чтобы он
    не попал клиенту/броузеру
    """
    class Silent:
        def write(self, line): pass
    save_stdout = sys.stdout
    sys.stdout = Silent()     # отправлять вывод в фиктивный объект,
    try:                      # который имеет метод write
        result = func(*args)  # попытаться вернуть результат функции
    finally:                  # и всегда восстанавливать stdout
        sys.stdout = save_stdout
    return result
```

```
def dumpstatepage(exhaustive=0):
    """
    для отладки: вызывается в начале сценария CGI
    для создания новой страницы с информацией о состоянии CGI
    """
    if exhaustive:
        cgi.test()            # вывести страницу с формой, окруж. и пр.
    else:
        pageheader(kind='state dump')
        form = cgi.FieldStorage()   # вывести только имена/значения
                                    # полей формы
        cgi.print_form(form)
        pagefooter()
    sys.exit()

def selftest(showastable=False):    # создает фиктивную веб-страницу
    links = [                       # [(text, url, {parms})]
         ('text1', urlroot + 'page1.cgi', {'a':1}),
         ('text2', urlroot + 'page1.cgi', {'a':2, 'b':'3'}),
         ('text3', urlroot + 'page2.cgi', {'x':'a b', 'y':'a<b&c', 'z':'?'}),
         ('te<>4', urlroot + 'page2.cgi', {'<x>':'', 'y':'<a>', 'z':None})]
    pageheader(kind='View')
    if showastable:
        pagelisttable(links)
    else:
        pagelistsimple(links)
    pagefooter()

if __name__ == '__main__':      # когда запускается, а не импортируется
    selftest(len(sys.argv) > 1) # разметка HTML выводится в stdout
```

Преимущества и недостатки сценариев CGI

Как было показано в этой главе, PyMailCGI все еще находится в стадии развития, но действует, как обещано: при установке на удаленном сервере, указав в броузере адрес URL главной страницы, я могу проверить почту и отослать ее из любого места, где окажусь, если только смогу найти компьютер с веб-броузером (и смогу мириться с ограничениями прототипа). Для этого подойдет практически любой компьютер и броузер: даже не обязательно должен быть установлен Python и с клиентского компьютера не потребуется обращаться к POP или SMTP-серверу. Это не относится к программе-клиенту PyMailGUI, которую мы написали в главе 14. Данное свойство особенно полезно в учреждениях, где имеется доступ к Веб, но ограничивается доступ к другим протоколам, таким как POP.

Но прежде чем бросаться всем в коллективное использование Интернета и совершенно позабыть традиционные настольные API, такие как tkinter, надлежит сказать несколько слов о ситуации в целом.

PyMailGUI и PyMailCGI

Помимо общей иллюстрации крупных приложений CGI примеры PyMailGUI и PyMailCGI были выбраны для этой книги, чтобы подчеркнуть некоторые компромиссы, встречающиеся при разработке веб-приложений. PyMailGUI и PyMailCGI выполняют примерно одинаковые функции, но в корне отличны по реализации:

PyMailGUI

Это традиционное «настольное» приложение с интерфейсом пользователя: оно целиком выполняется на локальном компьютере, для реализации пользовательского интерфейса вызывает функции библиотеки графического интерфейса, выполняемой в том же процессе, и взаимодействует с Интернетом через сокеты только при необходимости (например, для загрузки или отправки почты по требованию). Запросы пользователя немедленно направляются методам-обработчикам, выполняемым локально, а между запросами информация о состоянии автоматически сохраняется в совместно используемых переменных. Как уже говорилось, благодаря тому, что между событиями данные сохраняются в памяти, программа PyMailGUI способна сохранять сообщения в кэше – она загружает заголовки и выбранные сообщения только один раз, а при последующих операциях загружаются только заголовки вновь прибывших сообщений. Она обладает достаточным объемом информации для проверки синхронизации с почтовым ящиком на сервере. При выполнении удалений PyMailGUI может просто обновить в памяти кэш загруженных заголовков без необходимости повторно загружать весь список. Кроме того, благодаря тому, что программа PyMailGUI выполняется как единый процесс на локальном компьютере, она может использовать такие механизмы, как потоки выполнения, чтобы обеспечить возможность одновременного выполнения нескольких операций передачи почты (вы можете одновременно отправлять и загружать сообщения) и в ней проще реализовать поддержку дополнительных функциональных возможностей, таких как сохранение и просмотр файлов с почтой.

PyMailCGI

Как и все системы CGI, состоит из сценариев, располагающихся и выполняющихся на сервере и генерирующих разметку HTML для организации взаимодействий с пользователем через веб-броузер на компьютере клиента. Выполнение происходит только в контексте веб-броузера, а запросы пользователя обрабатываются путем выполнения удаленных сценариев CGI на сервере. Без явного использования механизмов сохранения информации, таких как базы данных, сценарии CGI не имеют возможность хранить информацию о состоянии подобно PyMailGUI – каждый обработчик запроса будет выполняться автономно и только с той информацией о состоянии, которая явно передается от предыдущих сценариев в виде скрытых полей форм

или параметров запросов в адресах URL. Вследствие этого в настоящей реализации программа PyMailCGI должна заново загружать все заголовки сообщений всякий раз, когда ей требуется отобразить список выбора, повторно загружая при этом сообщения, которые уже были загружены ранее в этом же сеансе, и в целом она не способна проверить синхронизацию с почтовым ящиком. Эту проблему можно решить путем реализации более сложных схем сохранения информации с применением cookies или баз данных, но ни одно из этих решений не отличается простотой, сравнимой с использованием обычной памяти процесса, как в PyMailGUI.

Веб-приложения и настольные приложения

Конечно, конкретные функциональные возможности этих систем отличаются – PyMailCGI реализует лишь подмножество возможностей PyMailGUI – но они близки достаточно, чтобы их можно было сравнивать. На базовом уровне в обеих системах для получения и отправки почты через сокеты используются модули Python поддержки протоколов POP и SMTP. Но в представляемых ими альтернативах реализации есть важные отличия, о которых следует знать при создании веб-систем:

Издержки производительности

Сети медленнее, чем CPU. В настоящей реализации скорость или полнота PyMailCGI не идут ни в какое сравнение с PyMailGui. В PyMailCGI каждый раз, когда пользователь щелкает на кнопке отправки формы, запрос передается через сеть (в случае использования имени сервера «localhost» запросы передаются другой программе, выполняющейся на том же компьютере, но эта конфигурация используется только во время тестирования). Более точно, каждый запрос пользователя влечет расходы, связанные с передачей по сети, вызов каждого обработчика может выливаться в запуск нового процесса или потока выполнения на сервере, параметры поступают в виде текстовых строк, требующих синтаксического анализа, а отсутствие на сервере информации о состоянии при переходе к новой странице означает, что почту приходится часто перезагружать или использовать механизмы сохранения информации о состоянии, которые существенно сложнее и медленнее, чем простое обращение к памяти.

Напротив, действия пользователя в PyMailGUI производят вызовы функций в том же процессе вместо передачи по сети и ветвления процессов, а информация о состоянии легко сохраняется в переменных процесса. Даже при наличии сверхбыстрого соединения с Интернетом CGI-система на стороне сервера проигрывает в скорости программе на стороне клиента. По правде говоря, некоторые операции tkinter тоже передаются обрабатывающей их библиотеке Tcl в виде строк, которые нужно анализировать. Со временем это может измениться; к тому же здесь производится сравнение CGI-сценариев с библиоте-

ками графического интерфейса в целом. Вызовы функций наверняка всегда будут превосходить в скорости сетевые взаимодействия.

Некоторые из этих узких мест могут быть устранены ценой увеличения сложности программы. Например, некоторыми веб-серверами используются потоки выполнения и пулы процессов, чтобы свести к минимуму количество операций создания процессов для сценариев CGI. Кроме того, часть информации о состоянии можно передавать вручную со страницы на страницу в скрытых полях формы, параметрах URL и cookies, а в промежутке между страницами состояние можно сохранять в базе данных, допускающей одновременный доступ, чтобы свести к минимуму необходимость повторной загрузки почты. Но нельзя пройти мимо того факта, что переправка событий сценариям через сеть происходит значительно медленнее, чем прямой вызов функций Python. Это может иметь большое значение для некоторых приложений, хотя и не для всех.

Издержки сложности

Работать с разметкой HTML неудобно. Поскольку PyMailCGI должна генерировать разметку HTML, чтобы взаимодействовать с пользователем через веб-броузер, она более сложна (или, по крайней мере, менее удобочитаема), чем PyMailGUI. В некотором смысле сценарии CGI встраивают код разметки HTML в программный код Python; механизмы шаблонов, такие как PSP, часто используют противоположный подход. В любом случае в конечном итоге получается смесь двух очень разных языков, поэтому создание интерфейса с помощью HTML в сценарии CGI может оказаться далеко не таким простым делом, как вызовы функций из библиотек графического интерфейса, таких как tkinter.

Посмотрите, например, сколько труда мы приложили для экранирования HTML и URL в примерах этой главы; такие ограничения заложены в природе HTML. Кроме того, внесение изменений в систему с целью сохранения списка загруженной почты в базе данных в промежутке между обращениями к страницам приведет к еще большему усложнению системы, основанной на CGI, (и, вероятнее всего, придется привлечь еще один язык, такой как SQL, пусть даже и в самых низкоуровневых функциях). Использование защищенного протокола HTTP могло бы уменьшить сложности, обусловленные необходимостью реализации шифрования, но с его введением появляются новые сложности настройки сервера.

Издержки функциональности

Язык HTML не обладает богатыми выразительными возможностями. Язык HTML служит переносимым способом определения простых страниц и форм, но слаб или бесполезен для описания более сложных интерфейсов пользователя. Поскольку сценарии CGI создают интерфейсы пользователя путем отправки разметки HTML броузеру, они весьма ограничены в отношении конструкций, использу-

мых интерфейсом пользователя. Попробуйте, например, реализовать программу обработки изображений и анимации в виде сценариев CGI: язык HTML не пригоден за пределами заполняемых форм и простых взаимодействий.

Можно создать сценарии CGI, генерирующие графические изображения. Эти изображения могут создаваться и сохраняться на сервере во временных файлах с именами, сконструированными из идентификатора сеанса и используемыми в тегах `<IMG>` в сгенерированной разметке HTML ответа. Для броузеров, поддерживающих такое понятие, как встроенные изображения, графику можно было бы встраивать непосредственно в теги HTML, в формате Base64 или подобном. Однако любой из этих приемов существенно сложнее, чем использование изображений в библиотеке tkinter. Кроме того, приложения обработки изображений и анимации, для которых большое значение имеет скорость реакции на действия пользователя, в принципе невозможно реализовать с применением таких протоколов, как CGI, требующих выполнения сетевых взаимодействий для каждой операции. Например, интерактивные сценарии обработки изображений и анимации, которые мы написали в конце главы 9, невозможно реализовать в виде обычных серверных сценариев.

Это как раз то ограничение, для преодоления которого были придуманы апплеты Java – программы, которые хранятся на сервере, но по требованию загружаются для выполнения у клиента и предоставляют доступ к полноценной библиотеке графического интерфейса для создания более богатых интерфейсов пользователя. Тем не менее программам, строго ограниченным стороной сервера, внутренне присущи ограничения языка HTML.

Кроме того, клиентские программы, такие как PyMailGUI, также обладают доступом к таким инструментам, как многопоточная модель выполнения, которую сложно имитировать в CGI-приложениях (потоки выполнения, порождаемые сценарием CGI, не могут существовать дольше самого сценария или дополнять отправленный им ответ). Модели веб-приложений с постоянными процессами, например FastCGI, могут предоставлять дополнительные возможности в этом направлении, но общая картина выглядит не так четко, как на стороне клиента.

Веб-разработчики прикладывают немалые усилия для имитации возможностей клиентских приложений – смотрите обсуждение модели RIA и HTML 5 далее – но эти усилия связаны с дополнительными сложностями, использованием модели программирования на стороне сервера почти на пределе возможностей и применением самых разнообразных веб-технологий.

Преимущества переносимости

Все, что вам нужно, – это броузер на стороне клиента. Преимущество программы PyMailCGI заключается в том, что она действует

в Сети и с ней можно работать на любом компьютере, где есть веб-броузер, независимо от наличия на нем Python и tkinter. Это значит, что Python должен быть установлен только на одном компьютере – на веб-сервере, где в действительности располагаются и выполняются сценарии. Фактически это самое значительное преимущество модели веб-приложений. Если известно, что пользователи вашей системы имеют веб-броузеры, установка ее становится тривиально простой. Вам все еще необходим Python на сервере, но удовлетворить это требование значительно проще.

Если вы помните, Python и tkinter тоже весьма переносимы – они выполняются на всех основных оконных системах (X11, Windows, Mac), но для выполнения клиентской программы Python/tkinter, такой как PyMailGUI, вам потребуются на машине клиента собственно Python и tkinter. Иное дело приложения, построенные как сценарии CGI: они будут работать в Macintosh, Linux, Windows и в любой другой системе, которая позволяет отображать веб-страницы HTML. В этом смысле HTML становится для веб-сценариев своего рода переносимым языком конструирования графических интерфейсов, интерпретируемым веб-броузером, который сам является графическим интерфейсом для отображения других графических интерфейсов. Вам даже не нужен исходный программный код или байт-код самих сценариев CGI – они выполняются на удаленном сервере, существующем где-то в сети, а не на компьютере, где работает броузер.

Требования к выполнению

Но вам нужен броузер. То есть сама природа веб-систем может сделать их бесполезными в некоторых средах. Несмотря на повсеместное распространение Интернета, многие приложения все еще выполняются в условиях, когда отсутствует броузер или нет доступа к Интернету. Возьмите, например, встроенные системы, системы реального времени или защищенные правительственные приложения. Хотя в *интрасетях* (локальных сетях без внешних соединений) веб-приложения также могут иногда выполняться, однако мне приходилось работать в компаниях, где у клиентов вообще отсутствовали веб-броузеры. С другой стороны, у таких клиентов бывает проще установить на локальных компьютерах системы типа Python, чем организовать поддержку внутренней или внешней сети.

Требования к администрированию

В действительности необходим еще и сервер. Системы на основе CGI вообще нельзя писать без доступа к веб-серверу. Кроме того, хранение программ на централизованном сервере создает довольно существенные административные издержки. Попросту говоря, в чистой архитектуре клиент/сервер клиенты проще, но сервер становится критически важным ресурсом и потенциально узким местом с точки зрения производительности. Если централизованный сервер выйдет из строя, вы, ваши служащие и клиенты можете оказаться лишены

возможности работать. Кроме того, если достаточно много клиентов пользуется одновременно общим сервером, издержки в скорости веб-систем становятся еще более явными. В промышленных системах можно использовать дополнительные методики, такие как балансировка нагрузки и использование отказоустойчивых серверов, но все это добавляет новые требования.

На самом деле можно утверждать, что смещение в сторону архитектуры веб-сервера является отчасти движением вспять – ко временам централизованных мейнфреймов и простых терминалов. Некоторые могли бы даже утверждать, что недавно появившаяся модель *облачных вычислений* отчасти является возвратом к более старым вычислительным моделям. Какой бы путь мы ни выбрали, разгрузка сервера и перенос хотя бы части обработки данных на клиентские компьютеры позволяет преодолеть это узкое место.

Другие подходы

Так как же лучше всего строить приложения для Интернета – как клиентские программы, общающиеся с Сетью, или как выполняемые на сервере программы, жизнь которых проходит в Сети? Естественно, однозначного ответа на этот вопрос нет, так как все зависит от конкретных ограничений каждого приложения. Более того, ответов может быть больше, чем здесь предложено. Модели программирования на стороне клиента и на стороне сервера имеют свои достоинства и недостатки, тем не менее, для большинства стандартных проблем Веб и CGI уже предложены стандартные решения. Например:

Решения для стороны клиента

Программы, выполняемые на стороне клиента и на стороне сервера, могут перемешиваться различными способами. Например, программы апплетов располагаются на сервере, но загружаются и выполняются как программы клиента, имея доступ к богатым библиотекам графических интерфейсов.

Другие технологии, такие как встраивание JavaScript или Python непосредственно в код разметки HTML, тоже поддерживают выполнение на стороне клиента и более богатые возможности графических интерфейсов. Такие сценарии располагаются в HTML на сервере, но после загрузки выполняются у клиента и имеют доступ к компонентам броузера через открытую объектную модель.

Новые расширения динамического HTML (Dynamic HTML, DHTML) предоставляют клиентским сценариям еще одну возможность изменять веб-страницы после их создания. А недавно появившаяся модель AJAX предлагает дополнительные способы добавления интерактивности и повышения отзывчивости веб-страниц и составляет основу модели полнофункциональных интернет-приложений (RIA), которая отмечается ниже. Все эти клиентские технологии добавля-

ют собственные сложности, но ослабляют некоторые ограничения, налагаемые обычным HTML.

Решения по сохранению информации о состоянии

В предыдущей главе мы подробно рассматривали средства и методы сохранения информации о состоянии, и нам еще предстоит в главе 17 познакомиться с полномасштабными базами данных для Python. Некоторые серверы веб-приложений (например, Zope) обеспечивают естественную поддержку сохранения информации о состоянии в промежутке между обращениями к страницам, предоставляя объектные базы данных с одновременным доступом. В некоторых из этих систем явно используются базы данных (например, Oracle или MySQL); в других используются файлы или хранилища объектов Python с соответствующей блокировкой. Кроме того, имеются механизмы объектно-реляционных отображений (Object Relational Mapper, ORM), такие как SQLObject, которые позволяют взаимодействовать с реляционными базами данных как с классами Python.

Сценарии могут также передавать информацию о состоянии через скрытые поля форм и параметры генерируемых URL, как это делается в PyMailCGI, либо сохранять ее на стороне клиента с помощью стандартного протокола cookies. Как мы узнали в главе 15, cookies представляют собой фрагменты информации, сохраняемые у клиента по запросу сервера, и отправляются обратно серверу при повторном посещении страницы (данные передаются в обоих направлениях в заголовках HTTP). Cookies сложнее, чем переменные программы, и являются чем-то спорным и необязательным, но они могут взять на себя часть задач по сохранению информации о состоянии.

Дополнительные возможности по сохранению информации предлагают альтернативные модели, такие как FastCGI и `mod_python`, – там, где они поддерживаются, приложения в модели FastCGI могут сохранять информацию в рамках долгоживущих процессов, а модуль `mod_python` предоставляет возможность сохранять данные сеанса в Apache.

Решения по созданию HTML

Сторонние расширения тоже могут отчасти уменьшить сложность встраивания HTML в CGI-сценарии Python, хотя и ценой некоторого снижения скорости выполнения. Например, система HTMLgen позволяет программам конструировать страницы как деревья объектов Python, которые «умеют» создавать разметку HTML. Имеются также другие фреймворки, предоставляющие объектно-ориентированные интерфейсы для создания потока ответа (например, объект ответа с методами). При использовании системы такого типа сценарии Python имеют дело только с объектами, а не с синтаксисом самого HTML.

Другие системы, такие как PHP, Python Server Pages (PSP), Zope DHTML и ZPT, и Active Server Pages предоставляют механизмы шаблонов, позволяющие встраивать в HTML программный код на языке сценариев, который выполняется на сервере, чтобы динамически соз-

давать или определять части разметки HTML, отправляемой клиенту в ответ на запрос. Все они позволяют избавить программный код Python от сложностей, связанных с созданием кода разметки HTML, и отделить логику от представления, но они могут привносить свои сложности, обусловленные необходимостью смешивания различных языков.

Разработка обобщенного пользовательского интерфейса

Чтобы охватить обе модели, некоторые системы пытаются максимально отделить логику от представления, благодаря чему выбор модели практически перестает играть существенное значение – полностью инкапсулируя детали отображения, одна и та же программа способна, в принципе, отобразить и традиционный графический интерфейс, и веб-страницы на основе HTML. Однако из-за существенных различий между архитектурами этот идеал труднодостижим, и он не устраняет существенные отличия между клиентской и серверной платформами. Такие проблемы, как сохранение информации о состоянии и необходимость взаимодействия с сетью, имеют более важное значение, чем создание окон и элементов управления, и могут гораздо существеннее влиять на программный код.

Другие системы могут пытаться достичь похожих целей за счет абстрагирования визуального представления – стандартное представление в формате XML, например, можно было бы использовать для создания и графического интерфейса, и разметки HTML. Однако здесь также решается только проблема отображения и никак не затрагиваются фундаментальные архитектурные различия между клиентскими и серверными подходами.

Новые технологии: RIA и HTML 5

Наконец, имеются более высокоуровневые подходы, такие как инструменты создания полнофункциональных интернет-приложений (Rich Internet Application, RIA), представленные в главах 7 и 12, которые могут предложить дополнительные функциональные возможности, отсутствующие в HTML, и способны приблизиться по своим возможностям к инструментам создания графических интерфейсов. С другой стороны, они могут еще больше усложнять разработку веб-приложений и добавлять в смесь дополнительные языки. Результат часто напоминает веб-эквивалент Вавилонской башни, для создания которой требуется одновременно писать на языках Python, HTML, SQL, JavaScript, на языке шаблонов, использовать прикладной интерфейс объектно-реляционного отображения и многое другое, и даже вложенные и встроенные друг в друга комбинации из них. В результате весь комплекс программного обеспечения получается более сложным, чем комбинация Python и библиотеки графического интерфейса.

Кроме того, современные полнофункциональные интернет-приложения унаследовали невысокую скорость реакции, свойственную сете-

вым системам в целом. Несмотря на то, что технология AJAX способна повысить интерактивность веб-страниц, тем не менее она предполагает использование сетевых взаимодействий, а не вызовы функций внутри процесса. Как ни странно, но полнофункциональные интернет-приложения подобно многим настольным приложениям могут также требовать для своей работы установки расширений броузера на стороне клиента. Новый стандарт HTML 5 может снять ограничения, связанные с расширениями, и несколько уменьшить сложность, но он несет в себе массу новых, своих собственных сложностей, которые мы не будем рассматривать здесь.

Очевидно, технология Интернета предполагает некоторые компромиссы и продолжает быстро развиваться. Тем не менее, это уместный контекст поставки многих, хотя и не всех, приложений. Как и во всяком проектном решении, вы сами должны выбрать подход. Хотя при размещении систем в Веб могут потеряться скорость, функциональность и увеличиться сложность, значимость таких потерь со временем, вероятно, уменьшится. Смотрите начало главы 12, где приводится дополнительная информация о некоторых системах, обещающих подобные изменения, и следите за дальнейшим ходом развития Интернета в Веб.

Рекомендуется для ознакомления: система PyErrata

Теперь, когда я поведал вам обо всех причинах, которые могли бы препятствовать созданию систем для Веб, я собираюсь высказать совершенно противоположное мнение и представить систему, которая просто обязана быть реализованной в виде веб-приложения. Во втором издании этой книги присутствовала глава, описывающая веб-сайт PyErrata – программу на языке Python, которая позволяла любому посетителю, с любого компьютера отправлять комментарии и отчеты об обнаруженных ошибках (или опечатках) через Интернет, используя лишь веб-броузер. Такая система должна хранить информацию на сервере, чтобы она была доступна произвольному клиенту.

Из-за нехватки места в книге эта глава была исключена из книги в третьем издании. Однако ее содержимое было включено в состав пакета примеров для книги как дополнительная литература. Файл с главой вы найдете в каталоге *PP4E\Internet\Web\PyErrata*[1] в пакете примеров для книги (подробнее о пакете примеров рассказывается в предисловии к книге).

[1] На английском языке. – *Прим. перев.*

PyErrata в некоторых отношениях проще примера PyMailCGI, представленного в этой главе. С точки зрения пользователя система PyErrata скорее иерархическая, чем линейная: взаимодействие с пользователем короче и порождает меньше страниц. Кроме того, в PyErrata мало данных состояния сохраняется в самих страницах – параметры URL передают состояние только в одном отдельном случае, и не генерируется скрытых полей форм.

С другой стороны, PyErrata вводит совершенно новое измерение: постоянное хранилище данных (persistent data storage). Состояние (сообщения об ошибках и комментарии) постоянно хранится этой системой на сервере, в плоских файлах или в базе данных, реализованной на основе модуля `shelve`. В обоих случаях присутствует призрак одновременного обновления, так как в одно и то же время обратиться к сайту может любое количество пользователей киберпространства. По этой причине в PyErrata используется также механизм блокировки файлов.

Я больше не поддерживаю веб-сайт, о котором рассказывается в этой дополнительной главе, а представленные в ней сведения несколько устарели. Например, теперь для блокировки файлов предпочтительнее использовать функцию `os.open`. Я мог бы также использовать другие современные системы хранения данных, такие как ZODB. Программный код в примерах к этой главе и код, входящий в состав пакета с примерами, написан для Python 2.X. И сам сайт лучше было бы реализовать как блог или Вики (понятия и названия, возникшие после того, как этот сайт был создан).

Однако программа PyErrata является дополнительным примером реализации веб-сайта на языке Python, и она достойно представляет веб-сайты, которые должны хранить информацию на сервере.

V

Инструменты и приемы

В этой части книги собраны дополнительные темы, касающиеся разработки приложений на языке Python. Большинство представленных здесь инструментов может быть использовано в самых разных областях прикладного программирования. Здесь вы найдете следующие главы:

Глава 17

Эта глава освещает часто используемые и развитые технологии Python хранения информации между запусками программы – файлы DBM, сериализация объектов, хранилища объектов и интерфейсы Python к базам данных SQL – и кратко знакомит с полноценными объектно-ориентированными базами данных (OODB), такими как ZODB, а также с механизмами объектно-реляционного отображения (ORM), такими как SQLObject и SQLAlchemy. В примерах работы с базами данных SQL будет использоваться поддержка SQLite, входящая в состав стандартной библиотеки Python, но сам прикладной интерфейс с легкостью можно перенести на использование более крупных систем, таких как MySQL.

Глава 18

В этой главе рассматриваются приемы реализации более сложных структур данных на языке Python – стеков, множеств, двоичных деревьев поиска, графов и других. В Python они принимают форму реализаций объектов.

Глава 19

В этой главе изучаются инструменты и приемы, используемые в языке Python для синтаксического анализа текстовой информации, – разбиение и объединение строк, поиск с применением регулярных выражений, анализ документов XML, анализ методом рекурсивного спуска и более сложные темы, связанные с языками.

Глава 20

В этой главе представлены приемы интеграции – расширение Python с помощью компилируемых библиотек и встраивание программного кода на языке Python в другие приложения. Несмотря на то, что основное внимание здесь будет уделяться связыванию Python с откомпилированным программным кодом на языке C, тем не менее мы также коснемся вопросов интеграции с Java, .NET и другими языками. В этой главе предполагается, что вы умеете читать программы на языке C, и она предназначена, главным образом, для разработчиков, занимающихся реализацией прикладных уровней интеграции.

Это последняя часть книги, посвященная чистому языку Python, и она интенсивно использует инструменты, представленные ранее в книге, особый упор делая на повторное использование программного кода. Например, калькулятор с графическим интерфейсом (PyCalc) служит для демонстрации понятий обработки языков и повторного использования программного кода.

17

Базы данных и постоянное хранение

«Дайте мне приказ стоять до конца, но сохранить данные»

До сих пор в этой книге мы использовали Python в системном программировании, для разработки графических интерфейсов и создания сценариев для Интернета – трех наиболее типичных областях применения Python, где он наиболее ярко проявляется как прикладной язык программирования в целом. В следующих четырех главах мы бросим беглый взгляд на другие важные темы программирования на языке Python: постоянное хранение данных, приемы работы со структурами данных, обработку текста и интеграцию языков Python/C.

Все эти темы имеют отношение не к прикладному программированию как таковому, а к смежным областям. Например, знания о базах данных, полученные в этой главе, можно применять при создании веб-приложений, при разработке настольных приложений с графическим интерфейсом и так далее. Умение обрабатывать текст также является универсальным навыком. Кроме того, хотя ни одна из четырех тем не освещается исчерпывающим образом (каждой из них вполне можно посвятить отдельную книгу), мы представим примеры работы Python в этих областях и подчеркнем основные идеи и инструменты. Если какая-либо из этих глав вызовет у вас интерес, то следует обратиться к дополнительным источникам.

Возможности постоянного хранения данных в Python

В этой главе наше внимание будет сосредоточено на *постоянно хранящихся* данных, которые продолжают существовать после завершения создавшей их программы. По умолчанию это не так для объектов, создаваемых сценариями: такие объекты, как списки, словари и даже экземпляры классов, находятся в памяти компьютера и исчезают, как только сценарий завершает работу. Чтобы заставить данные жить дольше, требуется предпринять особые меры. В программировании на языке Python есть по крайней мере шесть традиционных способов сохранения информации между запусками программы:

Плоские файлы

Обеспечивают хранение текста и байтов непосредственно на компьютере

Файлы DBM

Обеспечивают доступ к строкам, хранящимся в файлах, напоминающих словари, по ключу

Сериализованные объекты

Сериализованные объекты могут сохраняться в файлах и потоках

Файлы хранилищ (shelve)

Обеспечивают хранение сериализованных объектов в файлах DBM

Объектно-ориентированные базы данных (OODB)

Обеспечивают сохранение объектов в хранилищах, имеющих структуру словарей (ZODB, Durus)

Реляционные базы данных SQL (RDBMS)

Хранилища в виде таблиц, поддерживающие запросы SQL (SQLite, MySQL, PostgreSQL и другие)

Объектно-реляционные отображения (ORM)

Промежуточные механизмы, обеспечивающие отображение классов Python в реляционные таблицы (SQLObject, SQLAlchemy)

В некотором смысле интерфейсы Python к сетевым протоколам передачи объектов, таким как SOAP, XML-RPC и CORBA, также предоставляют возможность сохранения данных, но их описание выходит далеко за рамки этой главы. Интерес для нас здесь представляют приемы, позволяющие программам сохранять данные непосредственно и, обычно, на локальном компьютере. Некоторые серверы баз данных могут действовать на удаленном компьютере в сети, однако в целом это не имеет значения для большинства приемов, которые мы будем изучать здесь.

Мы изучали интерфейсы Python простых (или «плоских») файлов в главе 4 и с того момента пользуемся ими. Python предоставляет стандарт-

ный доступ к файловой системе stdio (через встроенную функцию open), а также на низком уровне – через дескрипторы файлов (с помощью встроенного модуля os). Для простых задач хранения данных многим сценариям ничего другого не требуется. Чтобы сохранить данные для использования при следующих запусках программы, нужно записать их в открытый файл на компьютере, а потом прочитать их обратно из этого файла. Как мы видели, в более сложных задачах Python поддерживает также другие интерфейсы, сходные с файлами, такие как каналы, очереди и сокеты.

Так как мы уже знакомы с плоскими файлами, я не буду больше рассказывать здесь о них. Оставшаяся часть главы знакомит с другими темами из списка, приведенного в начале раздела. В конце мы также познакомимся с программой с графическим интерфейсом для просмотра содержимого файлов хранилищ и файлов DBM. Но прежде нам нужно разобраться, что это за звери.

Примечание к четвертому изданию: В предыдущем издании этой книги использовался интерфейс mysql-python к системе управления реляционными базами данных MySQL, а также система управления объектно-ориентированными базами данных ZODB. Когда я занимался обновлением этой главы в июне 2010 года, ни один из них еще не был доступен в Python 3.X – версии Python, используемой в этом издании. По этой причине большая часть информации о ZODB была убрана из главы, а примеры работы с базами данных SQL были переориентированы на использование интерфейса SQLite к базе данных внутри процесса, который входит в состав стандартной библиотеки Python 3.X. Примеры использования ZODB и MySQL и обзоры из предыдущего издания по-прежнему доступны в пакете примеров к книге, как будет описано ниже. Однако благодаря переносимости интерфейса к базам данных SQL в языке Python, программный код, использующий SQLite, практически без изменений сможет работать с большинством других баз данных.

Файлы DBM

Плоские файлы удобно использовать для простых задач постоянного хранения данных, но обычно они связаны с последовательным режимом обработки. Несмотря на возможность произвольно перемещаться по файлам с помощью вызовов метода seek, плоские файлы мало что вносят в структуру данных помимо понятий байтов и текстовых строк.

Файлы DBM, стандартный инструмент в библиотеке Python для управления базами данных, улучшают это положение, предоставляя доступ к хранящимся строкам текста по ключу. Они реализуют представление хранящихся данных с произвольным доступом и одним ключом. Например, информация, относящаяся к объектам, может храниться в файле DBM с использованием уникального ключа для каждого объекта

и позднее может быть получена обратно с помощью того же самого ключа. Файлы DBM реализуются с помощью целого ряда базовых модулей (в том числе одного, написанного на языке Python), но если у вас есть Python, значит, есть и поддержка DBM.

Работа с файлами DBM

Хотя файловые системы DBM должны проделать некоторую работу по отображению сохраняемых данных в ключи для быстрого извлечения (технически для сохранения данных в файлах они обычно используют прием, называемый *хешированием*), сценариям не приходится беспокоиться о том, что происходит за кулисами. В действительности файлы DBM являются одним из простейших способов сохранения информации в Python – файлы DBM ведут себя настолько сходно со словарями, размещаемыми в памяти, что можно забыть о том, что в самом деле вы работаете с файлом. Например, если есть объект файла DBM:

- Операция индексирования по ключу извлекает данные из файла.
- Операция присвоения по индексу сохраняет данные в файле.

Объекты файлов DBM поддерживают также стандартные методы словарей, такие как выборка и проверка по списку ключей и удаление по ключу. Сама библиотека DBM скрыта за этой простой моделью. Ввиду простоты перейдем прямо к интерактивному примеру, в котором создается файл DBM и демонстрируется, как работает интерфейс:

```
C:\...\PP4E\Dbase> python
>>> import dbm                  # получить интерфейс: bsddb, gnu, ndbm, dumb
>>> file = dbm.open('movie', 'c')  # создать файл DBM с именем 'movie'
>>> file['Batman'] = 'Pow!'        # сохранить строку с ключом 'Batman'
>>> file.keys()                    # получить список ключей в файле
[b'Batman']
>>> file['Batman']                 # извлечь значение по ключу 'Batman'
b'Pow!'

>>> who = ['Robin', 'Cat-woman', 'Joker']
>>> what = ['Bang!', 'Splat!', 'Wham!']
>>> for i in range(len(who)):
...     file[who[i]] = what[i]     # добавить еще 3 "записи"
...
>>> file.keys()
[b'Cat-woman', b'Batman', b'Joker', b'Robin']
>>> len(file), 'Robin' in file, file['Joker']
(4, True, b'Wham!')
>>> file.close()                   # иногда требуется закрывать явно
```

На практике при импорте стандартного модуля `dbm` автоматически загружается тот интерфейс DBM, который доступен вашему интерпретатору Python (опробуются различные альтернативы в строго определенном порядке), а при открытии нового файла DBM создается один или более внешних файлов с именами, начинающимися со строки `'movie'`

(о деталях будет сказано чуть ниже). Но после импорта и открытия файл DBM фактически неотличим от словаря.

В сущности, объект с именем file в этом примере можно представить себе как словарь, отображаемый во внешний файл с именем movie. Единственные очевидные отличия состоят в том, что ключами могут быть только строки (а не произвольные неизменяемые объекты), и для доступа к данным необходимо открывать, а после изменения данных – закрывать файл.

Однако в отличие от обычных словарей, содержимое file сохраняется в перерыве между запусками программы Python. Если мы потом вернемся и заново запустим Python, наш словарь будет по-прежнему доступен. Файлы DBM похожи на словари, которые требуется открывать:

```
C:\...\PP4E\Dbase> python
>>> import dbm
>>> file = dbm.open('movie', 'c')  # открыть существующий файл DBM
>>> file['Batman']
b'Pow!'

>>> file.keys()                    # метод keys возвращает список ключей
[b'Cat-woman', b'Batman', b'Joker', b'Robin']

>>> for key in file.keys(): print(key, file[key])
...
b'Cat-woman' b'Splat!'
b'Batman' b'Pow!'
b'Joker' b'Wham!'
b'Robin' b'Bang!'
```

Обратите внимание, что при обращении к методу keys файлов DBM возвращается действительный список – здесь не показано, но метод values этих файлов, напротив, возвращает итерируемое представление, как при обращении к обычным словарям. Кроме того, файлы DBM всегда хранят ключи и значения в виде объектов bytes – интерпретация их в виде текста Юникода остается за клиентским приложением. При обращении к хранимым ключам или значениям в нашем программном коде мы можем использовать строки bytes или str – использование строк bytes позволяет сохранять ключи и значения в произвольных кодировках Юникода, тогда как при использовании объектов str они будут кодироваться в объекты bytes внутренней реализацией DBM с применением кодировки UTF-8.

Однако при необходимости мы всегда можем декодировать строки bytes в строки str для отображения, и подобно словарям файлы DBM имеют итератор по ключам. Кроме того, операции присваивания и удаления по ключу изменяют содержимое файла DBM, и после внесения изменений необходимо закрывать файл (это гарантирует сохранение изменений на диске):

```
>>> for key in file: print(key.decode(), file[key].decode())
...
Cat-woman Splat!
Batman Pow!
Joker Wham!
Robin Bang!

>>> file['Batman'] = 'Ka-Boom!'  # изменить значение ключа Batman
>>> del file['Robin']            # удалить запись Robin
>>> file.close()                 # закрыть после изменений
```

За исключением необходимости импортировать интерфейс и открывать/закрывать файл DBM, программам на языке Python не требуется ничего знать собственно о DBM. Модули DBM добиваются такой интеграции, перегружая операции индексирования и переадресуя их более простым библиотечным инструментам. Но глядя на этот программный код, вы никогда бы этого не узнали – файлы DBM выглядят как обычные словари Python, хранящиеся во внешних файлах. Произведенные в них изменения сохраняются неограниченно долго:

```
C:\...\PP4E\Dbase> python
>>> import dbm                   # открыть файл DBM еще раз
>>> file = dbm.open('movie', 'c')
>>> for key in file: print(key.decode(), file[key].decode())
...
Cat-woman Splat!
Batman Ka-Boom!
Joker Wham!
```

Как видите, проще уже некуда. В табл. 17.1 перечислены наиболее частые операции с файлами DBM. Если открыт такой файл, он обрабатывается так, как если бы был словарем Python, находящимся в памяти. Выборка элементов осуществляется индексированием объекта файла по ключу, а запись – путем присвоения по ключу.

Таблица 17.1. Операции с файлами DBM

Программный код Python	Действие	Описание
`import dbm`	Импорт	Получить реализацию DBM
`file=dbm.open('filename', 'c')`	Открытие	Создать или открыть существующий файл DBM для ввода-вывода
`file['key'] = 'value'`	Запись	Создать или изменить запись с ключом `key`
`value = file['key']`	Выборка	Загрузить значение из записи с ключом `key`
`count = len(file)`	Размер	Получить количество записей
`index = file.keys()`	Индекс	Получить список (не представление) ключей

Программный код Python	Действие	Описание
`found = 'key' in file`	Запрос	Проверить наличие записи с ключом `key`
`del file['key']`	Удаление	Удалить запись с ключом `key`
`for key in file:`	Итерации	Выполнить итерации по имеющимся ключам
`file.close()`	Закрытие	Закрыть вручную, требуется не всегда

Особенности DBM: файлы, переносимость и необходимость закрытия

Несмотря на интерфейс, напоминающий словари, файлы DBM в действительности отображаются в один или более внешних файлов. Например, при использовании интерфейса `dbm` в Python 3.1 для Windows создаются два файла – *movie.dir* и *movie.dat* – когда создается файл DBM с именем `movie`, а при последующих операциях открытия создается еще файл *movie.bak*. Если Python имеет доступ к другому базовому интерфейсу файлов с доступом по ключу, на компьютере могут появиться другие внешние файлы.

Технически модуль `dbm` является интерфейсом к той файловой системе типа DBM, которая имеется в Python.

- При открытии существующего файла DBM модуль `dbm` пытается определить создавшую его систему с помощью функции `dbm.whichdb`. Вывод делается на основе содержимого самой базы данных.
- При создании нового файла `dbm` пытается загрузить интерфейсы доступа к файлам по ключу в строго определенном порядке. Согласно документации он пытается загрузить интерфейс `dbm.bsd`, `dbm.gnu`, `dbm.ndbm` или `dbm.dumb` и использовать первый, который будет благополучно загружен. При их отсутствии Python автоматически использует универсальную реализацию с именем `dbm.dumb`, которая, в действительности, конечно же не является «тупой» («dumb»), но не отличается такой производительностью и надежностью, как другие реализации.

В будущих версиях Python этот порядок выбора реализации может измениться, и могут даже появиться дополнительные альтернативы. Однако обычно о таких вещах не приходится беспокоиться, если не удалять файлы, которые создает ваша система DBM, или не перемещать их между компьютерами с различными конфигурациями. Если вас волнует проблема *переносимости* файлов DBM (как будет показано дальше, то же самое относится к файлам хранилищ `shelve`), вам необходимо позаботиться о настройке компьютеров, чтобы на них были установлены одни и те же интерфейсы DBM, или положиться на интерфейс `dumb`. На-

пример, пакет поддержки Berkeley DB (он же bsddb), используемый интерфейсом dbm.bsd, достаточно широко распространен и обладает высокой степенью переносимости.

Обратите внимание, что файлы DBM иногда требуется закрывать явно, в формате, приведенном в последней строке в табл. 17.1. Некоторые файлы DBM не требуют вызова метода закрытия, но другим он нужен, чтобы записать изменения из буфера на диск. В таких системах файл может оказаться поврежден, если не выполнить закрытие. К несчастью, используемая по умолчанию поддержка DBM в старых версиях Python для Windows, dbhash (она же bsddb), является как раз той системой DBM, которая требует вызова метода закрытия во избежание потери данных. Как правило, всегда следует закрывать файлы DBM явно после внесения изменений и перед выходом из программы, чтобы обойти возможные проблемы – по сути, это операция подтверждения изменений («commit»). Данное правило распространяется и на хранилища shelve, с которыми мы познакомимся далее в этой главе.

Последние изменения: Не забывайте также передавать строку 'c' во втором аргументе в вызов dbm.open, чтобы заставить интерпретатор создать файл, если он еще не существует. Прежде этот аргумент подразумевался по умолчанию, но теперь это не так. Аргумент 'c' не требуется передавать при открытии файлов-хранилищ, создаваемых модулем shelve, обсуждаемых далее, – для них по-прежнему по умолчанию используется режим 'c' «открыть или создать», если отсутствует аргумент, определяющий режим открытия. Модулю dbm можно передавать также другие строки, определяющие режим открытия, включая 'n', в котором всегда создается новый файл, и 'r', определяющий доступ к файлу только для чтения, – по умолчанию используется режим создания нового файла. За дополнительной информацией обращайтесь к руководству по стандартной библиотеке Python.

Кроме того, в Python 3.X все элементы, ключи и значения сохраняются как строки bytes, а не str, как мы уже видели (что, как оказывается, удобно при работе с сериализованными объектами в хранилищах shelve, обсуждаемых ниже). В этой версии интерпретатора более не доступен компонент bsddb, бывший ранее стандартным, однако он доступен как независимое стороннее расширение, которое можно загрузить из Интернета, а при его отсутствии Python переходит на использование собственной реализации файлов DBM. Поскольку правила выбора базовой реализации DBM могут изменяться со временем, вы всегда должны обращаться за дополнительной информацией к руководствам по библиотеке Python, а также к исходному программному коду стандартного модуля dbm.

Сериализованные объекты

Вероятно, наибольшим ограничением файлов DBM является тип хранимых в них данных: данные, записываемые под ключом, должны быть

простой текстовой строкой. При необходимости сохранять в файле DBM объекты Python иногда можно вручную преобразовывать их в строки и обратно при записи и чтении (например, с помощью функций `str` и `eval`), но это полностью не решает проблемы. Для объектов Python произвольной сложности, таких как экземпляры классов, требуется нечто другое. Объекты экземпляров классов, к примеру, обычно нельзя впоследствии воссоздать из стандартного строкового представления. Кроме того, при реализации собственных методов преобразования в строку и воссоздания объекта из строки легко допустить ошибку, и такие инструменты не являются универсальным решением.

Модуль Python `pickle`, входящий в стандартную поставку системы Python, обеспечивает требуемое преобразование. Это своего рода универсальный инструмент, осуществляющий прямое и обратное преобразование, – модуль `pickle` может преобразовывать практически любые объекты Python, находящиеся в памяти, в формат одной линейной строки, пригодной для хранения в плоских файлах, пересылки через сокеты по сети и так далее. Это преобразование объекта в строку часто называют преобразованием в последовательную форму, или *сериализацией* – произвольные структуры данных, размещенные в памяти, отображаются в последовательный строковый формат.

Строковое представление объектов иногда также называют потоком байтов в соответствии с его линейным форматом. При сериализации сохраняются содержимое и структура оригинального объекта в памяти. При последующем воссоздании объекта из такой строки байтов создается новый объект, идентичный оригиналу по структуре и содержимому, однако он будет размещен в другой области памяти.

В результате при воссоздании объекта фактически создается его *копия* – говоря на языке Python, соблюдается условие `==`, но не `is`. Поскольку воссоздание объектов происходит, как правило, в совершенно новом процессе, это отличие обычно не имеет большого значения (хотя, как мы видели в главе 5, именно это обстоятельство обычно препятствует использованию сериализованных объектов для непосредственной передачи информации о состоянии между процессами).

Операция сериализации может применяться практически к любым типам данных в языке Python – числам, спискам, словарям, экземплярам классов, вложенным структурам и другим – благодаря чему она является универсальным способом сохранения данных. Поскольку в последовательной форме сохраняются фактические объекты Python, в библиотеке отсутствует какой-либо прикладной интерфейс баз данных для них; объекты сохраняются и при последующем извлечении обрабатываются с применением обычного синтаксиса языка Python.

Применение сериализации объектов

Сериализация может показаться сложной, когда сталкиваешься с ней впервые, но отрадно узнать, что Python скрывает все сложности преоб-

разования объектов в строки. На самом деле интерфейсы модуля `pickle` чрезвычайно просты в употреблении. Например, чтобы преобразовать объект в последовательную форму, можно создать объект `Pickler` и вызывать его методы или использовать функции из модуля для достижения того же эффекта:

`P = pickle.Pickler(`*`file`*`)`

Создаст новый объект `Pickler` для вывода в последовательном виде в открытый выходной файл *`file`*.

`P.dump(`*`object`*`)`

Запишет объект *`object`* в файл/поток объекта `Pickler`.

`pickle.dump(`*`object, file`*`)`

То же, что два последних вызова вместе: сериализует объект *`object`* и выведет его в открытый файл *`file`*.

`string = pickle.dumps(`*`object`*`)`

Вернет объект *`object`* в сериализованном представлении, в виде строки символов.

Воссоздание объекта из строки с сериализованным представлением выполняется похожим образом; доступны простой функциональный и объектно-ориентированный интерфейсы:

`U = pickle.Unpickler(`*`file`*`)`

Создаст объект `Unpickler`, осуществляющий обратное преобразование сериализованной формы объекта из открытого входного файла *`file`*.

`object = U.load()`

Прочитает объект из файла/потока объекта `Unpickler`.

`object = pickle.load(`*`file`*`)`

То же, что два последних вызова вместе: восстановит объект из открытого файла.

`object = pickle.loads(`*`string`*`)`

Прочитает объект из строки символов вместо файла.

`Pickler` и `Unpickler` — это экспортируемые классы. Во всех этих вызовах аргумент `file` является либо объектом открытого файла, либо любым объектом, в котором реализованы те же атрибуты, что и у объектов файла:

- Объект `Pickler` вызывает метод `write` файла, передавая в качестве аргумента строку.
- Объект `Unpickler` вызывает метод `read` файла со счетчиком байтов и метод `readline` без аргументов.

Любой объект, имеющий эти атрибуты, может быть передан в качестве параметра `file`. В частности, аргумент `file` может быть экземпляром

класса Python, предоставляющего методы read/write (то есть, поддерживающим *интерфейс* файлов). Благодаря этому можно произвольно отображать сериализованные потоки в объекты, находящиеся в памяти с классами. Например, класс io.BytesIO, имеющийся в стандартной библиотеке и обсуждавшийся в главе 3, предоставляет интерфейс, отображающий обращения к файлам на строки байтов в памяти, благодаря чему его экземпляры могут служить альтернативой использованию строковых функций dumps/loads из модуля pickler.

Можно также пересылать объекты Python через сеть, заключая сокеты в оболочку, выглядящую как файл и вызывающую методы сериализации у отправителя и методы обратного преобразования у получателя (подробнее об этом рассказывается в разделе «Придание сокетам внешнего вида файлов и потоков ввода-вывода» в главе 12). Фактически передача сериализованных объектов Python по сети между доверенными узлами является более простой альтернативой использованию сетевых транспортных протоколов, таких как SOAP и XML-RPC, при условии что с обоих концов соединения находится Python (сериализованные объекты передаются не в виде XML, а в формате, специфическом для Python).

Последние изменения: В Python 3.X сериализованные объекты всегда представлены в виде строк bytes, а не str независимо от запрошенного уровня протокола (даже при запросе старейшего протокола ASCII возвращается строка байтов). Вследствие этого файлы, используемые для сохранения сериализованных объектов Python, всегда должны открываться в двоичном режиме. Кроме того, в 3.X также автоматически выбирается и используется оптимизированная реализация модуля_pickle, если она присутствует. Подробнее обе эти темы описываются ниже.

Сериализация в действии

Несмотря на то, что сериализованные объекты могут переправляться самыми необычными способами, тем не менее, в наиболее обычном случае для сериализации объекта в плоский файл нужно открыть файл в режиме записи и вызвать функцию dump:

```
C:\...\PP4E\Dbase> python
>>> table = {'a': [1, 2, 3],
             'b': ['spam', 'eggs'],
             'c': {'name':'bob'}}
>>>
>>> import pickle
>>> mydb = open('dbase', 'wb')
>>> pickle.dump(table, mydb)
```

Обратите внимание на наличие здесь вложенных объектов – объект Pickler способен обрабатывать объекты произвольной структуры. От-

метьте также, что файл открывается в двоичном режиме. В Python 3.X это является обязательным условием, потому что сериализованные объекты всегда представлены в виде строки bytes. Чтобы выполнить обратное преобразование в другом сеансе или при последующих запусках приложения, достаточно просто открыть файл и вызвать load:

```
C:\...\PP4E\Dbase> python
>>> import pickle
>>> mydb = open('dbase', 'rb')
>>> table = pickle.load(mydb)
>>> table
{'a': [1, 2, 3], 'c': {'name': 'bob'}, 'b': ['spam', 'eggs']}
```

Восстановленный объект имеет то же самое содержимое и ту же структуру, что и оригинал, но он создается в другой области памяти[1]. Это относится и к объектам, воссозданным в том же процессе, и к объектам, воссозданным в другом процессе. Напомню, что для воссозданного объекта выполняется условие ==, но не is:

```
C:\...\PP4E\Dbase> python
>>> import pickle
>>> f = open('temp', 'wb')
>>> x = ['Hello', ('pickle', 'world')] # список с вложенным кортежем
>>> pickle.dump(x, f)
>>> f.close()                          # закрыть, чтобы записать на диск
>>>
>>> f = open('temp', 'rb')
>>> y = pickle.load(f)
>>> y
['Hello', ('pickle', 'world')]
>>>
>>> x == y, x is y                     # то же значение, но разные объекты
(True, False)
```

Чтобы еще больше упростить этот процесс, модуль в примере 17.1 заключает вызовы прямого и обратного преобразований объектов в функции, которые также открывают файлы, хранящие сериализованную форму объекта.

Пример 17.1. PP4E\Dbase\filepickle.py

```
"Утилиты сохранения и восстановления объектов из плоских файлов"
import pickle

def saveDbase(filename, object):
    "сохраняет объект в файле"
    file = open(filename, 'wb')
    pickle.dump(object, file)  # сохранить в двоичный файл
```

[1] При некоторых условиях не исключена вероятность, что объект будет воссоздан в той же области памяти, но это совпадение будет чисто случайным. – *Прим. перев.*

```
    file.close()                # подойдет любой объект, похожий на файл

def loadDbase(filename):
    "загружает объект из файла"
    file = open(filename, 'rb')
    object = pickle.load(file)  # загрузить из двоичного файла
    file.close()                # воссоздать объект в памяти
    return object
```

Теперь, чтобы сохранить и извлечь объект, просто вызывайте функции из этого модуля. В примере ниже они используются для манипулирования довольно сложной структурой со множественными ссылками на одни и те же вложенные объекты – вложенный список с именем L сохраняется в файле в единственном экземпляре:

```
C:\...\PP4E\Dbase> python
>>> from filepickle import *
>>> L = [0]
>>> D = {'x':0, 'y':L}
>>> table = {'A':L, 'B':D}       # присутствуют две ссылки на список L
>>> saveDbase('myfile', table)   # сериализовать в файл

C:\...\PP4E\Dbase> python
>>> from filepickle import *
>>> table = loadDbase('myfile')  # загрузить/воссоздать
>>> table
{'A': [0], 'B': {'y': [0], 'x': 0}}
>>> table['A'][0] = 1            # изменить совместно используемый объект
>>> saveDbase('myfile', table)   # перезаписать в файл

C:\...\PP4E\Dbase> python
>>> from filepickle import *
>>> print(loadDbase('myfile'))   # изменились оба списка L, как и ожидалось
{'A': [1], 'B': {'y': [1], 'x': 0}}
```

Помимо встроенных типов, таких как списки, кортежи и словари, использовавшихся в примерах до сих пор, сериализовать можно также *экземпляры классов*. Тем самым обеспечивается естественный способ связать поведение с хранимыми данными (методы классов обрабатывают атрибуты экземпляров) и простой способ миграции (изменения в классе автоматически будут подхвачены хранимыми экземплярами). Ниже приводится короткая демонстрация в интерактивной оболочке:

```
>>> class Rec:
        def __init__(self, hours):
            self.hours = hours
        def pay(self, rate=50):
            return self.hours * rate

>>> bob = Rec(40)
>>> import pickle
>>> pickle.dump(bob, open('bobrec', 'wb'))
```

```
>>>
>>> rec = pickle.load(open('bobrec', 'rb'))
>>> rec.hours
40
>>> rec.pay()
2000
```

Принцип действия этого механизма мы подробно рассмотрим, когда будем исследовать хранилища, создаваемые модулем shelve, далее в этой главе – как мы увидим позже, модуль pickle может использоваться непосредственно, но он также является базовым механизмом баз данных shelve и ZODB.

В целом Python может сериализовать почти все, что угодно, за исключением:

- Объектов компилированного программного кода: функций и классов, когда при сериализации известны только их имена, без имен модулей, что не позволяет позднее повторно импортировать их и автоматически подхватить изменения в файлах модулей.
- Экземпляры классов, не выполняющие правила импортируемости: если говорить кратко, классы должны быть доступны для импортирования при загрузке объекта (подробнее об этом будет рассказываться ниже, в разделе «Файлы хранилищ shelve»).
- Экземпляров некоторых встроенных и определяемых пользователем типов, написанных на языке C или зависящих от преходящих состояний операционной системы (например, объекты открытых файлов не могут быть сериализованы).

Если объект не может быть сериализован, возбуждается исключение PicklingError. Напомню, что мы еще вернемся к проблеме сериализуемости объектов и классов, когда будем знакомиться с хранилищами, создаваемыми модулем shelve.

Особенности сериализации: протоколы, двоичные режимы и модуль _pickle

В последних версиях Python в операцию сериализации было введено понятие *протоколов* – форматов хранения сериализованных данных. Чтобы определить желаемый протокол, необходимо передать дополнительный параметр функциям сериализации (его не требуется передавать функциям обратного преобразования: протокол определяется автоматически по сериализованным данным):

```
pickle.dump(object, file, protocol) # или именованный аргумент protocol=N
```

Сериализация данных может быть выполнена с применением текстового или двоичного протокола – двоичный протокол позволяет получить более эффективный формат, но при этом создаются нечитаемые человеком файлы. По умолчанию в Python 3.X используется исключительно

двоичный формат (известный также, как протокол 3). В текстовом режиме (протокол 0) сериализованные данные представляют собой печатаемый текст ASCII, который может читаться человеком (по сути, он представляет собой последовательность инструкций для машины стека), но в Python 3.X в любом случае получается объект bytes. Другие протоколы (протоколы 1 и 2) также создают сериализованные данные в двоичном формате.

В Python 3.X независимо от номера протокола сериализованные данные представляют собой объект bytes, а не str, и именно поэтому при сохранении и чтении их в плоских файлах требуется использовать двоичный режим (причины описываются в главе 4, если вы забыли). Аналогично, имитируя интерфейс объектов файлов, мы должны использовать объекты bytes:

```
>>> import io, pickle
>>> pickle.dumps([1, 2, 3])                    # по умолчанию=двоич. протокол
b'\x80\x03]q\x00(K\x01K\x02K\x03e.'
>>> pickle.dumps([1, 2, 3], protocol=0)        # протокол формата ASCII
b'(lp0\nL1L\naL2L\naL3L\na.'

>>> pickle.dump([1, 2, 3], open('temp','wb')) # даже если protocol=0, ASCII
>>> pickle.dump([1, 2, 3], open('temp','w'))  # при чтении необх. режим 'rb'
TypeError: must be str, not bytes
>>> pickle.dump([1, 2, 3], open('temp','w'), protocol=0)
TypeError: must be str, not bytes

>>> B = io.BytesIO()                # использовать двоичные потоки/буферы
>>> pickle.dump([1, 2, 3], B)
>>> B.getvalue()
b'\x80\x03]q\x00(K\x01K\x02K\x03e.'

>>> B = io.BytesIO()                # bytes и для формата ASCII
>>> pickle.dump([1, 2, 3], B, protocol=0)
>>> B.getvalue()
b'(lp0\nL1L\naL2L\naL3L\na.'

>>> S = io.StringIO()               # это не объект str
>>> pickle.dump([1, 2, 3], S)       # даже если protocol=0, ASCII
TypeError: string argument expected, got 'bytes'
>>> pickle.dump([1, 2, 3], S, protocol=0)
TypeError: string argument expected, got 'bytes'
```

За дополнительными сведениями о сериализации обращайтесь к руководству по библиотеке Python – там вы найдете описание дополнительных интерфейсов, которые могут использоваться классами для переопределения поведения этой операции, которые мы не будем рассматривать здесь ради экономии места. Обратите также внимание на модуль marshal, который тоже сериализует объекты, но может обрабатывать только простые типы объектов. Модуль pickle является более универсальным, чем модуль marshal, и обычно более предпочтителен.

Имеется еще один родственный модуль, `_pickle`, написанная на языке C оптимизация модуля `pickle`, который автоматически используется модулем `pickle`, если доступен – его не требуется выбирать вручную или использовать непосредственно. Модуль `shelve` наследует эту оптимизацию автоматически. Я еще не рассказывал о модуле `shelve`, но сейчас сделаю это.

Файлы shelve

Сериализация позволяет сохранять произвольные объекты в файлах и подобных им объектах, но это все же весьма неструктурированный носитель – он не обеспечивает непосредственного простого доступа к членам совокупностей сериализованных объектов. Можно добавлять структуры более высокого уровня, но они не являются внутренне присущими:

- Порой имеется возможность определить собственную организацию файлов сериализации более высокого уровня с помощью базовой файловой системы (например, можно записать каждый сериализованный объект в файл, имя которого уникально идентифицирует объект), но такая организация не является частью механизма сериализации и должна управляться вручную.
- Можно также записывать словари произвольно большого размера в файл сериализации и обращаться к ним по ключу после загрузки обратно в память, но при этом обратное восстановление из файла загружает весь словарь, а не только тот элемент, в котором мы заинтересованы.

Хранилища, создаваемые модулем `shelve`, позволяют определить некоторую структуру для совокупностей сериализованных объектов. В файлах этого типа, являющихся стандартной частью системы Python, произвольные объекты Python сохраняются по ключу для извлечения в дальнейшем. В действительности здесь не так много нового – файлы-хранилища являются простой комбинацией файлов DBM и объектов сериализации:

- Чтобы *сохранить* находящийся в памяти объект по ключу, модуль `shelve` сначала сериализует его в строку с помощью модуля `pickle`, а затем записывает эту строку в файл DBM по ключу с помощью модуля `dbm`.
- Чтобы *загрузить* обратно объект по ключу, модуль `shelve` сначала загружает по ключу строку с сериализованным объектом из файла DBM с помощью модуля `dbm`, а затем преобразует ее обратно в исходный объект с помощью модуля `pickle`.

Поскольку внутренняя реализация модуля `shelve` использует модуль `pickle`, она может сохранять те же объекты, что и `pickle`: строки, числа, списки, словари, рекурсивные объекты, экземпляры классов и другие. Поскольку внутренняя реализация модуля `shelve` использует модуль `dbm`,

она наследует все черты этого модуля, в том числе и его ограничения, касающиеся переносимости.

Использование хранилищ

Иными словами, модуль shelve служит всего лишь посредником – он сериализует и десериализует объекты, чтобы их можно было поместить в файлы DBM. В конечном итоге хранилища позволяют записывать в файлы по ключу почти любые объекты Python и позднее загружать их обратно по тому же ключу.

Однако сами сценарии никогда не видят всех этих взаимодействий. Подобно файлам DBM хранилища предоставляют интерфейс, напоминающий словарь, который нужно открыть. Фактически хранилища – это всего лишь постоянно хранимые словари с постоянно хранимыми объектами Python: содержимое словарей-хранилищ отображается в файлы на компьютере, благодаря чему они сохраняются между запусками программы. На словах все это звучит довольно сложно, но в программном коде оказывается просто. Чтобы получить доступ к хранилищу, импортируйте модуль и откройте свой файл:

```
import shelve
dbase = shelve.open("mydbase")
```

Модуль shelve откроет файл DBM с именем *mydbase* или создаст его, если он еще не существует (по умолчанию он использует режим 'c' открытия файлов DBM). Операция присваивания по ключу сохраняет объект:

```
dbase['key'] = object  # сохранить объект
```

Внутренне эта операция преобразует объект в сериализованный поток байтов и запишет его по ключу в файл DBM. Обращение к хранилищу по ключу загружает сохраненный объект:

```
value = dbase['key']   # извлечь объект
```

Внутренне эта операция обращения по индексу загрузит по ключу строку из файла DBM и развернет ее в объект в памяти, совпадающий с исходным объектом. Здесь также поддерживается большинство операций со словарями:

```
len(dbase)             # количество хранящихся элементов
dbase.keys()           # список ключей хранящихся элементов
```

И, за исключением некоторых тонких моментов, это все, что касается использования модуля shelve. Хранилища объектов обрабатываются с использованием обычного синтаксиса словарей Python, поэтому не нужно изучать новый интерфейс базы данных. Более того, объекты, сохраняемые в хранилищах и извлекаемые из них, являются обычными объектами Python. Для сохранения не требуется, чтобы они были экземплярами особых классов или типов. То есть система постоянного

хранения объектов Python является внешней по отношению к самим сохраняемым объектам. В табл. 17.2 сведены вместе эти и другие часто используемые операции с хранилищами.

Таблица 17.2. Операции с файлами хранилищ

Программный код Python	Действие	Описание
`import shelve`	Импорт	Получить интерфейс `bsddb`, `gdbm` и так далее в зависимости от того, что установлено
`file=shelve.open('filename')`	Открытие	Создать или открыть существующий файл DBM хранилища
`file['key'] = anyvalue`	Запись	Создать или изменить запись с ключом `key`
`value = file['key']`	Выборка	Загрузить значение из записи с ключом `key`
`count = len(file)`	Размер	Получить количество записей
`index = file.keys()`	Индекс	Получить список ключей (итерируемое представление)
`found = 'key' in file`	Запрос	Проверить наличие записи с ключом `key`
`del file['key']`	Удаление	Удалить запись с ключом `key`
`for key in file:`	Итерации	Выполнить итерации по имеющимся ключам
`file.close()`	Закрытие	Закрыть вручную, требуется не всегда

Так как хранилища тоже экспортируют интерфейс, подобный словарю, эта таблица почти идентична таблице операций с файлами DBM. Однако здесь имя модуля `dbm` заменяется на `shelve`, вызовы `open` не требуют второго аргумента `'c'`, а сохраняемые значения могут быть объектами практически любых типов, а не просто строками. Однако ключами все еще могут быть только строки (технически ключами могут быть только объекты `str`, которые автоматически преобразуются в тип `bytes` и обратно с применением кодировки UTF-8), и для надежности по-прежнему необходимо явно закрывать хранилища после проведенных изменений: хранилища внутренне используют модуль `dbm`, а некоторые базовые модули поддержки DBM требуют выполнять закрытие, чтобы избежать потери или повреждения данных.

Последние изменения: Функции открытия хранилища в модуле `shelve` теперь принимают необязательный аргумент `writeback` – если в нем передать значение `True`, все записи будут кэшироваться в памяти и записываться обратно на диск только при выполнении опе-

рации закрытия. Это устраняет необходимость вручную повторно присваивать модифицированные изменяемые объекты, чтобы вытолкнуть их на диск, но может приводить к непроизводительным потерям при большом количестве записей – кэш может занять огромный объем памяти, и операция закрытия в таких случаях будет выполняться достаточно медленно, так как необходимо будет записать в файл все извлеченные записи (интерпретатор не способен определить, какие из записей изменялись).

Помимо того что значениями могут быть не только простые строки, но и любые объекты, интерфейс хранилищ в Python 3.X имеет еще два тонких отличия от интерфейса файлов DBM. Во-первых, метод `keys` возвращает итерируемый объект *представления* (а не физический список). Во-вторых, значениями *ключей* всегда являются строки типа `str`, а не `bytes` – при извлечении, записи, удалении и в других операциях используемые ключи типа `str` кодируются в строки `bytes`, которые ожидает получить реализация DBM, с использованием кодировки UTF-8. Это означает, что в отличие от модуля `dbm`, нельзя в качестве ключей хранилища `shelve` использовать строки `bytes`, чтобы использовать произвольные кодировки.

Кроме того, ключи хранилищ декодируются из типа `bytes` в тип `str` с помощью кодировки UTF-8 всякий раз, когда они возвращаются функциями модуля `shelve` (например, при выполнении итераций по ключам). Хранимые *значения* всегда представлены объектами `bytes`, создаваемыми модулем `pickle` при сериализации объектов. Мы увидим эти особенности в действии, далее в этом разделе.

Сохранение объектов встроенных типов в хранилищах

Запустим интерактивный сеанс и поэкспериментируем с интерфейсами хранилищ. Как уже отмечалось, хранилища, по сути, являются постоянно хранимыми словарями объектов, которые необходимо открывать и закрывать:

```
C:\...\PP4E\Dbase> python
>>> import shelve
>>> dbase = shelve.open("mydbase")
>>> object1 = ['The', 'bright', ('side', 'of'), ['life']]
>>> object2 = {'name': 'Brian', 'age': 33, 'motto': object1}

>>> dbase['brian'] = object2
>>> dbase['knight'] = {'name': 'Knight', 'motto': 'Ni!'}
>>> dbase.close()
```

Здесь мы открываем хранилище и сохраняем две довольно сложных структуры данных в виде словаря и списка, просто присваивая их ключам хранилища. Поскольку модуль `shelve` внутренне использует модуль `pickle`, здесь можно использовать почти все – деревья вложенных объектов автоматически сериализуются в строки для хранения. Чтобы загрузить их обратно, достаточно просто открыть хранилище и обратиться к индексу:

```
C:\...\PP4E\Dbase> python
>>> import shelve
>>> dbase = shelve.open("mydbase")
>>> len(dbase)                          # количество записей
2

>>> dbase.keys()                        # индекс
KeysView(<shelve.DbfilenameShelf object at 0x0181F630>)

>>> list(dbase.keys())
['brian', 'knight']

>>> dbase['knight']                     # извлечь
{'motto': 'Ni!', 'name': 'Knight'}

>>> for row in dbase.keys():            # использовать метод .keys() необяз.
...     print(row, '=>')
...     for field in dbase[row].keys():
...         print(' ', field, '=', dbase[row][field])
...
brian =>
   motto = ['The', 'bright', ('side', 'of'), ['life']]
   age = 33
   name = Brian
knight =>
   motto = Ni!
   name = Knight
```

Вложенные друг в друга циклы в конце этого сеанса выполняют обход вложенных словарей – внешний просматривает хранилище, а внутренний – объекты, хранящиеся в нем (в обоих можно было бы использовать итераторы по ключам и опустить вызовы `.keys()`). Важно отметить, что для записи и выборки этих постоянных объектов, так же как для их обработки после загрузки, используется обычный синтаксис Python. Это данные Python, постоянно хранящиеся на диске.

Сохранение экземпляров классов в хранилищах

Более полезным типом объектов, которые можно хранить в хранилищах, являются экземпляры классов. Поскольку в их атрибутах записывается состояние, а унаследованные методы определяют поведение, постоянно хранимые объекты классов, в сущности, выполняют роль как *записей* базы данных, так и *программ* обработки баз данных. Для сериализации и сохранения экземпляров классов в плоских файлах и в других подобным им объектах (например, в сетевых сокетах) можно также использовать базовый модуль `pickle`, но высокоуровневый модуль `shelve` обеспечивает также возможность использовать доступ к хранилищу по ключу. Рассмотрим простой класс, представленный в примере 17.2, с помощью которого моделируются записи с информацией о сотрудниках гипотетического предприятия.

Пример 17.2. PP4E\Dbase\person.py (версия 1)

```
"объект с информацией о сотруднике: поля + поведение"

class Person:
    def __init__(self, name, job, pay=0):
        self.name = name
        self.job  = job
        self.pay  = pay            # действительные данные экземпляра

    def tax(self):
        return self.pay * 0.25     # вычисляется при вызове

    def info(self):
        return self.name, self.job, self.pay, self.tax()
```

Ничто в этом классе не говорит о том, что его экземпляры будут использоваться в качестве записей в базе данных – его можно импортировать и использовать независимо от внешнего хранилища. Однако его удобно использовать для создания записей в базе данных: из этого класса можно создавать постоянно хранимые объекты, просто создавая экземпляры как обычно и сохраняя их по ключу в открытом хранилище shelve:

```
C:\...\PP4E\Dbase> python
>>> from person import Person
>>> bob = Person('bob', 'psychologist', 70000)
>>> emily = Person('emily', 'teacher', 40000)
>>>
>>> import shelve
>>> dbase = shelve.open('cast')  # создать новое хранилище
>>> for obj in (bob, emily):   # сохранить объекты
...     dbase[obj.name] = obj  # использовать значение name в качестве ключа
...
>>> dbase.close()              # необходимо для bsddb
```

Здесь в качестве ключей в базе данных мы использовали атрибуты `name` экземпляров объектов. Когда позднее мы снова загрузим эти объекты в интерактивном сеансе Python или сценарии, они воссоздадутся в памяти в том виде, какой у них был в момент сохранения:

```
C:\...\PP4E\Dbase> python
>>> import shelve
>>> dbase = shelve.open('cast')  # открыть хранилище
>>>
>>> list(dbase.keys())           # в хранилище присутствуют оба объекта
['bob', 'emily']
>>> print(dbase['emily'])
<person.Person object at 0x0197EF70>
>>>
>>> print(dbase['bob'].tax())    # вызов: метода tax объекта с именем bob
17500.0
```

Обратите внимание, что вызов метода `tax` для объекта с именем «Bob» работает, несмотря на то, что мы не импортировали класс `Person`. Python достаточно сообразителен, чтобы снова связать объект с исходным классом после восстановления из сериализованной формы и сделать доступными все методы загруженных объектов.

Изменение классов хранимых объектов

Технически Python повторно импортирует класс для воссоздания его сохраненных экземпляров при их загрузке и восстановлении. Ниже описано, как это действует:

Запись

Когда Python сериализует экземпляр класса, чтобы сохранить его в хранилище, он сохраняет атрибуты и ссылку на класс экземпляра. Фактически сериализованные экземпляры класса в предыдущем примере записывают атрибуты `self`, присваивание которым выполняется в классе. В действительности Python сериализует и записывает словарь атрибутов `__dict__` экземпляра вместе с информацией об исходном файле модуля класса, чтобы позднее иметь возможность отыскать модуль класса – имена класса экземпляров и модуля, вмещающего этот класс.

Выборка

Когда Python восстанавливает экземпляр класса, извлеченный из хранилища, он воссоздает в памяти объект экземпляра, повторно импортируя класс, используя сохраненные строки с именами класса и модуля; присваивает сохраненный словарь атрибутов новому пустому экземпляру класса и связывает экземпляр с классом. Эти действия выполняются по умолчанию, но имеется возможность изменить этот процесс, определив специальные методы, которые будут вызываться модулем `pickle` при извлечении и сохранении информации об экземпляре (за подробностями обращайтесь к руководству по библиотеке Python).

Главное здесь то, что класс и хранимые экземпляры отделены друг от друга. Сам класс не хранится вместе со своими экземплярами, а находится в исходном файле модуля Python и импортируется заново, когда загружаются экземпляры.

Недостаток этой модели заключается в том, что класс должен быть доступен для импортирования, чтобы обеспечить возможность загрузки экземпляров из хранилища (подробнее об этом чуть ниже). А преимущество в том, что, модифицировав внешние классы в файлах модулей, можно изменить способ, которым интерпретируются и используются данные сохраненных объектов, не изменяя сами хранящиеся объекты. Это похоже на то, как если бы класс был программой, обрабатывающей хранящиеся записи.

Для иллюстрации предположим, что класс Person из предыдущего раздела был изменен, как показано в примере 17.3.

Пример 17.3. PP4E\Dbase\person.py (версия 2)

```
"""
объект с информацией о сотруднике: поля + поведение
изменения: метод tax теперь является вычисляемым атрибутом
"""


class Person:
    def __init__(self, name, job, pay=0):
        self.name = name
        self.job = job
        self.pay = pay              # действительные данные экземпляра

    def __getattr__(self, attr):    # в person.attr
        if attr == 'tax':
            return self.pay * 0.30  # вычисляется при попытке обращения
        else:
            raise AttributeError()  # другие неизвестные атрибуты

    def info(self):
        return self.name, self.job, self.pay, self.tax
```

В этой версии устанавливается новая ставка налога (30%), вводится метод __getattr__ перегрузки операции доступа к атрибуту класса и убран оригинальный метод tax. Поскольку при загрузке экземпляров из хранилища будет импортирована эта новая версия класса, они автоматически приобретут новую реализацию поведения – попытки обращения к атрибуту tax будут перехватываться и в ответ будет возвращаться вычисленное значение:

```
C:\...\PP4E\Dbase> python
>>> import shelve
>>> dbase = shelve.open('cast')     # открыть хранилище
>>>
>>> print(list(dbase.keys()))       # в хранилище присутствуют оба объекта
['bob', 'emily']
>>> print(dbase['emily'])
<person.Person object at 0x019AEE90>
>>>
>>> print(dbase['bob'].tax)         # больше не требуется вызывать tax()
21000.0
```

Так как класс изменился, к атрибуту tax теперь можно обращаться как к простому свойству, не вызывая его как метод. Кроме того, поскольку ставка налога в классе изменена, Бобу придется на этот раз платить больше. Конечно, этот пример искусственный, но при правильном использовании такое разделение классов и постоянно хранимых экземпляров может избавить от необходимости использовать традиционные программы обновления баз данных – чтобы добиться нового поведения,

в большинстве случаев можно просто изменить класс, а не каждый хранящийся экземпляр.

Ограничения модуля shelve

Обычно работа с хранилищами не вызывает затруднений, однако существуют некоторые шероховатости, о которых следует помнить.

Ключи должны быть строками (str)

Во-первых, несмотря на то, что сохранять можно любые объекты, ключи все же должны быть строками. Следующая инструкция не будет выполнена, если сначала вручную не преобразовать целое число 42 в строку 42:

```
dbase[42] = value      # ошибка, но str(42) работает
```

Этим хранилища отличаются от словарей, размещаемых в памяти, которые допускают использование в качестве ключей любых неизменяемых объектов, и обусловлено использованием файлов DBM. Как мы уже видели, в Python 3.X ключи могут быть только строками str, а не bytes, потому что внутренняя реализация хранилищ во всех случаях пытается их кодировать.

Объекты уникальны только по ключу

Хотя модуль shelve достаточно сообразителен, чтобы обнаруживать множественные случаи вложенного объекта и воссоздавать только один экземпляр при загрузке, это относится только к данной ситуации:

```
dbase[key] = [object, object]  # OK: сохраняется и извлекается
                               # только одна копия

dbase[key1] = object
dbase[key2] = object           # плохо?: две копии объекта в хранилище
```

После извлечения объектов по ключам key1 и key2 они будут указывать на независимые копии оригинального общего объекта – если этот объект относится к категории изменяемых, изменения в одном из них не будут отражены в другом. В действительности это обусловлено тем, что каждое присвоение ключу запускает независимую операцию сериализации – механизм сериализации обнаруживает повторяющиеся объекты, но только в рамках одного обращения к модулю pickle. Это может не коснуться вас на практике и преодолевается введением дополнительной логики, но следует помнить, что объект может дублироваться, если он соответствует нескольким ключам.

Обновления должны выполняться в режиме «загрузить–модифицировать–сохранить»

Так как объекты, загруженные из хранилища, не знают, что они были извлечены оттуда, то операции, изменяющие части загруженного объ-

екта, касаются только экземпляра, находящегося в памяти, а не данных в хранилище:

```
dbase[key].attr = value        # данные в хранилище не изменились
```

Чтобы действительно изменить объект в хранилище, нужно загрузить его в память, изменить и записать обратно в хранилище целиком, выполнив присваивание по ключу:

```
object = dbase[key]  # загрузить
object.attr = value  # модифицировать
dbase[key] = object  # записать обратно - хранилище изменилось
                     # (если при открытии не был указан аргумент writeback)
```

Как отмечалось выше, если методу shelve.open передать необязательный аргумент writeback, то выполнение последнего шага здесь не потребуется, – за счет автоматического кэширования извлеченных объектов и сохранения их на диске при закрытии хранилища. Но это может повлечь за собой существенный расход памяти и замедлить операцию закрытия.

Возможность одновременных обновлений не поддерживается

В настоящее время модуль shelve не поддерживает возможность одновременного обновления. Одновременное чтение допускается, но для записи программа должна получить исключительный доступ к хранилищу. Хранилище можно разрушить, если несколько процессов будут выполнять запись в него одновременно, а это, например, в серверных сценариях может происходить часто. Если изменять данные в хранилищах может потребоваться сразу нескольким процессам, оберните операции обновления данных в вызов функции os.open из стандартной библиотеки, чтобы заблокировать доступ к файлу и обеспечить исключительный доступ.

Переносимость базового формата DBM

Для сохранения объектов модуль shelve создает файлы с помощью базовой системы DBM, которые необязательно будут совместимы со всеми возможными реализациями DBM или версиями Python. Например, файл, созданный с помощью gdbm в Linux или библиотекой bsddb в Windows, может не читаться при использовании Python, установленного с другими модулями DBM.

Это все та же проблема переносимости, которую мы рассматривали при обсуждении файлов DBM выше. Как вы помните, когда создается файл DBM (а соответственно и файл хранилища модуля shelve), модуль dbm пытается импортировать все возможные модули системы DBM в предопределенном порядке и использует первый найденный модуль. Когда позднее модуль dbm открывает существующий файл, он пытается определить по содержимому файла, какая система DBM использовалась при его создании. При создании файла сначала делается попытка использо-

вать систему bsddb, доступную в Windows и во многих Unix-подобных системах, поэтому ваш файл DBM будет совместим со всеми платформами, где установлена версия Python с поддержкой BSD. То же относится к платформам, где установлена версия Python, использующая собственную реализацию dbm.dumb при отсутствии поддержки других форматов DBM. Однако если для создания файла DBM использовалась система, недоступная на целевой платформе, использовать этот файл будет невозможно.

Если обеспечение переносимости файлов DBM имеет большое значение для вас, примите меры, чтобы все версии Python, под управлением которых будет выполняться чтение ваших данных, использовали совместимые модули DBM. Если это невозможно, используйте для сохранения данных модуль pickle и плоские файлы (в обход модулей shelve и dbm) или одну из объектно-ориентированных баз данных, с которыми мы познакомимся далее в этой главе. Такие базы данных зачастую способны предложить полную поддержку *транзакций* в виде методов подтверждения изменений и автоматической отмены в случае ошибки.

Ограничения класса Pickler

Помимо указанных ограничений хранилищ модуля shelve, сохранение экземпляров классов в таких хранилищах вводит ряд правил, о которых необходимо знать. В действительности они налагаются модулем pickle, а не shelve, поэтому следуйте им, даже когда будете сохранять экземпляры классов непосредственно с помощью модуля pickle.

Классы должны быть доступны для импортирования

Объект класса Pickler при сериализации объекта экземпляра сохраняет только атрибуты экземпляра и затем заново импортирует класс, чтобы воссоздать экземпляр. По этой причине, когда объекты восстанавливаются из сериализованной формы, должна обеспечиваться возможность импортировать классы сохраненных объектов – они должны быть определены как невложенные, на верхнем уровне файла модуля, который должен находиться в пути поиска модулей в момент загрузки (например, в PYTHONPATH, или в файле *.pth*, или в текущем рабочем каталоге, или быть самим сценарием верхнего уровня).

Далее, при сериализации экземпляров классы должны ассоциироваться с действительным импортируемым модулем, а не со сценарием верхнего уровня (с именем модуля __main__), если только они не будут использоваться в сценарии верхнего уровня. Кроме того, нужно следить за тем, чтобы модули классов не перемещались после сохранения экземпляров. При восстановлении экземпляра Python должен иметь возможность найти модуль класса в пути поиска модулей по имени исходного модуля (включая префиксы путей в пакетах) и загрузить класс из этого модуля, используя первоначальное имя класса. Если модуль или класс будут перемещены или переименованы, интерпретатор не сможет отыскать класс.

В приложениях, обменивающихся сериализованными объектами через сокеты по сети, это ограничение можно удовлетворить, передавая определение класса вместе с хранимыми экземплярами – до того как извлечь сериализованные объекты, получатели могут просто сохранять класс в локальном файле модуля, находящегося в пути поиска. Где это неудобно или невозможно, вместо экземпляров классов можно передавать более простые сериализованные данные в виде вложенных друг в друга списков и словарей, так как для их воссоздания не требуется иметь исходный файл модуля с определением класса.

Изменения в классе должны обеспечивать обратную совместимость

Хотя Python позволяет изменить класс в то время, когда его экземпляры хранятся в хранилище, эти изменения должны быть обратно совместимыми с уже сохраненными объектами. Например, нельзя изменить класс так, чтобы он ждал атрибута, отсутствующего в уже хранящихся постоянных экземплярах, если только не изменить предварительно эти хранящиеся экземпляры или не реализовать дополнительные протоколы преобразования для класса.

Другие ограничения модуля pickle

Хранилища, создаваемые модулем `shelve`, также наследуют некоторые ограничения механизма сериализации, не относящиеся к классам. Как уже говорилось выше, некоторые виды объектов (например, открытые файлы и сокеты) не могут быть сериализованы и потому не могут быть записаны в хранилище.

В ранних версиях Python классы постоянно хранимых объектов также должны были иметь конструкторы без аргументов или обеспечивать значения по умолчанию для всех аргументов конструктора (подобно понятию конструктора копирования в C++). Это ограничение было снято в Python 1.5.2 – классы, конструкторы которых имеют аргументы без значений по умолчанию, теперь нормально обрабатываются при сериализации.[1]

---

[1] Интересно отметить: интерпретатор теперь не вызывает класс для воссоздания сериализованного экземпляра, а вместо этого создает обобщенный объект класса, вставляет атрибуты экземпляра и прямо устанавливает указатель `__class__` экземпляра, чтобы тот указывал на исходный класс. При этом не требуются значения по умолчанию, но это также означает, что конструкторы класса `__init__` больше не вызываются при восстановлении объектов, если не пользоваться дополнительными методами для принудительного вызова. Смотрите дополнительные подробности в руководстве по библиотеке, а также исходный программный код модуля `pickle` (*pickle.py* в библиотеке), если вас интересует, как это работает. Еще лучше, загляните в модуль `PyForm`, представленный далее в этой главе, – он делает нечто очень сходное со ссылками `__class__`, чтобы создать объект экземпляра из класса и словаря атрибутов, не вызывая конструктор `__init__` класса. В результате в классах для записей, которые просматривает `PyForm`, значения по умолчанию аргументов конструктора становятся необязательными, но идея та же самая.

Другие ограничения хранилищ модуля shelve

Наконец, несмотря на то, что хранилища позволяют сохранять объекты, они не являются в действительности объектно-ориентированными базами данных. В таких базах данных реализуются также такие функции, как немедленная автоматическая запись изменений, подтверждение и отмена транзакций, возможность безопасных одновременных изменений, декомпозиция объектов и отложенная выборка компонентов, основанные на генерируемых идентификаторах объектов. Части больших объектов загружаются в память только во время доступа к ним. Подобные функции можно реализовать и в хранилищах, но в этом нет никакой необходимости – система ZODB, среди прочих, предоставляет реализацию более полной системы объектно-ориентированной базы данных. Она построена поверх поддержки сериализации, встроенной в Python, но предлагает дополнительные функции для более развитых хранилищ данных. Дополнительная информация о ZODB приводится в следующем разделе.

Объектно-ориентированная база данных ZODB

ZODB, Zope Object Database, – это полнофункциональная, объектно-ориентированная база данных (OODB) специально для Python. ZODB можно рассматривать как более мощную альтернативу хранилищам модуля `shelve`, описываемым в предыдущем разделе. Она позволяет сохранять по ключу практически любые объекты Python подобно хранилищам `shelve`, но предлагает при этом ряд дополнительных особенностей в обмен на небольшое усложнение взаимодействия с ней.

ZODB – не единственная объектно-ориентированная база данных, доступная для программ на языке Python: система Durus в целом выглядит проще и во многом подобна ZODB. Однако, несмотря на некоторые преимущества, в настоящее время Durus не предлагает всех особенностей, имеющихся в ZODB, и используется не так широко (возможно просто потому, что является более новой). По этой причине концепции объектно-ориентированных баз данных будут рассматриваться в этом разделе на примере ZODB.

ZODB – это открытое, стороннее расширение для Python. Первоначально этот продукт разрабатывался как механизм базы данных для веб-сайтов в составе веб-фреймворка Zope, упоминавшегося в главе 12, но теперь он доступен в виде самостоятельного пакета. Его можно использовать во многих областях вне контекста Zope и Веб как универсальную систему управления базами данных.

ZODB не поддерживает язык запросов SQL, однако объекты, хранимые в этой базе данных, могут использовать всю мощь языка Python. Кроме того, в некоторых приложениях хранимые данные более естественно представлять в виде структурированных объектов Python. Реляцион-

ные системы, основанные на таблицах, часто вынуждены представлять такие данные в виде отдельных фрагментов, разбросанных по нескольким таблицам и связанных сложными, порой весьма медлительными в обработке соединениями по ключу, или как-то иначе отображать их в модель классов языка Python. Поскольку объектно-ориентированные базы данных сохраняют объекты Python непосредственно, они часто способны обеспечить более простую модель в виде систем, не требующих использования мощного языка SQL.

Принцип работы с базой данных ZODB очень близко напоминает работу с хранилищами, создаваемыми стандартным модулем `shelve`, описанными в предыдущем разделе. Так же как и модуль `shelve`, для реализации постоянно хранимых словарей с хранимыми объектами Python ZODB использует систему сериализации Python. Фактически база данных почти не имеет специализированного интерфейса – объекты сохраняются в ней за счет простого присваивания по ключу объекту корневого словаря ZODB или за счет встраивания их в объекты, хранящиеся в корне базы данных. И, как и в случае с модулем `shelve`, «записи» имеют вид обычных объектов Python, обрабатываемых с помощью обычного синтаксиса и инструментов Python.

Однако, в отличие от модуля `shelve`, ZODB добавляет особенности, имеющие большое значение для некоторых программ:

Возможность одновременного обновления

Вам не придется вручную блокировать доступ к файлам, чтобы избежать повреждения данных при наличии нескольких пишущих процессов, что необходимо при использовании модуля `shelve`.

Подтверждение и отмена транзакций

При аварийном завершении программы изменения не будут сохранены в базе данных, если они не были явно подтверждены.

Автоматическое сохранение изменений для некоторых типов объектов в памяти

Объекты в ZODB, порожденные от суперкласса постоянно хранимых объектов, достаточно сообразительны, чтобы знать, когда можно обновить содержимое в базе данных при присваивании значений их атрибутам.

Автоматическое кэширование объектов

Для большей эффективности объекты кэшируются в памяти и автоматически удаляются из кэша, когда надобность в них отпадает.

Платформонезависимое хранилище

Все данные ZODB хранит в одном плоском файле и поддерживает файлы большого размера, поэтому она не подвержена ограничениям на размер файлов и проблемам из-за различий в форматах DBM, свойственных модулю `shelve`. Как мы видели ранее в этой главе, хранилище, созданное в Windows с использованием `bsddb`, может ока-

заться недоступным для сценариев, выполняющихся в Linux и использующих gdbm.

Благодаря этим преимуществам ZODB определенно заслуживает внимания, если вам потребуется постоянно сохранять объекты Python в базе данных при промышленной эксплуатации. Единственное, чем придется заплатить за использование ZODB, – это небольшой объем дополнительного программного кода:

- Для организации доступа к базе данных ZODB необходим некоторый объем типового программного кода – это не просто вызов функции открытия.
- Классы должны наследовать суперкласс постоянно хранимых объектов, если вы пожелаете воспользоваться преимуществом автоматического обновления при их изменении – классы постоянно хранимых объектов в целом не являются полностью независимыми от базы данных, как и при использовании модуля shelve, хотя это в принципе возможно.

Во многих приложениях эти недостатки с лихвой окупаются дополнительными возможностями, предоставляемыми системой ZODB сверх того, что дает модуль shelve.

Сильно сокращенный учебник по ZODB

К сожалению, когда я писал эти строки, в июне 2010 года, еще отсутствовала версия ZODB для Python 3.X. Вследствие этого примеры и описание версии для Python 2.X, присутствовавшие в предыдущем издании, были убраны из этого раздела. Однако из уважения к пользователям Python 2.X, а также к читателям, использующим Python 3.X и ждущим, когда материализуется версия ZODB 3.X, я добавил материалы и примеры из прошлого издания, касающиеся ZODB, в пакет примеров для этого издания.

Дополнительную информацию о пакете с примерами вы найдете в предисловии, и ищите информацию о ZODB в следующих каталогах:

```
C:\...\Dbase\Zodb-2.x              # примеры использования ZODB из 3 издания
C:\...\Dbase\Zodb-2.x\Documentaion # Учебник по ZODB из 3 издания
```

Хотя я не умею предсказывать будущее, тем не менее ZODB для Python 3.X наверняка появится рано или поздно. Но пока ее нет, можно использовать другие объектно-ориентированные базы данных для Python 3.X, предлагающие некоторые дополнительные возможности.

Однако, чтобы дать вам некоторое представление о ZODB, мы кратко познакомимся с особенностями ее использования в Python 2.X. После установки поддержки ZODB в первую очередь необходимо создать базу данных:

```
...\PP4E\Dbase\Zodb-2.x> python
>>> from ZODB import FileStorage, DB
```

```
>>> storage = FileStorage.FileStorage(r'C:\temp\mydb.fs')
>>> db = DB(storage)
>>> connection = db.open()
>>> root = connection.root()
```

Это практически стандартный, «типовой» программный код, выполняющий подключение к базе данных ZODB: здесь импортируются необходимые инструменты, создаются объекты классов `FileStorage` и `DB`, а затем открывается база данных и создается *корневой объект*. Корневым объектом является постоянно хранимый словарь, в котором сохраняются другие объекты. Объект класса `FileStorage` отображает базу данных в плоский файл. Имеются также интерфейсы к другим типам хранилищ, таким как хранилища на основе реляционной базы данных.

Добавление объектов в базу данных ZODB выполняется так же просто, как при использовании модуля `shelve`. Поддерживаются почти любые объекты Python, включая кортежи, списки, словари, экземпляры классов и их комбинации. Как при использовании модуля `shelve`, чтобы сделать объект постоянно хранимым, достаточно просто присвоить его по ключу корневому объекту базы данных:

```
>>> object1 = (1, 'spam', 4, 'YOU')
>>> object2 = [[1, 2, 3], [4, 5, 6], [7, 8, 9]]
>>> object3 = {'name': ['Bob', 'Doe'],
               'age': 42,
               'job': ('dev', 'mgr')}

>>> root['mystr'] = 'spam' * 3
>>> root['mytuple'] = object1
>>> root['mylist'] = object2
>>> root['mydict'] = object3

>>> root['mylist']
[[1, 2, 3], [4, 5, 6], [7, 8, 9]]
```

Поскольку ZODB поддерживает возможность отмены транзакций, чтобы сохранить изменения в базе данных, их необходимо подтвердить. В конечном итоге объекты в сериализованном представлении сохраняются в файлы – здесь будут созданы три файла, включая файл с именем, которое было указано при открытии:

```
>>> import transaction
>>> transaction.commit()
>>> storage.close()

...\PP4E\Dbase\Zodb-2.x> dir /B c:\temp\mydb*
mydb.fs
mydb.fs.index
mydb.fs.tmp
```

Без заключительного подтверждения в этом сеансе ни одно из изменений не сохранилось бы. Именно такое поведение чаще всего бывает желательным – если программа завершится аварийно в середине проце-

дуры, выполняющей изменения, уже выполненные к этому моменту частичные изменения не сохранятся. Фактически ZODB поддерживает обычную для баз данных операцию отмены.

Извлечение хранимых объектов из базы данных ZODB в другом сеансе или программе выполняется так же просто: нужно открыть базу данных, как было показано выше, и обратиться по индексам к корневому объекту, чтобы извлечь объекты в память. Подобно хранилищам, создаваемым модулем shelve, корневой объект базы данных поддерживает интерфейс словарей – к нему можно обращаться с указанием индексов, использовать методы словарей, определять количество записей и так далее:

```
...\PP4E\Dbase\Zodb-2.x> python
>>> from ZODB import FileStorage, DB
>>> storage = FileStorage.FileStorage(r'C:\temp\mydb.fs')
>>> db = DB(storage)
>>> connection = db.open()
>>> root = connection.root()                    # соединиться

>>> len(root), root.keys()                      # размер, индекс
(4 ['mylist', 'mystr', 'mytuple', 'mydict'])

>>> root['mylist']                              # извлечь объекты
[[1, 2, 3], [4, 5, 6], [7, 8, 9]]
>>> root['mydict']
{'job': ('dev', 'mgr'), 'age': 42, 'name': ['Bob', 'Doe']}

>>> root['mydict']['name'][-1]                  # Фамилия Боба
'Doe'
```

Так как корневой объект базы данных выглядит как словарь, с ним можно работать как с обычным словарем, например выполнить итерации по списку ключей, чтобы обойти все записи:

```
>>> for key in root.keys():
        print('%s => %s' % (key.ljust(10), root[key]))

mylist     => [[1, 2, 3], [4, 5, 6], [7, 8, 9]]
mystr      => spamspamspam
mytuple    => (1, 'spam', 4, 'YOU')
mydict     => {'job': ('dev', 'mgr'), 'age': 42, 'name': ['Bob', 'Doe']}
```

Кроме того, подобно модулям pickle и shelve ZODB поддерживает возможность сохранения и извлечения экземпляров классов, правда при этом они должны наследовать суперкласс Persistent, реализующий необходимый протокол и перехватывающий операции изменения атрибутов, чтобы обеспечить автоматическое сохранение их на диск:

```
from persistent import Persistent
class Person(Persistent):
    def __init__(self, name, job=None, rate=0):
```

```
        self.name = name
        self.job = job
        self.rate = rate
    def changeRate(self, newrate):
        self.rate = newrate    # автоматически обновит базу данных
```

При изменении экземпляров классов, хранимых в ZODB, операции изменения атрибутов в памяти приводят к автоматическому сохранению изменений в базе данных. Другие типы операций изменения, такие как добавление элементов в список и присваивание словарям по ключу, все еще требуют повторно выполнять присваивание по оригинальному ключу корневому объекту, чтобы обеспечить принудительное сохранение изменений на диск (встроенные словари и списки не знают, что они являются постоянно хранимыми).

Поскольку пока еще не вышла версия ZODB для Python 3.X, это все, что можно сказать об этой базе данных в этой книге. За дополнительной информацией обращайтесь к ресурсам Интернета, посвященным проектам ZODB и Zope, а также ознакомьтесь с ресурсами в пакете с примерами, о которых говорилось выше. А теперь посмотрим, какие инструменты доступны программам на языке Python для работы с совершенно иной разновидностью баз данных – с реляционными базами данных и SQL.

Интерфейсы баз данных SQL

Модуль `shelve` и пакет ZODB, представленные в предыдущих разделах, являются мощными инструментами. Они дают возможность сценариям перемещать объекты Python в файлы с доступом по ключу и затем загружать их обратно – за один шаг при использовании модуля `shelve` и с помощью небольшого количества административного программного кода при использовании ZODB. Для приложений, где требуется сохранять высокоструктурированные данные, объектно-ориентированные базы данных являются весьма удобным и эффективным механизмом – они не требуют разбивать, а потом соединять части крупных объектов и позволяют хранить и обрабатывать объекты с применением привычного синтаксиса языка Python, поскольку они являются обычными объектами Python.

Тем не менее ни `shelve`, ни ZODB не являются полновесными системами баз данных – доступ к объектам (записям) происходит по единственному ключу и нет понятия запросов SQL. Хранилища, создаваемые модулем `shelve`, по сути представляют собой как бы базу данных с единственным индексом и отсутствием поддержки обработки других запросов. В принципе можно построить интерфейс с множественными индексами, чтобы хранить данные в нескольких хранилищах `shelve`, но это нетривиальная задача, требующая написания расширений вручную.

ZODB поддерживает некоторые виды поиска, не поддерживаемые модулем `shelve` (например, возможность каталогизации), а обход хранимых

объектов может выполняться с использованием всей мощи языка Python. Однако ни `shelve`, ни объектно-ориентированная база данных ZODB не предоставляют обобщенность запросов SQL. Кроме того, иногда, особенно для хранения данных, имеющих табличную организацию, реляционные базы данных подходят гораздо лучше.

Для программ, которые могут получить дополнительные преимущества от использования SQL, в языке Python имеется поддержка систем управления реляционными базами данных (СУРБД). Реляционные базы данных не исключают использование инструментов сохранения объектов, изучавшихся выше в этой главе, – например, в реляционной базе данных можно сохранять строки, представляющие сериализованные объекты Python, созданные с помощью модуля `pickle`. ZODB также поддерживает возможность отображения объектной базы данных в реляционное хранилище.

Однако базы данных, с которыми мы познакомимся в этом разделе, имеют иную организацию и обрабатываются совершенно иными способами:

- Они хранят данные в реляционных таблицах, состоящих из столбцов (а не в словарях с объектами Python, имеющими произвольную структуру).
- Для доступа к данным и использования отношений между ними они поддерживают язык запросов SQL (вместо простого обхода объектов Python).

Для некоторых приложений в конечном результате может получиться весьма мощная комбинация. Кроме того, некоторые системы управления базами данных SQL обеспечивают мощную поддержку хранения данных промышленного уровня.

В настоящее время имеются свободно распространяемые интерфейсы, позволяющие сценариям Python использовать не только свободные, но и коммерческие системы реляционных баз данных: MySQL, Oracle, Sybase, Informix, InterBase, PostgreSQL (Postgres), SQLite, ODBC и другие. Кроме того, сообщество Python определило спецификацию *API баз данных*, переносимым образом работающего с рядом пакетов баз данных. Сценарии, написанные для этого API, могут переходить на пакеты баз данных других поставщиков с минимальными изменениями в исходном программном коде или вообще без них.

Начиная с версии Python 2.5, в стандартную библиотеку Python была включена встроенная поддержка системы реляционных баз данных SQLite. Поскольку эта система поддерживает переносимый API баз данных, она может служить инструментом и для организации хранения данных, и для разработки прототипов – системы, разрабатываемые с использованием SQLite, будут работать практически без изменений с более полноценными базами данных, такими как MySQL или Oracle.

Кроме того, имеются такие популярные сторонние системы, как SQLObject и SQLAlchemy, реализующие механизмы объектно-реляционного отображения (Object Relational Mapper, ORM), которые добавляют объектный интерфейс к реляционным базам данных. Они моделируют таблицы на основе классов Python, где строки представляют экземпляры этих классов, а столбцы – атрибуты экземпляров. Поскольку механизмы ORM просто обертывают базы данных SQL в классы Python, мы обсудим их позднее в этой главе, а сейчас рассмотрим основы поддержки SQL в языке Python.

Обзор интерфейса SQL

Подобно ZODB и в отличие от модулей `pickle` и `shelve`, представленных выше, большинство баз данных SQL являются дополнительными расширениями, не являющимися частью самого Python. В настоящее время SQLite является единственным пакетом реляционных баз данных, включенных в состав Python. Кроме того, для понимания их интерфейсов требуется знание SQL. Поскольку в этой книге нет места для обучения языку SQL, в данном разделе приводится краткий обзор API – за дополнительными сведениями обращайтесь к справочникам по SQL и ресурсам с описанием API баз данных, указанным в следующем разделе.

Приятно отметить, что к базам данных SQL можно обращаться из Python с помощью простой и переносимой модели. Спецификация API баз данных в языке Python определяет интерфейс для связи с используемыми системами баз данных из сценариев Python. Интерфейсы баз данных конкретных поставщиков для Python могут не вполне соответствовать этому API, но все расширения баз данных для Python имеют лишь незначительные отклонения. В Python базы данных SQL основываются на нескольких понятиях:

Объекты соединений

Представляют соединение с базой данных, служат интерфейсом для операций отмены и подтверждения, предоставляют доступ к реализации пакета и создают объекты курсоров.

Объекты курсоров

Представляют одну команду SQL, посылаемую в виде строки, и могут использоваться для доступа к результатам, возвращаемым командой SQL.

Результаты запроса команды SQL `select`

Возвращаются в сценарии в виде последовательностей, содержащих последовательности (например, списков кортежей), представляющих таблицы записей баз данных. Внутри этих последовательностей записей значения полей являются обычными объектами Python, такими как строки, целые и вещественные числа (например, `[('bob',48)`,

('emily',47)]). Значения полей могут также быть объектами специального типа, хранящими такие значения, как дата и время, а значения NULL в базе данных возвращаются в виде объекта None.

Помимо этого API определяет стандартный набор типов исключений баз данных, конструкторы специальных типов баз данных и информационные вызовы верхнего уровня, поддерживающие многопоточную модель выполнения и проверку используемых символов замены в параметризованных запросах.

Например, чтобы установить соединение с базой данных через интерфейс для Oracle, совместимый с Python, необходимо установить расширение Python для поддержки Oracle, а затем выполнить инструкцию такого вида:

```
connobj = connect("user/password@system")
```

Содержимое строкового аргумента зависит от базы данных и производителя (например, в некоторых случаях может потребоваться указать дополнительную сетевую информацию или имя локального файла), но обычно это данные, с помощью которых производится регистрация в системе базы данных. Получив объект соединения, над ним можно выполнять различные операции, в том числе:

connobj.close()	*закрыть соединение сейчас (не в момент вызова __del__ объекта)*
connobj.commit()	*подтвердить текущие транзакции в базе данных*
connobj.rollback()	*откатить базу данных в начало текущих транзакций*

Одно из наиболее полезных действий с объектом соединения заключается в создании объекта курсора:

cursobj = connobj.cursor()	*возвратить новый объект курсора для выполнения команд SQL*

Объекты курсоров обладают множеством методов (например, метод close позволяет закрыть курсор раньше, чем будет выполнен его деструктор, а метод callproc вызывает хранимую процедуру), но самым важным, пожалуй, является следующий:

cursobj.execute(*sqlstring* [, *parmeters*])	*выполнить запрос SQL или команду*

Параметры *parameters* передаются в виде последовательности или отображения значений и подставляются в строку *sqlstring* с инструкцией SQL в соответствии с соглашениями о замене, действующими в модуле реализации интерфейса. С помощью метода execute можно выполнять различные инструкции SQL:

- инструкции определения данных DDL (например, CREATE TABLE)
- инструкции модификации данных DML (например, UPDATE или INSERT)
- инструкции запросов DQL (например, SELECT)

После выполнения инструкции SQL атрибут `rowcount` курсора содержит количество изменившихся (для инструкций DML) или извлеченных (для запросов DQL) строк, а атрибут `description` курсора содержит имена столбцов и их типы. Кроме того, в интерфейсах большинства поставщиков метод `execute` возвращает количество измененных или извлеченных строк. Для завершения выполнения запроса DQL необходимо вызвать один из следующих методов `fetch`:

`tuple = cursobj.fetchone()`	*получить следующую строку из результата запроса*
`listoftuple = cursobj.fetchmany([size])`	*получить следующую группу строк из результата*
`listoftuple = cursobj.fetchall()`	*получить все оставшиеся строки из результата*

После получения результатов с помощью метода `fetch` данные таблицы обрабатываются с помощью обычных операций последовательностей Python – например, можно обойти кортежи в списке результатов, полученном вызовом метода `fetchall` с помощью простого цикла `for` или выражения-генератора. Большинство интерфейсов баз данных Python позволяет также указывать значения, которые должны быть переданы в инструкции SQL, определяя символы подстановки и кортеж параметров. Например:

```
query = 'SELECT name, shoesize FROM spam WHERE job = ? AND age = ?'
cursobj.execute(query, (value1, value2))
results = cursobj.fetchall()
for row in results: ...
```

В данном случае интерфейс базы данных использует параметризованные инструкции (для оптимизации и удобства) и корректно передает параметры в базу данных независимо от их типов в языке Python. Обозначения параметров в интерфейсах некоторых баз данных могут быть иными (например, в интерфейсе для Oracle используются обозначения `:p1` и `:p2` или два `%s` вместо двух `?`). Во всяком случае, это не то же самое, что оператор Python форматирования строки `%`, а кроме того, применение параметров в запросах помогает обойти некоторые проблемы, связанные с безопасностью.

Наконец, если база данных поддерживает хранимые процедуры, их можно вызывать их с помощью метода `callproc` или передавая методу `execute` строку команды SQL `CALL` или `EXEC`. Метод `callproc` может генерировать таблицу с результатами, получить которую можно с помощью одного из методов `fetch`, и возвращает модифицированную копию входной последовательности – входные параметры остаются нетронутыми, а выходные и входные/выходные параметры замещаются новыми значениями. Описание дополнительных особенностей API, включая поддержку больших блоков двоичных данных, можно найти в документации API. А теперь рассмотрим, как реализовать работу с инструкциями SQL на языке Python.

Учебник по API базы данных SQL на примере SQLite

В этой книге недостаточно места, чтобы предоставить всеобъемлющее руководство по использованию API баз данных. Однако, чтобы получить некоторое представление об интерфейсе, рассмотрим несколько простых примеров. В этом учебнике мы будем использовать систему баз данных SQLite. SQLite является стандартной частью Python, поэтому вполне логично предполагать, что она поддерживается во всех установках Python. Механизм SQLite реализует законченную систему управления реляционными базами данных, однако он имеет вид библиотеки, подключаемой к процессу, а не сервера. В целом, такая организация больше подходит для создания хранилищ в программах, чем для удовлетворения потребностей хранения данных промышленного масштаба.

Однако благодаря переносимости API баз данных в языке Python другие пакеты популярных баз данных, таких как PostgreSQL, MySQL и Oracle, используются практически идентично – для сценариев, использующих стандартный программный код SQL, основные отличия будут заключаться в начальном вызове для подключения к базе данных, который для разных баз данных обычно требует передачи разных значений аргументов. Благодаря этому мы можем использовать систему SQLite и как инструмент создания прототипов при разработке приложений, и как простой способ приступить к работе с использованием API баз данных SQL языка Python в этой книге.

Как уже упоминалось выше, описание MySQL, представленное в третьем издании, было убрано из этого издания, потому что используемый интерфейс еще не был перенесен в Python 3.X. Однако примеры использования MySQL и обзор из третьего издания доступны в пакете примеров, в каталоге *C:\...\PP4E\Dbase\Sql\MySql-2.X* и в его подкаталоге *Documentation*. Примеры написаны для Python 2.X, но фрагменты программного кода, выполняющие операции с базой данных, в значительной степени могут работать в любой версии Python. Поскольку этот программный код также в значительной мере не зависит от типа используемой базы данных, он, вероятно, не будет иметь большого значения для читателей – сценарии, представленные в этой книге, должны работать и с другими пакетами баз данных, такими как MySQL, лишь с незначительными изменениями.

Введение

Независимо от типа базы данных, используемой в сценарии, основной интерфейс SQL в языке Python очень прост. Фактически, он вообще не является объектно-ориентированным – запросы и другие команды передаются базам данных в виде строк с инструкциями на языке SQL. Если вы знакомы с языком SQL, то вы уже владеете большинством навыков, необходимых для работы с реляционными базами данных в Python.

Это знание облегчает жизнь тем, кто относится к этой категории, и становится необходимым дополнительным условием для всех остальных.

Данное издание не является книгой по языку SQL, поэтому оставим описание подробностей об используемых здесь командах на долю других источников (издательством O'Reilly выпущена целая серия книг по этой теме). В действительности, мы будем использовать крошечные базы данных и умышленно применять лишь самые простые команды SQL – вам необходимо будет распространить все, что увидите здесь, на более реалистичные задачи, с которыми вам придется столкнуться. Этот раздел дает лишь краткий обзор приемов использования языка Python в соединении с базой данных SQL.

Тем не менее независимо от размера базы данных программный код Python, необходимый для работы с ней, является удивительно простым. Первое, что необходимо сделать, – это открыть соединение с базой данных и создать таблицу для хранения записей:

```
C:\...\PP4E\Dbase\Sql> python
>>> import sqlite3
>>> conn = sqlite3.connect('dbase1')   # при необходимости используйте
                                       # полный путь к файлу
```

В этом примере сначала импортируется интерфейс SQLite – это модуль `sqlite3` из стандартной библиотеки. Затем мы создаем объект соединения, передавая параметры, требуемые базой данных, – здесь передается имя локального файла, в котором будет храниться база данных. Именно для этого файла необходимо будет создавать резервные копии, чтобы сохранить базу данных. При необходимости этот вызов создаст новый файл или откроет существующий – интерфейсу SQLite также можно передать специальную строку «:memory:», чтобы создать временную базу данных в памяти.

Если сценарий придерживается использования стандартного программного кода SQL, аргументы метода `connect` – это единственное, что будет отличаться при использовании разных систем баз данных. Например, в интерфейсе MySQL этот метод принимает доменное имя сетевого хоста, имя пользователя и пароль, которые передаются в виде именованных аргументов, а интерфейс Oracle, который демонстрировался в примере выше, определяет более специфический синтаксис строки подключения. Однако кроме этого платформозависимого вызова остальная часть API в значительной степени не зависит от типа базы данных.

Создание баз данных и таблиц

Далее, создадим курсор для отправки инструкций SQL серверу баз данных и создадим с его помощью первую таблицу:

```
>>> curs = conn.cursor()
>>>
>>> tblcmd = 'create table people (name char(30), job char(10), pay int(4))'
>>> curs.execute(tblcmd)
```

Последняя команда в этом примере создаст в базе данных таблицу с именем «people» – идентификаторы name, job и pay определяют столбцы в этой таблице и их типы с использованием синтаксических конструкций «тип(размер)» – две строки и целое число. Типы данных могут быть более сложными, чем в нашем примере, но мы пока не будем обращать внимание на такие подробности (обращайтесь к руководствам по языку SQL). В интерфейсе SQLite роль базы данных играет файл, поэтому здесь отсутствует понятие создания и выбора базы данных внутри него, как в некоторых других системах. Теперь у нас имеется в текущем рабочем каталоге простой плоский файл с именем *data1*, который содержит двоичные данные и нашу таблицу people.

Добавление записей

К настоящему моменту мы подключились к базе данных (в SQLite это означает – просто открыли локальный файл) и создали таблицу. Теперь запустим новый интерактивный сеанс Python и создадим несколько записей. Существует три основных способа, основанных на инструкциях, которые мы можем использовать здесь: вставлять записи по одной, вставить сразу несколько записей в одной инструкции или задействовать цикл Python. Ниже приводится простейшее решение (я опустил вывод результатов некоторых вызовов, которые не имеют отношения к обсуждаемой теме):

```
C:\...\PP4E\Dbase\Sql> python
>>> import sqlite3
>>> conn = sqlite3.connect('dbase1')
>>> curs = conn.cursor()
>>> curs.execute('insert into people values (?, ?, ?)', ('Bob', 'dev', 5000))
>>> curs.rowcount
1
>>> sqlite3.paramstyle
'qmark'
```

Здесь создается объект курсора, как уже было показано выше, чтобы получить возможность отправлять инструкции SQL серверу баз данных. Команда insert в языке SQL добавляет в таблицу единственную запись. После вызова метода execute в атрибуте rowcount курсора возвращается количество записей, созданных или затронутых последней выполненной инструкцией. В некоторых модулях реализации интерфейсов баз данных это же число доступно в виде значения, возвращаемого методом execute, но это не оговаривается в спецификации API баз данных и не реализовано в SQLite. Иными словами, не следует полагаться на это, если необходимо, чтобы сценарий для работы с базой данных был способен взаимодействовать с другими системами баз данных.

Значения параметров для подстановки в инструкцию SQL обычно передаются в виде последовательности (например, в виде списка или кортежа). Обратите внимание на вызов метода paramstyle, который сообщает

стиль обозначения параметров в строке с инструкцией. В данном случае qmark означает, что параметры внутри инструкций обозначаются знаками вопроса ?. Другие модули баз данных могут использовать такие стили, как format (параметры обозначаются, как %s), числовые индексы или именованные параметры, – за дополнительной информацией обращайтесь к описанию API базы данных.

Для добавления в одной инструкции сразу несколько записей используется метод executemany и последовательность записей (например, список списков). По своему действию вызов этого метода напоминает вызов метода execute для каждой записи в аргументе и в действительности может быть реализован именно так, однако интерфейсы баз данных могут также использовать приемы, специфичные для конкретной базы данных, чтобы ускорить выполнение:

```
>>> curs.executemany('insert into people values (?, ?, ?)',
...          [ ('Sue', 'mus', '70000'),
...            ('Ann', 'mus', '60000')])
>>> curs.rowcount
2
```

Последняя инструкция добавит сразу две записи. Немногим больше работы требуется выполнить, чтобы добиться того же результата, вставляя по одной записи в цикле:

```
>>> rows = [['Tom', 'mgr', 100000],
...         ['Kim', 'adm', 30000],
...         ['pat', 'dev', 90000]]
>>> for row in rows:
...     curs.execute('insert into people values (? , ?, ?)', row)
...
>>> conn.commit()
```

Подобное смешивание Python и SQL открывает весьма интересные возможности. Обратите внимание на последнюю команду – чтобы сохранить изменения в базе данных, необходимо всегда вызывать метод commit объекта соединения. В противном случае, когда соединение будет закрыто, изменения будут потеряны. В действительности, пока не будет вызван метод commit, ни одна из добавленных записей не будет видна из других соединений с базой данных.

Технически спецификация API требует, чтобы при закрытии (вручную, вызовом метода close или автоматически в момент утилизации объекта сборщиком мусора) объект соединения автоматически вызывал метод rollback с целью отменить неподтвержденные изменения. Для систем баз данных, не поддерживающих операции подтверждения и отмены транзакций, эти методы могут не выполнять никаких действий. SQLite реализует оба метода, commit и rollback; последний из них откатывает любые изменения до момента последнего вызова метода commit.

Выполнение запросов

Итак, к настоящему моменту мы добавили шесть записей в таблицу базы данных. Выполним запрос SQL, чтобы посмотреть, что у нас получилось:

```
>>> curs.execute('select * from people')
>>> curs.fetchall()
[('Bob', 'dev', 5000), ('Sue', 'mus', 70000), ('Ann', 'mus', 60000), ('Tom',
'mgr', 100000), ('Kim', 'adm', 30000), ('pat', 'dev', 90000)]
```

Здесь с помощью объекта курсора выполняется инструкция SQL `select`, которая отбирает все записи, и вызывается метод `fetchall` курсора, чтобы извлечь их. Записи возвращаются сценарию в виде последовательности последовательностей. В данном модуле это список кортежей – внешний список представляет таблицу результатов, вложенные кортежи представляют записи, а содержимое вложенных кортежей – столбцы данных. Поскольку все эти данные являются обычными данными Python, после получения результатов запроса их можно обрабатывать с помощью обычного программного кода Python. Например, чтобы сделать вывод более удобочитаемым, выполним цикл по результатам:

```
>>> curs.execute('select * from people')
>>> for row in curs.fetchall():
...     print(row)
...
('Bob', 'dev', 5000)
('Sue', 'mus', 70000)
('Ann', 'mus', 60000)
('Tom', 'mgr', 100000)
('Kim', 'adm', 30000)
('pat', 'dev', 90000)
```

В цикле также удобно использовать операцию распаковывания кортежей для выборки значений столбцов в итерациях. Ниже демонстрируется простой форматированный вывод значений двух столбцов:

```
>>> curs.execute('select * from people')
>>> for (name, job, pay) in curs.fetchall():
...     print(name, ':', pay)
...
Bob : 5000
Sue : 70000
Ann : 60000
Tom : 100000
Kim : 30000
pat : 90000
```

Поскольку результатом запроса является последовательность, для ее обработки можно использовать мощные операции над последовательностями и инструменты итераций, имеющиеся в Python. Например,

чтобы отобрать значения только из столбца name, можно выполнить более специализированный запрос SQL и получить список кортежей:

```
>>> curs.execute('select name from people')
>>> names = curs.fetchall()
>>> names
[('Bob',), ('Sue',), ('Ann',), ('Tom',), ('Kim',), ('pat',)]
```

Или использовать для выборки желаемых полей генератор списков – используя программный код Python, мы получаем более полный контроль над данными и их форматированием:

```
>>> curs.execute('select * from people')
>>> names = [rec[0] for rec in curs.fetchall()]
>>> names
['Bob', 'Sue', 'Ann', 'Tom', 'Kim', 'pat']
```

Использовавшийся до сих пор метод fetchall извлекает сразу все результаты запроса в виде единой последовательности (в случае отсутствия результатов возвращается пустая последовательность). Это удобно, но такой способ может оказаться достаточно медленным, чтобы временно заблокировать вызывающую программу при большом объеме результатов или необходимости передавать значительные объемы данных по сети, когда взаимодействие выполняется с удаленным сервером (подобные операции в графическом интерфейсе можно было бы производить в параллельном потоке выполнения). Чтобы избежать этого, можно извлекать данные по одной записи или пакетами записей с помощью методов fetchone и fetchmany. Метод fetchone возвращает следующую запись из результатов или None по достижении конца таблицы:

```
>>> curs.execute('select * from people')
>>> while True:
...     row = curs.fetchone()
...     if not row: break
...     print(row)
...
('Bob', 'dev', 5000)
('Sue', 'mus', 70000)
('Ann', 'mus', 60000)
('Tom', 'mgr', 100000)
('Kim', 'adm', 30000)
('pat', 'dev', 90000)
```

Метод fetchmany возвращает последовательность записей из результатов, но не всю таблицу – можно явно указать количество записей, извлекаемых при каждом обращении, или положиться на значение по умолчанию, которое определяется атрибутом arraysize курсора. Каждый вызов возвращает не более указанного числа записей из результатов, или пустую последовательность по достижении конца таблицы:

```
>>> curs.execute('select * from people')
>>> while True:
```

```
...     rows = curs.fetchmany()  # size=N необязательный аргумент
...     if not rows: break
...     for row in rows:
...         print(row)
...
('Bob', 'dev', 5000)
('Sue', 'mus', 70000)
('Ann', 'mus', 60000)
('Tom', 'mgr', 100000)
('Kim', 'adm', 30000)
('pat', 'dev', 90000)
```

Для этого модуля таблица результатов будет исчерпана, как только метод `fetchone` или `fetchmany` вернет значение `False`. Спецификация API баз данных требует, чтобы метод `fetchall` возвращал «все оставшиеся записи», поэтому перед извлечением новых данных, как правило, необходимо снова вызвать метод `execute`, чтобы получить результаты:

```
>>> curs.fetchone()
>>> curs.fetchmany()
[]
>>> curs.fetchall()
[]
```

Естественно, есть возможность не только извлекать таблицу целиком – в Python нам доступна вся мощь языка SQL:

```
>>> curs.execute('select name, job from people where pay > 60000')
>>> curs.fetchall()
[('Sue', 'mus'), ('Tom', 'mgr'), ('pat', 'dev')]
```

Последний запрос извлекает поля `name` и `job` для тех сотрудников, которые зарабатывают более $60 000 в год. Следующий фрагмент реализует аналогичную операцию, но передает значение, ограничивающее выбор, в виде параметра и указывает порядок следования результатов:

```
>>> query = 'select name, job from people where pay >= ? order by name'
>>> curs.execute(query, [60000])
>>> for row in curs.fetchall(): print(row)
...
('Ann', 'mus')
('Sue', 'mus')
('Tom', 'mgr')
('pat', 'dev')
```

Выполнение обновлений

Объекты курсоров также используются для отправки команд SQL, изменяющих, удаляющих и вставляющих новые данные. Мы уже видели инструкцию `insert`, а теперь откроем новый сеанс и попробуем выполнить некоторые изменения – мы начинаем с тем же набором данных, который имелся в предыдущем разделе:

```
C:\...\PP4E\Dbase\Sql> python
>>> import sqlite3
>>> conn = sqlite3.connect('dbase1')
>>> curs = conn.cursor()
>>> curs.execute('select * from people')
>>> curs.fetchall()
[('Bob', 'dev', 5000), ('Sue', 'mus', 70000), ('Ann', 'mus', 60000), ('Tom',
'mgr', 100000), ('Kim', 'adm', 30000), ('pat', 'dev', 90000)]
```

Инструкция SQL `update` изменяет записи – следующий пример запишет новое значение 65000 столбца `pay` в трех записях (Bob, Ann и Kim), потому что зарплата этих сотрудников не превышает $60 000. Как обычно, атрибут `rowcount` содержит количество изменившихся записей:

```
>>> curs.execute('update people set pay=? where pay <= ?', [65000, 60000])
>>> curs.rowcount
3
>>> curs.execute('select * from people')
>>> curs.fetchall()
[('Bob', 'dev', 65000), ('Sue', 'mus', 70000), ('Ann', 'mus', 65000),
('Tom', 'mgr', 100000), ('Kim', 'adm', 65000), ('pat', 'dev', 90000)]
```

Инструкция SQL `delete` удаляет записи в соответствии с необязательным условием (чтобы удалить все записи, достаточно опустить условие). Следующий пример удалит запись о сотруднике с именем Bob, а также все другие записи, в которых поле `pay` имеет значение не меньше $90,000:

```
>>> curs.execute('delete from people where name = ?', ['Bob'])
>>> curs.execute('delete from people where pay >= ?',(90000,))
>>> curs.execute('select * from people')
>>> curs.fetchall()
[('Sue', 'mus', 70000), ('Ann', 'mus', 65000), ('Kim', 'adm', 65000)]

>>> conn.commit()
```

Наконец, не забывайте подтверждать изменения перед завершением сценария или сеанса Python, если вы предполагали сохранить их. Без подтверждения вызов метода `rollback` объекта соединения из метода `close` или деструктора `__del__` откатит все неподтвержденные изменения. Объекты соединений автоматически закрываются при утилизации сборщиком мусора, который вызывает деструктор `__del__`, выполняющий откат изменений, – сборка мусора выполняется автоматически при завершении программы, если не раньше.

Создание словарей записей

Теперь, когда мы познакомились с основами, двинемся дальше и применим их для решения более крупных задач. Спецификация API SQL определяет, что результаты запросов должны возвращаться в виде последовательностей, содержащих последовательности. Одной из наибо-

лее типичных особенностей API, о которой часто забывают, является возможность получить записи в более структурированном виде – в виде словаря или экземпляра класса, например, ключи или атрибуты которых соответствуют именам столбцов. Механизмы ORM, с которыми мы познакомимся в конце этой главы, отображают записи в экземпляры классов, но, поскольку это Python, подобные трансформации совсем несложно реализовать другими способами. Кроме того, API уже дает нам все необходимые инструменты.

Использование описаний таблиц

Например, спецификация API баз данных определяет, что после выполнения запроса методом `execute` атрибут `description` курсора должен содержать имена и (для некоторых баз данных) типы столбцов в таблице с результатами. Чтобы увидеть, что возвращается в этом атрибуте, продолжим эксперименты с базой данных, находящейся в том состоянии, в каком мы оставили ее в предыдущем разделе:

```
>>> curs.execute('select * from people')
>>> curs.description
(('name', None, None, None, None, None, None), ('job', None, None, None,
None, None, None), ('pay', None, None, None, None, None, None))

>>> curs.fetchall()
[('Sue', 'mus', 70000), ('Ann', 'mus', 65000), ('Kim', 'adm', 65000)]
```

Формально значением атрибута `description` является последовательность последовательностей с описаниями столбцов, следующих друг за другом. Описание поля `type_code` можно найти в спецификации API баз данных – оно отображается в объекты на верхнем уровне модуля интерфейса базы данных, но модуль `sqlite3` реализует только поле с именем столбца:

```
(name, type_code, display_size, internal_size, precision, scale, null_ok)
```

Теперь мы в любой момент сможем использовать эти метаданные, когда потребуется вывести метки столбцов, например при форматированном выводе записей (для начала необходимо повторно выполнить запрос, поскольку результаты прошлого запроса уже были извлечены):

```
>>> curs.execute('select * from people')
>>> colnames = [desc[0] for desc in curs.description]
>>> colnames
['name', 'job', 'pay']

>>> for row in curs.fetchall():
...     for name, value in zip(colnames, row):
...         print(name, '\t=>', value)
...    print()
...
name   => Sue
```

```
job    => mus
pay    => 70000

name   => Ann
job    => mus
pay    => 65000

name   => Kim
job    => adm
pay    => 65000
```

Обратите внимание, что для выравнивания вывода здесь был использован символ табуляции – более удачное решение состоит в том, чтобы определить максимальную длину имени поля (как это сделать, будет показано в примере ниже).

Конструирование словарей записей

Мы можем немного усовершенствовать программный код форматированного вывода, предусмотрев создание словаря для каждой записи, роль ключей в котором будут выполнять имена полей – нам нужно лишь заполнить словарь в процессе обхода:

```
>>> curs.execute('select * from people')
>>> colnames = [desc[0] for desc in curs.description]
>>> rowdicts = []
>>> for row in curs.fetchall():
...     newdict = {}
...    for name, val in zip(colnames, row):
...        newdict[name] = val
...    rowdicts.append(newdict)
...
>>> for row in rowdicts: print(row)
...
{'pay': 70000, 'job': 'mus', 'name': 'Sue'}
{'pay': 65000, 'job': 'mus', 'name': 'Ann'}
{'pay': 65000, 'job': 'adm', 'name': 'Kim'}
```

Однако поскольку это Python, существуют более мощные способы конструирования словарей записей. Например, конструктор словарей принимает объединенную последовательность пар имя/значение и на ее основе создает словарь:

```
>>> curs.execute('select * from people')
>>> colnames = [desc[0] for desc in curs.description]
>>> rowdicts = []
>>> for row in curs.fetchall():
...     rowdicts.append( dict(zip(colnames, row)) )
...
>>> rowdicts[0]
{'pay': 70000, 'job': 'mus', 'name': 'Sue'}
```

И, наконец, можно использовать генератор списков для объединения словарей в список – получившийся программный код не только компактнее, но и, возможно, выполняется быстрее, чем исходная версия:

```
>>> curs.execute('select * from people')
>>> colnames = [desc[0] for desc in curs.description]
>>> rowdicts = [dict(zip(colnames, row)) for row in curs.fetchall()]
>>> rowdicts[0]
{'pay': 70000, 'job': 'mus', 'name': 'Sue'}
```

При переходе к словарям мы потеряли порядок следования полей в записях – если вернуться назад, к результатам, полученным с помощью `fetchall`, можно заметить, что поля `name`, `job` и `pay` записей в результатах следуют в том же порядке, в каком они были определены при создании таблицы. Поля в нашем словаре следуют в псевдослучайном порядке, что вообще характерно для отображений Python. Пока поля извлекаются по ключу, в этом нет никаких проблем. Таблицы по-прежнему поддерживают свой порядок следования полей и операция создания словарей выполняется безукоризненно – благодаря тому, что кортежи с описанием полей в результате следуют в том же порядке, что и поля в кортежах записей, возвращаемых запросами.

Мы оставим задачу преобразования кортежей записей в экземпляры классов в качестве самостоятельного упражнения, однако я дам две подсказки: модуль `collections` из стандартной библиотеки Python реализует такие необычные типы данных, как именованные кортежи и упорядоченные словари; и имеется возможность обеспечить возможность обращения к полям как к атрибутам, а не ключам, для чего достаточно просто создать пустой экземпляр класса и присвоить значения его атрибутам с помощью функции Python `setattr`. Кроме того, классы являются естественным местом размещения наследуемого программного кода, такого как стандартные методы отображения. В действительности, это именно то, что предоставляют нам механизмы ORM, описываемые ниже.

Автоматизация операций с помощью сценариев и модулей

До настоящего момента мы использовали Python как своеобразный клиент SQL командной строки – мы вводили и выполняли запросы в интерактивном режиме. Однако программный код, который мы запускали выше, можно использовать в качестве основы реализации операций с базами данных в сценариях. При работе в интерактивной оболочке требуется заново вводить такие конструкции, как многострочные циклы, что может быть весьма утомительным занятием. С помощью сценариев мы можем автоматизировать нашу работу.

Для демонстрации возьмем последний пример из предыдущего раздела и создадим вспомогательный модуль – в примере 17.4 приводится реализация модуля многократного использования, который знает, как преобразовывать результаты запросов из кортежей с записями в словари.

Пример 17.4. PP4E\Dbase\Sql\makedicts.py

```
"""
преобразует список кортежей записей в список словарей,
роль ключей в которых играют имена полей
это не утилита командной строки: при запуске из командной строки
выполняется жестко определенный программный код самотестирования
"""

def makedicts(cursor, query, params=()):
    cursor.execute(query, params)
    colnames = [desc[0] for desc in cursor.description]
    rowdicts = [dict(zip(colnames, row)) for row in cursor.fetchall()]
    return rowdicts

if __name__ == '__main__':          # самотестирование
    import sqlite3
    conn = sqlite3.connect('dbase1')
    cursor = conn.cursor()
    query = 'select name, pay from people where pay < ?'
    lowpay = makedicts(cursor, query, [70000])
    for rec in lowpay: print(rec)
```

Как обычно, этот файл можно запустить из командной строки как самостоятельный сценарий, чтобы выполнить программный код самотестирования:

```
...\PP4E\Dbase\Sql> makedicts.py
{'pay': 65000, 'name': 'Ann'}
{'pay': 65000, 'name': 'Kim'}
```

Также можно импортировать его как модуль и вызывать его функции из другого контекста, например из интерактивного сеанса. Поскольку это модуль, он превратился в инструмент баз данных многократного использования:

```
...\PP4E\Dbase\Sql> python
>>> from makedicts import makedicts
>>> from sqlite3 import connect
>>> conn = connect('dbase1')
>>> curs = conn.cursor()
>>> curs.execute('select * from people')
>>> curs.fetchall()
[('Sue', 'mus', 70000), ('Ann', 'mus', 65000), ('Kim', 'adm', 65000)]

>>> rows = makedicts(curs, "select name from people where job = 'mus'")
>>> rows
[{'name': 'Sue'}, {'name': 'Ann'}]
```

Наша утилита способна обрабатывать запросы произвольной сложности – они просто передаются модулю связи с сервером баз данных или библиотекой. Предложение `order by` в следующем примере выполняет сортировку результатов по полю `name`:

```
>>> query = 'select name, pay from people where job = ? order by name'
>>> musicians = makedicts(curs, query, ['mus'])
>>> for row in musicians: print(row)
...
{'pay': 65000, 'name': 'Ann'}
{'pay': 70000, 'name': 'Sue'}
```

Объединяем все вместе

Теперь мы знаем, как создавать базы данных и таблицы, как вставлять записи в таблицы, как запрашивать содержимое таблиц и как извлекать имена столбцов. Для справки и для демонстрации того, как можно комбинировать эти приемы, в примере 17.5 они объединены в один сценарий.

Пример 17.5. PP4E\Dbase\Sql\testdb.py

```
from sqlite3 import connect
conn = connect('dbase1')
curs = conn.cursor()
try:
    curs.execute('drop table people')
except:
    pass                        # не существует
curs.execute('create table people (name char(30), job char(10), pay
int(4))')

curs.execute('insert into people values (?, ?, ?)', ('Bob', 'dev', 50000))
curs.execute('insert into people values (?, ?, ?)', ('Sue', 'dev', 60000))

curs.execute('select * from people')
for row in curs.fetchall():
    print(row)

curs.execute('select * from people')
colnames = [desc[0] for desc in curs.description]
while True:
    print('-' * 30)
    row = curs.fetchone()
    if not row: break
    for (name, value) in zip(colnames, row):
        print('%s => %s' % (name, value))

conn.commit()                   # сохранить вставленные записи
```

Если что-то в этом сценарии вам покажется непонятным, обращайтесь за разъяснениями к предыдущим разделам в этом учебнике. При запуске он создает базу данных с двумя записями и выводит ее содержимое в стандартный поток вывода:

```
C:\...\PP4E\Dbase\Sql> testdb.py
('Bob', 'dev', 50000)
```

```
('Sue', 'dev', 60000)
------------------------------
name => Bob
job => dev
pay => 50000
------------------------------
name => Sue
job => dev
pay => 60000
------------------------------
```

Данный пример предназначен лишь для демонстрации интерфейса баз данных. В нем жестко определены имена в базе данных, и при каждом запуске он заново создает базу данных. Мы можем превратить этот программный код в более универсальные инструменты, организовав его в виде компонентов многократного использования, как будет показано далее в этом разделе. Но сначала рассмотрим приемы загрузки данных в наши базы данных.

Загрузка таблиц базы данных из файлов

Одна из замечательных особенностей языка Python в области баз данных заключается в том, что он позволяет объединять мощь языка запросов SQL с мощью многоцелевого языка программирования Python. Они естественным образом дополняют друг друга.

Загрузка с помощью SQL и Python

Предположим, что нам необходимо загрузить данные в таблицу из плоского файла, каждая строка которого представляет запись в базе данных и содержит значения полей, разделенные запятыми. В примерах 17.6 и 17.7 приводятся два таких файла с данными, которые мы будем использовать здесь.

Пример 17.6. PP4E\Dbase\Sql\data.txt

```
bob,devel,50000
sue,music,60000
ann,devel,40000
tim,admin,30000
kim,devel,60000
```

Пример 17.7. PP4E\Dbase\Sql\data2.txt

```
bob,developer,80000
sue,music,90000
ann,manager,80000
```

В настоящее время в некоторых системах баз данных, таких как MySQL, имеется удобная инструкция SQL, позволяющая быстро загружать такие таблицы. Инструкция `load data` анализирует и загружает данные из текстового файла, находящегося на стороне клиента или сервера.

В следующем примере первая команда удаляет все записи в таблице, а затем мы разбиваем инструкцию SQL на несколько строк, используя тот факт, что Python автоматически объединяет смежные строковые литералы:

```
# Используется MySQL (в настоящее время интерфейс к этой базе данных
# доступен только для Python 2.X)
```

...сначала выполняется регистрация в базе данных MySQL...

```
>>> curs.execute('delete from people')      # все записи
>>> curs.execute(
...     "load data local infile 'data.txt' "
...         "into table people fields terminated by ','")

>>> curs.execute('select * from people')
>>> for row in curs.fetchall(): print(row)
...
('bob', 'devel', 50000L)
('sue', 'music', 60000L)       # длинные целые в Python 2.X
('ann', 'devel', 40000L)
('tim', 'admin', 30000L)
('kim', 'devel', 60000L)
>>> conn.commit()
```

Этот прием действует именно так, как и ожидалось. Но что если нам потребуется использовать другую систему баз данных, такую как SQLite, в которой отсутствует такая инструкция SQL? Или может быть вам просто потребуется выполнить что-то особенное, чего не позволяет эта инструкция MySQL. Не волнуйтесь – чтобы достичь того же результата при использовании SQLite и Python 3.X, потребуется написать совсем немного простого программного кода на языке Python (в выводе ниже опущены некоторые строки, не имеющие отношения к обсуждаемой теме):

```
C:\...\PP4E\Dbase\Sql> python
>>> from sqlite3 import connect
>>> conn = connect('dbase1')
>>> curs = conn.cursor()

>>> curs.execute('delete from people')      # очистить таблицу
>>> curs.execute('select * from people')
>>> curs.fetchall()
[]

>>> file = open('data.txt')
>>> rows = [line.rstrip().split(',') for line in file]
>>> rows[0]
['bob', 'devel', '50000']

>>> for rec in rows:
...     curs.execute('insert into people values (?, ?, ?)', rec)
...
```

```
>>> curs.execute('select * from people')
>>> for rec in curs.fetchall(): print(rec)
...
('bob', 'devel', 50000)
('sue', 'music', 60000)
('ann', 'devel', 40000)
('tim', 'admin', 30000)
('kim', 'devel', 60000)
```

Здесь используется генератор списков, который собирает в список результаты разбиения всех строк в файле после удаления из них символов перевода строки, и итераторы файлов, выполняющие построчное чтение содержимого файлов. Цикл `for` в этом примере делает то же самое, что и инструкция `load` базы данных MySQL, но он может работать с базами данных разных типов, включая SQLite. Похожий результат можно также получить с помощью метода `executemany`, показанного выше, однако цикл `for`, использованный здесь, в целом является более универсальным.

Python и SQL

Фактически в вашем распоряжении имеется целый язык Python для обработки результатов запроса к базе данных, а программный код Python даже небольшого объема часто способен не только продублировать, но и превзойти возможности SQL. Например, в языке SQL имеются специальные агрегатные функции, вычисляющие такие значения, как сумма и среднее арифметическое:

```
>>> curs.execute("select sum(pay), avg(pay) from people where job = 'devel'")
>>> curs.fetchall()
[(150000, 50000.0)]
```

Перекладывая обработку данных на программный код Python, иногда можно упростить запросы SQL и реализовать более сложную логику (хотя при этом, возможно, придется пожертвовать любыми оптимизациями производительности запросов, которые может предложить база данных). Вычисление суммы зарплат и среднего значения на языке Python можно реализовать с помощью простого цикла:

```
>>> curs.execute("select name, pay from people where job = 'devel'")
>>> result = curs.fetchall()
>>> result
(('bob', 50000L), ('ann', 40000L), ('kim', 60000L))

>>> tot = 0
>>> for (name, pay) in result: tot += pay
...
>>> print('total:', tot, 'average:', tot / len(result)) # используйте //
total: 150000 average: 50000.0                          # для деления
                                                        # с усечением
```

Также для вычисления суммы, максимального и среднего значений можно задействовать более сложные инструменты, такие как генераторы и выражения-генераторы, как показано ниже:

```
>>> print(sum(rec[1] for rec in result))   # выражение-генератор
150000
>>> print(sum(rec[1] for rec in result) / len(result))
50000.0
>>> print(max(rec[1] for rec in result))
60000
```

Подход на основе языка Python является более универсальным, но потребность в нем не столь очевидна, пока не возникает необходимость в реализации более сложной логики обработки. Например, ниже приводятся чуть более сложные генераторы списков, которые отбирают из результатов имена сотрудников, зарплата которых выше или ниже среднего значения:

```
>>> avg = sum(rec[1] for rec in result) / len(result)
>>> print([rec[0] for rec in result if rec[1] > avg])
['kim']
>>> print([rec[0] for rec in result if rec[1] < avg])
['ann']
```

Подобного рода задачи можно решать также с применением дополнительных возможностей языка SQL, таких как вложенные запросы, но рано или поздно мы достигнем уровня сложности, когда универсальная природа языка Python и, возможно, его переносимость, станут более привлекательными. Для сравнения ниже приводится эквивалентное решение на языке SQL:

```
>>> query = ("select name from people where job = 'devel' and "
...               "pay > (select avg(pay) from people where job = 'devel')")
>>> curs.execute(query)
>>> curs.fetchall()
[('kim',)]

>>> query = ("select name from people where job = 'devel' and "
...                "pay < (select avg(pay) from people where job =
'devel')")
>>> curs.execute(query)
>>> curs.fetchall()
[('ann',)]
```

Это наверняка не самые сложные запросы SQL, с которыми вам придется столкнуться, но за этой гранью код SQL может стать намного более сложным. Кроме того, в отличие от языка Python, SQL ограничивается решением задач, связанных лишь с базой данных. Представьте запрос, который сравнивает значения столбцов с данными, полученными из Интернета или введенными пользователем в графическом интерфейсе, – это простая операция в Python, с поддержкой Интернета и графических интерфейсов, которая выходит за пределы такого узкоспециализиро-

ванного языка, как SQL. Объединяя Python и SQL, вы получаете все самое лучшее от обоих языков и можете выбирать, когда какой из них лучше подходит для достижения ваших целей.

При использовании Python вы также получаете доступ к уже написанным утилитам: ваш набор инструментов для работы с базами данных может безгранично расширяться новыми функциями, модулями и классами. Для иллюстрации ниже представлена более удобочитаемая реализация тех же операций с применением модуля преобразования записей в словари, написанного нами выше:

```
>>> from makedicts import makedicts
>>> recs = makedicts(curs, "select * from people where job = 'devel'")
>>> print(len(recs), recs[0])
3 {'pay': 50000, 'job': 'devel', 'name': 'bob'}

>>> print([rec['name'] for rec in recs])
['bob', 'ann', 'kim']
>>> print(sum(rec['pay'] for rec in recs))
150000

>>> avg = sum(rec['pay'] for rec in recs) / len(recs)
>>> print([rec['name'] for rec in recs if rec['pay'] > avg])
['kim']
>>> print([rec['name'] for rec in recs if rec['pay'] >= avg])
['bob', 'kim']
```

Аналогично тип `set` в языке Python предоставляет такие операции, как пересечение, объединение и разность, которые могут служить альтернативами других операций SQL (в интересах экономии места мы оставим эту тему для самостоятельного изучения). Дополнительные расширения Python для работы с базами данных можно найти среди инструментов, созданных сторонними разработчиками. Например, существует множество пакетов, добавляющих объектно-ориентированные возможности к интерфейсу баз данных – механизмы ORM, которые мы рассмотрим ближе к концу этой главы.

Вспомогательные сценарии SQL

К настоящему моменту в нашем путешествии по SQL и интерфейсу баз данных мы дошли до точки, когда становится неудобно использовать интерактивную оболочку – в начале каждого сеанса и перед началом каждого теста нам приходится снова и снова вводить один и тот же типовой программный код. Кроме того, этот программный код мог бы использоваться в других программах. Давайте преобразуем наш программный код в сценарии, автоматизирующие часто выполняемые задачи и обеспечивающие возможность многократного использования.

Для иллюстрации всей мощи, которой обладает комбинация Python/SQL, в этом разделе представлены вспомогательные сценарии, выполняющие типичные задачи, которые часто приходится реализовывать

во время разработки. Дополнительно большинство этих файлов могут использоваться и как сценарии командной строки, и как модули функций, которые можно импортировать и использовать в других программах. Большинство сценариев в этом разделе также позволяют передавать им имя файла базы данных в аргументе командной строки, что дает возможность использовать в процессе разработки различные базы данных для различных целей – изменение одной из них не будет затрагивать другие.

Сценарии загрузки таблиц

Прежде чем увидеть сценарии в действии, познакомимся с их реализацией – вы можете свободно перепрыгивать вперед и назад, сопоставляя программный код с его поведением. Первым рассмотрим в примере 17.8 способ (не самый лучший) организации в виде сценария логики загрузки таблиц из предыдущего раздела.

Пример 17.8. PP4E\Dbase\Sql\loaddb1.py

```
"""
загружает таблицу из текстового файла со значениями, разделенными запятыми;
его эквивалентом является следующая непереносимая инструкция SQL:
load data local infile 'data.txt' into table people fields terminated by
',' "
"""


import sqlite3
conn = sqlite3.connect('dbase1')
curs = conn.cursor()

file = open('data.txt')
rows = [line.rstrip().split(',') for line in file]
for rec in rows:
    curs.execute('insert into people values (?, ?, ?)', rec)

conn.commit() # подтвердить изменения, если БД поддерживает транзакции
conn.close()  # close, __del__ вызовут откат, если изменения не подтверждены
```

В таком виде представленный в примере 17.8 сценарий является сценарием верхнего уровня, предназначенным для использования в строго определенном случае. Однако, приложив совсем немного усилий, его можно разложить на функции, которые можно импортировать и использовать в различных ситуациях. В примере 17.9 представлен намного более полезный модуль и сценарий командной строки.

Пример 17.9. PP4E\Dbase\Sql\loaddb.py

```
"""
загружает таблицу из текстового файла со значениями, разделенными запятыми:
обобщенная версия, готовая к многократному использованию
Функции доступны для импортирования;
```

```
порядок использования: loaddb.py dbfile? datafile? table?
"""


def login(dbfile):
    import sqlite3
    conn = sqlite3.connect(dbfile) # создать или открыть файл БД
    curs = conn.cursor()
    return conn, curs

def loaddb(curs, table, datafile, conn=None, verbose=True):
    file = open(datafile)                           # x,x,x\nx,x,x\n
    rows = [line.rstrip().split(',') for line in file] # [[x,x,x], [x,x,x]]
    rows = [str(tuple(rec)) for rec in rows]        # ["(x,x,x)","(x,x,x)"]
    for recstr in rows:
        curs.execute('insert into ' + table + ' values ' + recstr)
    if conn: conn.commit()
    if verbose: print(len(rows), 'rows loaded')

if __name__ == '__main__':
    import sys
    dbfile, datafile, table = 'dbase1', 'data.txt', 'people'
    if len(sys.argv) > 1: dbfile   = sys.argv[1]
    if len(sys.argv) > 2: datafile = sys.argv[2]
    if len(sys.argv) > 3: table    = sys.argv[3]
    conn, curs = login(dbfile)
    loaddb(curs, table, datafile, conn)
```

Обратите внимание, как здесь используются два генератора списков для конструирования строк со значениями записей для инструкции `insert` (применяемые преобразования приводятся в комментариях). Мы могли бы также использовать метод `executemany`, как это делали ранее, но нам требуется обеспечить универсальность и избежать применения жестко определенных шаблонов вставляемых полей – эта функция может использоваться для таблиц с любым количеством столбцов.

В этом файле также определена функция `login`, автоматизирующая начальное подключение к базе данных, – после ввода последовательности из четырех команд достаточное количество раз эта последовательность выглядит отличным кандидатом на оформление ее в виде функции. Кроме того, при этом снижается избыточность программного кода – в будущем при переходе на другую базу данных логику регистрации придется изменить только в одном месте, при условии, что повсюду используется эта функция `login`.

Сценарий вывода таблицы

После загрузки данных нам может потребоваться вывести их. Сценарий в примере 17.10 позволяет отображать результаты по мере их получения – он выводит таблицу целиком в простом виде (ее можно анализировать с помощью дополнительных инструментов) или форматированном (с помощью утилиты создания словарей записей, написанной

нами ранее). Обратите внимание, что здесь вычисляется максимальный размер имен полей, чтобы реализовать выравнивание в выражении-генераторе; ширина поля вывода в выражении форматирования строки определяется символом звездочки (*).

Пример 17.10. PP4E\Dbase\Sql\dumpdb.py

```
"""
отображает содержимое таблицы в виде простых кортежей
или в форматированном виде с именами полей
порядок использования из командной строки:
dumpdb.py dbname? table? [-] (dash=formatted display)
"""

def showformat(recs, sept=('-' * 40)):
    print(len(recs), 'records')
    print(sept)
    for rec in recs:
        maxkey = max(len(key) for key in rec)     # макс. длина ключа
        for key in rec:                           # или: \t
            print('%-*s => %s' % (maxkey, key, rec[key]))  # -ljust, *длина
        print(sept)

def dumpdb(cursor, table, format=True):
    if not format:
        cursor.execute('select * from ' + table)
        while True:
            rec = cursor.fetchone()
            if not rec: break
            print(rec)
    else:
        from makedicts import makedicts
        recs = makedicts(cursor, 'select * from ' + table)
        showformat(recs)

if __name__ == '__main__':
    import sys
    dbname, format, table = 'dbase1', False, 'people'
    cmdargs = sys.argv[1:]
    if '-' in cmdargs:         # форматировать, если '-' в арг. ком. строки
        format = True          # имя БД в другом аргументе ком. строки
        cmdargs.remove('-')
    if cmdargs: dbname = cmdargs.pop(0)
    if cmdargs: table  = cmdargs[0]

    from loaddb import login
    conn, curs = login(dbname)
    dumpdb(curs, table, format)
```

Раз уж мы заговорили об этом, напишем еще несколько вспомогательных сценариев для инициализации и очистки базы данных, чтобы нам не приходилось всякий раз вводить одни и те же строки в интерактив-

ной оболочке, когда нам потребуется начать все сначала. Сценарий в примере 17.11 полностью удаляет и воссоздает заново содержимое базы данных, чтобы вернуть ее в начальное состояние (мы вручную выполняли эту операцию в начале учебника).

Пример 17.11. PP4E\Dbase\Sql\makedb.py

```
"""
физически удаляет и воссоздает файлы базы данных
порядок использования: makedb.py dbname? tablename?
"""


import sys
if input('Are you sure?').lower() not in ('y', 'yes'):
    sys.exit()

dbname = (len(sys.argv) > 1 and sys.argv[1]) or 'dbase1'
table  = (len(sys.argv) > 2 and sys.argv[2]) or 'people'

from loaddb import login
conn, curs = login(dbname)
try:
    curs.execute('drop table ' + table)
except:
    print('database table did not exist')

command = 'create table %s (name char(30), job char(10), pay int(4))' % table
curs.execute(command)
conn.commit()          # подтверждение здесь может быть необязательным
print('made', dbname, table)
```

Следующий сценарий, представленный в примере 17.12, очищает базу данных, удаляя все строки в таблице, вместо полного удаления таблицы и повторного ее создания. Для тестирования можно использовать любой из подходов. Небольшое предупреждение: атрибут `rowcount` в SQLite не всегда отражает количество удаленных строк – подробности смотрите в руководстве по библиотеке.

Пример 17.12. PP4E\Dbase\Sql\cleardb.py

```
"""
удаляет все строки в таблице, но не удаляет таблицу в базе данных
порядок использования: cleardb.py dbname? tablename?
"""


import sys
if input('Are you sure?').lower() not in ('y', 'yes'):
    sys.exit()

dbname = sys.argv[1] if len(sys.argv) > 1 else 'dbase1'
table  = sys.argv[2] if len(sys.argv) > 2 else 'people'
```

```
from loaddb import login
conn, curs = login(dbname)
curs.execute('delete from ' + table)
#print(curs.rowcount, 'records deleted') # соед. будет закрыто
                                         # методом __del__
conn.commit()                            # иначе строки не будут удалены
```

Наконец, в примере 17.13 представлен инструмент командной строки, который выполняет запрос и выводит таблицу с результатами в форматированном виде. Это довольно короткий сценарий, потому что большую часть его задач мы уже автоматизировали. В значительной мере он просто объединяет существующие инструменты. Такова мощь повторного использования программного кода в Python.

Пример 17.13. PP4E\Dbase\Sql\querydb.py

```
"""
выполняет строку запроса, выводит результаты в форматированном виде
пример: querydb.py dbase1 "select name, job from people where pay > 50000"
"""


import sys
database, querystr = 'dbase1', 'select * from people'
if len(sys.argv) > 1: database = sys.argv[1]
if len(sys.argv) > 2: querystr = sys.argv[2]

from makedicts import makedicts
from dumpdb import showformat
from loaddb import login

conn, curs = login(database)
rows = makedicts(curs, querystr)
showformat(rows)
```

Использование сценариев

Далее приводится листинг сеанса, в котором эти сценарии запускаются из командной строки, чтобы проиллюстрировать их действие. Большинство файлов содержат функции, которые могут импортироваться и вызываться из других программ – при запуске сценарии просто отображают аргументы командной строки в аргументы функций. Для начала инициализируем тестовую базу данных и загрузим ее таблицу из текстового файла:

```
...\PP4E\Dbase\Sql> makedb.py testdb
Are you sure?y
database table did not exist
made testdb people

...\PP4E\Dbase\Sql> loaddb.py testdb data2.txt
3 rows loaded
```

Затем проверим результат наших действий с помощью утилиты вывода (используйте аргумент -, чтобы обеспечить форматированный вывод):

```
...\PP4E\Dbase\Sql> dumpdb.py testdb
('bob', 'developer', 80000)
('sue', 'music', 90000)
('ann', 'manager', 80000)

...\PP4E\Dbase\Sql> dumpdb.py testdb -
3 records
----------------------------------------
pay  => 80000
job  => developer
name => bob
----------------------------------------
pay  => 90000
job  => music
name => sue
----------------------------------------
pay  => 80000
job  => manager
name => ann
----------------------------------------
```

Сценарий производит исчерпывающий вывод – чтобы просмотреть определенные записи, передайте сценарию строку запроса в командной строке (следующие командные строки были разбиты, чтобы уместить их по ширине страницы):

```
...\PP4E\Dbase\Sql> querydb.py testdb
                            "select name, job from people where pay = 80000"
2 records
----------------------------------------
job  => developer
name => bob
----------------------------------------
job  => manager
name => ann
----------------------------------------

...\PP4E\Dbase\Sql> querydb.py testdb
                               "select * from people where name = 'sue'"
1 records
----------------------------------------
pay  => 90000
job  => music
name => sue
----------------------------------------
```

Теперь очистим базу данных и вновь наполним ее данными из файла. Сценарий очистки стирает все записи, но он не выполняет полную инициализацию базы данных:

```
...\PP4E\Dbase\Sql> cleardb.py testdb
Are you sure?y

...\PP4E\Dbase\Sql> dumpdb.py testdb -
0 records
----------------------------------------

...\PP4E\Dbase\Sql> loaddb.py testdb data.txt
5 rows loaded

...\PP4E\Dbase\Sql> dumpdb.py testdb
('bob', 'devel', 50000)
('sue', 'music', 60000)
('ann', 'devel', 40000)
('tim', 'admin', 30000)
('kim', 'devel', 60000)
```

В заключение ниже приводится пример выполнения трех запросов с этими новыми данными: они отбирают имена разработчиков, должности с зарплатой выше определенного уровня и записи с уровнем зарплаты выше заданного уровня, отсортированные по должности. Разумеется, эти операции можно было бы выполнить в интерактивном сеансе Python, но здесь мы бесплатно получаем в свое распоряжение значительную часть настроек и типового программного кода:

```
...\PP4E\Dbase\Sql> querydb.py testdb
                                "select name from people where job = 'devel'"
3 records
----------------------------------------
name => bob
----------------------------------------
name => ann
----------------------------------------
name => kim
----------------------------------------

...\PP4E\Dbase\Sql> querydb.py testdb
                                "select job from people where pay >= 60000"
2 records
----------------------------------------
job => music
----------------------------------------
job => devel
----------------------------------------

...\PP4E\Dbase\Sql> querydb.py testdb
                       "select * from people where pay >= 60000 order by job"
2 records
----------------------------------------
pay  => 60000
job  => devel
```

```
name => kim
---------------------------------------
pay  => 60000
job  => music
name => sue
---------------------------------------
```

Прежде чем двинуться дальше, немного дополнительной информации: сценарии в этом разделе иллюстрируют преимущества повторного использования программного кода, соответствуют поставленной цели (хотя бы частично представляют интерфейс баз данных) и служат моделью готовых утилит для работы с базами данных. Но они все еще не настолько универсальны, насколько могли бы быть. Например, поддержка сортировки могла бы быть полезным расширением сценария вывода. Мы могли бы обобщить эти сценарии еще больше, добавив поддержку дополнительных возможностей, тем не менее, рано или поздно нам может потребоваться вернуться к вводу команд SQL на стороне клиента – отчасти потому, что SQL является языком и должен поддерживать достаточный уровень общности. Дальнейшее расширение этих сценариев я оставляю в качестве самостоятельного упражнения. Изменяйте их реализацию по своему усмотрению, в конце концов, это – Python.

Ресурсы SQL

Несмотря на простоту примеров, которые мы видели в этом разделе, приемы, использованные в них, легко можно распространить на другие базы данных и приложения. Например, системы баз данных SQL, такие как MySQL, можно использовать на веб-сайтах, которые мы изучали в предыдущей части книги, для сохранения информации о состоянии страниц, а также других данных. Благодаря тому, что система MySQL (среди прочих) поддерживает возможность управления крупными базами данных и одновременное изменение информации несколькими клиентами, ее вполне естественно использовать для реализации веб-сайтов.

Об интерфейсах баз данных можно рассказать больше, чем это сделано здесь, но дополнительную документацию по API легко получить из Интернета. Чтобы отыскать полную спецификацию API баз данных, выполните поиск в Интернете по фразе «Python Database API».[1] Вы найдете формальное определение API, которое является простым текстовым файлом с описанием PEP (Python Enhancement Proposal – предложение по развитию Python), на основе которого велось обсуждение API.

Наилучшим ресурсом информации по расширениям баз данных на сегодняшний день является, вероятно, домашняя страница группы интересов Python по базам данных (SIG). Зайдите на страницу *http://www.python.org*, щелкните на ссылке `Community` вверху и перейдите на стра-

[1] Выполнив поиск по строке «API-спецификация баз данных языка Python», можно найти спецификацию на русском языке. – *Прим. перев.*

ницу группы по базам данных, или выполните поиск по сайту. Там вы найдете документацию по API (официально поддерживаемую), ссылки на модули расширений баз данных конкретных поставщиков и многое другое. И, как обычно, за дополнительными инструментами и расширениями сторонних разработчиков обращайтесь на веб-сайт PyPI или выполните поиск в Интернете.

ORM: механизмы объектно-реляционного отображения

В этой главе мы познакомились с объектно-ориентированными базами данных, которые позволяют хранить объекты Python, а также с базами данных SQL, хранящими данные в таблицах. Как оказывается, существует еще один класс систем, которые служат мостом между мирами объектов и таблиц, и которые я упоминал ранее в этой главе: механизмы ORM позволяют подружить модель классов Python с таблицами в реляционных базах данных. Они объединяют мощь систем реляционных баз данных с простотой объектно-ориентированного синтаксиса Python; они позволяют использовать базы данных SQL и сохранять в них данные, которые в сценариях выглядят как объекты Python.

В настоящее время среди открытых систем сторонних разработчиков имеются два лидера, реализующих такое отображение: SQLObject и SQLAlchemy. Обе являются слишком сложными системами, чтобы их можно было достаточно полно рассмотреть в этой книге, поэтому ищите соответствующую документацию в Интернете для получения более полного представления о них (в настоящее время имеются даже книги, специально посвященные SQLAlchemy). Кроме того, к моменту написания этих слов ни одна из них еще не была полностью готова для использования в Python 3.X, поэтому я не могу в этой книге привести примеры работы с ними.

Однако, чтобы дать вам некоторое представление о модели ORM, ниже кратко рассказывается о том, как можно использовать систему SQLObject для создания и обработки записей в базе данных. В двух словах, SQLObject отображает:

- Классы Python в таблицы базы данных.
- Экземпляры классов Python в записи внутри таблиц.
- Атрибуты экземпляров в столбцы записей.

Например, чтобы создать таблицу, необходимо определить класс, атрибуты которого будут определять столбцы, и вызвать метод создания (следующий фрагмент взят из более полного примера, который приводится на веб-сайте SQLObject):

```
from sqlobject import *
sqlhub.processConnection = connectionForURI('sqlite:/:memory:')
```

```
class Person(SQLObject): # класс: описывает таблицу
    first = StringCol()  # атрибуты класса: описывают столбцы
    mid   = StringCol(length=1, default=None)
    last  = StringCol()

Person.createTable()      # создать таблицу в базе данных
```

После создания экземпляра в базу данных автоматически добавляется новая запись, а операции обращения к атрибутам автоматически отображаются в операции с соответствующими столбцами записи в таблице:

```
p = Person(first='Bob', last='Smith') # новый экземпляр: создает
                                      # новую запись
p                                     # выведет все атрибуты по именам

p.first                               # атрибуты: извлечет значение столбца
p.mid = 'M'                           # атрибуты: обновит запись
```

Существующие записи/экземпляры могут извлекаться вызовом методов, и имеется возможность присваивать значения сразу нескольким столбцам/атрибутам в одной операции:

```
p2 = Person.get(1)        # извлечь существующую запись/экземпляр: p2 это p
p.set(first='Tom', last='Jones') # изменить 2 атрибута/поля в одной операции
```

Кроме того, имеется возможность выбирать записи по значениям столбцов, создав объект запроса и выполнив его:

```
ts = Person.select(Person.q.first=='Tom') # запрос: выборка
                                          # по значению столб.
list(ts)                              # выполнить запрос: список экземпл.
tjs = Person.selectBy(first='Tom', last='Jones') # альтернативная форма
                                                 # запроса (по И)
```

Естественно, в этих примерах мы коснулись лишь малой части имеющихся функциональных возможностей. Однако даже на этом уровне сложности невозможно не заметить главную хитрость – механизм SQLObject автоматически выполняет все запросы SQL, необходимые для извлечения, сохранения и запроса таблиц и записей, подразумеваемых реализацией классов Python здесь. Все это позволяет использовать мощь промышленных реляционных баз данных, используя при этом знакомый синтаксис классов Python для обработки хранимых данных в сценариях Python.

Порядок использования ORM SQLAlchemy, конечно же, существенно отличается, но его функциональные возможности и получаемый конечный результат аналогичны. За дополнительной информацией о механизмах ORM для Python обращайтесь через свои поисковые системы к Интернету. Вы можете также поближе познакомиться с такими системами и их задачами в некоторых крупных веб-фреймворках. Например, в состав Django входит механизм ORM, являющийся вариацией на эту тему.

PyForm: просмотр хранимых объектов (внешний пример)

Вместо того чтобы заниматься дополнительными деталями интерфейса баз данных, которые можно легко найти в Интернете, в завершение этой главы я направлю вас к дополнительному примеру, демонстрирующему, как можно соединить технологии графических интерфейсов, с которыми мы познакомились ранее в книге, с технологиями постоянного хранения, представленными в этой главе. Этот пример называется PyForm – приложение с графическим интерфейсом на основе tkinter, предназначенное для просмотра и редактирования таблиц записей:

- Таблицы, которые просматривает программа, могут быть хранилищами `shelve`, файлами DBM, словарями в памяти или любыми другими объектами, которые имеют вид словаря.
- Записи в просматриваемых таблицах могут быть экземплярами классов, простыми словарями, строками или любыми другими объектами, которые могут быть преобразованы в словари и обратно.

Этот пример демонстрирует создание графического интерфейса и организацию постоянного хранения, однако он также иллюстрирует технологию проектирования программ на языке Python. Чтобы сохранить реализацию простой и не зависящей от типов данных, графический интерфейс PyForm написан в предположении, что таблицы имеют вид словаря словарей. Для поддержки различных типов таблиц и записей PyForm полагается на отдельные классы-оболочки, преобразующие таблицы и записи в предполагаемый протокол:

- На верхнем уровне таблицы трансляция происходит просто – хранилища модуля `shelve`, файлы DBM и словари в оперативной памяти имеют одинаковый интерфейс, основанный на ключах.
- Для вложенных записей в графическом интерфейсе предполагается, что хранимые элементы тоже имеют интерфейс словарей, но операции со словарями перехватываются классами, чтобы сделать записи совместимыми с протоколом PyForm. Записи, хранящиеся в виде строк, преобразуются в объекты словарей и обратно при выборке и сохранении. Записи, хранящиеся в виде экземпляров классов, транслируются в словари атрибутов и обратно. Более специальные виды трансляции можно добавить в новых классах-оболочках таблиц.

В результате PyForm можно использовать для просмотра и редактирования большого числа разных типов таблиц, несмотря на предполагаемый интерфейс словаря. При просмотре с помощью PyForm хранилищ `shelve` и файлов DBM изменения в таблицах, производимые в графическом интерфейсе, оказываются постоянными – они сохраняются в файлах хранилищ. При просмотре хранилища `shelve` с экземплярами классов PyForm становится, по существу, клиентом с графическим интерфейсом для простой объектной базы данных, созданной с помощью стан-

дартных средств Python, обеспечивающих постоянное хранение. Чтобы просмотреть и изменить хранимые объекты с помощью PyForm, например, достаточно использовать следующий программный код:

```
import shelve
from formgui import FormGui            # после создания хранилища
db = shelve.open('../data/castfile')   # повторно открыть файл shelve
FormGui(db).mainloop()                 # и перейти к просмотру существующего
                                       # хранилища словарей
```

Для просмотра и изменения хранилища с экземплярами импортированного класса `Actor` можно использовать такой программный код:

```
from PP4E.Dbase.testdata import Actor
from formgui import FormGui              # выполнять в каталоге TableBrowser
from formtable import ShelveOfInstance

testfile = '../data/shelve'                # имя внешнего файла
table = ShelveOfInstance(testfile, Actor)  # обернуть хранилище
                                           # в объект Table
FormGui(table).mainloop()
table.close()                              # для некоторых dbm необходимо
                                           # явно вызывать метод close
```

На рис. 17.1 показано, как выглядит графический интерфейс приложения, выполняющегося под управлением Python 3.1 в Windows 7, при просмотре хранилища экземпляров классов. Этот сеанс работы с PyForm был запущен командой, описанной в программном коде самопроверки модуля `formtable`: `formtable.py shelve 1`, без аргумента `1` (или с аргументом `0`), чтобы избежать повторной инициализации хранилища в начале сеанса и сохранить изменения.

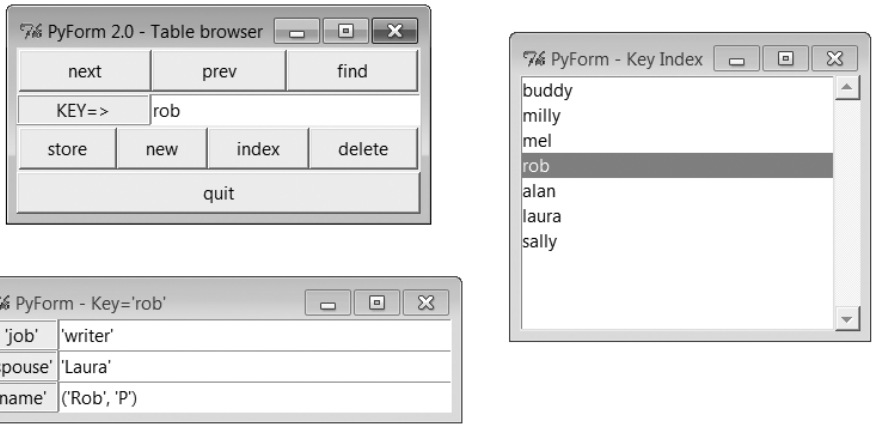


Рис. 17.1. *PyForm отображает объекты класса Actor, находящиеся в хранилище shelve*

Приложение PyForm также можно запустить из программы PyDemos, которую мы рассматривали в главе 10, однако в этом случае оно не будет сохранять изменения. Запустите пример на своем компьютере, чтобы получить более полное представление о том, как он действует. Хотя программа PyForm и не является универсальным средством просмотра хранимых объектов Python, тем не менее, она может служить простым интерфейсом к объектным базам данных.

Из-за нехватки места в этом издании я опущу исходный программный код этого примера и его описание. Чтобы поближе познакомиться с PyForm, смотрите содержимое следующего каталога в пакете с примерами к этой книге, информация о котором есть в предисловии:

```
C:\...\PP4E\Dbase\TableBrowser
```

Загляните, в частности, в подкаталог *Documentation*, где вы найдете файл PDF с обзором PyForm из третьего издания книги. Исходный программный код PyForm был адаптирован для работы под управлением Python 3.X, однако в обзоре приводится программный код для 2.X из третьего издания. А теперь перейдем к следующей главе и к следующей теме, касающейся инструментов: реализации структур данных.

18

Структуры данных

«Розы – красные, фиалки – голубые; списки изменяемы, а также и класс Foo»

Структуры данных составляют основу большинства программ, хотя программисты на языке Python могут зачастую совсем не беспокоиться об этом. Их спокойствие оправданно, потому что Python предоставляет богатый набор встроенных и оптимизированных типов, облегчающих работу со структурированными данными: списки, строки, кортежи, словари, множества и другие. В простых системах этих типов обычно достаточно. Словари, например, делают многие классические алгоритмы поиска ненужными в Python, а списки избавляют от большой части работы, необходимой для поддержки коллекций в языках более низкого уровня. Теми и другими настолько просто пользоваться, что обычно не приходится задумываться над ними.

Но в более сложных приложениях может потребоваться вводить собственные, более сложные типы, чтобы справиться с дополнительными требованиями или особыми ситуациями. В этой главе мы рассмотрим несколько реализаций более сложных структур данных: множеств, стеков, графов и других. Как будет показано, структуры данных принимают в Python вид новых типов объектов, интегрированных в модель типов языка. То есть объекты, которые создаются на языке Python, становятся законченными *типами данных* – для использующих их сценариев они выглядят точно так же, как списки, числа и словари.

Хотя примеры в этой главе иллюстрируют более сложную технику программирования, в них также подчеркнута поддержка в Python создания *повторно используемого* программного обеспечения. Реализации объектов, написанные с помощью классов и модулей, естественным об-

разом становятся полезными компонентами и могут быть использованы в любой импортирующей их программе. В сущности, мы будем создавать *библиотеки* инструментов для работы со структурами данных, даже не планируя этого делать.

Кроме того, большинство примеров в этой главе представляют собой чистый программный код на языке Python (и, по крайней мере, для тех, кто читал все по порядку, некоторые из них могут показаться относительно простыми в сравнении с примерами из предыдущих глав), однако они также представляют собой основу для обсуждения проблем производительности и могут служить подсказками к главе 20. С самой общей точки зрения новые объекты Python могут быть реализованы на Python или интегрированном языке, таком как C. При определении типов в языке C используются шаблоны, похожие на используемые здесь.

Наконец, мы также увидим, что зачастую вместо частных решений в этой области могут использоваться встроенные возможности Python. Хотя реализация собственных структур данных иногда бывает просто необходима и они могут обеспечивать определенные удобства с точки зрения сопровождения программного кода и его развития, тем не менее в языке Python они могут не играть такой доминирующей роли, как в языках, менее дружественных по отношению к программистам.

Реализация стеков

Стек является распространенной и простой структурой данных, используемой в целом ряде приложений: обработке языков, поиске на графах и так далее. Например, вычисление выражений в калькуляторе с графическим интерфейсом, который будет представлен в следующей главе, в значительной степени опирается на манипулирование стеками. Языки программирования в целом обычно реализуют вызовы функций как операции со стеком, на котором запоминается местоположение в программе, откуда следует продолжить выполнение после возврата из функции. Стеки также могут использоваться при синтаксическом анализе документов XML: они естественным образом подходят для отслеживания движения по произвольно вложенным конструкциям.

Вкратце, стек представляет собой коллекцию объектов, построенную по принципу «последним пришел, первым ушел»: элемент, добавленный в коллекцию последним, всегда оказывается следующим удаляемым. В отличие от очередей, которые мы использовали для организации взаимодействий с потоками выполнения и которые добавляют и удаляют объекты с противоположных концов, все операции со стеком происходят на его вершине. Клиенты используют следующие операции со стеками:

- Проталкивание элементов на вершину стека (pushing).
- Выталкивание элементов с вершины стека (popping).

В зависимости от требований клиентов, могут также иметься средства решения таких задач, как проверка наличия элементов в стеке, извлечение верхнего элемента без его выталкивания, обход элементов стека, проверка наличия в стеке некоторого элемента и так далее.

Встроенные возможности

В Python для реализации стека часто достаточно простого списка: так как списки могут изменяться непосредственно, это позволяет добавлять или удалять элементы с начала (слева) или с конца (справа). В табл. 18.1 приводятся различные встроенные операции, которые могут использоваться для реализации стека на основе списка Python, в зависимости от того, является ли «вершина» стека первым или последним узлом списка. В этой таблице строка `'b'` является верхним элементом стека.

Таблица 18.1. Стеки в виде списков

Операция	Вершина в конце списка	Вершина в начале списка	Вершина в начале списка
New	`stack=['a', 'b']`	`stack=['b', 'a']`	`stack=['b', 'a']`
Push	`stack.append('c')`	`stack.insert(0,'c')`	`stack[0:0]=['c']`
Pop	`top = stack[-1]`; `del stack[-1]`	`top = stack[0]`; `del stack[0]`	`top = stack[0]`; `stack[:1] = []`

Интересно отметить, что со временем в языке Python появился еще более удобный метод `pop` списков, предназначенный в паре с методом `append` для реализации стеков и других распространенных структур данных, таких как очереди, что привело к появлению еще более простых способов реализации, перечисленных в табл. 18.2.

Таблица 18.2. Стеки в виде списков, альтернативные реализации операций

Операция	Вершина в конце списка	Вершина в начале списка
New	`stack=['a', 'b']`	`stack=['b', 'a']`
Push	`stack.append('c')`	`stack.insert(0,'c')`
Pop	`top = stack.pop()`	`top = stack.pop(0)`

По умолчанию метод `pop` извлекает последний элемент со смещением –1, и затем удаляет его из списка. При вызове с аргументом метод `pop` удаляет из списка и возвращает элемент с указанным смещением – вызов `list.pop(-1)` эквивалентен вызову `list.pop()`. Операции, выполняющие изменения на месте, такие как `append`, `insert`, `del` и `pop`, не создают новый список, поэтому они действуют быстро (производительность операций может зависеть от того, какой конец списка считается «вершиной»

стека, а это, в свою очередь, зависит от текущей реализации списков, а также от способов измерения производительности, которые мы исследуем позднее). Очереди реализуются похожим образом, но выталкивание элементов производится с противоположного конца списка.

Имеются также и другие встроенные схемы реализации. Например, инструкция `del stack[:1]` является еще одним способом удаления первого элемента стека на основе списка. В зависимости от того, какой конец списка считается вершиной стека, для извлечения и удаления элемента на вершине могут использоваться следующие операции над последовательностями (ценой создания каждый раз нового объекта списка):

```
# вершиной является начало списка
top, stack = stack[0], stack[1:]    # Python 1.X+
top, *stack = stack                 # Python 3.X

# вершиной является конец списка
stack, top = stack[:-1], stack[-1]  # Python 1.X+
*stack, top = stack                 # Python 3.X
```

Зачем же тогда реализовывать другие операции при таком богатстве встроенных операций со стеком? С одной стороны, они служат простым и привычным контекстом для исследования понятий структур данных в этой книге. Однако, что более важно, имеется более практический повод. Реализация на основе списка действует, и относительно быстро, но при этом основанные на стеках программы привязываются к избранному представлению стека: все операции со стеком будут жестко определены. Если позднее потребуется изменить представление стека или расширить набор его базовых операций, мы попадем в неприятную ситуацию – придется исправлять все программы, использующие стеки, и все инструкции, реализующие доступ к ним.

Например, чтобы добавить логику, которая контролирует количество операций над стеком, выполняемых программой, пришлось бы добавлять программный код рядом с каждой операцией со стеком. В большой системе решение такой задачи может оказаться нетривиальным. Как будет показано в главе 20, если стеки окажутся узким местом с точки зрения эффективности системы, можно перейти на стеки, реализованные на языке C. Как правило, жестко определенные операции над встроенными структурами данных требуют «ручного» вмешательства при внесении изменений в будущем.

Как мы увидим позднее, встроенные типы, такие как списки, в действительности являются объектами, напоминающими классы, от которых можно порождать подклассы, чтобы настраивать их поведение. Однако это лишь частичное решение – не предполагая изменения в будущем и не создавая экземпляры подкласса, мы все еще можем столкнуться с проблемами, если будем использовать встроенные операции над списками непосредственно и если позднее нам потребуется, чтобы они выполняли какие-либо дополнительные действия.

Модуль stack

Более удачным решением таких проблем может оказаться *инкапсуляция*, то есть обертывание, реализаций стеков в интерфейсы с помощью инструментов Python – для организации повторного использования программного кода. Пока клиенты продолжают использовать интерфейсы, мы легко можем изменить реализацию этих интерфейсов произвольным образом и избежать необходимости изменять все инструкции обращения к ним. Начнем с реализации стека в виде модуля, содержащего список Python вместе с функциями, действующими над ним. В примере 18.1 представлена одна из возможных реализаций.

Пример 18.1. PP4E\Dstruct\Basic\stack1.py

```
"модуль реализации стека"

stack = []                          # при первом импортировании
class error(Exception): pass        # локальные исключения, stack1.error

def push(obj):
    global stack                    # 'global', чтобы иметь возм. изменять
    stack = [obj] + stack           # добавить элемент в начало

def pop():
    global stack
    if not stack:
        raise error('stack underflow') # возбудить локальное исключение
    top, *stack = stack               # удалить элемент в начале
    return top

def top():
    if not stack:                     # возбудить локальное исключение
        raise error('stack underflow') # или позволить возбудить IndexError
    return stack[0]

def empty():      return not stack     # стек пуст?
def member(obj):  return obj in stack  # элемент имеется в стеке?
def item(offset): return stack[offset] # элемент стека по индексу
def length():     return len(stack)    # количество элементов на стеке
def dump():       print('<Stack:%s>' % stack)
```

Этот модуль создает объект списка (`stack`) и экспортирует функции для управления доступом к нему. Стек объявляется глобальным в функциях, которые его изменяют, но не в тех, которые только читают его. В модуле также объявлен объект ошибки (`error`), с помощью которого можно перехватывать исключения, возбуждаемые локально в этом модуле. Некоторые ошибки стека являются встроенными исключениями: метод `item` вызывает `IndexError` при выходе индексов за пределы списка.

Большинство функций в модуле `stack` просто передают выполнение операции встроенному списку, представляющему стек. В действительно-

сти модуль служит просто оболочкой, в которую заключен список Python. Но этот дополнительный слой логики интерфейса обеспечивает независимость клиентов от фактической реализации стека, поэтому в дальнейшем можно будет изменить стек, не затрагивая его клиентов.

Как всегда, лучший способ разобраться в таком программном коде – посмотреть на него в деле. Ниже приводится листинг интерактивного сеанса, иллюстрирующий интерфейсы модуля, – он реализует стек, способный принимать произвольные объекты Python:

```
C:\...\PP4E\Dstruct\Basic> python
>>> import stack1
>>> stack1.push('spam')
>>> stack1.push(123)
>>> stack1.top()
123
>>> stack1.stack
[123, 'spam']
>>> stack1.pop()
123
>>> stack1.dump()
<Stack:['spam']>
>>> stack1.pop()
'spam'
>>> stack1.empty()
True
>>> for c in 'spam': stack1.push(c)
...
>>> while not stack1.empty():
...     print(stack1.pop(), end=' ')
...
m a p s
>>>
>>> stack1.pop()
stack1.error: stack underflow
```

Другие операции действуют аналогично, но главное, на что здесь нужно обратить внимание, – все операции со стеком являются *функциями* модуля. Например, можно совершить обход стека, но нужно использовать счетный цикл и вызывать функцию обращения по индексу (`item`). Ничто не мешает клиенту обращаться (и даже изменять) стек `stack1.stack` непосредственно, но при этом теряется весь смысл таких интерфейсов, как показано ниже:

```
>>> for c in 'spam': stack1.push(c)
...
>>> stack1.dump()
<Stack:['m', 'a', 'p', 's']>
>>>
>>> for i in range(stack1.length()):
... print(stack1.item(i), end=' ')
...
```

```
m a p s
>>>
```

Класс Stack

Наибольшим, пожалуй, недостатком стека, основанного на модуле, является поддержка только одного объекта стека. Все клиенты модуля `stack` фактически пользуются одним и тем же стеком. Иногда такая особенность нужна: стек может служить объектом памяти, совместно используемой несколькими модулями. Но для реализации стека как настоящего типа данных необходимо использовать классы.

Для иллюстрации определим полнофункциональный *класс* стека. Класс `Stack`, представленный в примере 18.2, определяет новый тип данных с разнообразным поведением. Как и модуль, для хранения помещаемых на стек объектов класс использует список Python. Но на этот раз каждый экземпляр имеет собственный список. В классе определены как «обычные» методы, так и специальные методы с особыми именами, реализующие стандартные операции над типом данных. Комментарии в программном коде описывают специальные методы.

Пример 18.2. PP4E\Dstruct\Basic\stack2.py

```
"класс стека, позволяющий создавать множество экземпляров"

class error(Exception): pass    # при импортировании: локальное исключение

class Stack:
    def __init__(self, start=[]):  # self - объект экземпляра
        self.stack = []         # start - любая последовательность: stack...
        for x in start: self.push(x)
        self.reverse()          # переупорядочивает операции push
                                # в обратном порядке

    def push(self, obj):                 # методы: подобно модулю + self
        self.stack = [obj] + self.stack  # вершина в начале списка

    def pop(self):
        if not self.stack: raise error('underflow')
        top, *self.stack = self.stack
        return top

    def top(self):
        if not self.stack: raise error('underflow')
            return self.stack[0]

    def empty(self):
        return not self.stack              # instance.empty()

    # методы перегрузки операторов
    def __repr__(self):
```

```
        return '[Stack:%s]' % self.stack   # print, repr(),..

    def __eq__(self, other):
        return self.stack == other.stack   # '==', '!='?

    def __len__(self):
        return len(self.stack)             # len(instance), not instance

    def __add__(self, other):
        return Stack(self.stack + other.stack) # instance1 + instance2

    def __mul__(self, reps):
        return Stack(self.stack * reps)    # instance * reps

    def __getitem__(self, offset):         # смотрите также __iter__
        return self.stack[offset]          # instance[i], [i:j], in, for

    def __getattr__(self, name):
        return getattr(self.stack, name)   # instance.sort()/reverse()/..
```

Теперь можно создавать отдельные экземпляры обращением к имени класса Stack как к функции. Во многих отношениях операции в классе Stack реализованы точно так же, как в модуле stack из примера 18.1. Но здесь доступ к стеку выполняется через аргумент self, объект экземпляра. Каждый экземпляр имеет свой атрибут stack, который ссылается на собственный список экземпляра. Кроме того, экземпляры стеков создаются и инициализируются в методе конструктора __init__, а не при импортировании модуля. Создадим несколько стеков и посмотрим, как все это действует на практике:

```
>>> from stack2 import Stack
>>> x = Stack()                # создать объект стека, поместить
                               # в него данные
>>> x.push('spam')
>>> x.push(123)
>>> x                          # __repr__ выведет содержимое стека
[Stack:[123, 'spam']]

>>> y = Stack()                # два независимых объекта стека
>>> y.push(3.1415)             # они не используют совместно данные
>>> y.push(x.pop())
>>> x, y
([Stack:['spam']], [Stack:[123, 3.1415]])

>>> z = Stack()                # третий независимый объект стека
>>> for c in 'spam': z.push(c)
...
>>> while z:                   # __len__ проверит истинность стека
...     print(z.pop(), end=' ')
...
```

```
m a p s
>>>

>>> z = x + y                    # __add__ реализует операцию + над стеком
>>> z                            # хранит три объекта разных типов
[Stack:['spam', 123, 3.1415]]

>>> for item in z:               # __getitem__ используется в итерациях
...     print(item, end=' ')
...
spam 123 3.1415
>>>

>>> z.reverse()                  # вызов __getattr__ передается списку
>>> z
[Stack:[3.1415, 123, 'spam']]
```

Подобно спискам и словарям класс `Stack` определяет методы и операторы для обработки операций выражений и обращения к атрибутам. Кроме того, он определяет специальный метод `__getattr__` для перехвата обращений к атрибутам, не определенным в классе, и передачи их обернутому объекту списка (для поддержки методов списка: `sort`, `append`, `reverse` и так далее). Многие операции модуля превратились в операции в классе. В табл. 18.3 показаны эквивалентные операции модуля и класса (колонки 1 и 2) и приводится метод класса, который выполняет каждую из них (колонка 3).

Таблица 18.3. Сравнение операций модуля/класса

Операции в модуле	**Операции в классе**	**Метод класса**
`module.empty()`	`not instance`	`__len__`
`module.member(x)`	`x in instance`	`__getitem__`
`module.item(i)`	`instance[i]`	`__getitem__`
`module.length()`	`len(instance)`	`__len__`
`module.dump()`	`print(instance)`	`__repr__`
`range()` *счетчик циклов*	`for x in instance`	`__getitem__`
выполнение итераций вручную	`instance + instance`	`__add__`
`module.stack.reverse()`	`instance.reverse()`	`__getattr__`
`module.push/pop/top`	`instance.push/pop/top`	`push/pop/top`

В сущности, классы позволяют расширять набор встроенных типов Python с помощью многократно используемых типов, реализуемых в модулях Python. Основанные на классах типы могут использоваться так же, как встроенные типы: в зависимости от того, какие методы операций в них определены, классы могут реализовывать числа, отображе-

ния и последовательности и быть изменяемыми или нет. Основанные на классах типы могут также занимать положение, промежуточное между этими категориями.

Индивидуальная настройка: мониторинг производительности

Как мы видели, классы поддерживают возможность создания нескольких экземпляров и лучше интегрируются с объектной моделью Python благодаря определению методов операторов. Одной из других важных причин использования классов является возможность дальнейшего расширения и настройки. Реализуя стеки с помощью класса, можно в дальнейшем добавлять подклассы, точнее определяющие реализацию в соответствии с возникшими требованиями. В действительности, эта причина часто является основной, из-за которой предпочтение отдается собственным классам, а не встроенным альтернативам.

Допустим, например, что мы стали использовать класс `Stack` из примера 18.2, но столкнулись с проблемами производительности. Один из способов выявить узкие места состоит в том, чтобы оснастить структуры данных логикой, ведущей статистику использования, которую можно проанализировать после выполнения клиентских приложений. Так как `Stack` является классом, такую новую логику можно добавить в подклассе, не трогая исходный модуль стека (или его клиентов). Подкласс в примере 18.3 расширяет класс `Stack`, вводя слежение за суммарной частотой операций push/pop и регистрируя максимальный размер каждого экземпляра.

Пример 18.3. PP4E\Dstruct\Basic\stacklog.py

```
"расширяет стек возможностью сбора статистики об использовании данных"

from stack2 import Stack      # расширяет импортируемый класс Stack

class StackLog(Stack):        # подсчитывает операции push/pop, макс. размер
    pushes = pops = 0               # разделяемые/статические члены класса
    def __init__(self, start=[]):   # могут быть переменными модуля
        self.maxlen = 0
        Stack.__init__(self, start)

    def push(self, object):
        Stack.push(self, object)                  # выполнить операцию push
        StackLog.pushes += 1                      # общая статистика
        self.maxlen = max(self.maxlen, len(self)) # статистика экземпляра

    def pop(self):
        StackLog.pops += 1                       # общий счетчик
        return Stack.pop(self)                   # не 'self.pops': экземпляр

    def stats(self):
```

```
        return self.maxlen, self.pushes, self.pops    # вернуть счетчики
                                                      # экземляра
```

Этот подкласс действует так же, как и оригинальный класс `Stack`: в него просто добавлена логика мониторинга. Новый метод `stats` возвращает кортеж со статистикой экземпляра:

```
>>> from stacklog import StackLog
>>> x = StackLog()
>>> y = StackLog()                  # создать два объекта стека
>>> for i in range(3): x.push(i)    # и поместить в них объекты
...
>>> for c in 'spam': y.push(c)
...
>>> x, y                            # вызов унаследованного метода __repr__
([Stack:[2, 1, 0]], [Stack:['m', 'a', 'p', 's']])
>>> x.stats(), y.stats()
((3, 7, 0), (4, 7, 0))
>>>
>>> y.pop(), x.pop()
('m', 2)
>>> x.stats(), y.stats()            # моя макс. длина, все операции push,
((3, 7, 2), (4, 7, 2))              # все операции pop
```

Обратите внимание на использование атрибутов *класса* для регистрации суммарных количеств операций помещения в стек и выталкивания из стека и атрибутов *экземпляра* для хранения максимальной длины каждого экземпляра. Добавляя атрибуты к разным объектам, можно расширять или сужать область их действия.

Оптимизация: стеки в виде деревьев кортежей

Одной из замечательных особенностей обертывания объектов в классы является возможность изменения базовой реализации без нарушения работы остальной части программы. Например, в будущем можно выполнить оптимизацию с минимальным воздействием – интерфейс сохранится неизменным, даже если изменится внутреннее устройство. Существуют разные способы реализации стеков, обладающих различной эффективностью. До сих пор для обмена данными с нашими стеками использовались операции получения среза и конкатенации. Это довольно неэффективно: обе операции создают копии заключенного в оболочку объекта списка. Для больших стеков такой способ будет отнимать много времени.

Одним из способов добиться ускорения является полное изменение базовой структуры данных. Например, помещаемые на стек объекты можно хранить в двоичном дереве кортежей: каждый элемент можно записать как пару `(object, tree)`, где `object` – это элемент, помещенный на стек, а `tree` – это либо другой кортеж, определяющий остальной стек, либо `None` для обозначения пустого стека. Стек элементов `[1,2,3,4]` внутренне будет храниться как дерево кортежей `(1,(2,(3,(4,None))))`.

Такое представление, основанное на кортежах, аналогично понятию списков в семействе языков Lisp: объект слева – это car (голова списка), а остальная часть дерева справа – это cdr (остальная часть списка). Так как при помещении элемента в стек или снятии со стека мы добавляем или удаляем только верхний кортеж, использование такой структуры позволяет избежать копирования всего стека. Для больших стеков преимущество может оказаться весьма существенным. Эти идеи реализованы в следующем классе, представленном в примере 18.4.

Пример 18.4. PP4E\Dstruct\Basic\stack3.py

```
"оптимизация за счет использования дерева кортежей"

class Stack:
    def __init__(self, start=[]):      # инициализируется любой последоват.
        self.stack = None              # даже другими (быстрыми) стеками
        for i in range(-len(start), 0):
            self.push(start[-i - 1])   # втолкнуть в обратном порядке

    def push(self, node):              # дерево растет 'вверх/влево'
        self.stack = node, self.stack  # новый корневой кортеж: (node, tree)

    def pop(self):
        node, self.stack = self.stack  # удалить корневой кортеж
        return node                    # TypeError, если пустой

    def empty(self):
        return not self.stack          # 'None'?

    def __len__(self):                 # для операций: len, not
        len, tree = 0, self.stack
        while tree:
            len, tree = len+1, tree[1] # обойти правые поддеревья
        return len

    def __getitem__(self, index):      # для операций: x[i], in, for
        len, tree = 0, self.stack
        while len < index and tree:    # обход/подсчет узлов
            len, tree = len+1, tree[1]
        if tree:
            return tree[0]             # IndexError, при выходе за границы
        else:
            raise IndexError()         # остановка для 'in' и 'for'

    def __repr__(self):
        return '[FastStack:' + repr(self.stack) + ']'
```

Метод __getitem__ этого класса обрабатывает операции обращения по индексу, проверки in и итерации в цикле for, как и прежде (в случае отсутствия метода __iter__), но в этой версии нужно выполнить обход дерева, чтобы найти узел по индексу. Обратите внимание, что это не под-

класс первоначального класса Stack. Так как здесь почти все операции реализованы иначе, от наследования пользы мало. Но клиенты, использующие только общие для обоих классов операции, могут использовать их взаимозаменяемым образом – чтобы перейти на другую реализацию, нужно лишь импортировать класс стека из другого модуля. Ниже приводится листинг сеанса использования этой версии стека – если придерживаться только операций проталкивания, выталкивания, индексирования и итераций, эта версия, по существу, неотличима от первоначальной:

```
>>> from stack3 import Stack
>>> x = Stack()
>>> y = Stack()
>>> for c in 'spam': x.push(c)
...
>>> for i in range(3): y.push(i)
...
>>> x
[FastStack:('m', ('a', ('p', ('s', None))))]
>>> y
[FastStack:(2, (1, (0, None)))]

>>> len(x), x[2], x[-1]
(4, 'p', 'm')
>>> x.pop()
'm'
>>> x
[FastStack:('a', ('p', ('s', None)))]
>>>
>>> while y: print(y.pop(), end=' ')
...
2 1 0
>>>
```

Оптимизация: непосредственная модификация списка в памяти

В предыдущем разделе мы пытались повысить скорость выполнения операций проталкивания в стек и выталкивания со стека, применив иную организацию данных, однако скорость работы стека можно также повысить, вернувшись к объекту списка Python и воспользовавшись его изменяемостью. Поскольку списки могут изменяться непосредственно в памяти, их можно модифицировать быстрее, чем во всех предыдущих примерах. Операции непосредственной модификации списков, такие как append, могут приводить к осложнениям, когда к списку обращаются из нескольких мест. Но так как список внутри объекта стека не предназначается для прямого доступа, то здесь, вероятно, мы защищены.

Модуль, представленный в примере 18.5, демонстрирует один из способов реализации стека с непосредственными изменениями в памяти.

Для простоты некоторые методы перегрузки операторов опущены. Новый метод pop, используемый здесь, эквивалентен доступу по индексу и удалению элемента со смещением -1 (здесь вершина находится в конце списка). По сравнению с непосредственным использованием встроенных списков, этот класс имеет более низкую производительность, что обусловлено дополнительными вызовами методов, но он поддерживает дальнейшее усовершенствование, инкапсулируя операции со стеком.

Пример 18.5. PP4E\Dstruct\Basic\stack4.py

```
" оптимизация за счет непосредственного изменения списка в памяти"

class error(Exception): pass      # при импортировании: локальное исключение

class Stack:
    def __init__(self, start=[]): # self - объект экземпляра
        self.stack = []         # start - любая последовательность: stack...
        for x in start: self.push(x)

    def push(self, obj):          # методы: подобно модулю + self
        self.stack.append(obj)    # вершина в конце списка

    def pop(self):
        if not self.stack: raise error('underflow')
        return self.stack.pop()   # подобно извлечению и удалению stack[-1]

    def top(self):
        if not self.stack: raise error('underflow')
        return self.stack[-1]

    def empty(self):
        return not self.stack     # instance.empty()

    def __len__(self):
        return len(self.stack)    # len(instance), not instance

    def __getitem__(self, offset):
        return self.stack[offset] # instance[offset], in, for

    def __repr__(self):
        return '[Stack:%s]' % self.stack
```

Эта версия работает, как оригинал в модуле stack2, – просто замените stack2 на stack4 в предыдущем примере интерактивного сеанса, и вы получите представление о его работе. Единственное заметное отличие в том, что элементы стека выводятся в обратном порядке (то есть вершиной стека является конец списка):

```
>>> from stack4 import Stack
>>> x = Stack()
>>> x.push('spam')
>>> x.push(123)
```

```
>>> x
[Stack:['spam', 123]]
>>>
>>> y = Stack()
>>> y.push(3.1415)
>>> y.push(x.pop())
>>> x, y
([Stack:['spam']], [Stack:[3.1415, 123]])
>>> y.top()
123
```

Хронометраж усовершенствований

Объект стека, модифицируемый непосредственно в памяти, представленный в прошлом разделе, работает, вероятно, быстрее, чем первоначальная версия и версия на основе дерева кортежей, но единственный способ действительно выяснить, насколько быстрее, – это провести хронометраж альтернативных реализаций.[1] Поскольку такая операция может потребоваться неоднократно, определим сначала общий модуль функций хронометража в Python. Встроенный модуль `time` в примере 18.6 предоставляет функцию `clock`, с помощью которой можно получить текущее время CPU в виде секунд, выраженных числом с плавающей точкой, а функция `timer.test` просто вызывает нужную функцию `reps` раз и возвращает количество истекших секунд, вычитая время конца из времени начала.

Пример 18.6. PP4E\Dstruct\Basic\timer.py

```
"универсальный инструмент хронометража"
def test(reps, func, *args):    # или лучший из N? см. "Изучаем Python"
    import time
    start = time.clock()        # текущее время CPU в секундах
    for i in range(reps):       # вызвать функцию reps раз
        func(*args)             # отбросить возвращаемое значение
    return time.clock() - start # время конца - время начала
```

[1] Поскольку Python является чрезвычайно динамичным языком, предположения об относительной производительности тех или иных алгоритмов в равной степени могут оказаться как верными, так и ошибочными. Кроме того, достоверность этих предположений может изменяться со временем. Уж можете мне поверить. В других книгах мне пришлось убрать множество рекомендаций, касающихся производительности, только потому, что в более поздних версиях Python они стали ошибочными из-за того, что одни операции были оптимизированы больше, чем другие. Измерение производительности в языке Python является нетривиальной задачей, которую следует выполнять постоянно. В целом, основное внимание следует уделять удобочитаемости программного кода, а о производительности беспокоиться в последнюю очередь, но старайтесь всегда собирать сведения, которые пригодятся для оптимизации.

Существуют также другие способы хронометража, включая подход «лучший из N» и собственный модуль Python `timeit`, но для наших целей вполне достаточно будет и этого модуля. Если вам интересно познакомиться с более удачными решениями, обращайтесь к четвертому изданию книги «Изучаем Python», где приводятся более крупные примеры по этой теме, или поэкспериментируйте с собственными решениями.

Далее, определим управляющий сценарий теста, как показано в примере 18.7. Он принимает три аргумента командной строки: количество операций проталкивания в стек, выталкивания со стека и обращения по индексу (будем менять эти аргументы для проверки разных ситуаций). При запуске в виде самостоятельного сценария он создает 200 экземпляров исходного и оптимизированного классов стека и выполняет заданное количество операций с каждым стеком. Операции проталкивания в стек и выталкивания со стека изменяют стек; при индексировании происходит только выборка значений из него.

Пример 18.7. PP4E\Dstruct\Basic\stacktime.py

```
"сравнение производительности альтернативных реализаций стека"

import stack2 # стек на основе списка: [x]+y
import stack3 # стек на основе дерева кортежей: (x,y)
import stack4 # стек, выполняющий модификацию списка в памяти: y.append(x)
import timer  # вспомогательная функция хронометража

rept = 200
from sys import argv
pushes, pops, items = (int(arg) for arg in argv[1:])

def stackops(stackClass):
    x = stackClass('spam')              # создать объект стека
    for i in range(pushes): x.push(i)   # применить его методы
    for i in range(items): t = x[i]     # 3.X: range - генератор
    for i in range(pops): x.pop()
                                        # или mod = __import__(n)
for mod in (stack2, stack3, stack4):    # rept*(push+pop+ix)
    print('%s:' % mod.__name__, end=' ')
    print(timer.test(rept, stackops, getattr(mod, 'Stack')))
```

Результаты в Python 3.1

Ниже приводятся некоторые результаты, которые были получены с помощью управляющего сценария. В трех тестах было получено время в секундах для трех реализаций: оригинальной, на основе кортежей и на основе непосредственного изменения списка в памяти. Для каждой разновидности стека тест создает 200 экземпляров стека и выполняет примерно 120 000 операций со стеком (200 повторов × (200 операций проталкивания в стек + 200 обращений по индексу + 200 операций выталкивания со стека)) в течение указанного в результатах времени. Эти результаты были получены на очень медлительном ноутбуке, работаю-

щем под управлением Windows 7, в Python 3.1. Как обычно, у вас могут получиться иные результаты.

```
C:\...\PP4E\Dstruct\Basic> python stacktime.py 200 200 200
stack2: 0.838853884098
stack3: 2.52424649244
stack4: 0.215801718938

C:\...\PP4E\Dstruct\Basic> python stacktime.py 200 50 200
stack2: 0.775219065818
stack3: 2.539294115
stack4: 0.156989574341

C:\...\PP4E\Dstruct\Basic> python stacktime.py 200 200 50
stack2: 0.743521212289
stack3: 0.286850521181
stack4: 0.156262000363

C:\...\PP4E\Dstruct\Basic> python stacktime.py 200 200 0
stack2: 0.721035029026
stack3: 0.116366779208
stack4: 0.141471921584
```

Если внимательно посмотреть, то можно заметить по результатам, что стек, основанный на кортежах (stack3), показывает лучшую производительность, когда производится больше операций проталкивания в стек и выталкивания со стека, но оказывается хуже, когда производится много обращений по индексу. Операция обращения по индексу происходит очень быстро для встроенных списков (stack2 и stack4), но очень медленно для деревьев кортежей – класс Python вынужден вручную выполнять обход дерева.

Стеки, основанные на непосредственной модификации списка в памяти (stack4), *почти* всегда действуют быстрее всего, кроме случаев, когда вообще не выполняется никаких операций обращения по индексу – в последнем тесте кортежи (stack3) победили с небольшим перевесом. При отсутствии операций обращения по индексу, как в последнем тесте, реализации на основе кортежей и на основе непосредственной модификации списка в памяти оказываются примерно в шесть и в пять раз быстрее, чем простая реализация на основе списка соответственно. Поскольку операции проталкивания в стек и выталкивания со стека являются основными операциями, необходимыми клиентам от стека, кортежи являются претендентом на победу, несмотря на слабую производительность при обращении по индексу.

Конечно, мы говорим о долях секунды после многих десятков тысяч операций. Во многих приложениях пользователи могут не почувствовать разницы. Однако если программа обращается к стеку миллионы раз, эта разница может стать весьма существенной.

Дополнительно об анализе производительности

Два последних примечания, касающихся производительности. Несмотря на то, что абсолютные значения изменяются со временем, с появлением новых версий Python и компьютеров, на которых выполняется тестирование, относительные результаты остаются практически неизменными. То есть стек на основе кортежей побеждает при отсутствии операций обращения по индексу. Однако любые результаты измерения производительности в таком динамично развивающемся языке, как Python, могут измениться со временем, поэтому обязательно выполняйте подобные тесты у себя, чтобы получить более точные результаты.

Во-вторых, часто существует больше способов измерения производительности, чем простой хронометраж. Чтобы получить более полное представление, обращайтесь к описанию `profile` в руководстве по стандартной библиотеке (и его оптимизированной альтернативы `cProfile`). Профилировщики выполняют программный код Python, попутно собирают информацию о производительности и выводят отчет по завершении выполнения программного кода. Это наиболее надежный способ обнаружения узких мест в программном коде перед началом работ по оптимизации с применением других синтаксических конструкций, алгоритмов и структур данных или переноса части приложения на язык C.

Однако в случае простого измерения производительности наш модуль хронометража предоставляет все необходимые нам данные. Мы будем повторно использовать его для измерения более ощутимого повышения скорости при выполнении альтернативных реализаций множеств в следующем разделе.

Реализация множеств

Другой часто используемой структурой данных является *множество* – коллекция объектов, поддерживающая следующие операции:

Пересечение

Создает новое множество, состоящее из общих элементов.

Объединение

Создает новое множество, состоящее из элементов, принадлежащих хотя бы одному операнду.

Принадлежность

Проверяет, содержится ли элемент в множестве.

В зависимости от предполагаемого использования полезными могут оказаться и другие операции, такие как разность и проверка на подмножество. Множества удобно использовать для работы с более абстрактными сочетаниями групп. Например, если есть группа инженеров и группа авторов, можно выбрать тех, кто занимается обоими видами деятельности, определив пересечение двух множеств. В объединении таких мно-

жеств будут содержаться оба типа специалистов, но каждый из них будет включен только один раз. Эта последняя особенность делает множества идеальным инструментом для удаления из коллекций дубликатов элементов – чтобы отфильтровать дубликаты, достаточно просто преобразовать коллекцию в множество и обратно.

Фактически мы уже использовали подобные операции в предыдущих главах. Например, в PyMailGUI, в главе 14, операции определения пересечения, объединения и разности множеств использовались для управления множеством активных операций загрузки почты, а операция преобразования в множество использовалась для удаления повторяющихся адресов получателей в разных ситуациях.

Встроенные возможности

Если вы знакомы с основами языка Python, вы должны знать, что в Python имеется встроенная поддержка множеств, как и стеков. Однако множества поддерживаются еще более непосредственным способом – тип данных set в Python предоставляет стандартный набор оптимизированных операций над множествами. Встроенный тип set достаточно прост в использовании: объекты множеств создаются обращением к имени типа set как к функции, которой передается итерируемый объект или последовательность компонентов для включения в множество, или выполнением выражения генератора множества:

```
>>> x = set('abcde')          # из итерируемого объекта/последовательности
>>> y = {c for c in 'bdxyz'}  # то же самое с помощью генератора множеств
>>> x
{'a', 'c', 'b', 'e', 'd'}
>>> y
{'y', 'x', 'b', 'd', 'z'}
```

После создания множества становятся доступными все обычные операции. Ниже демонстрируются наиболее типичные из них:

```
>>> 'e' in x           # принадлежность
True
>>> x - y              # разность
{'a', 'c', 'e'}
>>> x & y              # пересечение
{'b', 'd'}
>>> x | y              # объединение
{'a', 'c', 'b', 'e', 'd', 'y', 'x', 'z'}
```

Интересно отметить, что встроенные множества, подобно встроенным словарям, являются неупорядоченными коллекциями и требуют, чтобы включаемые в них элементы были хешируемыми (неизменяемыми). Создать множество из словаря возможно, но только лишь потому, что конструктор set будет использовать итератор словаря, который в каждой итерации возвращает очередной ключ (он игнорирует значения ключей):

```
>>> x = set(['spam', 'ham', 'eggs']) # последовательность неизменяемых элем.
>>> x
{'eggs', 'ham', 'spam'}
>>> x = {'spam', 'ham', 'eggs'}  # литерал множества, если элементы известны
>>> x
{'eggs', 'ham', 'spam'}

>>> x = set([['spam', 'ham'], ['eggs']]) # изменяемые элементы не могут
TypeError: unhashable type: 'list'       # включаться в множество

>>> x = set({'spam':[1, 1], 'ham': [2, 2], 'eggs':[3, 3]})
>>> x
{'eggs', 'ham', 'spam'}
```

Существуют также другие операции, которые мы не будем описывать здесь, – подробности ищите в книгах, посвященных основам языка, в таких как «Изучаем Python». Например, встроенные множества поддерживают такие операции, как проверка на надмножество, и имеют две разновидности: изменяемые и фиксированные (фиксированные множества являются хешируемыми и могут использоваться для создания множества множеств). Кроме того, генераторы множеств обладают более широкими возможностями, чем показано здесь, и множества являются естественным инструментом удаления дубликатов:

```
>>> y = {c.upper() * 4 for c in 'spamham'} # генератор множеств
>>> y
{'SSSS', 'AAAA', 'MMMM', 'HHHH', 'PPPP'}
>>>
>>> list(set([1, 2, 3, 1, 2]))             # удалит дубликаты из списка
[1, 2, 3]
```

Однако, как и в случае со стеками, встроенный тип `set` может не полностью удовлетворять наши потребности. Кроме того, собственная реализация множества может оказаться идеальным примером реализации нестандартных структур данных в языке Python. Хотя по своей производительности конечный результат может оказаться неконкурентоспособным в сравнении со встроенными множествами, тем не менее, попытка создать собственную реализацию может оказаться весьма поучительной и интересной.

Кроме того, подобно стекам наша собственная реализация множества будет опираться на использование других встроенных типов. Списки, кортежи и строки Python близки к понятию множества: оператор `in` проверяет принадлежность, `for` производит итерации и так далее. Здесь мы введем операции, не поддерживаемые непосредственно последовательностями Python. Идея состоит в *расширении* встроенных типов в соответствии с особыми требованиями.

Функции множеств

Как и прежде, начнем с основанного на функциях менеджера множеств. Но на этот раз вместо управления совместно используемым объектом множества в модуле определим функции, реализующие операции множеств над передаваемыми им последовательностями Python (пример 18.8).

Пример 18.8. PP4E\Dstruct\Basic\inter.py

```
"операции множеств над двумя последовательностями"

def intersect(seq1, seq2):
    res = []                    # начать с пустого списка
    for x in seq1:              # просмотр первой последовательности
        if x in seq2:
            res.append(x)       # добавить общие элементы в конец
    return res

def union(seq1, seq2):
    res = list(seq1)            # создать копию seq1
    for x in seq2:              # добавить новые элементы в seq2
        if not x in res:
            res.append(x)
    return res
```

Эти функции работают с последовательностями любого типа – списками, строками, кортежами и другими итерируемыми объектами, удовлетворяющими протоколам, предполагаемым этими функциями (циклы `for`, проверки принадлежности `in`). На самом деле их можно использовать даже с объектами разных типов: последние две команды в следующем фрагменте вычисляют пересечение и объединение списка и кортежа. Как обычно в Python, значение имеет *интерфейс* объекта, а не конкретный тип:

```
C:\...\PP4E\Dstruct\Basic> python
>>> from inter import *
>>> s1 = "SPAM"
>>> s2 = "SCAM"
>>> intersect(s1, s2), union(s1, s2)
(['S', 'A', 'M'], ['S', 'P', 'A', 'M', 'C'])
>>> intersect([1,2,3], (1,4))
[1]
>>> union([1,2,3], (1,4))
[1, 2, 3, 4]
```

Обратите внимание, что результат здесь всегда является списком независимо от типов передаваемых последовательностей. Мы могли бы обойти это ограничение путем преобразования типов или используя класс (что мы и сделаем чуть ниже). Но преобразования типов становятся неясными, если операнды имеют разные типы. В какой тип нужно преобразовывать?

Поддержка нескольких операндов

Если мы намерены использовать функции intersect и union в качестве универсальных инструментов, то полезным расширением будет поддержка нескольких аргументов (то есть больше двух). Функции, представленные в примере 18.9, используют возможность использования списков аргументов переменной длины, чтобы вычислить пересечение и объединение произвольного количества операндов.

Пример 18.9. PP4E\Dstruct\Basic\inter2.py

```
"операции множеств над несколькими последовательностями"

def intersect(*args):
    res = []
    for x in args[0]:                  # просмотр первого списка
        for other in args[1:]:         # для всех остальных аргументов
            if x not in other: break   # этот элемент во всех последов.?
        else:
            res.append(x)          # добавить общие элементы в конец
    return res

def union(*args):
    res = []
    for seq in args:               # для всех аргументов-последовательностей
        for x in seq:              # для всех элементов в аргументах
            if not x in res:
                res.append(x)      # добавить новые элементы в результат
    return res
```

Функции с несколькими операндами действуют над последовательностями так же, как исходные функции, но поддерживают три и более операндов. Обратите внимание на использование предложения else в цикле for в функции intersect для определения общих элементов. Заметьте также, что последние два примера в следующем листинге сеанса оперируют списками, содержащими составные аргументы: проверки in в функциях intersect и union для определения результатов сравнения с совокупностью рекурсивно применяют проверку равенства к узлам последовательностей на необходимую глубину:

```
C:\...\PP4E\Dstruct\Basic> python
>>> from inter2 import *
>>> s1, s2, s3 = 'SPAM', 'SLAM', 'SCAM'
>>> intersect(s1, s2)
['S', 'A', 'M']
>>> intersect(s1, s2, s3)
['S', 'A', 'M']
>>> intersect(s1, s2, s3, 'HAM')
['A', 'M']

>>> union(s1, s2), union(s1, s2, s3)
(['S', 'P', 'A', 'M', 'L'], ['S', 'P', 'A', 'M', 'L', 'C'])
```

```
>>> intersect([1, 2, 3], (1, 4), range(5))   # 3.X: использ. range допустимо
[1]
>>> s1 = (9, (3.14, 1), "bye", [1, 2], "mello")
>>> s2 = [[1, 2], "hello", (3.14, 0), 9]
>>> intersect(s1, s2)
[9, [1, 2]]
>>> union(s1, s2)
[9, (3.14, 1), 'bye', [1, 2], 'mello', 'hello', (3.14, 0)]
```

Классы множеств

Функции множеств, представленные в предыдущем разделе, могут оперировать различными объектами, но они не столь дружественны, как подлинные объекты. Помимо всего прочего сценариям может потребоваться вручную следить за последовательностями, передаваемыми в эти функции. Классы могут действовать успешнее: класс, представленный в примере 18.10, реализует объект множества, который может содержать объекты любого типа. Подобно классам стеков они в сущности являются оболочками вокруг списков Python с дополнительными операциями множеств.

Пример 18.10. PP4E\Dstruct\Basic\set.py

```
"""
настраиваемый класс множества, допускающий возможность создания
нескольких экземпляров
"""


class Set:
    def __init__(self, value = []): # при создании объекта
        self.data = []              # управляет локальным списком
        self.concat(value)

    def intersect(self, other):     # other - последовательность любого типа
        res = []                    # self - объект экземпляра
        for x in self.data:
            if x in other:
                res.append(x)
        return Set(res)             # вернуть новое множество

    def union(self, other):
        res = self.data[:]          # создать копию собственного списка
        for x in other:
            if not x in res:
                res.append(x)
        return Set(res)

    def concat(self, value):        # value: список, строка, множество...
        for x in value:             # отфильтровать дубликаты
            if not x in self.data:
                self.data.append(x)
```

```
    def __len__(self):          return len(self.data)
    def __getitem__(self, key): return self.data[key]
    def __and__(self, other):   return self.intersect(other)
    def __or__(self, other):    return self.union(other)
    def __repr__(self):         return '<Set:' + repr(self.data) + '>'
```

Класс Set используется так же, как класс Stack, рассмотренный выше в этой главе: мы создаем экземпляры и применяем к ним операторы последовательностей плюс особые операции множеств. Пересечение и объединение можно вызывать как методы или с помощью операторов & и |, обычно используемых со встроенными целочисленными объектами. Поскольку теперь в выражениях можно записывать операторы в строку (например, x & y & z), очевидно, нет нужды поддерживать здесь несколько операндов в методах intersect/union (хотя эта модель создает внутри выражения временные объекты, что может привести к потере производительности). Как и для всех объектов, можно использовать класс Set внутри программы или тестировать его интерактивно, как показано ниже:

```
>>> from set import Set
>>> users1 = Set(['Bob', 'Emily', 'Howard', 'Peeper'])
>>> users2 = Set(['Jerry', 'Howard', 'Carol'])
>>> users3 = Set(['Emily', 'Carol'])
>>> users1 & users2
<Set:['Howard']>
>>> users1 | users2
<Set:['Bob', 'Emily', 'Howard', 'Peeper', 'Jerry', 'Carol']>
>>> users1 | users2 & users3
<Set:['Bob', 'Emily', 'Howard', 'Peeper', 'Carol']>
>>> (users1 | users2) & users3
<Set:['Emily', 'Carol']>
>>> users1.data
['Bob', 'Emily', 'Howard', 'Peeper']
```

Оптимизация: перевод множеств на использование словарей

Первой проблемой, с которой можно столкнуться, начав использовать класс Set, оказывается производительность: его вложенные циклы for и проверки in замедляют выполнение с экспоненциальной скоростью. Библиотечные классы должны быть реализованы с максимально возможной эффективностью, иначе такая медлительность может оказаться весьма существенной для некоторых приложений.

Один из способов оптимизировать производительность множеств состоит в переходе к использованию словарей вместо списков для внутреннего хранения множеств – элементы могут храниться в виде ключей словаря, все значения в котором равны None. Так как время поиска в словарях постоянно и невелико, проверку в списке оператором in в первоначальном множестве в этой схеме можно заменить прямой выборкой из

словаря. Говоря традиционным языком, при переходе к использованию словарей мы заменяем медлительный линейный поиск быстрыми хеш-таблицами. Программист мог бы пояснить, что повторяющееся сканирование списков в поисках пересечений является *экспоненциальным* алгоритмом, а выборка из словарей – *линейным*.

Эту идею реализует модуль в примере 18.11. Его класс является подклассом оригинального множества и переопределяет методы, связанные с внутренним представлением, но наследует остальные. Унаследованные методы перегрузки операторов & и | вызывают здесь новые методы intersect и union, а унаследованный метод len действует со словарями без изменений. Пока клиенты класса Set не зависят от порядка элементов в множестве, они могут непосредственно переключиться на эту версию, просто изменив имя модуля, из которого импортируется класс Set, – имя класса остается прежним.

Пример 18.11. PP4E\Dstruct\Basic\fastset.py

```
"оптимизация за счет применения линейного алгоритма сканирования
по словарям"

import set
                            # fastset.Set расширяет set.Set
class Set(set.Set):
    def __init__(self, value = []):
        self.data = {}       # управляет локальным словарем
        self.concat(value)   # хеширование: время поиска изменяется линейно

    def intersect(self, other):
        res = {}
        for x in other:         # other: последовательность или Set
            if x in self.data:  # поиск по хеш-таблицам; 3.X
                res[x] = None
        return Set(res.keys())  # новый экземпляр Set на основе словаря

    def union(self, other):
        res = {}                # other: последовательность или Set
        for x in other:         # сканировать каждое множество только 1 раз
            res[x] = None
        for x in self.data.keys(): # '&' и '|' попадают сюда
            res[x] = None          # так они создают новые быстрые множества
        return Set(res.keys())

    def concat(self, value):
        for x in value: self.data[x] = None

    # inherit and, or, len
    def __getitem__(self, ix):
        return list(self.data.keys())[ix]      # 3.X: вызов list() необходим

    def __repr__(self):
        return '<Set:%r>' % list(self.data.keys()) # то же самое
```

Эта версия действует точно так же, как предыдущая, хотя ее внутренняя реализация радикально отличается:

```
>>> from fastset import Set
>>> users1 = Set(['Bob', 'Emily', 'Howard', 'Peeper'])
>>> users2 = Set(['Jerry', 'Howard', 'Carol'])
>>> users3 = Set(['Emily', 'Carol'])
>>> users1 & users2
<Set:['Howard']>
>>> users1 | users2
<Set:['Howard', 'Peeper', 'Jerry', 'Carol', 'Bob', 'Emily']>
>>> users1 | users2 & users3
<Set:['Peeper', 'Carol', 'Howard', 'Bob', 'Emily']>
>>> (users1 | users2) & users3
<Set:['Carol', 'Emily']>
>>> users1.data
{'Peeper': None, 'Bob': None, 'Howard': None, 'Emily': None}
```

Главным функциональным отличием этой версии является порядок следования элементов в множестве: поскольку словари упорядочиваются произвольным образом, порядок будет отличаться от исходного. Более того, порядок следования может быть разным даже в разных версиях Python (фактически такая разница наблюдается между Python 2.X и 3.X в третьем и четвертом изданиях этой книги). Например, можно хранить в множествах составные объекты, но порядок элементов в этой версии отличается:

```
>>> import set, fastset
>>> a = set.Set([(1,2), (3,4), (5,6)])
>>> b = set.Set([(3,4), (7,8)])
>>> a & b
<Set:[(3, 4)]>
>>> a | b
<Set:[(1, 2), (3, 4), (5, 6), (7, 8)]>
>>> a = fastset.Set([(1,2), (3,4), (5,6)])
>>> b = fastset.Set([(3,4), (7,8)])
>>> a & b
<Set:[(3, 4)]>
>>> a | b
<Set:[(1, 2), (5, 6), (3, 4), (7, 8)]>
>>> b | a
<Set:[(1, 2), (5, 6), (3, 4), (7, 8)]>
```

Как бы то ни было, но от множеств не требуется упорядоченность, поэтому здесь нет никаких проблем. А вот отклонение, которое может иметь значение, состоит в том, что в этой версии нельзя хранить *нехешируемые* объекты. Это следует из того, что ключи словаря должны быть неизменяемыми. Так как значения хранятся в ключах, множества, основанные на словарях, могут содержать только такие элементы, как кортежи, строки, числа и объекты классов, объявленные как неизменяемые. Изменяемые объекты, такие как списки и словари, нельзя

использовать непосредственно в таких множествах на основе словарей, но можно в множествах оригинального класса. Однако в качестве составных элементов множеств вполне допустимо использовать кортежи, потому что они являются неизменяемыми – при их использовании интерпретатор будет вычислять хеш-значения и проверять равенство ключей, как обычно:

```
>>> set.Set([[1, 2],[3, 4]])
<Set:[[1, 2], [3, 4]]>
>>> fastset.Set([[1, 2],[3, 4]])
TypeError: unhashable type: 'list'

>>> x = fastset.Set([(1, 2), (3, 4)])
>>> x & fastset.Set([(3, 4), (1, 5)])
<Set:[(3, 4)]>
```

Хронометраж работы в Python 3.1

Добились ли мы оптимизации? Напомню еще раз, что ожидания не всегда соответствуют действительности, хотя алгоритмическая сложность данной реализации выглядит веским основанием. Чтобы убедиться, в примере 18.12 приводится сценарий, сравнивающий производительность классов множеств. Он повторно использует модуль `timer` из примера 18.6, с помощью которого ранее сравнивались различные реализации стеков (наш программный код может реализовывать различные объекты, но это никак не влияет на порядок измерения времени).

Пример 18.12. PP4E\Dstruct\Basic\settime.py

```
"сравнение производительности альтернативных реализаций множеств"
import timer, sys
import set, fastset

def setops(Class):              # 3.X: использование range вполне допустимо
    a = Class(range(50))        # множество из 50 целых чисел
    b = Class(range(20))        # множество из 20 целых чисел
    c = Class(range(10))
    d = Class(range(5))
    for i in range(5):
        t = a & b & c & d       # 3 пересечения
        t = a | b | c | d       # 3 объединения

if __name__ == '__main__':
    rept = int(sys.argv[1])
    print('set => ', timer.test(rept, setops, set.Set))
    print('fastset =>', timer.test(rept, setops, fastset.Set))
```

Функция `setops` создает четыре множества и пять раз обрабатывает их операторами пересечения и объединения. Число повторений всего процесса определяется аргументом командной строки. Точнее, при каждом вызове `setops` создается 34 экземпляра `Set` (4 + [5 × (3 + 3)]) и выполняются методы `intersect` и `union` по 15 раз каждый (5 × 3) в теле цикла `for`.

На этот раз на том же самом ноутбуке с Windows 7 и Python 3.1 был достигнут резкий прирост производительности:

```
C:\...\PP4E\Dstruct\Basic> python settime.py 50
set => 0.637593916437
fastset => 0.20435049302

C:\...\PP4E\Dstruct\Basic> python settime.py 100
set => 1.21924758303
fastset => 0.393896570828

C:\...\PP4E\Dstruct\Basic> python settime.py 200
set => 2.51036677716
fastset => 0.802708664223
```

Результаты во многом зависят от общего быстродействия компьютера и могут отличаться в разных версиях Python. Но, по крайней мере в этом контрольном примере, реализация множества на основе словаря (`fastset`) оказалась *втрое* быстрее, чем реализация множества на основе списка (`set`). На самом деле такое троекратное ускорение вполне объяснимо. Словари Python уже являются оптимизированными хеш-таблицами, усовершенствовать которые еще больше весьма затруднительно. Пока не появится свидетельств того, что основанные на словарях множества все же слишком медлительны, наша работа здесь закончена.

Для сравнения, результаты, полученные в Python 2.4 в предыдущем издании этой книги, показывали во всех тестах шестикратный прирост скорости реализации `fastset` по сравнению с реализацией `set`. Отсюда можно сделать вывод, что в Python 3.X либо увеличилась скорость выполнения итераций, либо замедлилась скорость работы словарей. В еще более старой версии Python 1.5.2, во втором издании книги, были получены точно такие же относительные результаты, как в современной версии Python 3.1. В любом случае, это подчеркивает тот факт, что вам обязательно следует самостоятельно провести тестирование на своем компьютере и со своей версией Python – нынешние результаты тестирования производительности Python легко могут стать основой для анекдотов в будущем.

Алгебра отношений для множеств (внешний пример)

Если вас интересует реализация дополнительных операций с множествами на языке Python, посмотрите следующие файлы в пакете примеров для книги:

PP4E\Dstruct\Basic\rset.py

Реализация класса `RSet`

PP4E\Dstruct\Basic\reltest.py

Сценарий для тестирования класса `RSet` – ожидаемый вывод этого сценария приводится в файле *reltest.results.txt*

Подкласс `RSet`, объявленный в файле *rset.py*, добавляет в множества, основанные на словарях, основные операции алгебры отношений. В нем предполагается, что элементы множеств являются отображениями (записями) с одним элементом в столбце (поле). Класс `RSet` наследует все оригинальные операции класса `Set` (итерацию, пересечение, объединение, операторы `&` и `|`, фильтрацию дубликатов и прочие) и добавляет новые операции в виде методов:

Select

Возвращает множество узлов, в которых поле равно заданному значению.

Bagof

Отбирает узлы, удовлетворяющие строке выражения.

Find

Отбирает кортежи согласно сравнению, полю и значению.

Match

Отыскивает в двух множествах узлы с одинаковыми значениями в общих полях.

Product

Вычисляет декартово произведение: формирует кортежи из двух множеств.

Join

Собирает кортежи из двух множеств, в которых некоторое поле имеет одинаковое значение.

Project

Извлекает указанные поля из кортежей в таблицу.

Difference

Удаляет кортежи одного множества из другого.

Эти операции не поддерживаются встроенными объектами множеств языка Python и наглядно демонстрируют, почему может потребоваться реализовать собственный тип множества. Хотя я и адаптировал этот программный код для работы под управлением Python 3.X, тем не менее, я не проводил сколько-нибудь глубокий его анализ при подготовке к включению в это издание, потому что сейчас я предпочел бы реализовать его как подкласс встроенного типа `set`, а не как полностью собственную реализацию множества. По случайному совпадению это ведет нас к следующей теме.

Создание подклассов встроенных типов

Прежде чем перейти к другим классическим структурам данных, необходимо коснуться еще одного аспекта, имеющегося в истории со стеками и множествами. В последних версиях Python имеется возможность

создавать подклассы от встроенных типов данных, таких как списки и словари, с целью расширения их возможностей. То есть благодаря тому, что типы данных теперь сами являются классами, доступными для наследования, мы можем создавать уникальные типы данных, являющиеся расширенными версиями встроенных типов, определяя подклассы, наследующие встроенные инструменты. Это особенно верно для Python 3.X, где термины «тип» и «класс» стали настоящим синонимами.

Для демонстрации в примере 18.13 приводится основная часть модуля, содержащего версии наших объектов стека и множества, реализованные в предыдущих разделах, переработанные в списки с дополнительными методами. Для разнообразия здесь был также немного упрощен метод объединения множеств и из него был убран избыточный цикл.

Пример 18.13. PP4E\Dstruct\Basic\typesubclass.py

```
"расширение встроенных типов вместо создания совершенно новой реализации"

class Stack(list):
    "список с дополнительными методами"
    def top(self):
        return self[-1]

    def push(self, item):
        list.append(self, item)

    def pop(self):
        if not self:
            return None            # чтобы избежать исключения
        else:
            return list.pop(self)

class Set(list):
    "список с дополнительными методами и операторами"
    def __init__(self, value=[]):  # при создании объекта
        list.__init__(self)
        self.concat(value)

    def intersect(self, other):    # other - последовательность любого типа
        res = []                   # self - объект экземпляра
        for x in self:
            if x in other:
                res.append(x)
        return Set(res)            # вернуть новый объект типа Set

    def union(self, other):
        res = Set(self)            # новое множество с копией своего списка
        res.concat(other)          # вставить уникальные элементы из other
        return res

    def concat(self, value):       # value: список, строка, Set...
        for x in value:            # отфильтровать дубликаты
```

```
        if not x in self:
            self.append(x)

# методы len, getitem, iter унаследованы, используется метод repr списка
def __and__(self, other): return self.intersect(other)
def __or__(self, other):  return self.union(other)
def __str__(self):        return '<Set:' + repr(self) + '>'

class FastSet(dict):
    pass                          # не содержит значительных упрощений
```

...программный код самотестирования опущен: смотрите файл в пакете примеров...

Стек и множество здесь реализованы по сути так же, как и прежде, но вместо встраивания списка и управления им эти объекты сами являются расширенными списками. Они добавляют несколько дополнительных методов, но основную функциональность наследуют от списков.

Этот прием позволяет уменьшить объем необходимого обертывающего программного кода, но он также делает доступной функциональность, которая может быть несвойственна некоторым типам. Например, в данной реализации имеется возможность сортировать содержимое стека, вставлять данные в середину и менять порядок следования элементов на обратный, потому что методы, реализующие эти операции, были унаследованы от встроенного списка. В большинстве случаев эти операции не имеют смысла для подобных структур данных, поэтому решение из предыдущих разделов в виде классов-оберток, запрещающих несвойственные операции, может оказаться более предпочтительным.

Подробности, касающиеся классов, наследующих встроенные типы, смотрите в программном коде самопроверки в оставшейся части файла с реализацией и в файле с ожидаемым выводом. Так как эти объекты используются точно так же, как и оригинальные версии стеков и множеств, дальнейшее исследование особенностей их использования я оставляю на роль самостоятельного упражнения.

Наследование встроенных типов может найти другие применения, пожалуй, более полезные, чем было продемонстрировано выше. Представьте, например, очередь или упорядоченный словарь. Очередь могла бы быть реализована как подкласс списка с методами извлечения из очереди и помещения в нее, которые добавляли бы новые элементы в конец и удаляли имеющиеся из начала. Словарь можно было бы реализовать как подкласс словаря с дополнительным списком ключей, который сортируется при вставке новых элементов и по запросу. Такой подход отлично подходит для реализации типов, напоминающих встроенные, но может оказаться непригодным для создания таких радикально отличающихся от них структур данных, как представленные в следующих двух разделах.

Двоичные деревья поиска

Двоичные деревья представляют собой структуры данных, упорядочивающие вставляемые узлы: элементы, меньшие узла, записываются в его левое поддерево, а элементы, б льшие узла, помещаются в правое. Самые нижние поддеревья пусты. Вследствие такой структуры двоичные деревья естественным образом поддерживают быстрый рекурсивный обход и, соответственно, быстрый поиск для широкого круга применений – по крайней мере, в идеальном случае при каждом переходе к поддереву пространство поиска сокращается вдвое.

Встроенные возможности

В данном случае Python также поддерживает операции поиска с помощью встроенных инструментов. Словари, например, предоставляют оптимизированные инструменты поиска по таблицам, реализованные на языке C. Встроенная операция доступа к элементам словаря по ключу, вероятнее всего, будет выполняться быстрее, чем эквивалентная реализация поиска на языке Python:

```
>>> x = {}                                          # пустой словарь
>>> for i in [3, 1, 9, 2, 7]: x[i] = None           # вставка
...
>>> x
{7: None, 1: None, 2: None, 3: None, 9: None}
>>>
>>> for i in range(8): print((i, i in x), end=' ') # просмотр
...
(0,False) (1,True) (2,True) (3,True) (4,False) (5,False) (6,False) (7,True)
```

Поскольку словари встроены в язык, они доступны всегда и обычно всегда оказываются быстрее структур данных, реализованных на языке Python. Часто аналогичную функциональность могут предложить встроенные множества – при этом достаточно просто представить себе множество как словарь, не имеющий значений:

```
>>> x = set()                                       # пустое множество
>>> for i in [3, 1, 9, 2, 7]: x.add(i)              # вставка
...
>>> x
{7, 1, 2, 3, 9}
>>> for i in range(8): print((i, i in x), end=' ') # просмотр
...
(0,False) (1,True) (2,True) (3,True) (4,False) (5,False) (6,False) (7,True)
```

Существует множество способов добавления элементов в множества и словари – оба типа удобно использовать для проверки наличия ключа, но словари дополнительно позволяют присваивать значения ключам:

```
>>> v = [3, 1, 9]

>>> {k for k in v}                                  # генератор множеств
```

```
{1, 3, 9}
>>> set(v)                          # конструктор множества
{1, 3, 9}

>>> {k: k+100 for k in v}           # генератор словарей
{1: 101, 3: 103, 9: 109}
>>> dict(zip(v, [99] * len(v)))     # конструктор словаря
{1: 99, 3: 99, 9: 99}
>>> dict.fromkeys(v, 99)            # метод словаря
{1: 99, 3: 99, 9: 99}
```

Тогда зачем заниматься реализацией собственных структур данных для поиска, если встроенные типы обеспечивают такую гибкость? В некоторых приложениях этого может и не потребоваться, но в данном случае может оказаться более целесообразным применить оптимизированные алгоритмы поиска в деревьях. Например, использование сбалансированных деревьев позволяет повысить скорость поиска даже в самых тяжелых случаях и превзойти по быстродействию обобщенные алгоритмы поиска, используемые в словарях и множествах. Кроме того, здесь действуют те же побудительные мотивы, которые подтолкнули нас к созданию собственных реализаций стеков и множеств, – заключая доступ к деревьям в интерфейс на основе классов, мы упрощаем возможность расширения и изменения в будущем.

Реализация двоичных деревьев

Двоичные деревья названы так из-за подразумеваемой древовидной структуры ссылок на поддеревья. Обычно их узлы реализуются как тройки значений: (ЛевоеПоддерево, ЗначениеУзла, ПравоеПоддерево). Помимо этого реализация деревьев ничем более не ограничивается. Здесь мы будем использовать подход на основе классов:

- BinaryTree – главный объект, который инициализирует и контролирует фактическое дерево.
- EmptyNode – пустой объект, совместно используемый всеми пустыми поддеревьями (внизу).
- BinaryNode – объекты, представляющие непустые узлы дерева, имеющие значение и два поддерева.

Чтобы не писать отдельные функции поиска, двоичные деревья строят из «интеллектуальных» объектов (экземпляров классов), которые умеют обрабатывать запросы вставки/поиска и вывода и передавать их объектам поддеревьев. Фактически это еще один пример действия отношения композиции в ООП: одни узлы дерева вкладываются в другие и запросы на поиск передаются вложенным поддеревьям. Единственный экземпляр пустого класса совместно используется всеми пустыми поддеревьями двоичного дерева, а при вставке лист EmptyNode заменяется на BinaryNode. Как это реализуется в программном коде, показано в примере 18.14.

Пример 18.14. PP4E\Dstruct\Classics\btree.py

```
"двоичное дерево поиска, не имеющее значений"

class BinaryTree:
    def __init__(self):      self.tree = EmptyNode()
    def __repr__(self):      return repr(self.tree)
    def lookup(self, value): return self.tree.lookup(value)
    def insert(self, value): self.tree = self.tree.insert(value)

class EmptyNode:
    def __repr__(self):
        return '*'
    def lookup(self, value):                    # достигнут конец снизу
        return False
    def insert(self, value):
        return BinaryNode(self, value, self)    # добавить новый узел снизу

class BinaryNode:
    def __init__(self, left, value, right):
        self.data, self.left, self.right = value, left, right

    def lookup(self, value):
        if self.data == value:
            return True
        elif self.data > value:
            return self.left.lookup(value)          # искать слева
        else:
            return self.right.lookup(value)         # искать справа

    def insert(self, value):
        if self.data > value:
            self.left = self.left.insert(value)     # вставить слева
        elif self.data < value:
            self.right = self.right.insert(value)   # вставить справа
        return self

    def __repr__(self):
        return ('( %s, %s, %s )' %
                (repr(self.left), repr(self.data), repr(self.right)))
```

Как обычно, экземпляр `BinaryTree` может содержать объекты любого типа, поддерживающего предполагаемый протокол интерфейса, – в данном случае операторы сравнения > и <. В их число входят экземпляры классов с методами `__lt__` и `__gt__`. Поэкспериментируем с интерфейсами этого модуля. В следующем сеансе в новое дерево добавляются пять целых чисел, а затем выполняется поиск значений 0...7, как мы делали это раньше с применением словарей и множеств:

```
C:\...\PP4E\Dstruct\Classics> python
>>> from btree import BinaryTree
>>> x = BinaryTree()
```

```
>>> for i in [3, 1, 9, 2, 7]: x.insert(i)
...
>>> for i in range(8): print((i, x.lookup(i)), end=' ')
...
(0,False) (1,True) (2,True) (3,True) (4,False) (5,False) (6,False) (7,True)
```

Чтобы посмотреть, как растет это дерево, добавьте вызов функции `print` после каждой операции вставки. Узлы деревьев выводят себя как триплеты, а пустые узлы – как `*`. Результат отражает вложенность деревьев:

```
>>> y = BinaryTree()
>>> y
*
>>> for i in [3, 1, 9, 2, 7]:
...     y.insert(i); print(y)
...
( *, 3, * )
( ( *, 1, * ), 3, * )
( ( *, 1, * ), 3, ( *, 9, * ) )
( ( *, 1, ( *, 2, * ) ), 3, ( *, 9, * ) )
( ( *, 1, ( *, 2, * ) ), 3, ( ( *, 7, * ), 9, * ) )
```

В конце этой главы мы увидим другой способ визуализации таких деревьев в графическом интерфейсе с именем PyTree (при желании вы можете перелистать книгу прямо сейчас). Значениями узлов в этом объекте дерева могут быть любые допускающие сравнение объекты Python. Например, ниже приводятся деревья строк:

```
>>> z = BinaryTree()
>>> for c in 'badce': z.insert(c)
...
>>> z
( ( *, 'a', * ), 'b', ( ( *, 'c', * ), 'd', ( *, 'e', * ) ) )

>>> z = BinaryTree()
>>> for c in 'abcde': z.insert(c)
...
>>> z
( *, 'a', ( *, 'b', ( *, 'c', ( *, 'd', ( *, 'e', * ) ) ) ) )

>>> z = BinaryTree()
>>> for c in 'edcba': z.insert(c)
...
>>> z
( ( ( ( ( *, 'a', * ), 'b', * ), 'c', * ), 'd', * ), 'e', * )
```

Обратите внимание на последний результат: если вставляемые в двоичное дерево элементы уже упорядочены, то получается *линейная* структура и утрачивается способность деления пространства поиска двоичных деревьев (дерево растет только вправо или только влево). Это наихудшая ситуация, и на практике двоичные деревья ее избегают, разде-

ляя значения. Но если вы хотите заняться этой темой дальше, почитайте книгу по структурам данных, в которой описывается технология балансировки деревьев для обеспечения их максимальной густоты.

Деревья с ключами и значениями

Обратите также внимание, что для простоты эти деревья хранят только значения, а поиск возвращает «истину» или «ложь». На практике иногда требуется хранить вместе ключ и ассоциируемое значение (и что-то еще) в каждом узле дерева. Для потенциальных лесорубов среди читателей в примере 18.15 показано, как выглядит такой объект дерева.

Пример 18.15. PP4E\Dstruct\Classics\btreevals.py

```
"двоичное дерево поиска, хранящее значения вместе с ключами"

class KeyedBinaryTree:
    def __init__(self):         self.tree = EmptyNode()
    def __repr__(self):         return repr(self.tree)
    def lookup(self, key):      return self.tree.lookup(key)
    def insert(self, key, val): self.tree = self.tree.insert(key, val)

class EmptyNode:
    def __repr__(self):
        return '*'
    def lookup(self, key):                      # достигнут конец снизу
        return None
    def insert(self, key, val):
        return BinaryNode(self, key, val, self)  # добавить новый узел снизу

class BinaryNode:
    def __init__(self, left, key, val, right):
        self.key, self.val = key, val
        self.left, self.right = left, right

    def lookup(self, key):
        if self.key == key:
            return self.val
        elif self.key > key:
            return self.left.lookup(key)              # искать слева
        else:
            return self.right.lookup(key)             # искать справа

    def insert(self, key, val):
        if self.key == key:
            self.val = val
        elif self.key > key:
            self.left = self.left.insert(key, val)    # вставить слева
        elif self.key < key:
            self.right = self.right.insert(key, val)  # вставить справа
        return self
```

```
    def __repr__(self):
        return ('( %s, %s=%s, %s )' %
                (repr(self.left), repr(self.key), repr(self.val),
                 repr(self.right)))

if __name__ == '__main__':
    t = KeyedBinaryTree()
    for (key, val) in [('bbb', 1), ('aaa', 2), ('ccc', 3)]:
        t.insert(key, val)
    print(t)
    print(t.lookup('aaa'), t.lookup('ccc'))
    t.insert('ddd', 4)
    t.insert('aaa', 5)          # изменить значение ключа
    print(t)
```

Ниже показаны результаты выполнения программного кода самотестирования – на этот раз узлы просто содержат больше информации:

```
C:\...\PP4E\Dstruct\Classics> python btreevals.py
( ( *, 'aaa'=2, * ), 'bbb'=1, ( *, 'ccc'=3, * ) )
2 3
( ( *, 'aaa'=5, * ), 'bbb'=1, ( *, 'ccc'=3, ( *, 'ddd'=4, * ) ) )
```

Фактически результат напоминает ключи и значения во встроенном словаре, но собственная древовидная структура, подобная этой, может поддерживать особые случаи использования и алгоритмы, а также дальнейшее развитие программного кода. Однако, чтобы познакомиться со структурами данных, еще дальше отстоящих от встроенных типов, нам необходимо перейти к следующему разделу.

Поиск на графах

Многие задачи, которые порой приходится решать в реальной жизни и в программировании, можно представить в виде графа, являющегося множеством состояний с переходами («дугами»), ведущими из одного состояния в другое. Например, планирование маршрута путешествия в действительности является замаскированной задачей поиска на графе – состояния представляют те места, где вы хотите побывать, а дуги представляют транспортные маршруты между ними. Программа, отыскивающая оптимальный маршрут путешествия, решает задачу поиска на графе. По сути, таковыми являются великое множество программ, осуществляющих обход гиперссылок в Веб.

В этом разделе представлены простые программы на языке Python, осуществляющие поиск путей в направленном циклическом графе между начальным и конечным состоянием. Графы являются более общими конструкциями, чем деревья, потому что ссылки смогут указывать на произвольные узлы – даже на уже посещавшиеся (отсюда название *циклический*). Кроме того, в Python нет никакой встроенной поддержки этого типа данных. Несмотря на то, что программы поиска на гра-

фах могут, в конечном счете, использовать встроенные типы, тем не менее, конкретные процедуры поиска являются достаточно специфическими, чтобы имело смысл создавать собственные реализации.

Граф, используемый для тестирования программ поиска, представленных в этом разделе, изображен на рис. 18.1. Стрелки на концах дуг указывают разрешенные маршруты (например, из *A* можно попасть в *B*, *E* и *G*). Алгоритмы поиска будут выполнять обход этого графа методом поиска в глубину, не допуская возврата в ту же точку, чтобы избежать зацикливания. Если вообразить, что это карта, на которой узлы представляют города, а дуги представляют дороги, то этот пример может показаться более осмысленным.

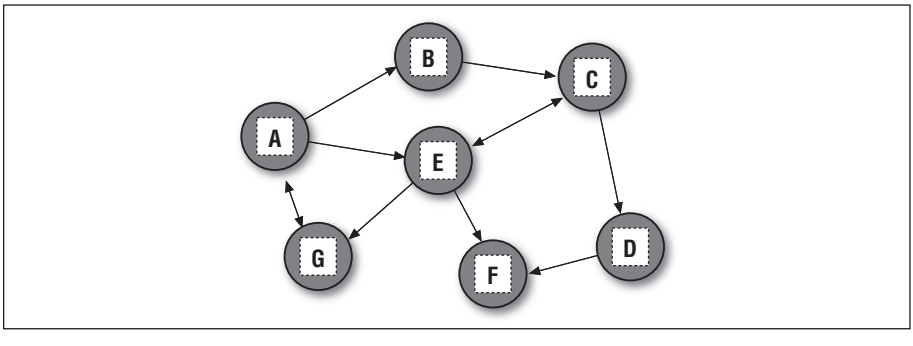


Рис. 18.1. *Направленный граф*

Реализация поиска на графе

Первое, что нужно сделать, это выбрать способ представления данного графа в сценарии на языке Python. Одно из решений заключается в использовании встроенных типов данных и функций поиска. Сценарий, представленный в примере 18.16, создает контрольный граф просто как словарь: каждое состояние является ключом словаря со списком ключей узлов, в которые можно из него попасть (то есть его дуг). В этом файле также определена функция, с помощью которой мы выполним несколько поисков на графе.

Пример 18.16. PP4E\Dstruct\Classics\gtestfunc.py

```
"представление графа в виде словаря"

Graph = {'A': ['B', 'E', 'G'],
         'B': ['C'],            # направленный циклический граф
         'C': ['D', 'E'],       # хранится в виде словаря
         'D': ['F'],            # 'ключ' ведет к [узлам]
         'E': ['C', 'F', 'G'],
         'F': [ ],
         'G': ['A'] }
```

```
def tests(searcher):                  # функция поиска
    print(searcher('E', 'D', Graph))  # отыскивает все пути от 'E' до 'D'
    for x in ['AG', 'GF', 'BA', 'DA']:
        print(x, searcher(x[0], x[1], Graph))
```

Теперь напишем два модуля, реализующие фактические алгоритмы поиска. Они не зависят от формы графа, в котором производится поиск (он передается им в качестве аргумента). Первая функция поиска, представленная в примере 18.17, для обхода графа использует *рекурсию*.

Пример 18.17. PP4E\Dstruct\Classics\gsearch1.py

```
"находит все пути на графе из начальной точки в конечную"

def search(start, goal, graph):
    solns = []
    generate([start], goal, solns, graph)  # выборка всех путей
    solns.sort(key=lambda x: len(x))       # сортировка по длине пути
    return solns

def generate(path, goal, solns, graph):
    state = path[-1]
    if state == goal:                     # если конечная точка присутствует
        solns.append(path)                # добавить в список solns
    else:                                 # проверить все дуги
        for arc in graph[state]:          # пропустить циклы в пути
            if arc not in path:
                generate(path + [arc], goal, solns, graph)

if __name__ == '__main__':
    import gtestfunc
    gtestfunc.tests(search)
```

Вторая функция поиска, представленная в примере 18.18, использует рассматривавшуюся выше реализацию *стека* путей, которые должны быть продолжены, на основе дерева кортежей.

Пример 18.18. PP4E\Dstruct\Classics\gsearch2.py

```
"поиск на графе, вместо рекурсии использует стек путей"

def search(start, goal, graph):
    solns = generate(([start], []), goal, graph)
    solns.sort(key=lambda x: len(x))
    return solns

def generate(paths, goal, graph):          # возвращает список solns
    solns = []                             # использует стек кортежей
    while paths:
        front, paths = paths               # вытолкнуть верхний путь
        state = front[-1]
        if state == goal:
            solns.append(front)            # конечная точка присутствует
```

```
        else:
            for arc in graph[state]:       # добавить все продолжения
                if arc not in front:
                    paths = (front + [arc]), paths
    return solns

if __name__ == '__main__':
    import gtestfunc
    gtestfunc.tests(search)
```

Обе функции запоминают посещавшиеся узлы, чтобы избежать циклов. Если продолжение оказывается в текущем пути, то возникает цикл. Полученный список решений сортируется по увеличению длины с помощью метода списка `sort` и его необязательного аргумента `key`, выполняющего преобразования исходных значений. Чтобы проверить модули поиска, просто запустите их; программный код самотестирования вызывает подготовленный тест поиска в модуле `gtestfunc`:

```
C:\...\PP4E\Dstruct\Classics> python gsearch1.py
[['E', 'C', 'D'], ['E', 'G', 'A', 'B', 'C', 'D']]
AG [['A', 'G'], ['A', 'E', 'G'], ['A', 'B', 'C', 'E', 'G']]
GF [['G', 'A', 'E', 'F'], ['G', 'A', 'B', 'C', 'D', 'F'],
    ['G', 'A', 'B', 'C', 'E', 'F'], ['G', 'A', 'E', 'C', 'D', 'F']]
BA [['B', 'C', 'E', 'G', 'A']]
DA []

C:\...\PP4E\Dstruct\Classics> python gsearch2.py
[['E', 'C', 'D'], ['E', 'G', 'A', 'B', 'C', 'D']]
AG [['A', 'G'], ['A', 'E', 'G'], ['A', 'B', 'C', 'E', 'G']]
GF [['G', 'A', 'E', 'F'], ['G', 'A', 'E', 'C', 'D', 'F'],
    ['G', 'A', 'B', 'C', 'E', 'F'], ['G', 'A', 'B', 'C', 'D', 'F']]
BA [['B', 'C', 'E', 'G', 'A']]
DA []
```

Эти модули выводят списки возможных путей через контрольный граф – я перенес две строки, чтобы облегчить его чтение (здесь также можно было бы использовать модуль Python `pprint` форматированного вывода). Обратите внимание, что обе функции во всех тестах находят одинаковые пути, но порядок нахождения может быть различным. Порядок путей, возвращаемых модулем `gsearch2`, зависит от того, как и когда продолжения добавляются в его стек пути – попробуйте проследить порядок следования путей в выводе, чтобы понять, как это происходит.

Перевод графов на классы

Использование словарей для представления графов эффективно: поиск смежных узлов выполняется быстрой операцией хеширования. Но для некоторых приложений более оправданными могут оказаться другие представления. Например, для моделирования узлов в сети могут также использоваться классы, как и для двоичных деревьев в более раннем примере. Узлы, реализованные в виде классов, могут содержать до-

полнительную информацию, полезную при более сложном поиске. Они могут также участвовать в иерархиях наследования и приобретать дополнительные особенности поведения. Для иллюстрации этой идеи в примере 18.19 показана альтернативная реализация поиска на графе – используемый алгоритм ближе всего к gsearch1.

Пример 18.19. PP4E\Dstruct\Classics\graph.py

```
"конструирует граф из объектов, способных выполнять поиск"

class Graph:
    def __init__(self, label, extra=None):
        self.name = label                  # узлы = объекты экземпляров
        self.data = extra                  # граф = связанные объекты
        self.arcs = []

    def __repr__(self):
        return self.name

    def search(self, goal):
        Graph.solns = []
        self.generate([self], goal)
        Graph.solns.sort(key=lambda x: len(x))
        return Graph.solns

    def generate(self, path, goal):
        if self == goal:                   # класс == цель
            Graph.solns.append(path)       # или self.solns: то же самое
        else:
            for arc in self.arcs:
                if arc not in path:
                    arc.generate(path + [arc], goal)
```

В этой версии графы представляются как сеть вложенных объектов экземпляров класса. Каждый узел в графе содержит список объектов узлов, к которым он ведет (arcs) и в котором он умеет выполнять поиск. Метод generate выполняет обход объектов в графе. Но на этот раз переходы прямо доступны в списке arcs каждого узла – нет необходимости индексировать (или передавать) словарь, чтобы найти связанные объекты.

Для проверки в примере 18.20 представлен модуль, который снова строит контрольный граф, на этот раз с помощью связанных экземпляров класса Graph. Обратите внимание на вызов функции exec в коде самотестирования: он выполняет динамически создаваемые строки, чтобы осуществить семь присваиваний (A=Graph('A'), B=Graph('B') и так далее).

Пример 18.20. PP4E\Dstruct\Classics\gtestobj1.py

```
"создает граф на основе класса и выполняет контрольный поиск"

from graph import Graph

# этот фрагмент не работает внутри инструкции def в 3.1: B - не определено
```

```
for name in "ABCDEFG":                        # создать сначала объект
    exec("%s = Graph('%s')" % (name, name)) # метка = имя переменной

A.arcs = [B, E, G]
B.arcs = [C]                    # теперь настроить связи:
C.arcs = [D, E]                 # вложенный список экземпляра
D.arcs = [F]
E.arcs = [C, F, G]
G.arcs = [A]

A.search(G)
for (start, stop) in [(E,D), (A,G), (G,F), (B,A), (D,A)]:
    print(start.search(stop))
```

Этот сценарий выполняет тот же самый обход графа, но на этот раз всю работу выполняют методы объектов:

```
C:\...\PP4E\Dstruct\Classics> python gtestobj1.py
[[E, C, D], [E, G, A, B, C, D]]
[[A, G], [A, E, G], [A, B, C, E, G]]
[[G, A, E, F], [G, A, B, C, D, F], [G, A, B, C, E, F], [G, A, E, C, D, F]]
[[B, C, E, G, A]]
[]
```

Результат получается такой же, как при использовании функций, но имена узлов выводятся без кавычек: узлы в списках путей теперь являются экземплярами класса `Graph`, а метод `__repr__` класса подавляет вывод кавычек. Перед тем как двинуться дальше, рассмотрим еще один сценарий проверки графов, представленный в примере 18.21, – набросайте узлы и дуги на бумаге, если проследить пути вам труднее, чем интерпретатору Python.

Пример 18.21. PP4E\Dstruct\Classics\gtestobj2.py

```
from graph import Graph

S = Graph('s')
P = Graph('p')      # граф spam
A = Graph('a')      # создать объекты узлов
M = Graph('m')

S.arcs = [P, M]     # S ведет к P и M
P.arcs = [S, M, A]  # дуги: встроенные объекты
A.arcs = [M]
print(S.search(M))  # найти все пути из S в M
```

Этот сценарий находит в графе три пути между узлами S и M. Мы лишь слегка коснулись здесь этой академической, но весьма полезной области знаний. Поэкспериментируйте с графами самостоятельно и поищите в других книгах дополнительные темы, связанные с ними (например, *поиск в ширину* и по *первому наилучшему совпадению*):

```
C:\...\PP4E\Dstruct\Classics> python gtestobj2.py
[[s, m], [s, p, m], [s, p, a, m]]
```

Перестановки последовательностей

Следующая наша тема в обзоре структур данных – реализация функциональных возможностей для работы с последовательностями, отсутствующих во встроенных объектах Python. Функции, объявленные в примере 18.22, реализуют несколько способов перестановки элементов последовательностей:

`permute`

строит список всех возможных перестановок элементов последовательности

`subset`

строит список всех возможных перестановок указанной длины

`combo`

действует, как `subset`, но порядок не имеет значения: перестановки с одинаковыми элементами удаляются.

Эти результаты полезны в ряде алгоритмов: при поиске, статистическом анализе и так далее. Например, один из способов отыскать оптимальный порядок следования элементов состоит в том, чтобы поместить их в список, сгенерировать всевозможные перестановки и просто по очереди проверить их. Все три функции используют общие приемы извлечения срезов последовательностей, чтобы список с результатами содержал последовательности того же типа, что и переданная в качестве аргумента (например, при перестановке строки мы получаем список строк).

Пример 18.22. PP4E\Dstruct\Classics\permcomb.py

```
"операции перестановки элементов последовательностей"

def permute(list):
    if not list:                # перестановка в любых последовательностях
        return [list]           # пустая последовательность
    else:
        res = []
        for i in range(len(list)):
            rest = list[:i] + list[i+1:]  # удалить текущий узел
            for x in permute(rest):       # выполнить перестановку остальных
                res.append(list[i:i+1] + x)  # добавить узел в начало
        return res

def subset(list, size):
    if size == 0 or not list:                # здесь порядок имеет значение
        return [list[:0]]                    # пустая последовательность
    else:
        result = []
        for i in range(len(list)):
            pick = list[i:i+1]              # срез последовательности
            rest = list[:i] + list[i+1:]    # сохранить часть [:i]
            for x in subset(rest, size-1):
```

```
                result.append(pick + x)
        return result

def combo(list, size):
    if size == 0 or not list:   # здесь порядок не имеет значения
        return [list[:0]]       # xyz == yzx
    else:
        result = []
        for i in range(0, (len(list) - size) + 1): # если осталось достаточно
            pick = list[i:i+1]
            rest = list[i+1:]                      # отбросить часть [:i]
            for x in combo(rest, size - 1):
                result.append(pick + x)
        return result
```

Все три функции могут оперировать любыми объектами последовательностей, поддерживающими операции `len`, извлечения среза и конкатенации. Например, можно применить функцию `permute` к экземплярам некоторых классов стеков, определенных в начале этой главы (поэкспериментируйте с ними самостоятельно).

Ниже приводится несколько примеров использования наших функций перестановки. Перестановка элементов списка позволяет найти все способы, которыми могут быть упорядочены элементы. Например, для списка из четырех элементов возможны 24 перестановки ($4 \times 3 \times 2 \times 1$). После выбора одного из четырех элементов для первой позиции остается только три, из которых можно выбрать элемент для второй позиции, и так далее. Порядок имеет значение: `[1,2,3]` не то же самое, что `[1,3,2]`, поэтому в результате появляются оба варианта перестановки:

```
C:\...\PP4E\Dstruct\Classics> python
>>> from permcomb import *
>>> permute([1, 2, 3])
[[1, 2, 3], [1, 3, 2], [2, 1, 3], [2, 3, 1], [3, 1, 2], [3, 2, 1]]
>>> permute('abc')
['abc', 'acb', 'bac', 'bca', 'cab', 'cba']
>>> permute('help')
['help', 'hepl', 'hlep', 'hlpe', 'hpel', 'hple', 'ehlp', 'ehpl', 'elhp',
 'elph', 'ephl', 'eplh', 'lhep', 'lhpe', 'lehp', 'leph', 'lphe', 'lpeh',
 'phel', 'phle', 'pehl', 'pelh', 'plhe', 'pleh']
```

Функция `combo` возвращает похожие результаты, но здесь накладывается ограничение фиксированной длины, а порядок не имеет значения: `abc` — то же самое, что `acb`, поэтому только один элемент добавляется в множество результатов:

```
>>> combo([1, 2, 3], 3)
[[1, 2, 3]]
>>> combo('abc', 3)
['abc']
>>> combo('abc', 2)
['ab', 'ac', 'bc']
```

```
>>> combo('abc', 4)
[]
>>> combo((1, 2, 3, 4), 3)
[(1, 2, 3), (1, 2, 4), (1, 3, 4), (2, 3, 4)]
>>> for i in range(0, 6): print(i, combo("help", i))
...
0 [``]
1 [`h`, `e`, `l`, `p`]
2 [`he`, `hl`, `hp`, `el`, `ep`, `lp`]
3 [`hel`, `hep`, `hlp`, `elp`]
4 [`help`]
5 []
```

Наконец, функция subset представляет просто перестановки фиксированной длины – порядок здесь имеет значение, поэтому результат получается больше по объему, чем для функции combo. На самом деле вызов subset с длиной последовательности идентичен вызову функции permute:

```
>>> subset([1, 2, 3], 3)
[[1, 2, 3], [1, 3, 2], [2, 1, 3], [2, 3, 1], [3, 1, 2], [3, 2, 1]]
>>> subset('abc', 3)
[`abc`, `acb`, `bac`, `bca`, `cab`, `cba`]
>>> for i in range(0, 6): print(i, subset("help", i))
...
0 [``]
1 [`h`, `e`, `l`, `p`]
2 [`he`, `hl`, `hp`, `eh`, `el`, `ep`, `lh`, `le`, `lp`, `ph`, `pe`, `pl`]
3 [`hel`, `hep`, `hle`, `hlp`, `hpe`, `hpl`, `ehl`, `ehp`, `elh`, `elp`,
   `eph`, `epl`, `lhe`, `lhp`, `leh`, `lep`, `lph`, `lpe`, `phe`, `phl`,
   `peh`, `pel`, `plh`, `ple`]
4 [`help`, `hepl`, `hlep`, `hlpe`, `hpel`, `hple`, `ehlp`, `ehpl`, `elhp`,
   `elph`, `ephl`, `eplh`, `lhep`, `lhpe`, `lehp`, `leph`, `lphe`, `lpeh`,
   `phel`, `phle`, `pehl`, `pelh`, `plhe`, `pleh`]
5 [`help`, `hepl`, `hlep`, `hlpe`, `hpel`, `hple`, `ehlp`, `ehpl`, `elhp`,
   `elph`, `ephl`, `eplh`, `lhep`, `lhpe`, `lehp`, `leph`, `lphe`, `lpeh`,
   `phel`, `phle`, `pehl`, `pelh`, `plhe`, `pleh`]
```

Эти функции реализуют достаточно компактные алгоритмы (и кому-то может показаться, что для полного их понимания необходимо пройти через «момент истины»), но они не настолько замысловаты, чтобы вы не смогли проследить их работу на простых примерах. Кроме того, они представляют также класс операций, требующих реализации собственных структур данных, в отличие от тех, что будут продемонстрированы в последнем пункте нашего обзора структур данных.

Обращение и сортировка последовательностей

Функции перестановки элементов последовательностей из предыдущего раздела могут пригодиться в самых разных приложениях, но существуют еще более фундаментальные операции, которые могут пока-

заться первыми кандидатами для реализации. Обращение и сортировка коллекций значений, например, являются типичными операциями в широком круге программ. Чтобы продемонстрировать приемы и примеры реализации и чтобы завершить сквозную тему этой главы, рассмотрим эти операции по очереди.

Реализация обращения

Обращение последовательностей может быть реализовано рекурсивно или итеративно, а также в виде функций или методов классов. В примере 18.23 представлена первая пробная реализация двух простых функций обращения.

Пример 18.23. PP4E\Dstruct\Classics\rev1.py

```
def reverse(list):                  # рекурсивная
    if list == []:
        return []
    else:
        return reverse(list[1:]) + list[:1]

def ireverse(list):                 # итеративная
    res = []
    for x in list: res = [x] + res
    return res
```

Обе функции обращения корректно обрабатывают списки. Но если попробовать обратить последовательности, не являющиеся списками (строки, кортежи и другие), функция `ireverse` всегда будет возвращать в качестве результата список независимо от того, какая последовательность ей передана:

```
>>> ireverse("spam")
['m', 'a', 'p', 's']
```

Что еще хуже, рекурсивная версия `reverse` вообще работает только со списками – с другими последовательностями она входит в бесконечный цикл. Такое поведение обусловлено весьма тонкой причиной: когда `reverse` добирается до пустой строки (`""`), последняя оказывается не равной пустому списку (`[]`), поэтому выбирается предложение `else`. Но срез пустой последовательности снова возвращает пустую последовательность (индексы увеличиваются): снова повторяется предложение `else` с пустой строкой без возбуждения исключения. В итоге функция попадает в бесконечный цикл, вызывая себя снова и снова, пока Python не исчерпает всю память.

Версии, представленные в примере 18.24, исправляют обе проблемы, используя обобщенную технику работы с последовательностями, подобную той, что использовалась в функциях перестановки элементов из предыдущего раздела:

- reverse использует оператор not для обнаружения конца последовательности и возвращает саму пустую последовательность, а не константу пустого списка. Поскольку пустая последовательность имеет тип исходного аргумента, операция + всегда строит последовательность правильного типа по мере развертывания рекурсии.
- ireverse использует то обстоятельство, что операция извлечения среза последовательности возвращает последовательность того же типа. Она сначала инициализирует результат срезом [:0] – новым пустым срезом того же типа, что и аргумент. Затем с помощью операций извлечения среза извлекаются последовательности из одного узла, добавляемые в начало результата, а не в константу списка.

Пример 18.24. PP4E\Dstruct\Classics\rev2.py

```
def reverse(list):
    if not list:                            # пустая? (не всегда [])
        return list                         # последов. того же типа
    else:
        return reverse(list[1:]) + list[:1]  # добавить передн. элемент
                                            # в конец

def ireverse(list):
    res = list[:0]                      # пустая, того же типа
    for i in range(len(list)):
        res = list[i:i+1] + res         # добавить каждый элемент в начало
    return res
```

Внесение этих изменений позволяет новым функциям работать с любыми последовательностями и возвращать новые последовательности того же типа, что и исходные. Если передать строку, в качестве результата получим новую строку. На самом деле они переворачивают любую последовательность объектов, которая поддерживает извлечение срезов, конкатенацию и len, – даже экземпляров классов Python и типов языка C. Иными словами, они могут переворачивать любые объекты, поддерживающие интерфейс последовательностей. Ниже приводится пример их работы со списками, строками и кортежами:

```
>>> from rev2 import *
>>> reverse([1, 2, 3]), ireverse([1, 2, 3])
([3, 2, 1], [3, 2, 1])
>>> reverse("spam"), ireverse("spam")
('maps', 'maps')
>>> reverse((1.2, 2.3, 3.4)), ireverse((1.2, 2.3, 3.4))
((3.4, 2.3, 1.2), (3.4, 2.3, 1.2))
```

Реализация сортировки

Еще одним главным занятием многих систем является сортировка: упорядочение элементов коллекций в соответствии с некоторым ограничением. Сценарий в примере 18.25 определяет простую функцию сортировки на языке Python, которая упорядочивает список объектов

по полю. Поскольку операция индексирования в языке Python является универсальной операцией, поле может быть индексом или ключом – эта функция может сортировать списки последовательностей или отображений.

Пример 18.25. PP4E\Dstruct\Classics\sort1.py

```
def sort(list, field):
    res = []                                # всегда возвращает список
    for x in list:
        i = 0
        for y in res:
            if x[field] <= y[field]: break  # узел списка попадает сюда?
            i += 1
        res[i:i] = [x]                      # вставить в результат
    return res

if __name__ == '__main__':
    table = [ {'name':'john', 'age':25}, {'name':'doe', 'age':32} ]
    print(sort(table, 'name'))
    print(sort(table, 'age'))
    table = [ ('john', 25), ('doe', 32) ]
    print(sort(table, 0))
    print(sort(table, 1))
```

Ниже приводится результат работы программного кода самотестирования этого модуля. В первых двух тестах выполняется сортировка словарей, в последних – кортежей:

```
C:\...\PP4E\Dstruct\Classics> python sort1.py
[{'age': 32, 'name': 'doe'}, {'age': 25, 'name': 'john'}]
[{'age': 25, 'name': 'john'}, {'age': 32, 'name': 'doe'}]
[('doe', 32), ('john', 25)]
[('john', 25), ('doe', 32)]
```

Добавление функций сравнения

Поскольку функцию можно передавать в качестве аргумента, как любой другой объект, легко предусмотреть передачу дополнительной функции сравнения. В следующей версии, в примере 18.26, второй аргумент принимает функцию, которая возвращает значение `true`, если ее первый аргумент должен быть помещен перед вторым. По умолчанию используется `lambda`-выражение, обеспечивающее порядок по возрастанию. Кроме того, эта функция сортировки возвращает новую последовательность того же типа, что и переданная ей, применяя такую же технику извлечения срезов, как в инструментах обращения последовательностей – если сортируется кортеж узлов, назад тоже возвращается кортеж.

Пример 18.26. PP4E\Dstruct\Classics\sort2.py

```
def sort(seq, func=(lambda x,y: x <= y)):  # по умолчанию: по возрастанию
    res = seq[:0]                          # возвращает послед. того же типа
    for j in range(len(seq)):
```

```
        i = 0
        for y in res:
            if func(seq[j], y): break
            i += 1
        res = res[:i] + seq[j:j+1] + res[i:]   # послед. может быть
                                               # неизменяемой
    return res

if __name__ == '__main__':
    table = ({'name':'doe'}, {'name':'john'})
    print(sort(list(table), (lambda x, y: x['name'] > y['name'])))
    print(sort(tuple(table), (lambda x, y: x['name'] <= y['name'])))
    print(sort('axbyzc'))
```

На этот раз записи в таблице упорядочиваются с помощью переданной функции сравнения полей:

```
C:\...\PP4E\Dstruct\Classics> python sort2.py
[{'name': 'john'}, {'name': 'doe'}]
({'name': 'doe'}, {'name': 'john'})
abcxyz
```

Эта версия обходится вообще без понятия поля и позволяет переданной функции при необходимости работать с индексами. В результате данная версия становится еще более универсальной, например ее также можно использовать для сортировки строк.

Структуры данных в сравнении со встроенными типами: заключение

Теперь, показав вам эти алгоритмы обращения и сортировки последовательностей, я должен также сообщить, что не во всех случаях они дают оптимальное решение. Все эти примеры служат целям обучения, но нужно сказать, что встроенные списки и функции часто позволяют добиться тех же результатов более простым способом:

Встроенные инструменты сортировки

В языке Python имеются два встроенных инструмента сортировки, которые действуют настолько быстро, что в большинстве случаев вам придется сильно попотеть, чтобы превзойти их. Чтобы воспользоваться методом `sort` списков для сортировки итерируемых объектов произвольных типов, для начала можно преобразовать их в списки, если это необходимо:

```
temp = list(sequence)
temp.sort()
...использовать элементы в temp...
```

Напротив, встроенная функция `sorted` может оперировать любыми итерируемыми объектами без необходимости преобразования в спи-

сок и возвращает новый отсортированный список, благодаря чему ее можно использовать в самых разных контекстах. Поскольку она не изменяет исходный объект, вам также не придется беспокоиться по поводу возможных побочных эффектов, вызванных изменением оригинального списка:

```
for item in sorted(iterable):
    ...использовать item...
```

Чтобы реализовать нестандартную сортировку, просто передайте именованный аргумент key встроенному инструменту сортировки – он должен не выполнять сравнение, а отображать значения в ключи сортировки, но результат получается такой же (например, смотрите упорядочивание по длине результатов поиска на графах выше):

```
>>> L = [{'n':3}, {'n':20}, {'n':0}, {'n':9}]
>>> L.sort(key=lambda x: x['n'])
>>> L
[{'n': 0}, {'n': 3}, {'n': 9}, {'n': 20}]
```

Оба инструмента сортировки принимают также логический флаг reverse, который обеспечивает сортировку по убыванию вместо сортировки по возрастанию – нет никакой необходимости переворачивать последовательность после сортировки. Базовая процедура сортировки в языке Python настолько хороша, что в документации к ней даже говорится, что она обладает «сверхъестественной производительностью» – весьма неплохо для сортировки.

Встроенные инструменты обращения последовательностей

Наши функции обращения обычно излишни по той же причине – язык Python предоставляет быстрые инструменты обращения, выполняющие преобразования непосредственно на месте и реализующие итеративный алгоритм, и вам, вероятно, правильнее будет использовать их, когда это возможно:

```
>>> L = [2, 4, 1, 3, 5]
>>> L.reverse()
>>> L
[5, 3, 1, 4, 2]

>>> L = [2, 4, 1, 3, 5]
>>> list(reversed(L))
[5, 3, 1, 4, 2]
```

Этот тезис специально несколько раз повторялся в данной главе, чтобы подчеркнуть ключевые моменты в работе Python: несмотря на большое количество исключений из правил, обычно лучше не изобретать колесо, если в этом нет явной необходимости. В конце концов, встроенные инструменты часто являются более удачным выбором.

Поймите правильно: иногда действительно бывают нужны объекты, которые добавляют новые возможности к встроенным типам или вы-

полняют что-то более индивидуальное. Например, классы множеств, с которыми мы познакомились выше, вводят инструменты, не поддерживаемые сегодня в Python непосредственно; двоичные деревья могут поддерживать алгоритмы, обеспечивающие более высокую эффективность, чем словари и множества; а реализация стека на основе деревьев кортежей оказалась действительно более быстрой, чем основанная на встроенных списках, в стандартных схемах применения. Иногда также возникает потребность в реализации собственных графов и функций перестановки.

Кроме того, как мы уже видели, инкапсуляция в классах позволяет изменять и расширять внутреннее устройство объекта, не трогая остальной части системы. Несмотря на то, что для решения большинства тех же проблем можно создавать подклассы от встроенных типов, тем не менее, в конечном итоге получаются те же самые нестандартные структуры данных.

Однако поскольку в Python есть набор встроенных, гибких и оптимизированных типов данных, реализация собственных структур данных часто не настолько необходима, как в хуже оснащенных низкоуровневых языках программирования. Прежде чем начать писать новый тип данных, обязательно спросите себя, не будет ли встроенный тип или функция лучше согласовываться с идеями Python.

Дополнительные структуры данных вы также найдете в относительно новом модуле `collections`, входящем в стандартную библиотеку Python. Как упоминалось в предыдущей главе, этот модуль реализует именованные кортежи, упорядоченные словари и многое другое. Он описан в руководстве по стандартной библиотеке Python, но исходный программный код этого модуля, как и многих других модулей в стандартной библиотеке, может также служить источником дополнительных примеров.

PyTree: универсальное средство просмотра деревьев объектов

Эта глава ориентировалась на использование командной строки. В завершение я хочу показать программу, в которой объединены технология создания графического интерфейса, изучавшаяся ранее, и некоторые идеи структур данных, с которыми мы познакомились здесь.

Эта программа называется PyTree и служит универсальным средством просмотра древовидных структур данных, написанным на языке Python с применением библиотеки tkinter. Программа PyTree рисует на экране узлы дерева в виде прямоугольников, соединенных стрелками. Она также умеет пересылать в дерево щелчки мыши на изображениях узлов, чтобы вызывать специфические для дерева действия. Так как PyTree позволяет визуализировать структуру дерева, генерируемого

набором параметров, это интересный способ исследования основанных на деревьях алгоритмов.

PyTree поддерживает произвольные типы деревьев, «обертывая» фактические деревья в объекты графического интерфейса. Объекты интерфейса реализуют стандартный протокол путем обмена данными с базовым объектом дерева. Для целей данной главы программа PyTree оснащена средствами отображения двоичных деревьев поиска; она также может выводить деревья синтаксического разбора выражений, что понадобится в следующей главе. Просмотр других типов деревьев осуществляется реализацией классов-оболочек, обеспечивающих взаимодействия с новыми типами деревьев.

Взаимодействие с графическим интерфейсом, используемым в программе PyTree, в полной мере освещалось ранее в этой книге. Так как она написана на языке Python с использованием библиотеки tkinter, то должна работать под управлением Windows, Unix и Mac. На верхнем уровне инструмент просмотра используется, как показано ниже:

```
root = Tk()                            # создать окно программы просмотра
bwrapper = BinaryTreeWrapper()         # добавить: строку ввода, кнопки
pwrapper = ParseTreeWrapper()          # создать объекты-обертки
viewer   = TreeViewer(bwrapper, root)  # запуск просмотра двоичного дерева

def onRadio():
    if var.get() == 'btree':
        viewer.setTreeType(bwrapper)   # изменить объект-обертку
    elif var.get() == 'ptree':
        viewer.setTreeType(pwrapper)
```

На рис. 18.2 изображено окно программы PyTree, выполняющейся под управлением Python 3.1 в Windows 7, созданное сценарием *treeview.py*, запущенным без аргументов. Программу PyTree можно также запустить щелчком на кнопке в панели запуска PyDemos, с которой мы познакомились в главе 10. Как обычно, чтобы получить более полное представление о ее работе, попробуйте запустить эту программу на своем компьютере. На этом снимке экрана изображено дерево только одного типа, однако программа PyTree способна отображать деревья произвольных типов и может даже переключаться между ними в процессе выполнения.

Как и в случае с программой PyForm, представленной в предыдущей главе, исходный программный код и описание этого примера для экономии места в этом издании были вынесены в пакет примеров. Дополнительную информацию о программе PyTree ищите в каталоге:

```
C:\...\PP4E\Dstruct\TreeView
```

Кроме того, как и в случае с программой PyForm, каталог *Documentation* содержит оригинальное описание этого примера из третьего издания – текущая реализация PyTree адаптирована для выполнения под управлением Python 3.X, но из третьего издания – нет.

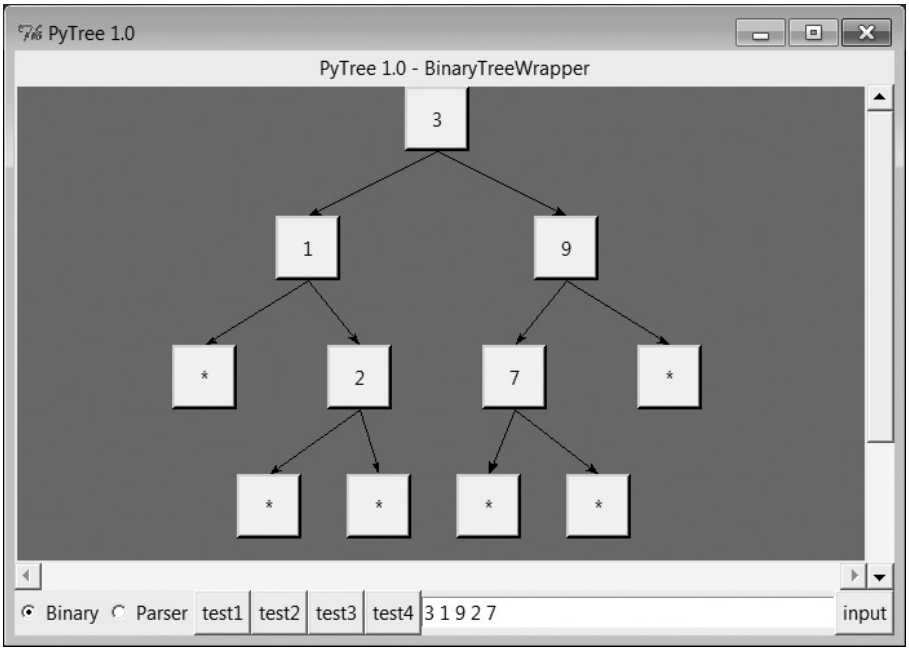


Рис. 18.2. *Просмотр бинарного дерева поиска в программе PyTree (кнопка test1)*

Как уже упоминалось, PyTree предусматривает возможность отображения двоичных деревьев поиска и деревьев синтаксического анализа выражений, что пригодится нам в следующей главе. При отображении последних PyTree превращается в своего рода визуальный калькулятор – вы можете генерировать деревья произвольных выражений и вычислять различные их части щелчком на отображаемых узлах. Но я не могу рассказать или показать вам что-то еще о деревьях этого типа, пока вы не перейдете к главе 19.

19

Текст и язык

«Пилите, Шура, пилите!»

Обработка текстовой информации в той или иной форме является одной из наиболее частых задач, которые приходится выполнять приложениям. Под этим может подразумеваться все, что угодно, – от просмотра текстового файла по колонкам до анализа инструкций языка, определяемого формальной грамматикой. Такую обработку обычно называют *синтаксическим анализом* (или парсингом) – разбором структуры текстовой строки. В этой главе мы исследуем способы обработки языковой и текстовой информации, а попутно во врезках будут кратко освещаться некоторые концепции разработки на языке Python. В процессе мы познакомимся с методами строк, приемами сопоставления с шаблонами, средствами синтаксического анализа разметки XML и HTML и другими инструментами.

Часть материала является достаточно сложной, но я привожу примеры небольшого объема, чтобы не удлинять главу. Например, синтаксический анализ методом рекурсивного спуска иллюстрируется простым примером, показывающим, как он может быть реализован на языке Python. Мы также увидим, что часто нет необходимости писать специальные инструменты анализа для всех задач обработки языков в Python. Обычно вместо этого можно экспортировать прикладной интерфейс и использовать его в программе Python, а иногда достаточно одного вызова встроенной функции. Завершится эта глава представлением PyCalc – калькулятора с графическим интерфейсом, написанного на языке Python и являющегося последним крупным примером программ в этой книге. Как мы увидим, написание калькуляторов состоит в основном в жонглировании стеками во время лексического анализа текста.

Стратегии обработки текста в Python

В целом существует множество различных способов обработки текста и синтаксического анализа в Python:

Выражения

Выражения со встроенными объектами строк

Методы

Вызовы методов встроенных объектов строк

Шаблоны

Сопоставление с шаблонами регулярных выражений

Программы синтаксического анализа (парсеры): разметка

Анализ текста в формате XML и HTML

Программы синтаксического анализа (парсеры): грамматики

Нестандартные механизмы синтаксического анализа (парсеры), собственные и сгенерированные

Встраивание

Выполнение программного кода Python с помощью встроенных функций `eval` и `exec`

Для простых задач часто достаточно встроенного строкового объекта Python. Строки в языке Python поддерживают операции обращения по индексу, конкатенации, извлечения среза и могут обрабатываться с помощью строковых методов и встроенных функций. Однако основное внимание в этой главе будет уделяться инструментам более высокого уровня и приемам анализа текстовой информации. Мы кратко рассмотрим все перечисленные способы по очереди. Приступим.

Некоторые читатели могли прийти в эту главу, рассчитывая получить информацию о поддержке Юникода, однако эта тема не будет рассматриваться здесь. Сведения о поддержке Юникода в Python вы найдете в обсуждении строковых инструментов в главе 2, в обсуждении кодировок и различий между текстовыми и двоичными файлами в главе 4 и в главе 9, в обсуждении поддержки текста в библиотеке tkinter. Кроме того, поддержка Юникода упоминается в различных темах, посвященных Интернету и базам данных (например, кодировки в электронных письмах).

Поскольку поддержка Юникода является одной из основных особенностей языка, все перечисленные главы также отсылают к обсуждению этой темы в четвертом издании книги «Изучаем Python» (*http://oreilly.com/catalog/9780596158071/*). Большинство инструментов, обсуждаемых в этой главе, включая строковые методы и регулярные выражения, автоматически поддерживают Юникод просто потому, что тип строки `str` в Python 3.X представляет строки Юникода, включая набор символов ASCII и его расширения.

Строковые методы

Первую остановку в нашем обзоре средств обработки и синтаксического анализа текста мы сделаем на наиболее простых из них: объекты строк в языке Python обладают массой инструментов для обработки текста и служат первой линией обороны в этой области. Как вы уже наверняка знаете, конкатенация, извлечение среза, форматирование и другие операции над строками являются рабочими лошадками в большинстве программ (я включил в эту категорию и новейший метод `format`, так как в действительности он является просто альтернативной версией оператора `%` форматирования):

```
>>> 'spam eggs ham'[5:10]            # извлечение среза: подстрока
'eggs '
>>> 'spam ' + 'eggs ham'             # конкатенация (и *, len(), [ix])
'spam eggs ham'
>>> 'spam %s %s' % ('eggs', 'ham')   # выражение форматирования: подстановка
'spam eggs ham'
>>> 'spam {} {}'.format('eggs', 'ham') # метод форматирования: альтернатива %
'spam eggs ham'
>>> 'spam = "%-5s", %+06d' % ('ham', 99)  # более сложное форматирование
'spam = "ham ", +00099'
>>> 'spam = "{0:<5}", {1:+06}'.format('ham', 99)
'spam = "ham ", +00099'
```

Эти операции рассматриваются в источниках, посвященных основам языка, таких как «Изучаем Python». Однако в этой главе нас интересуют более мощные инструменты: помимо операторов строковых выражений объекты строк в языке Python поддерживают самые разнообразные утилиты обработки текста, реализованные как *методы*. Мы познакомились с некоторыми из них в главе 2 и с тех пор постоянно использовали их. Например, экземпляр `str` встроенного строкового типа предоставляет следующие операции в виде методов объектов:

`str.find(`*`substr`*`)`

Выполняет поиск подстроки.

`str.replace(`*`old, new`*`)`

Выполняет подстановку подстроки.

`str.split(`*`delimiter`*`)`

Разбивает строку по указанному разделителю или пробельным символам.

`str.join(`*`iterable`*`)`

Объединяет подстроки, вставляя разделители между ними.

`str.strip()`

Удаляет ведущие и завершающие пробельные символы.

`str.rstrip()`

Удаляет только завершающие пробельные символы, если они имеются.

`str.rjust(width)`

Выравнивает строку по правому краю в поле фиксированной ширины.

`str.upper()`

Переводит все символы в верхний регистр.

`str.isupper()`

Проверяет – все ли символы в строке являются символами верхнего регистра.

`str.isdigit()`

Проверяет – все ли символы в строке являются цифрами.

`str.endswith(substr-or-tuple)`

Проверяет присутствие подстроки (или одного из вариантов в кортеже) в конце данной строки.

`str.startswith(substr-or-tuple)`

Проверяет присутствие подстроки (или одного из вариантов в кортеже) в начале данной строки.

Этот список является достаточно показательным, но неполным, и некоторые из представленных методов принимают дополнительные необязательные аргументы. Чтобы получить полный перечень строковых методов, выполните вызов `dir(str)` в интерактивной оболочке Python, а чтобы получить краткую справку по тому или иному методу, выполните вызов `help(str.method)`. Кроме того, исчерпывающий список можно найти в руководстве по стандартной библиотеке Python и в справочниках, таких как «Python Pocket Reference».

Более того, в современных версиях Python все обычные строковые методы могут применяться к строкам обоих типов, `bytes` и `str`. Последний из них позволяет применять методы к тексту, содержащему произвольные символы Юникода, просто потому, что строки типа `str` являются текстом Юникода, даже если он состоит только из символов ASCII. Первоначально эти методы были реализованы в виде функций в модуле `string`, но на сегодняшний день они доступны только в виде методов. Модуль `string` все еще присутствует в стандартной библиотеке, потому что он содержит ряд предопределенных констант (например, `string.ascii_uppercase`), а также реализацию интерфейса подстановки `Template`, появившуюся в версии Python 2.4, – один из инструментов, рассматриваемых в следующем разделе.

Обработка шаблонов с помощью операций замены и форматирования

Посредством краткого обзора рассмотрим строковые методы в контексте некоторых наиболее типичных случаев их использования. Как мы уже видели при создании HTML-страниц переадресации в главе 6, строковый метод replace зачастую вполне справляется с ролью инструмента обработки шаблонов – мы можем вычислять значения и вставлять их в фиксированные позиции строк простыми вызовами методов:

```
>>> template = '---$target1---$target2---'
>>> val1 = 'Spam'
>>> val2 = 'shrubbery'
>>> template = template.replace('$target1', val1)
>>> template = template.replace('$target2', val2)
>>> template
'---Spam---shrubbery---'
```

Кроме того, когда мы создавали HTML-страницы ответов в сценариях CGI в главах 15 и 16, мы видели, что оператор % форматирования строк также является мощным инструментом обработки шаблонов, особенно в соединении со словарями – достаточно просто заполнить словарь требуемыми значениями и выполнить множественную подстановку в строку HTML:

```
>>> template = """
... ---
... ---%(key1)s---
... ---%(key2)s---
... """
>>>
>>> vals = {}
>>> vals['key1'] = 'Spam'
>>> vals['key2'] = 'shrubbery'
>>> print(template % vals)
---
---Spam---
---shrubbery---
```

Начиная с версии Python 2.4, в модуле string появилась реализация интерфейса Template, который по сути является упрощенной и ограниченной версией только что показанной схемы форматирования на основе словаря, но который предоставляет некоторые дополнительные методы, более простые в использовании:

```
>>> vals
{'key2': 'shrubbery', 'key1': 'Spam'}

>>> import string
>>> template = string.Template('---$key1---$key2---')
>>> template.substitute(vals)
```

```
'---Spam---shrubbery---'

>>> template.substitute(key1='Brian', key2='Loretta')
'---Brian---Loretta---'
```

За дополнительной информацией об этом расширении обращайтесь к руководству по библиотеке. Несмотря на то, что сам строковый тип не поддерживает обработку текста с применением шаблонов, с которой мы познакомимся далее в этой главе, тем не менее, его инструменты обладают возможностями, достаточно широкими для большинства задач.

Анализ текста с помощью методов split и join

С учетом основной направленности этой главы особенно полезными для анализа текста оказываются встроенные инструменты Python для разбиения и соединения строк по элементам:

`str.split(delimiter?, maxsplits?)`

Разбивает строку на список подстрок, по пробельным символам (табуляция, пробел, перевод строки) или по явно указанной строке-разделителю. Параметр `maxsplits`, если указан, ограничивает количество выполняемых разбиений.

`delimiter.join(iterable)`

Объединяет последовательность или итерируемый объект подстрок (например, список, кортеж, генератор), добавляя между ними строку-разделитель, относительно которой был выполнен вызов.

Эти два метода являются наиболее мощными среди строковых методов. Как мы видели в главе 2, метод `split` разбивает строку в список подстрок, а метод `join` соединяет их вместе:

```
>>> 'A B C D'.split()
['A', 'B', 'C', 'D']
>>> 'A+B+C+D'.split('+')
['A', 'B', 'C', 'D']
>>> '--'.join(['a', 'b', 'c'])
'a--b--c'
```

Несмотря на свою простоту, они позволяют решать весьма сложные задачи обработки текста. Более того, строковые методы действуют очень быстро, потому что они реализованы на языке C. Например, чтобы быстро заменить в файле каждый символ табуляции четырьмя точками, достаточно просто передать этот файл на стандартный ввод следующего сценария:

```
from sys import *
stdout.write(('.' * 4).join(stdin.read().split('\t')))
```

Метод `split` здесь разбивает входную строку по символам табуляции, а метод `join` объединяет полученные подстроки, вставляя точки там, где раньше были символы табуляции. В данном случае комбинация вы-

зовов двух методов эквивалентна использованию более простого вызова строкового метода, выполняющего глобальную замену, как показано ниже:

```
stdout.write(stdin.read().replace('\t', '.' * 4))
```

Как будет показано в следующем разделе, операции разбиения строк вполне достаточно для решения многих задач обработки текста.

Суммирование по колонкам в файле

Рассмотрим несколько практических применений операций разбиения и объединения строк. Во многих областях довольно часто возникает необходимость просмотра файлов по колонкам. Допустим, например, что имеется вывод какой-то системы в файл в виде колонок чисел и требуется сложить числа в каждой колонке. В Python эту задачу можно решить с помощью операции разбиения строк, как показано в примере 19.1. Python предоставляет также дополнительный бонус, позволяя сделать это решение инструментом многократного пользования, оформив его в виде доступной для импортирования функции.

Пример 19.1. PP4E\Lang\summer.py

```
#!/usr/local/bin/python

def summer(numCols, fileName):
    sums = [0] * numCols                # создать список, заполненный нулями
    for line in open(fileName):         # просмотр строк в файле
        cols = line.split()             # разбить на колонки
        for i in range(numCols):        # по пробелам/табуляциям
            sums[i] += eval(cols[i])    # суммировать значения в sums
    return sums

if __name__ == '__main__':
    import sys
    print(summer(eval(sys.argv[1]), sys.argv[2])) # '% summer.py cols file'
```

Обратите внимание, что здесь для чтения строк мы использовали итератор файла вместо явного вызова метода `readlines` (в главе 4 говорилось, что итераторы файлов позволяют избежать необходимости загружать файл в память целиком). Сам файл является временным объектом, который будет автоматически закрыт сборщиком мусора.

Как обычно для корректно оформленных сценариев этот модуль можно *импортировать* и вызывать его функцию, а можно *запускать* его как инструмент командной строки. Сценарий *summer.py* вызывает метод `split`, чтобы создать список строк, представляющих ряды колонок, и `eval` – для преобразования строк в колонках в числа. Ниже приводится пример входного файла, в котором для разделения колонок используются пробелы и табуляции, и результат применения сценария к этому файлу:

```
C:\...\PP4E\Lang> type table1.txt
1       5       10    2   1.0
2       10      20    4   2.0
3       15      30    8    3
4       20      40   16   4.0

C:\...\PP4E\Lang> python summer.py 5 table1.txt
[10, 50, 100, 30, 10.0]
```

Обратите также внимание, что так как сценарий суммирования использует функцию `eval` для преобразования текста из файла в числа, то в файле можно было бы хранить произвольные выражения на языке Python:

```
C:\...\PP4E\Lang> type table2.txt
2      1+1         1<<1           eval("2")
16     2*2*2*2     pow(2,4)       16.0
3      len('abc')  [1,2,3][2]     {'spam':3}['spam']

C:\...\PP4E\Lang> python summer.py 4 table2.txt
[21, 21, 21, 21.0]
```

Суммирование с помощью функции zip и генераторов

Мы еще вернемся к функции `eval` далее в этой главе, когда будем исследовать приемы вычисления выражений. Но иногда она оказывается слишком мощным инструментом. Если нет уверенности, что выполняемые таким способом строки не содержат злонамеренный программный код, может потребоваться выполнять их, ограничив доступ к ресурсам компьютера, или использовать более строго инструменты преобразований. Взгляните на следующую версию функции `summer` (если у вас появится желание поэкспериментировать с ней, вы найдете ее в файле *summer2.py* в пакете примеров):

```
def summer(numCols, fileName):
    sums = [0] * numCols
    for line in open(fileName):        # использует итератор файла
        cols = line.split(',')         # поля разделяются запятыми
        nums = [int(x) for x in cols]  # ограниченный инстр. преобразования
        both = zip(sums, nums)         # отказ от вложенного цикла for
        sums = [x + y for (x, y) in both] # 3.X: zip возвращает итератор
    return sums
```

Используя для преобразования строк функцию `int`, эта версия поддерживает только числа, а не произвольные и, возможно, небезопасные выражения. Первые четыре строки в этой версии напоминают оригинал, однако для разнообразия эта версия исходит из предположения, что данные отделяются запятыми, а не пробельными символами, и использует генератор списков и функцию `zip`, чтобы избавиться от вложенного цикла `for`. Кроме того, эта версия существенно сложнее оригинала и по этой причине выглядит менее предпочтительной с точки зрения сопро-

вождения. Если что-то в этом программном коде вам не понятно, попробуйте добавить вызовы функции print после каждой инструкции, чтобы следить за результатами всех операций. Ниже показан результат работы этой функции:

```
C:\...\PP4E\Lang> type table3.txt
1,5,10,2,1
2,10,20,4,2
3,15,30,8,3
4,20,40,16,4

C:\...\PP4E\Lang> python summer2.py 5 table3.txt
[10, 50, 100, 30, 10]
```

Суммирование с помощью словарей

Функция summer вполне работоспособна, но ее можно обобщить еще больше – сделав номера колонок ключами словаря, а не смещениями в списке, можно вообще избавиться от необходимости использовать номера колонок. Помимо того, что словари позволяют ассоциировать данные с описательными метками, а не с числовыми позициями, они часто оказываются более гибкими, чем списки, особенно когда число колонок в файле не является фиксированным. Например, предположим, что необходимо вычислить суммы колонок данных в текстовом файле, когда число колонок заранее неизвестно или оно не является фиксированным:

```
C:\...\PP4E\Lang> python
>>> print(open('table4.txt').read())
001.1 002.2 003.3
010.1 020.2 030.3 040.4
100.1 200.2 300.3
```

В этой ситуации нет возможности заранее определить список сумм фиксированной длины, потому что количество колонок в разных строках может отличаться. Операция разбиения по пробельным символам извлекает колонки, а функция float преобразует их в числа, но список фиксированного размера не так-то просто приспособить для хранения сумм (по крайней мере, без дополнительного программного кода, управляющего его размером). Здесь удобнее использовать словари, потому что появляется возможность использовать позиции колонок в качестве ключей, вместо абсолютных смещений в списке. Это решение демонстрируется в следующем листинге интерактивного сеанса (определение функции находится также в файле *summer3.py* в пакете примеров):

```
>>> sums = {}
>>> for line in open('table4.txt'):
...     cols = [float(col) for col in line.split()]
...     for pos, val in enumerate(cols):
...         sums[pos] = sums.get(pos, 0.0) + val
...
>>> for key in sorted(sums):
```

```
...     print(key, '=', sums[key])
...
0 = 111.3
1 = 222.6
2 = 333.9
3 = 40.4

>>> sums
{0: 111.3, 1: 222.6, 2: 333.90000000000003, 3: 40.4}
```

Интересно отметить, что в этом программном коде используются инструменты, которыми Python обрастал с годами, – итераторы файлов и словарей, генераторы, метод `dict.get`, встроенные функции `enumerate` и `sorted` отсутствовали в Python в первые годы его развития. Похожие примеры можно найти в главе 9, в обсуждении механизма библиотеки tkinter компоновки по сетке, где для работы с таблицами также используется функция `eval`. Логика вычисления сумм, реализованная в той главе, является вариацией на эту же тему – она получает количество колонок из первой строки в файле с данными и выполняет суммирование для отображения в графическом интерфейсе.

Синтаксический анализ строк правил и обратное преобразование

Разбиение строк удобно использовать для деления текста на колонки, но эта операция также может использоваться как более универсальный инструмент синтаксического анализа – разбивая строку несколько раз, по разным разделителям, можно выполнить разбор более сложного текста. Анализ такого рода можно проводить с помощью более мощных инструментов, таких как регулярные выражения, с которыми мы познакомимся далее в этой главе, однако анализ на основе операции разбиения строки проще реализуется в виде быстро создаваемых прототипов, и, вполне возможно, он будет выполняться быстрее.

В примере 19.2 демонстрируется один из способов применения операций разбиения и объединения строк для синтаксического разбора предложений из простого языка. Он взят из основанной на правилах оболочки экспертной системы (holmes), которая написана на Python и включена в пакет примеров для книги (подробнее о holmes чуть ниже). Строки правил в системе holmes имеют вид:

```
"rule <id> if <test1>, <test2>... then <conclusion1>, <conclusion2>..."
```

Проверки и выводы являются конъюнкциями выражений («,» означает «и»). Каждое выражение является списком слов или переменных, разделенных пробелами; переменные начинаются с символа `?`. Перед использованием правила оно переводится во внутренний формат – словарь с вложенными списками. Чтобы отобразить правило, оно переводится обратно в строковый формат. Например, для вызова:

```
rules.internal_rule('rule x if a ?x, b then c, d ?x')
```

преобразование в функции internal_rule происходит, как показано ниже:

```
string = 'rule x if a ?x, b then c, d ?x'
i = ['rule x', 'a ?x, b then c, d ?x']
t = ['a ?x, b', 'c, d ?x']
r = ['', 'x']
result = {'rule':'x', 'if':[['a','?x'], ['b']], 'then':[['c'], ['d','?x']]}
```

Сначала выполняется разбиение по if, затем по then и наконец по rule. В результате получаются три подстроки, разделенные по ключевым словам. Подстроки проверок и выводов разбиваются сначала по «,», а затем по пробелам. Для обратного преобразования в исходную строку используется метод join. В примере 19.2 приводится конкретная реализация этой схемы.

Пример 19.2. PP4E\Lang\rules.py

```
def internal_rule(string):
    i = string.split(' if ')
    t = i[1].split(' then ')
    r = i[0].split('rule ')
    return {'rule': r[1].strip(), 'if':internal(t[0]), 'then':internal(t[1])}

def external_rule(rule):
    return ('rule '  + rule['rule']           +
            ' if '   + external(rule['if'])   +
            ' then ' + external(rule['then']) + '.')

def internal(conjunct):
    res = []                            # 'a b, c d'
    for clause in conjunct.split(','): # -> ['a b', ' c d']
        res.append(clause.split())     # -> [['a','b'], ['c','d']]
    return res

def external(conjunct):
    strs = []
    for clause in conjunct:            # [['a','b'], ['c','d']]
        strs.append(' '.join(clause))  # -> ['a b', 'c d']
    return ', '.join(strs)             # -> 'a b, c d'
```

В настоящее время для получения некоторых строк-выводов можно было бы использовать генераторы списков и выражения-генераторы. Функции internal и external, например, можно было бы реализовать так (смотрите файл *rules2.py*):

```
def internal(conjunct):
    return [clause.split() for clause in conjunct.split(',')]

def external(conjunct):
    return ', '.join(' '.join(clause) for clause in conjunct)
```

В результате требуемые вложенные списки и строка обрабатываются за один шаг. Эта реализация, возможно, будет работать быстрее – я оставляю читателю право самому решить, какой из вариантов более сложен для понимания. Как обычно, компоненты этого модуля можно протестировать в интерактивной оболочке:

```
>>> import rules
>>> rules.internal('a ?x, b')
[['a', '?x'], ['b']]

>>> rules.internal_rule('rule x if a ?x, b then c, d ?x')
{'then': [['c'], ['d', '?x']], 'rule': 'x', 'if': [['a', '?x'], ['b']]}

>>> r = rules.internal_rule('rule x if a ?x, b then c, d ?x')
>>> rules.external_rule(r)
'rule x if a ?x, b then c, d ?x.'
```

Это максимум того, что может предоставить подобный разбор путем разбиения строк по лексемам. Он не предоставляет прямой поддержки рекурсивной вложенности компонентов и довольно примитивно обрабатывает синтаксические ошибки. Но для задач анализа инструкций простых языков разбиения строк может оказаться достаточно, по крайней мере, для создания прототипов и экспериментальных систем. Потом всегда можно добавить более мощный механизм анализа правил или заново реализовать правила в виде программного кода или классов Python.

Подробнее об оболочке экспертной системы holmes

Как на практике можно применять эти правила? Как уже упоминалось, анализатор правил, с которым мы только что познакомились, входит в состав оболочки экспертной системы holmes. Holmes – довольно старая система, написанная в 1993 году, еще до появления версии Python 1.0. Она осуществляет прямой и обратный логический вывод согласно задаваемым правилам. Например, правило

```
rule pylike if ?X likes coding, ?X likes spam then ?X likes Python
```

можно использовать для проверки того, что любит ли некто Python (*обратный* вывод, от «then» к «if»), и для вывода о том, что некто любит Python, из множества известных фактов (*прямой* вывод, от «if» к «then»). Для логического вывода может использоваться несколько правил, состоящих из набора предложений, представленных в виде конъюнкции; правила, указывающие на одно заключение, представляют альтернативы. Система holmes также осуществляет попутно простой поиск по шаблону для присваивания значений переменным, имеющимся в правилах (например, `?X`), и может объяснить свои выводы.

Урок 1: делай прототип и переноси

Если вас волнует проблема производительности, старайтесь всюду использовать методы строковых объектов вместо таких инструментов, как регулярные выражения. Хотя это утверждение может стать неверным с развитием Python, тем не менее, обычно строковые методы действуют значительно быстрее, потому что они выполняют меньший объем работы.

Фактически сейчас мы касаемся истории развития Python. Современные строковые методы начинались как обычные функции Python в оригинальном модуле `string`. Позднее, благодаря широкому их использованию, они были оптимизированы и реализованы на языке C. При импорте модуля `string` происходила внутренняя замена большей части его содержимого функциями, импортированными из модуля расширения `strop`, написанного на языке C; методы модуля `strop` оказываются в 100–1000 раз быстрее, чем их эквиваленты на Python.

В результате была резко повышена производительность программ-клиентов модуля `string` без каких-либо изменений в интерфейсе. То есть клиенты модуля мгновенно стали быстрее без всякой модификации, за счет использования нового модуля на языке C. Аналогичный подход был применен к модулю `pickle`, с которым мы познакомились в главе 17, – появившиеся позднее модули `cPickle` в Python 2.X и `_pickle` в Python 3.X сохранили совместимость, но стали значительно быстрее.

Это служит хорошим уроком для тех, кто занимается разработкой на языке Python: модули можно сначала быстро разрабатывать на Python, а затем переносить на язык C, если требуется повысить их быстродействие. Поскольку интерфейсы к модулям Python и расширениям на языке C идентичны (оба типа модулей можно импортировать; оба содержат атрибуты в виде вызываемых функций), модули, перенесенные на язык C, являются обратно совместимыми со своими прототипами на языке Python. Для клиентов единственным результатом трансляции таких модулей является повышение производительности.

Обычно нет нужды переписывать все модули на языке C перед поставкой приложения клиенту: можно отобрать те модули, которые являются критическими для быстродействия (такие как `string` и `pickle`), и транслировать их, а остальные оставить на Python. Для выявления модулей, трансляция которых на язык C может дать наибольший выигрыш, используйте технику хронометража и профилирования, описанную в главе 18. Модули расширения, написанные на языке C, будут представлены в следующей главе.

Система holmes служила доказательством широких возможностей Python в целом, когда такие доказательства еще были необходимы, и последняя проверка показала, что она без внесения изменений способна работать под управлением Python 2.X. Однако ее реализация уже не отражает передовые методы программирования на языке Python. По этой причине я отказался от поддержки этой системы. В действительности она устарела настолько сильно и не развивалась так долго, что я решил не возвращаться к ней в этом издании вообще. Оригинальная реализация этой системы для Python 0.X включена в состав пакета примеров к книге, но она не была адаптирована для работы под управлением Python 3.X и не соответствует современным возможностям языка Python.

Таким образом, holmes можно признать системой, отжившей свое. Она утратила свой статус и не будет более обсуждаться здесь. Чтобы найти более современные инструменты искусственного интеллекта для Python, попробуйте поискать в Интернете. Это весьма интересная тема для исследования, при этом систему holmes, вероятно, лучше оставить в туманных далях истории Python (чтобы возродить ее, нужно обладать достаточным запасом смелости).

Поиск по шаблонам регулярных выражений

Операции разбиения и объединения строк представляют собой простой способ обработки текста, если он соответствует предполагаемому формату. Для решения более общих задач анализа текста, когда структура данных определена не так жестко, Python предоставляет средства сопоставления с регулярными выражениями. В частности, для текстовых данных, связанных с такими прикладными областями, как Интернет и базы данных, высокая гибкость регулярных выражений может оказаться очень полезной.

Регулярные выражения являются простыми строками, определяющими *шаблоны* для сопоставления с другими строками. Вы указываете шаблон и строку и спрашиваете, соответствует ли строка вашему шаблону. После нахождения совпадения части строки, соответствующие частям шаблона, становятся доступными сценарию. Таким образом, поиск соответствия не только дает ответ «да/нет», но и позволяет выбирать подстроки.

Строки шаблонов регулярных выражений могут быть весьма сложными (скажем честно – они могут выглядеть устрашающе). Но после того как вы освоитесь с ними, они смогут заменить собой подпрограммы поиска в строках, которые пришлось бы писать вручную, – обычно один шаблон способен выполнить работу десятков строк программного кода, выполняющего сканирование строк, и может оказаться намного быстрее. Регулярные выражения позволяют кратко описать ожидаемую структуру текста и извлекать ее части.

Модуль re

В Python регулярные выражения не входят в состав синтаксиса самого языка, но поддерживаются стандартным библиотечным модулем `re`, который нужно импортировать. Модуль определяет функции для немедленного сопоставления, компиляции строк шаблонов в объекты шаблонов, сопоставления этих объектов со строками и выбора совпавших подстрок после обнаружения соответствия. Он также предоставляет инструменты для выполнения разбиения, замены и других операций на основе регулярных выражений.

Модуль `re` обеспечивает богатый синтаксис шаблонов регулярных выражений, образцом при создании которого послужил способ представления шаблонов в языке Perl (регулярные выражения являются особенностью Perl, достойной подражания). Например, модуль `re` поддерживает понятия именованных групп, классов символов, а также поиск *минимального* (nongreedy) совпадения – модуль `re` поддерживает операторы шаблонов регулярных выражений, соответствующие минимально возможному числу совпадающих символов (другие операторы шаблонов регулярных выражений всегда соответствуют самой длинной возможной подстроке). Кроме того, модуль `re` неоднократно оптимизировался и в Python 3.X поддерживает возможность сопоставления с обоими типами строк, `bytes` и `str`. Таким образом, реализация регулярных выражений в Python поддерживает шаблоны в стиле языка Perl, но работа с ними происходит через интерфейс модуля.

Прежде чем двинуться дальше, я должен предупредить, что регулярные выражения являются сложным инструментом, который невозможно рассмотреть здесь во всех подробностях. Если эта область вас заинтересовала, то вам будет полезна книга «Mastering Regular Expressions» Джеффри Фридла (Jeffrey E.F. Friedl), выпущенная издательством O'Reilly.[1] Мы не сможем охватить здесь тему конструирования шаблонов настолько полно, чтобы превратить вас в эксперта в этой области. Однако если вы уже научились писать шаблоны, интерфейс модуля для выполнения операций сопоставления покажется вам простым. Интерфейс модуля настолько прост в использовании, что мы сразу перейдем к изучению примеров и только потом погрузимся в обсуждение подробностей.

Первые примеры

Существует два основных способа реализации сопоставлений: с помощью функций и с помощью методов предварительно скомпилированных объектов шаблонов. Второй способ, основанный на предварительно скомпилированных шаблонах, обеспечивает более высокую скорость выполнения, когда один и тот же шаблон применяется многократно – ко всем строкам в текстовом файле, например.

---

[1] Джеффри Фридл «Регулярные выражения. 3-е издание», СПб.: Символ-Плюс, 2008.

Для демонстрации попробуем выполнить поиск в следующих строках (полный листинг интерактивного сеанса, который приводится в этом разделе, можно найти в файле *re-interactive.txt*):

```
>>> text1 = 'Hello spam...World'
>>> text2 = 'Hello spam...other'
```

При сопоставлении, выполняемом следующим программным кодом, не производится предварительная компиляция шаблона: сопоставление выполняется немедленно и отыскивает все символы между словами «Hello» и «World» в текстовых строках:

```
>>> import re
>>> matchobj = re.match('Hello(.*)World', text2)
>>> print(matchobj)
None
```

Когда сопоставление завершается неудачей, как в данном примере (строка `text2` не содержит слово «World»), обратно возвращается объект `None`, который при проверке инструкцией `if` интерпретируется как «ложь».

В строке шаблона, использованной здесь в качестве первого аргумента функции `re.match`, слова «Hello» и «World» соответствуют сами себе, а конструкция `(.*)` означает – любой символ (`.`), повторяющийся ноль или более раз (`*`). Наличие круглых скобок сообщает интерпретатору, что он должен сохранить часть строки, совпавшую с этой конструкцией, в виде группы – подстроки, доступной после операции сопоставления. Чтобы увидеть, как это происходит, рассмотрим пример, где сопоставление завершается успехом:

```
>>> matchobj = re.match('Hello(.*)World', text1)
>>> print(matchobj)
<_sre.SRE_Match object at 0x009D6520>

>>> matchobj.group(1)
' spam...'
```

Когда сопоставление завершается успехом, обратно возвращается *объект соответствия*, который предоставляет интерфейс для извлечения совпавших подстрок – вызов `group(1)` возвращает первый, самый левый фрагмент испытуемой строки, совпавший с частью шаблона, заключенной в круглые скобки (с нашей конструкцией `(.*)`). Как уже упоминалось, операция сопоставления не просто дает ответ «да/нет» – заключая части шаблонов в круглые скобки, можно также получать совпавшие подстроки. В данном случае мы получили текст, находящийся между словами «Hello» и «World». Группу с номером `0` – всю строку, совпавшую с полным шаблоном, – удобно использовать, когда необходимо убедиться, что шаблон охватил весь текст испытуемой строки.

Для работы с предварительно скомпилированными шаблонами предоставляется похожий интерфейс, но под шаблоном в этом случае подразумевается *объект шаблона*, полученный вызовом функции `compile`:

```
>>> pattobj = re.compile('Hello(.*)World')
>>> matchobj = pattobj.match(text1)
>>> matchobj.group(1)
' spam...'
```

Напомню, что предварительная компиляция необходима для повышения скорости работы, когда один и тот же шаблон используется многократно, что обычно происходит при построчном сканировании содержимого файлов. Ниже демонстрируется немного более сложный пример, иллюстрирующий обобщенность шаблонов. Этот шаблон допускает наличие нуля и более символов пробела или табуляции в начале строки (`[ \t]*`), пропускает один или более таких символов после слова «Hello» (`[ \t]+`), сохраняет символы в середине (`(.*)`) и допускает, что завершающее слово может начинаться с заглавной или строчной буквы (`[Ww]`). Как видите, шаблоны способны обрабатывать самые разнообразные данные:

```
>>> patt = '[ \t]*Hello[ \t]+(.*)[Ww]orld'
>>> line = ' Hello spamworld'
>>> mobj = re.match(patt, line)
>>> mobj.group(1)
'spam'
```

Обратите внимание, что в последнем фрагменте мы сопоставляли шаблон типа `str` со строкой типа `str`. Мы можем также сопоставлять строки типа `bytes`, чтобы обработать, например, кодированный текст, но мы не можем смешивать строки разных типов (ограничение, которое действует в Python повсюду, – интерпретатор не имеет информации о кодировке, необходимой для преобразования строк байтов в текст Юникода):

```
>>> patt = b'[ \t]*Hello[ \t]+(.*)[Ww]orld' # строки bytes допустимы
>>> line = b' Hello spamworld'              # возвращает группу типа bytes
>>> re.match(patt, line).group(1)           # но нельзя смешивать str/bytes
b'spam'

>>> re.match(patt, ' Hello spamworld')
TypeError: can't use a bytes pattern on a string-like object
```

(TypeError: нельзя смешивать шаблоны типа bytes со строковыми объектами)

```
>>> re.match('[ \t]*Hello[ \t]+(.*)[Ww]orld', line)
TypeError: can't use a string pattern on a bytes-like object
```

(TypeError: нельзя смешивать строковые шаблоны с объектами типа bytes)

В дополнение к инструментам, продемонстрированным в этих примерах, существуют инструменты для сканирования вперед, в поисках одного совпадения (`search`), всех совпадений (`findall`), разбиения и замены по шаблону и так далее. Все они имеют аналогичные реализации в виде функций модуля и методов скомпилированных шаблонов. В следую-

щем разделе приводятся несколько примеров, демонстрирующих основные возможности.

Строковые операции и шаблоны

Обратите внимание, что в предыдущем примере была предусмотрена возможность пропуска необязательных пробельных символов и использования заглавных и строчных букв. Это позволяет выделить основную причину, по которой может оказаться предпочтительнее использовать шаблоны, – они поддерживают более обобщенный способ обработки текста, чем строковые методы. Ниже приводится еще один показательный пример: мы уже видели, что строковые методы способны разбивать строки и выполнять замену подстрок, но они не смогут проделать эту работу, если в качестве разделителей используются различные строки:

```
>>> 'aaa--bbb--ccc'.split('--')
['aaa', 'bbb', 'ccc']
>>> 'aaa--bbb--ccc'.replace('--', '...') # строковые методы используют
'aaa...bbb...ccc'                        # фиксированные строки

>>> 'aaa--bbb==ccc'.split(['--', '=='])
TypeError: Can't convert 'list' object to str implicitly
(TypeError: Невозможно неявно преобразовать объект 'list' в str)
>>> 'aaa--bbb==ccc'.replace(['--', '='], '...')
TypeError: Can't convert 'list' object to str implicitly
(TypeError: Невозможно неявно преобразовать объект 'list' в str)
```

Шаблоны не только справляются с подобными задачами, но и дают возможность напрямую определять альтернативы, благодаря особому синтаксису шаблонов. В следующем примере конструкция `--|==` соответствует *либо* строке `--`, *либо* строке `==`. Конструкции `[-=]` соответствует либо символ `-`, либо `=` (множество символов). А конструкция `(?:)` может использоваться для группировки вложенных частей шаблона без создания группы, сохраняющей подстроку (функция `split` интерпретирует группы особым образом):

```
>>> import re
>>> re.split('--', 'aaa--bbb--ccc')
['aaa', 'bbb', 'ccc']
>>> re.sub('--', '...', 'aaa--bbb--ccc')    # случай с единственным
                                            # разделителем
'aaa...bbb...ccc'

>>> re.split('--|==', 'aaa--bbb==ccc')      # разбить по -- или ==
['aaa', 'bbb', 'ccc']
>>> re.sub('--|==', '...', 'aaa--bbb==ccc') # заменить -- или ==
'aaa...bbb...ccc'

>>> re.split('[-=]', 'aaa-bbb=ccc')      # односимвольные альтернативы
['aaa', 'bbb', 'ccc']
```

```
>>> re.split('(--)|(==)', 'aaa--bbb==ccc')     # результат разбиения
['aaa', '--', None, 'bbb', None, '==', 'ccc']  # включает группы
>>> re.split('(?:--)|(?:==)', 'aaa--bbb==ccc') # часть выражения, не группы
['aaa', 'bbb', 'ccc']
```

Операция разбиения тоже может извлекать простые подстроки при использовании фиксированных разделителей, но шаблоны способны обрабатывать окружающий контекст подобно скобкам, помечать части как необязательные, пропускать пробельные символы и многое другое. Проверки `\s*` в следующем примере обозначают ноль или более пробельных символов (задан класс символов); `\s+` – означают один или более таких символов; `/?` соответствует необязательному символу слэша; `[a-z]` – любому строчному символу (задан диапазон); `(.*?)` обозначает необходимость сохранения подстроки, состоящей из нуля или более любых символов – но не более, чем необходимо для совпадения с оставшейся частью шаблона (минимально); а метод `groups` используется для извлечения сразу всех подстрок, совпавших с фрагментами шаблона, заключенными в круглые скобки:

```
>>> 'spam/ham/eggs'.split('/')
['spam', 'ham', 'eggs']

>>> re.match('(.*)/(.*)/(.*)', 'spam/ham/eggs').groups()
('spam', 'ham', 'eggs')

>>> re.match('<(.*)>/<(.*)>/<(.*)>', '<spam>/<ham>/<eggs>').groups()
('spam', 'ham', 'eggs')

>>> re.match('\s*<(.*)>/?<(.*)>/?<(.*)>', ' <spam>/<ham><eggs>').groups()
('spam', 'ham', 'eggs')

>>> 'Hello pattern world!'.split()
['Hello', 'pattern', 'world!']
>>> re.match('Hello\s*([a-z]*)\s+(.*?)\s*!', 'Hellopattern world !').groups()
('pattern', 'world')
```

На практике иногда существует более одного совпадения. Метод `findall` предоставляет возможность разделывать объекты строк в пух и прах – он отыскивает все совпадения с шаблоном и возвращает все совпавшие подстроки (или список кортежей, при наличии нескольких групп). Метод `search` похож на него, но он прекращает сопоставление после первого же совпадения – он напоминает метод `match` с начальным сканированием вперед. В следующем примере строковый метод обнаруживает только одну заданную строку, а шаблоны позволяют отыскать и извлечь текст в скобках, находящийся в любом месте строки, даже при наличии дополнительного текста:

```
>>> '<spam>/<ham>/<eggs>'.find('ham')     # отыскать смещение подстроки
8
>>> re.findall('<(.*?)>', '<spam>/<ham>/<eggs>') # отыскать все
['spam', 'ham', 'eggs']                          # совпадения/группы
```

```
>>> re.findall('<(.*?)>', '<spam> / <ham><eggs>')
['spam', 'ham', 'eggs']

>>> re.findall('<(.*?)>/?<(.*?)>', '<spam>/<ham> ... <eggs><cheese>')
[('spam', 'ham'), ('eggs', 'cheese')]
>>> re.search('<(.*?)>/?<(.*?)>',
              'todays menu: <spam>/<ham>...<eggs><s>').groups()
('spam', 'ham')
```

При использовании `findall` особенно удобно использовать оператор `(?s)`, чтобы обеспечить совпадение `.` с символами *конца строки* в многострочном тексте – без этого `.` будет совпадать с любыми другими символами *кроме* символов конца строки. Ниже демонстрируется поиск двух смежных строк в скобках с произвольным текстом между ними, с пропуском и без пропуска символов конца строки:

```
>>> re.findall('<(.*?)>.*<(.*?)>', '<spam> \n <ham>\n<eggs>')   # останов
                                                               # на \n
[]
>>> re.findall('(?s)<(.*?)>.*<(.*?)>', '<spam> \n <ham>\n<eggs>')  # максим.
[('spam', 'eggs')]
>>> re.findall('(?s)<(.*?)>.*?<(.*?)>', '<spam> \n <ham>\n<eggs>') # миним.
[('spam', 'ham')]
```

Чтобы сделать крупные шаблоны более удобочитаемыми, группам с совпавшими подстроками можно даже присваивать *имена*, используя синтаксис `<?P<name>`, и после сопоставления извлекать их по именам, однако эта особенность имеет ограниченное применение для `findall`. В следующем примере выполняется поиск строк с символами «слов» (`\w`), разделенных символами `/`; это не более чем операция разбиения строки, но в результате получаются именованные группы, и обе функции, `search` и `findall`, выполняют сканирование вперед:

```
>>> re.search('(?P<part1>\w*)/(?P<part2>\w*)', '...aaa/bbb/ccc]').groups()
('aaa', 'bbb')
>>> re.search('(?P<part1>\w*)/(?P<part2>\w*)', '...aaa/bbb/ccc]').groupdict()
{'part1': 'aaa', 'part2': 'bbb'}

>>> re.search('(?P<part1>\w*)/(?P<part2>\w*)', '...aaa/bbb/ccc]').group(2)
'bbb'
>>> re.search('(?P<part1>\w*)/(?P<part2>\w*)',
              '...aaa/bbb/ccc]').group('part2')
'bbb'

>>> re.findall('(?P<part1>\w*)/(?P<part2>\w*)', '...aaa/bbb ccc/ddd]')
[('aaa', 'bbb'), ('ccc', 'ddd')]
```

Наконец, несмотря на то, что базовых операций, таких как извлечение среза и разбиение, часто бывает вполне достаточно, шаблоны позволяют получить более гибкое решение. В следующем примере используется конструкция `[^ ]`, соответствующая любому символу, *отличному* от

символа, следующего за ^, и выполняется экранирование дефиса в наборе альтернатив внутри [] с использованием \- , благодаря чему он не воспринимается как разделитель в определении диапазона символов. Здесь выполняются эквивалентные операции извлечения срезов, разбиения и сопоставления, а также более обобщенная операция сопоставления, решающая задачу, непосильную для других двух способов:

```
>>> line = 'aaa bbb ccc'
>>> line[:3], line[4:7], line[8:11] # срез извлекается
('aaa', 'bbb', 'ccc')               # по фиксированным смещениям
>>> line.split()                    # разбиение выполняется по фиксированным
['aaa', 'bbb', 'ccc']               # разделителям

>>> re.split(' +', line)            # обобщенное разбиение по разделителям
['aaa', 'bbb', 'ccc']
>>> re.findall('[^ ]+', line)       # поиск строк, не являющихся
                                    # разделителями
['aaa', 'bbb', 'ccc']

>>> line = 'aaa...bbb-ccc / ddd.-/e&e*e' # обработка обобщенного текста
>>> re.findall('[^ .\-/]+', line)
['aaa', 'bbb', 'ccc', 'ddd', 'e&e*e']
```

Если в прошлом вам не приходилось использовать регулярные выражения, от представленных примеров ваша голова может пойти кругом (или вообще отключиться!). Поэтому, прежде чем перейти к другим примерам, познакомимся с некоторыми подробностями, лежащими в основе модуля re и его шаблонов регулярных выражений.

Использование модуля re

Модуль re содержит функции, которые могут искать совпадения с шаблонами сразу или создавать компилированные объекты шаблонов для выполнения поиска в будущем. Объекты шаблонов (и вызовы функций поиска модуля) в свою очередь генерируют объекты соответствия, которые содержат информацию о найденных совпадениях и соответствующих подстроках. Следующие несколько разделов описывают интерфейсы модуля и некоторые операторы, которые можно использовать для написания шаблонов.

Функции модуля

Верхний уровень модуля предоставляет функции поиска, замены, предварительной компиляции и так далее:

compile(pattern [, flags])

Компилирует строку pattern шаблона регулярного выражения в объект регулярного выражения для выполнения поиска в будущем. Значение аргумента flags смотрите в справочном руководстве или в книге «Python Pocket Reference».

`match(pattern, string [, flags])`

Если ноль или более символов в начале строки `string` соответствуют строке шаблона `pattern`, возвращает соответствующий экземпляр объекта соответствия, или `None`, если соответствие не найдено. По своему действию напоминает функцию `search` с шаблоном, который начинается с оператора ^.

`search(pattern, string [, flags])`

Отыскивает в строке `string` место, соответствующее шаблону `pattern`, и возвращает объект соответствия или `None`, если соответствие не найдено.

`findall(pattern, string [, flags])`

Возвращает список строк, соответствующих всем неперекрывающимся совпадениям шаблона `pattern` в строке `string`. При наличии групп в шаблонах возвращает список групп и возвращает список кортежей, если в шаблоне имеется более одной группы.

`finditer(pattern, string [, flags])`

Возвращает итератор, выполняющий обход всех неперекрывающихся совпадений шаблона `pattern` в строке `string`.

`split(pattern, string [, maxsplit, flags])`

Разбивает строку `string` по совпадениям с шаблоном `pattern`. Если в шаблоне используются сохраняющие круглые скобки (`()`), то возвращается также текст всех групп в шаблоне.

`sub(pattern, repl, string [, count, flags])`

Возвращает строку, полученную заменой (первых `count`) самых левых неперекрывающихся совпадений с шаблоном `pattern` (строкой или объектом шаблона) в строке `string` на `repl` (которая может быть строкой с обратными ссылками на совпавшие группы или функцией, принимающей единственный объект соответствия и возвращающей строку замены).

`subn(pattern, repl, string [, count, flags])`

То же, что и `sub`, но возвращает кортеж: (новая-строка, число-сделанных-подстановок).

`escape(string)`

Возвращает строку `string`, в которой все не алфавитно-цифровые символы экранированы символом обратного слэша, чтобы ее можно было компилировать как строковый литерал.

Компилированные объекты шаблонов

Объекты шаблонов предоставляют аналогичные методы, но при этом строка шаблона хранится в самом объекте. Функция `re.compile`, представленная в предыдущем разделе, полезна для оптимизации шаблонов, сопоставление с которыми может выполняться неоднократно (ком-

пилированные шаблоны действуют быстрее). Объекты шаблонов, возвращаемые функцией re.compile, обладают следующими методами:

```
match(string [, pos] [, endpos])
search(string [, pos] [, endpos])
findall(string [, pos [, endpos]])
finditer(string [, pos [, endpos]])
split(string [, maxsplit])
sub(repl, string [, count])
subn(repl, string [, count])
```

Они действуют точно так же, как и одноименные функции модуля re, но шаблон хранится в самом объекте, а аргументы pos и endpos указывают позицию начала и конца поиска соответствия в строке.

Объекты соответствия

Наконец, если функция или метод match или search обнаруживают совпадение, они возвращают объект соответствия (None, если поиск не увенчался успехом). Объекты соответствия экспортируют ряд собственных атрибутов, в том числе:

group(g)
group([g1, g2, ...])

Возвращает подстроки, совпавшие с группой (или группами) в круглых скобках в шаблоне. Нумерация групп начинается с 1 – группа с номером 0 содержит всю строку, совпавшую с полным шаблоном. Если при вызове передается несколько номеров групп, возвращает кортеж. При вызове без аргументов по умолчанию возвращается группа с номером 0.

groups()

Возвращает кортеж всех подстрок, соответствующих группам (для групп с номерами 1 и выше).

groupdict()

Возвращает словарь, содержащий все совпавшие именованные группы (смотрите описание конструкции (?P<name>R) ниже).

start([group]) end([group])

Индексы начала и конца подстроки, совпавшей с группой group (или всей найденной строки при вызове без аргумента group).

span([group])

Возвращает кортеж из двух элементов: (start(group), end(group)).

expand([template])

Выполняет подстановку содержимого групп – смотрите руководство по библиотеке Python.

Шаблоны регулярных выражений

Строки регулярных выражений строятся путем конкатенации односимвольных форм регулярных выражений, представленных в табл. 19.1. Обычно для каждой формы отыскивается самое длинное совпадение, за исключением минимальных операторов. В таблице символ R означает любое регулярное выражение, C обозначает символ, а N обозначает цифру.

Таблица 19.1. Синтаксис шаблонов re

Оператор	Описание
.	Соответствует любому символу (включая перевод строки, если установлен флаг DOTALL или перед точкой стоит конструкция (?s))
^	Соответствует началу строки (каждой строки в режиме MULTILINE)
$	Соответствует концу строки (каждой строки в режиме MULTILINE)
C	Любой неспециальный (или экранированный обратным слэшем) символ соответствует самому себе
R*	Ноль или более совпадений с предшествующим регулярным выражением R (как можно больше)
R+	Одно или более совпадений с предшествующим регулярным выражением R (как можно больше)
R?	Ноль или одно совпадение с предшествующим регулярным выражением R (необязательное)
R{m}	Точное число m совпадений с предшествующим регулярным выражением R: a{5} соответствует подстроке 'aaaaa'
R{m,n}	От m до n совпадений с предшествующим регулярным выражением R
R*?, R+?, R??, R{m,n}?	То же, что *, + и ?, но соответствует минимально возможному числу совпадений; известны как *минимальные* (нежадные) квантификаторы (в отличие от прочих, ищут и поглощают минимально возможное количество символов)
[...]	Определяет набор символов: например, [a-zA-Z] соответствует любой букве (альтернативы; символ - обозначает диапазон)
[^...]	Определяет дополняющий набор символов: соответствует символам, отсутствующим в наборе

Оператор	Описание
\	Экранирует специальные символы (например, *?+|()) и вводит специальные последовательности, представленные в табл. 19.2
\\	Соответствует литералу \ (в шаблоне записывается как \\\\ или используйте r'\\')
\N	Соответствует содержимому группы с номером N: '(.+) \1' совпадет со строкой "42 42"
R|R	Альтернатива: соответствие левому или правому выражению R
RR	Конкатенация: соответствие обоим выражениям R
(R)	Соответствует любому выражению R внутри () и создает группу (сохраняет совпавшую подстроку)
(?:R)	То же, что и (R), но просто отделяет часть регулярного выражения и не создает сохраняющую группу
(?=R)	Опережающая проверка: соответствует, если имеется совпадение R с символами строки, следующими далее, но не поглощает их (например, X (?=Y) соответствует символу X, только если за ним следует символ Y)
(?!R)	Соответствует, если выражение R не соответствует символам, следующим далее; проверка, обратная по отношению к (?=R)
(?P<name>R)	Соответствует любому регулярному выражению R в () и создает именованную группу
(?P=name)	Соответствует тексту, найденному предшествующей группой с именем name
(?#...)	Комментарий; игнорируется
(?letter)	Флаг режима; letter может иметь значение a, i, L, m, s, u, x (смотрите руководство по библиотеке)
(?<=R)	Ретроспективная проверка: соответствует, если текущей позиции в строке предшествует совпадение с выражением R, завершающееся в текущей позиции
(?<!R)	Соответствует, если текущей позиции в строке не предшествует совпадение с выражением R; проверка, обратная по отношению к (?<=R)
(?(id/name/yespattern| nopattern)	Пытается отыскать совпадение с шаблоном yespattern, если существует группа с номером id или именем name; иначе используется шаблон nopattern

Внутри шаблонов можно сочетать диапазоны и символы. Например, [a-zA-Z0-9_]+ соответствует самой длинной строке из одного или более символов букв, цифр или подчеркиваний. Специальные символы преобра-

зуются как обычно в строках Python: `[\t ]*` соответствует нулю или более табуляций и пробелов (то есть пропускаются пробельные символы).

Конструкция группировки с помощью круглых скобок, `(R)`, позволяет извлекать совпавшие подстроки после успешного поиска соответствия. Часть строки, соответствующая выражению в круглых скобках, сохраняется в нумерованном регистре. Доступ к ней после успешного поиска производится с помощью метода `group` объекта соответствия.

Помимо конструкций, описанных в этой таблице, в шаблонах можно также использовать специальные последовательности, представленные в табл. 19.2. В соответствии с правилами Python для строк иногда нужно удваивать символ обратного слэша (\\) или пользоваться необработанными строками Python (`r'...'`), чтобы сохранить обратные слэши в шаблоне. Интерпретатор игнорирует обратные слэши в обычных строках, если следующий за ним символ не имеет специального значения. Некоторые последовательности в табл. 19.2 учитывают особенности Юникода, когда в сопоставлении участвуют строки `str`, а не `bytes`, а с целью эмуляции поведения для строк `bytes` можно использовать флаг ASCII; за подробностями обращайтесь к руководству по библиотеке Python.

Таблица 19.2. Специальные последовательности re

Последовательность	Описание
`\`*число*	Соответствует содержимому группы с номером *число* (нумерация начинается с 1)
`\A`	Соответствует только началу строки
`\b`	Пустая строка на границе слова
`\B`	Пустая строка не на границе слова
`\d`	Любая десятичная цифра (`[0-9]` для ASCII)
`\D`	Любой символ, не являющийся десятичной цифрой (`[^0-9]` для ASCII)
`\s`	Любой пробельный символ (`[ \t\n\r\f\v]` для ASCII)
`\S`	Любой непробельный символ (`[^ \t\n\r\f\v]` для ASCII)
`\w`	Любой алфавитно-цифровой символ (`[a-zA-Z0-9_]` для ASCII)
`\W`	Любой не алфавитно-цифровой символ (`[^a-zA-Z0-9_]` для ASCII)
`\Z`	Соответствует только концу строки

Большинство стандартных экранированных последовательностей, которые поддерживаются строковыми литерами Python, также принимаются механизмом анализа регулярных выражений: `\a`, `\b`, `\f`, `\n`, `\r`, `\t`,

\v, \x и \\. В руководстве по стандартной библиотеке Python вы найдете описание этих экранированных последовательностей и дополнительные сведения о синтаксисе шаблонов в целом. Но чтобы продемонстрировать типичное использование интерфейсов модуля re, вернемся к программному коду.

Дополнительные примеры шаблонов

Чтобы помочь вам лучше разобраться в рассматриваемой теме, ниже приводятся несколько коротких примеров сценариев, демонстрирующих простые, но показательные формы шаблонов. Комментарии в примере 19.3 описывают осуществляемые операции; смотрите по табл. 19.1, какие операторы здесь используются. Если что-то для вас все равно останется непонятным, попробуйте выполнить инструкции из этих сценариев в интерактивной оболочке и вместо start() вызывать метод group(0), чтобы видеть, какие строки соответствуют шаблонам.

Пример 19.3. PP4E\Lang\re-basics.py

```
"""
литералы, множества, диапазоны, альтернативы
и экранированные последовательности
все тесты здесь выводят число 2: смещение, где найдено совпадение
"""

import re                              # единственный, необходимый сейчас

pattern, string = "A.C.", "xxABCDxx"  # неспец. символы соотв. сами себе
matchobj = re.search(pattern, string) # '.' означает один любой символ
if matchobj:                        # search возвращ. объект соотв. или None
    print(matchobj.start())         # start - смещение совпадения

pattobj = re.compile("A.*C.*")        # 'R*' - ноль или более совпадений
matchobj = pattobj.search("xxABCDxx") # compile возвращает объект шаблона
if matchobj:                          # patt.search возвращ. объект соотв.
    print(matchobj.start())

# наборы выбора
print(re.search(" *A.C[DE][D-F][^G-ZE]G\t+ ?", "..ABCDEFG\t..").start())

# альтернативы: R1|R2 означает R1 или R2
print(re.search("(A|X)(B|Y)(C|Z)D", "..AYCD..").start()) # проверяем каждый
                                                         # символ
print(re.search("(?:A|X)(?:B|Y)(?:C|Z)D","..AYCD..").start()) # без сохр.
print(re.search("A|XB|YC|ZD", "..AYCD..").start())      # совпадет только A!
print(re.search("(A|XB|YC|ZD)YCD", "..AYCD..").start()) # лишь первый символ

# границы слов
print(re.search(r"\bABCD", "..ABCD ").start()) # \b означает границу слова
print(re.search(r"ABCD\b", "..ABCD ").start()) # используйте r'...'
                                               # для экранирования '\'
```

Обратите внимание на наличие нескольких способов выполнить поиск совпадения с использованием `re`: вызовом функций поиска модуля и посредством создания компилированных объектов шаблонов. В любом случае в результате можно получить объект соответствия. Все вызовы функции `print` в этом сценарии выводят результат `2` – смещение в строке, по которому было найдено соответствие с шаблоном. Например, в первом тесте шаблону `A.C.` соответствует подстрока `ABCD` со смещением `2` в проверяемой строке (то есть после начальных символов `xx`):

```
C:\...\PP4E\Lang> python re-basic.py
2
```
...8 других двоек опущено...

В примере 19.4 части строк шаблонов, заключенные в круглые скобки, образуют *группы* – части строки, которым они соответствуют, будут доступны после выполнения поиска.

Пример 19.4. PP4E\Lang\re-groups.py

```
"""
группы: извлечение подстрок, соответствующих частям
регулярных выражений в '()'
группы обозначаются номерами позиций, но конструкция (?P<name>R) позволяет
присваивать им символические имена
"""


import re

patt = re.compile("A(.)B(.)C(.)")                  # сохраняет 3 подстроки
mobj = patt.match("A0B1C2")                        # группы '()', 1..n
print(mobj.group(1), mobj.group(2), mobj.group(3)) # group() возвр. подстр.

patt = re.compile("A(.*)B(.*)C(.*)")               # сохраняет 3 подстроки
mobj = patt.match("A000B111C222")                  # groups() возвр. все гр.
print(mobj.groups())

print(re.search("(A|X)(B|Y)(C|Z)D", "..AYCD..").groups())
print(re.search("(?P<a>A|X)(?P<b>B|Y)(?P<c>C|Z)D", "..AYCD..").groupdict())

patt = re.compile(r"[\t ]*#\s*define\s*([a-z0-9_]*)\s*(.*)")
mobj = patt.search(" # define spam 1 + 2 + 3")    # поиск инструкций #define
print(mobj.groups())                              # \s - пробельный символ
```

Например, в первом тесте имеются три группы `(.)`, каждая из которых соответствует одному символу и сохраняет найденный символ – метод `group` возвращает найденные совпадения. Во втором тесте группы `(.*)` соответствуют любому числу символов и сохраняют их. В третьем и четвертом тестах показано, как можно группировать альтернативы по позициям и именам, а последний шаблон здесь соответствует строкам с инструкциями `#define` языка C – подробнее об этом чуть ниже:

```
C:\...\PP4E\Lang> python re-groups.py
0 1 2
('000', '111', '222')
('A', 'Y', 'C')
{'a': 'A', 'c': 'C', 'b': 'Y'}
('spam', '1 + 2 + 3')
```

Наконец, кроме поиска совпадений и извлечения подстрок, в модуле re есть средства замены строк (пример 19.5).

Пример 19.5. PP4E\Lang\re-subst.py

```
"замена: замещает совпадения с шаблоном в строке"

import re
print(re.sub('[ABC]', '*', 'XAXAXBXBXCXC'))
print(re.sub('[ABC]_', '*', 'XA-XA_XB-XB_XC-XC_'))     # замещает символ + _
print(re.sub('(.) spam', 'spam\\1', 'x spam, y spam')) # обратная ссылка
                                                       # на группу (или r'')
def mapper(matchobj):
    return 'spam' + matchobj.group(1)

print(re.sub('(.) spam', mapper, 'x spam, y spam'))    # функция отображения
```

В первом тесте заменяются все символы, присутствующие в наборе, – вслед за ними должен следовать символ подчеркивания. Последние два теста иллюстрируют обратные ссылки на группы и функцию отображения, используемую при выполнении операции замены. Обратите внимание, что конструкцию \1 необходимо экранировать, как \\1, из-за правил оформления строк в языке Python – точно так же можно было бы использовать r'spam\1'. Дополнительные примеры замены и разбиения строк смотрите в тестах в интерактивной оболочке в этом разделе выше.

```
C:\...\PP4E\Lang> python re-subst.py
X*X*X*X*X*X*X*
XA-X*XB-X*XC-X*
spamx, spamy
spamx, spamy
```

Поиск совпадений с шаблонами в файлах заголовков C

В завершение обратимся к более практичному примеру: сценарий в примере 19.6 объединяет представленные выше операторы шаблонов для решения более практичной задачи. С помощью регулярных выражений он отыскивает строки #define и #include в файлах заголовков на языке C и извлекает их составляющие. Обобщенность этих шаблонов позволяет обнаруживать целый ряд строк; группы в шаблонах (части, заключенные в круглые скобки) используются для извлечения из строк соответствующих подстрок после нахождения соответствия.

Пример 19.6. PP4E\Lang\cheader.py

```
"Сканирует файлы заголовков C и извлекает части инструкций #define
и #include"

import sys, re
pattDefine = re.compile(                    # компилировать в объект шаблона
    '^#[\t ]*define[\t ]+(\w+)[\t ]*(.*)') # "# define xxx yyy..."
                                            # \w аналог [a-zA-Z0-9_]
pattInclude = re.compile(
    '^#[\t ]*include[\t ]+[<"]([\w\./]+)') # "# include <xxx>..."

def scan(fileobj):
    count = 0
    for line in fileobj:                   # сканировать построчно: итератор
        count += 1
        matchobj = pattDefine.match(line) # None, если совпадение не найдено
        if matchobj:
            name = matchobj.group(1)      # подстроки для групп (...)
            body = matchobj.group(2)
            print(count, 'defined', name, '=', body.strip())
            continue
        matchobj = pattInclude.match(line)
        if matchobj:
            start, stop = matchobj.span(1)    # start/stop смещения (...)
            filename = line[start:stop]       # вырезать из строки
            print(count, 'include', filename) # то же,
                                              # что и matchobj.group(1)

if len(sys.argv) == 1:
    scan(sys.stdin)              # без аргументов: читать stdin
else:
    scan(open(sys.argv[1], 'r')) # аргумент: имя файла
```

Для проверки просканируем с помощью этого сценария текстовый файл, представленный в примере 19.7.

Пример 19.7. PP4E\Lang\test.h

```
#ifndef TEST_H
#define TEST_H

#include <stdio.h>
#include <lib/spam.h>
# include    "Python.h"

#define DEBUG
#define HELLO 'hello regex world'
# define SPAM     1234

#define EGGS sunny + side + up
#define ADDER(arg) 123 + arg
#endif
```

Обратите внимание на пробелы после # в некоторых строках – регулярные выражения достаточно гибки, чтобы учитывать такие отклонения от нормы. Ниже показан результат работы этого сценария – он выбирает строки #include и #define и их части. Для каждой найденной строки в файле выводится ее номер, тип и найденные подстроки:

```
C:\...\PP4E\Lang> python cheader.py test.h
2 defined TEST_H =
4 include stdio.h
5 include lib/spam.h
6 include Python.h
8 defined DEBUG =
9 defined HELLO = 'hello regex world'
10 defined SPAM = 1234
12 defined EGGS = sunny + side + up
13 defined ADDER = (arg) 123 + arg
```

Еще один пример использования регулярных выражений можно найти в файле *pygrep1.py* в пакете примеров к книге. Он реализует простую утилиту «grep» поиска строк в файлах по шаблону, но не вошел в книгу из-за экономии места. Как будет показано далее, регулярные выражения иногда можно использовать для извлечения информации из текстовых файлов в формате XML и HTML, что является темой следующего раздела.

Синтаксический анализ XML и HTML

Помимо объектов строк и регулярных выражений в Python имеется поддержка синтаксического анализа некоторых специфических и часто используемых типов форматированного текста. В частности, в Python имеются готовые механизмы синтаксического анализа для XML и HTML, которые можно использовать и адаптировать для решения наших собственных задач обработки текста.

Что касается XML, в состав стандартной библиотеки Python включена поддержка анализа этого формата, а в самом проекте Python имеется отдельная группа разработчиков, занимающаяся проблемами XML. XML (eXtensible Markup Language – расширяемый язык разметки) – это язык разметки, основанный на тегах, предназначенный для описания структурированных данных самых разных типов. Помимо всего прочего он играет роль стандартной базы данных и представления содержимого во многих контекстах. Будучи объектно-ориентированным языком сценариев, Python прекрасно подходит для реализации основного назначения XML – обмена структурированными документами.

Язык XML основан на синтаксисе тегов, знакомом разработчикам веб-страниц, и используется для описания и упаковывания данных. Пакет xml, входящий в состав стандартной библиотеки Python, включает инструменты *анализа* данных в документах XML, основанные на обеих стандартных моделях, SAX и DOM, синтаксического анализа, а также

специфический для Python пакет ElementTree. Несмотря на то, что иногда для извлечения информации из документов XML также можно использовать регулярные выражения, тем не менее, они легко могут давать неверные результаты из-за наличия неожиданного текста и не имеют прямой поддержки произвольно вложенных конструкций языка XML (подробнее об этом ограничении будет рассказываться ниже, когда мы приступим к теме синтаксического анализа языков в целом).

В двух словах, реализация модели SAX предоставляет возможность создавать свои подклассы с методами, которые будут вызываться в процессе синтаксического анализа, а реализация модели DOM обеспечивает доступ к дереву объектов, представляющему (обычно) уже проанализированный документ. Парсеры SAX по сути являются конечными автоматами и в процессе анализа должны сохранять (возможно, на стеке) данные, полученные в процессе анализа. Парсеры DOM выполняют обход деревьев объектов, используя циклы, атрибуты и методы, определяемые стандартом DOM. Модель ElementTree является примерным аналогом DOM и позволяет писать более простой программный код. Она также может использоваться для генерации текста в формате XML из объектного представления.

Помимо этих инструментов синтаксического анализа в Python также имеется пакет `xmlrpc` поддержки протокола XML-RPC (протокол вызова удаленных процедур, который передает объекты, представленные в формате XML, по протоколу HTTP) на стороне клиента и сервера, а также стандартный парсер HTML, `html.parser`, который действует похожим образом и будет представлен далее в этой главе. Сторонние модули также предлагают множество инструментов для обработки документов XML. Большинство из них функционирует независимо от Python, что обеспечивает большую гибкость выпуска новых версий. Начиная с версии Python 2.3, в качестве основного механизма, управляющего синтаксическим анализом, стал использоваться парсер Expat.

Анализ XML

Обработка документов XML – это большая и сложная тема, обсуждение которой далеко выходит за рамки этой книги. Тем не менее, мы рассмотрим простой пример синтаксического анализа документа XML, представленного в примере 19.8. В этом файле определен список из нескольких книг о Python, выпущенных издательством O'Reilly, содержащий номера ISBN в виде атрибутов, а также названия, даты публикации и имена авторов в виде вложенных тегов (приношу извинения авторам, чьи книги не попали в этот список, отобранный совершенно случайно, – книг так много!).

Пример 19.8. PP4E\Lang\Xml\books.xml

```
<catalog>
    <book isbn="0-596-00128-2">
```

```
        <title>Python &amp; XML</title>
        <date>December 2001</date>
        <author>Jones, Drake</author>
    </book>
    <book isbn="0-596-15810-6">
        <title>Programming Python, 4th Edition</title>
        <date>October 2010</date>
        <author>Lutz</author>
    </book>
    <book isbn="0-596-15806-8">
        <title>Learning Python, 4th Edition</title>
        <date>September 2009</date>
        <author>Lutz</author>
    </book>
    <book isbn="0-596-15808-4">
        <title>Python Pocket Reference, 4th Edition</title>
        <date>October 2009</date>
        <author>Lutz</author>
    </book>
    <book isbn="0-596-00797-3">
        <title>Python Cookbook, 2nd Edition</title>
        <date>March 2005</date>
        <author>Martelli, Ravenscroft, Ascher</author>
    </book>
    <book isbn="0-596-10046-9">
        <title>Python in a Nutshell, 2nd Edition</title>
        <date>July 2006</date>
        <author>Martelli</author>
    </book>
    <!-- плюс еще много, много книг о Python, которые должны бы быть здесь -->
</catalog>
```

Рассмотрим для примера способы извлечения из этого файла номеров ISBN и соответствующих им названий, воспользовавшись всеми четырьмя основными инструментами Python, имеющимися в нашем распоряжении – шаблонами и парсерами SAX, DOM и ElementTree.

Анализ с помощью регулярных выражений

В некоторых ситуациях для анализа документов XML можно использовать регулярные выражения, с которыми мы познакомились выше. Они не являются полноценными инструментами синтаксического анализа и не слишком надежны и точны в ситуациях, когда в документе может присутствовать произвольный текст (текст в атрибутах тегов в особенности может сбить их с толку). Однако там, где они применимы, они предоставляют наиболее простой способ. Сценарий в примере 19.9 демонстрирует, как можно реализовать анализ файла XML из примера 19.8 с применением модуля `re`, о котором рассказывалось в предыдущем разделе. Как и все четыре примера, которые будут представлены в этом разделе, он просматривает содержимое файла XML в поис-

ках номеров ISBN и связанных с ними названий и сохраняет их в виде ключей и значений в словаре Python.

Пример 19.9. PP4E\Lang\Xml\rebook.py

```
"""
Анализ XML: регулярные выражения (ненадежное и неуниверсальное решение)
"""

import re, pprint
text    = open('books.xml').read()                  # str, если шаблон str
pattern = '(?s)isbn="(.*?)".*?<title>(.*?)</title>' # *?=минимальный
found   = re.findall(pattern, text)             # (?s)=точке соответств. /n
mapping = {isbn: title for (isbn, title) in found}  # словарь из списка
pprint.pprint(mapping)                              # кортежей
```

Вызов метода `re.findall` в этом сценарии отыскивает все интересующие нас вложенные теги, извлекает их содержимое и возвращает список кортежей, включающих содержимое двух групп в круглых скобках. Модуль Python `pprint` позволяет вывести созданный словарь в удобно отформатированном виде. Этот сценарий работает, но только пока содержимое файла не отклоняется от ожидаемого формата до такой степени, что анализ пойдет по неверному пути. Кроме того, мнемоника XML, обозначающая «&» в названии первой книги, не преобразуется автоматически:

```
C:\...\PP4E\Lang\Xml> python rebook.py
{'0-596-00128-2': 'Python &amp; XML',
 '0-596-00797-3': 'Python Cookbook, 2nd Edition',
 '0-596-10046-9': 'Python in a Nutshell, 2nd Edition',
 '0-596-15806-8': 'Learning Python, 4th Edition',
 '0-596-15808-4': 'Python Pocket Reference, 4th Edition',
 '0-596-15810-6': 'Programming Python, 4th Edition'}
```

Анализ с помощью парсера SAX

Точнее и надежнее извлекать данные позволяют более полноценные инструменты анализа XML, имеющиеся в языке Python. Так, в примере 19.10 представлена реализация процедуры анализа на основе модели SAX: представленный класс реализует методы обратного вызова, которые будут вызываться в процессе анализа, а программный код на верхнем уровне создает и запускает парсер.

Пример 19.10. PP4E\Lang\Xml\saxbook.py

```
"""
Анализ XML: SAX - это прикладной интерфейс на основе методов обратного
вызова, перехватывающих события, возникающие в процессе разбора документа
"""

import xml.sax, xml.sax.handler, pprint
```

```
class BookHandler(xml.sax.handler.ContentHandler):
    def __init__(self):
        self.inTitle = False       # обрабатывает события парсера XML
        self.mapping = {}          # модель конечного автомата

    def startElement(self, name, attributes):
        if name == "book":         # если встречен открывающий тег book,
            self.buffer = ""       # сохранить ISBN в ключе словаря
            self.isbn = attributes["isbn"]
        elif name == "title":      # если встречен открывающий тег title,
            self.inTitle = True    # установить флаг нахождения в теге title

    def characters(self, data):
        if self.inTitle:           # вызывается при получении текста из тега
            self.buffer += data    # если тег title, добавить текст в буфер

    def endElement(self, name):
        if name == "title":
            self.inTitle = False                  # закрывающий тег title
            self.mapping[self.isbn] = self.buffer # сохранить текст в словаре

parser  = xml.sax.make_parser()
handler = BookHandler()
parser.setContentHandler(handler)
parser.parse('books.xml')
pprint.pprint(handler.mapping)
```

Модель SAX является наиболее эффективной, но малопонятной на первый взгляд, потому что класс должен запоминать, анализ какого участка выполняется в текущий момент. Например, когда обнаруживается открывающий тег `title`, мы устанавливаем флаг состояния и инициализируем буфер – при обнаружении каждого символа внутри тега `title` мы добавляем его в буфер, пока не будет встречен закрывающий тег `title`. В результате содержимое тега `title` сохраняется в виде строки. Сама модель очень проста, но управлять ею иногда бывает очень сложно. Например, когда могут встречаться произвольно вложенные теги, информацию о состоянии, возвращаемую методами, вызываемыми при обработке вложенных тегов, может потребоваться сохранять в стеке.

Перед началом синтаксического анализа мы создаем объект парсера `parser`, устанавливаем экземпляр нашего класса как обработчик событий и запускаем анализ. По мере просмотра содержимого файла XML при встрече различных компонентов автоматически будут вызываться методы нашего класса. По завершении анализа мы снова используем модуль Python `pprint` для вывода результатов – объект `mapping` словаря находится в экземпляре обработчика. Результат получился практически точно такой же, но обратите внимание, что мнемоника, обозначающая символ «&», на этот раз была корректно преобразована – парсер SAX действительно выполняет анализ XML, а не отыскивает совпадения в тексте:

```
C:\...\PP4E\Lang\Xml> python saxbook.py
{'0-596-00128-2': 'Python & XML',
 '0-596-00797-3': 'Python Cookbook, 2nd Edition',
 '0-596-10046-9': 'Python in a Nutshell, 2nd Edition',
 '0-596-15806-8': 'Learning Python, 4th Edition',
 '0-596-15808-4': 'Python Pocket Reference, 4th Edition',
 '0-596-15810-6': 'Programming Python, 4th Edition'}
```

Анализ с помощью парсера DOM

Модель DOM синтаксического анализа XML является, пожалуй, наиболее простой для понимания – мы просто выполняем обход дерева объектов, полученного в результате анализа, – но она может оказаться менее эффективной при анализе больших документов, когда документ анализируется заранее и целиком сохраняется в памяти. Кроме того, модель DOM поддерживает произвольный доступ к частям документов, вложенные циклы по известным структурам и рекурсивный обход произвольно вложенных фрагментов, тогда как модель SAX ограничивается линейной процедурой анализа. В примере 19.11 приводится сценарий, реализующий анализ на основе модели DOM, эквивалентный реализации парсера SAX из предыдущего раздела.

Пример 19.11. PP4E\Lang\Xml\dombook.py

```
"""
Анализ XML: модель DOM позволяет получить представление всего документа
в виде объекта, доступного приложению для обхода
"""


import pprint
import xml.dom.minidom
from xml.dom.minidom import Node

doc = xml.dom.minidom.parse("books.xml")   # загрузить документ в объект
                                           # для предварительного анализа
mapping = {}
for node in doc.getElementsByTagName("book"):   # обход объекта DOM
    isbn = node.getAttribute("isbn")            # с применением DOM API
    L = node.getElementsByTagName("title")
    for node2 in L:
        title = ""
        for node3 in node2.childNodes:
            if node3.nodeType == Node.TEXT_NODE:
                title += node3.data
        mapping[isbn] = title

# словарь mapping теперь содержит те же значения, что и в примере
# использования SAX
pprint.pprint(mapping)
```

Этот сценарий выводит те же результаты, которые были получены при использовании парсера SAX. Однако здесь они были получены в результате обхода объекта дерева документа после завершения анализа и за счет вызова методов и обращения к атрибутам, определяемых спецификациями стандарта DOM. В этом заключается одновременно и достоинство, и недостаток модели DOM – его прикладной интерфейс не зависит от конкретного языка программирования, но он может показаться неинтуитивным и слишком многословным некоторым программистам на Python, приученным к более простым моделям:

```
C:\...\PP4E\Lang\Xml> python dombook.py
{'0-596-00128-2': 'Python & XML',
 '0-596-00797-3': 'Python Cookbook, 2nd Edition',
 '0-596-10046-9': 'Python in a Nutshell, 2nd Edition',
 '0-596-15806-8': 'Learning Python, 4th Edition',
 '0-596-15808-4': 'Python Pocket Reference, 4th Edition',
 '0-596-15810-6': 'Programming Python, 4th Edition'}
```

Анализ с помощью парсера ElementTree

Четвертый инструмент, популярный пакет ElementTree, входит в состав стандартной библиотеки и позволяет анализировать и генерировать документы XML. Как парсер он по сути является более питонической реализацией модели DOM – он точно так же преобразует документ в дерево объектов, но предоставляет более легковесный прикладной интерфейс навигации по дереву, потому что он разрабатывался специально для языка Python.

Пакет ElementTree предоставляет простые в использовании инструменты синтаксического анализа, изменения и создания документов XML. При анализе и создании документов он представляет их в виде деревьев объектов «элементов». Каждый элемент в дереве имеет имя тега, словарь атрибутов, текстовое значение и последовательность дочерних элементов. Объекты элементов, созданные при анализе, можно просматривать с помощью обычных циклов Python, когда их структура известна, и с помощью приема рекурсивного спуска, когда возможна произвольная вложенность элементов.

Изначально система ElementTree была сторонним расширением, но позднее значительная ее часть была включена в состав стандартной библиотеки Python в виде пакета `xml.etree`. Пример 19.12 демонстрирует, как можно использовать этот пакет для анализа нашего файла со списком книг.

Пример 19.12. PP4E\Lang\Xml\etreebook.py

```
"""
Анализ XML: ElementTree (etree) предоставляет API на основе языка Python
для анализа/создания документов XML
"""
```

```
import pprint
from xml.etree.ElementTree import parse

mapping = {}
tree = parse('books.xml')
for B in tree.findall('book'):
    isbn = B.attrib['isbn']
    for T in B.findall('title'):
        mapping[isbn] = T.text
pprint.pprint(mapping)
```

Этот сценарий выводит те же результаты, что и сценарии, использующие парсеры SAX и DOM, но программный код, необходимый для извлечения данных из файла, выглядит на этот раз намного проще:

```
C:\...\PP4E\Lang\Xml> python etreebook.py
{'0-596-00128-2': 'Python & XML',
 '0-596-00797-3': 'Python Cookbook, 2nd Edition',
 '0-596-10046-9': 'Python in a Nutshell, 2nd Edition',
 '0-596-15806-8': 'Learning Python, 4th Edition',
 '0-596-15808-4': 'Python Pocket Reference, 4th Edition',
 '0-596-15810-6': 'Programming Python, 4th Edition'}
```

Прочие темы, связанные с XML

Естественно, язык Python обладает более обширной поддержкой XML, чем следует из этих простых примеров. Однако ради экономии места в книге я приведу лишь ссылки на ресурсы, посвященные XML, где можно найти дополнительные примеры:

Стандартная библиотека

В первую очередь обязательно ознакомьтесь с руководством по стандартной библиотеке Python, где дается более подробное описание стандартных инструментов поддержки XML. Загляните в разделы с описанием модулей `re`, `xml.sax`, `xml.dom` и `xml.etree`, где подробнее рассказывается о примерах в этом разделе.

Инструменты PyXML SIG

Дополнительные инструменты для обработки документов XML и документацию можно также найти на веб-странице специальной заинтересованной группы (Special Interest Group, SIG) по проблемам XML на сайте *http://www.python.org*. Эта группа занимается вопросами интеграции технологий XML в язык Python и разрабатывает бесплатные инструменты обработки XML, независимые от самого языка Python. Большая часть поддержки XML в стандартной библиотеке реализована членами этой группы.

Сторонние инструменты

Вы можете также найти бесплатные сторонние инструменты поддержки XML для языка Python, следуя по ссылкам, которые приво-

дятся на веб-странице специальной заинтересованной группы по проблемам XML. Особое внимание обратите на свободно распространяемый пакет 4Suite, предоставляющий интегрированные инструменты обработки XML, включая реализацию таких открытых технологий, как DOM, SAX, RDF, XSLT, XInclude, XPointer, XLink и XPath.

Документация

Существует множество разнообразных книг, специально посвященных обработке текста и документов XML на языке Python. Издательство O'Reilly также предлагает книгу, посвященную теме обработки XML на языке Python, «Python & XML» (*http://oreilly.com/catalog/9780596001285/*), написанную Кристофером Джонсом (Christopher A. Jones) и Фредом Дрейком (Fred L. Drake, Jr).

Как обычно, не забудьте с помощью своей любимой поисковой системы ознакомиться с самыми последними разработками в этой области.

Анализ HTML

В стандартной библиотеке Python имеется также модуль `html.parser`, поддерживающий анализ разметки HTML, хотя и с ограниченными возможностями, который можно использовать для извлечения информации из веб-страниц. Помимо всего прочего этот парсер можно использовать для обработки веб-страниц, получаемых с помощью модуля `urllib.request`, с которым мы познакомились в части книги, посвященной программированию для Интернета, для извлечения простого текста из сообщений электронной почты в формате HTML и других нужд.

Модуль `html.parser` имеет прикладной интерфейс, напоминающий модель XML SAX, представленную в предыдущем разделе: он предоставляет механизм синтаксического анализа, который можно наследовать в своих подклассах для перехвата событий обнаружения тегов и их содержимого, возникающих в процессе анализа. В отличие от модели SAX, мы должны не предоставлять свой класс-обработчик, а расширять класс парсера непосредственно. Ниже приводится короткий интерактивный пример, демонстрирующий основы (весь программный код для этого раздела я скопировал в файл *htmlparser.py*, который находится в пакете примеров, на случай, если у вас появится желание поэкспериментировать самостоятельно):

```
>>> from html.parser import HTMLParser
>>> class ParsePage(HTMLParser):
...     def handle_starttag(self, tag, attrs):
...         print('Tag start:', tag, attrs)
...     def handle_endtag(self, tag):
...         print('tag end: ', tag)
...     def handle_data(self, data):
...         print('data......', data.rstrip())
...
```

Теперь создадим текстовую строку с разметкой HTML веб-страницы – мы жестко определяем ее здесь, но с тем же успехом ее можно загрузить из файла или получить с веб-сайта с помощью модуля `urllib.request`:

```
>>> page = """
... <html>
... <h1>Spam!</h1>
... <p>Click this <a href="http://www.python.org">python</a> link</p>
... </html>"""
```

Наконец, запустим анализ, передав текст экземпляру парсера, – при обнаружении тегов HTML будут вызываться методы обратного вызова, которым в качестве аргументов будут передаваться имена тегов и последовательности атрибутов:

```
>>> parser = ParsePage()
>>> parser.feed(page)
data......
Tag start: html []
data......
Tag start: h1 []
data...... Spam!
tag end:   h1
data......
Tag start: p []
data...... Click this
Tag start: a [('href', 'http://www.python.org')]
data...... python
tag end:   a
data...... link
tag end:   p
data......
tag end:   html
```

Как видите, в процессе анализа по событиям вызываются методы парсера. Как и при использовании модели SAX XML, в процессе анализа ваш класс парсера должен сохранять информацию о состоянии в своих атрибутах, если он должен делать что-то более конкретное, чем просто выводить имена тегов, атрибуты и содержимое. При этом реализация извлечения содержимого определенных тегов может состоять просто из проверки имен тегов и установки флагов состояния. Кроме того, в процессе анализа достаточно просто можно было бы реализовать конструирование дерева объектов, отражающего структуру страницы.

Обработка мнемоник HTML (еще раз)

Ниже приводится еще один пример анализа разметки HTML: в главе 15 мы использовали простой метод, экспортируемый этим модулем, чтобы преобразовать экранированные последовательности HTML (мнемоники) в строках, встроенных в HTML-страницу ответа:

```
>>> import cgi, html.parser
>>> s = cgi.escape("1<2 <b>hello</b>")
>>> s
'1&lt;2 &lt;b&gt;hello&lt;/b&gt;'
>>>
>>> html.parser.HTMLParser().unescape(s)
'1<2 <b>hello</b>'
```

Этот прием работает для преобразования экранированных последовательностей HTML, но это не все. Когда мы рассматривали это решение, я давал понять, что существует более универсальный подход. Теперь, когда вы познакомились с моделью методов обратного вызова парсера HTML, есть смысл рассмотреть более характерный для Python способ обработки мнемоник в процессе анализа – достаточно просто перехватывать события обнаружения мнемоник в подклассе парсера и выполнять необходимые преобразования:

```
>>> class Parse(html.parser.HTMLParser):
...     def handle_data(self, data):
...         print(data, end='')
...     def handle_entityref(self, name):
...         map = dict(lt='<', gt='>')
...         print(map[name], end='')
...
>>> p = Parse()
>>> p.feed(s); print()
1<2 <b>hello</b>
```

Или, что еще лучше, использовать модуль Python `html.entities`, чтобы избежать необходимости определять таблицу преобразований мнемоник HTML в символы. Этот модуль определяет значительно большее количество мнемоник, чем простой словарь в предыдущем примере, и включает все мнемоники, с которыми можно столкнуться при анализе текста HTML:

```
>>> s
'1&lt;2 &lt;b&gt;hello&lt;/b&gt;'
>>>
>>> from html.entities import entitydefs
>>> class Parse(html.parser.HTMLParser):
...     def handle_data(self, data):
...         print(data, end='')
...     def handle_entityref(self, name):
...         print(entitydefs[name], end='')
...
>>> P = Parse()
>>> P.feed(s); print()
1<2 <b>hello</b>
```

Строго говоря, модуль `html.entities` способен отображать имена мнемоник в кодовые пункты Юникода и наоборот – таблица, используемая здесь, просто преобразует целочисленные кодовые пункты в символы

с помощью функции chr. Дополнительные подробности ищите в описании этого модуля, а также в его исходном программном коде в стандартной библиотеке Python.

Извлечение простого текста из разметки HTML (еще раз)

Теперь, когда вы понимаете принцип действия класса парсера HTML из стандартной библиотеки Python, модуль извлечения простого текста, использовавшийся в приложении PyMailGUI (пример 14.8) в главе 14, также, вероятно, будет более понятен вам (тогда это была необходимая ссылка вперед, которую мы, наконец, можем закрыть).

Вместо того чтобы снова повторять реализацию модуля, я отсылаю вас для самостоятельного изучения приведенного там примера, а также его программного кода самотестирования и тестовых входных файлов, – как к еще одному примеру реализации синтаксического анализа разметки HTML на языке Python. Он является немного более сложной версией примеров, приводившихся здесь, которая определяет большее количество типов тегов в своих методах обратного вызова.

Из-за ограниченности места в книге мы вынуждены завершить дальнейшее изучение приемов синтаксического анализа HTML. Как обычно, знания того, что подобная возможность существует, уже достаточно, чтобы начать осваивать ее самостоятельно. За дополнительными подробностями о прикладном интерфейсе обращайтесь к руководству по библиотеке Python. А для получения дополнительной информации о поддержке HTML ищите в Интернете пакеты парсеров HTML для версии 3.X сторонних разработчиков, подобные тем, что упоминались в главе 14.

Дополнительные инструменты синтаксического анализа

Если у вас есть некоторая подготовка в теории синтаксического анализа, то вы должны знать, что ни регулярных выражений, ни операций разбиения строк недостаточно для работы с более сложными грамматиками языков (грубо говоря, у них нет «памяти», необходимой настоящим грамматикам), и поэтому они неспособны обеспечить поддержку языковых конструкций, произвольно вложенных друг в друга, например вложенные инструкции if в языках программирования. Фактически именно этим обусловлена необходимость использования парсеров XML и HTML, представленных в предыдущем разделе: оба являются языками, допускающими произвольную вложенность синтаксических конструкций, анализ которых в целом невозможно реализовать с применением одних только регулярных выражений.

С теоретической точки зрения регулярные выражения в действительности предназначены только для реализации первой стадии анализа – разделения текста на компоненты, – которая известна как *лексический анализ*. Несмотря на то, что шаблоны часто можно использовать для

извлечения данных из текста, для истинного синтаксического анализа языков требуется нечто большее. В Python существует целый ряд способов восполнить этот недостаток:

Python как инструмент синтаксического анализа

В большинстве приложений сам язык Python способен заменить нестандартные языки и парсеры – программный код, введенный пользователем, может быть передан интерпретатору Python для выполнения с помощью таких инструментов, как функции `eval` и `exec`. При расширении системы с помощью собственных модулей пользовательский программный код получает доступ не только ко всему богатству языка Python, но и к любым другим специфическим расширениям, необходимым приложению. В некотором смысле, такие системы встраивают Python в Python. Поскольку это вполне обычная роль языка Python, мы рассмотрим данный подход далее в этой главе.

Собственные парсеры: созданные вручную или с привлечением других инструментов

Однако для решения некоторых сложных задач анализа синтаксиса языков все еще может потребоваться использовать полноценные парсеры. Такие парсеры всегда можно написать вручную, но так как Python поддерживает интеграцию с инструментами на языке C, можно написать инструмент интеграции с традиционными генераторами парсеров, такими как yacc и bison, создающими парсеры на основе определения грамматики языка. Но еще лучше использовать уже существующие инструменты интеграции – интерфейсы к таким распространенным генераторам парсеров свободно доступны в мире открытого программного обеспечения (воспользуйтесь поисковой системой, чтобы найти более свежую информацию и ссылки).

Кроме того, в Интернете можно найти множество систем синтаксического анализа специально для языка Python. Среди них: PLY – реализация инструментов синтаксического анализа lex и yacc на Python и для Python; система kwParsing – генератор парсеров, написанный на языке Python; PyParsing – библиотека классов на языке Python, которая позволяет быстро создавать парсеры, действующие по алгоритму рекурсивного спуска; и инструментальный набор SPARK – легковесная система, использующая алгоритм Earley для решения технических проблем, связанных с генерированием парсеров LALR (если вы не знаете, что это такое, то вам, вероятно, не стоит и беспокоиться).

Особый интерес для данной главы представляет YAPPS (Yet Another Python Parser System – еще одна система синтаксического анализа Python) – генератор парсеров, написанный на языке Python. С помощью заданных правил грамматики он генерирует программный код Python в доступном для чтения виде, реализующий парсер, использующий метод рекурсивного спуска. Парсеры, генерируемые системой YAPPS, весьма похожи на парсеры выражений, реализованные вручную и представленные в следующем разделе (которыми они

и инспирированы). YAPPS создает парсеры LL(1), не такие мощные, как парсеры LALR, но достаточные для многих задач синтаксического анализа языков. Для получения подробной информации о YAPPS смотрите страницу *http://theory.stanford.edu/~amitp/Yapps* или ищите дополнительную информацию в Интернете.

Урок 2: не нужно изобретать колесо

По поводу генераторов парсеров: для использования некоторых из этих инструментов в программах Python вам потребуется интегрирующий их модуль расширения. В таких ситуациях первым делом нужно проверить, не существует ли уже необходимое общедоступное расширение. В особенности для подобных стандартных инструментов весьма вероятно, что кто-то уже написал интегрирующий модуль, который можно использовать в готовом виде, а не писать с самого начала новый.

Конечно, не каждый может подарить все свои модули расширения обществу, но библиотека имеющихся компонентов, которые можно брать бесплатно, растет, и есть сообщество специалистов, к которым можно обратиться с вопросом. Посетите сайт PyPI по адресу *http://www.python.org*, где можно найти ссылки на программные ресурсы для языка Python, или поищите в Интернете. При наличии во всем мире на момент написания книги миллиона пользователей Python в разделе прототипов можно найти немало.

Если колесо не подходит: в этой книге мы видели несколько примеров, когда модули из стандартной библиотеки не соответствовали поставленной задаче или имели ограниченную работоспособность. Например, в главе 13, когда мы занимались проблемами пакета `email` в Python 3.1, нам потребовалось реализовать свои собственные обходные решения. В таких случаях вам может потребоваться писать собственный программный код поддержки инфраструктуры. При нарушении программных зависимостей все еще может торжествовать ситуация «еще не реализовано».

Однако в целом лучше стремиться использовать стандартную поддержку, предоставляемую языком Python в большинстве случаев, даже если это потребует писать собственные обходные решения. В примере с использованием пакета `email` исправление его проблем выглядит намного проще, чем создание собственного парсера электронной почты с самого начала, – эта задача слишком объемная, чтобы даже пытаться решить ее в этой книге. Подход «батарейки входят в комплект», используемый в Python, может оказаться удивительно производительным, даже когда эти «батарейки» приходится подзаряжать.

Обработка текстов на естественных языках

Еще более сложные задачи синтаксического анализа естественных языков требуют использования приемов искусственного интеллекта, таких как семантический анализ и средства самообучения. Например, Natural Language Toolkit, или NLTK, – комплект программ и библиотек на языке Python с открытыми исходными текстами для символического и статистического анализа текстов на естественных языках. Этот комплект применяет к текстовым данным лингвистические методики обработки и может использоваться в разработке программных систем распознавания естественных языков. Подробнее эта тема раскрывается в книге издательства O'Reilly «Natural Language Processing with Python», в которой среди всего прочего исследуются приемы использования NLTK в Python. Разумеется, далеко не в каждой системе пользователи должны задавать вопросы на естественном языке, но существует множество прикладных областей, где такая возможность была бы полезна.

Несмотря на безусловную полезность, генераторы парсеров и инструменты анализа естественных языков слишком сложны, чтобы хоть с какой-то пользой их можно было рассматривать в этой книге. За дополнительной информацией об инструментах синтаксического анализа, доступных для использования в программах на языке Python, обращайтесь на сайт *http://python.org/* или выполните поиск в Интернете. А теперь, продолжая тему этой главы, перейдем к исследованию наиболее простого подхода, иллюстрирующего основные концепции, – *синтаксического анализа методом рекурсивного спуска*.

Парсеры, написанные вручную

Несмотря на изобилие инструментов в рассматриваемой области, тем не менее, универсальные возможности языка программирования Python позволяют вручную создавать парсеры для решения нестандартных задач синтаксического анализа. Например, *синтаксический анализ методом рекурсивного спуска* – довольно хорошо известная методика анализа языковой информации. Хотя они и не отличаются широтой возможностей, как некоторые другие инструменты, тем не менее, парсеры, реализующие синтаксический анализ методом рекурсивного спуска, вполне можно применять для решения разнообразных задач синтаксического анализа.

Для иллюстрации в этом разделе будет разработан специальный парсер для простой грамматики – он разбирает и вычисляет строки арифметических выражений. Несмотря на то, что основной темой здесь является синтаксический анализ, этот пример также демонстрирует возможности Python как универсального языка программирования. Хотя Python часто используется в качестве интерфейса или языка быстрой раз-

работки приложений, он полезен и для таких вещей, которые обычно пишутся на системных языках, таких как C или C++.

Грамматика выражений

Грамматику, которую будет распознавать наш парсер, можно описать так:

```
goal -> <expr> END                      [number, variable, ( ]
goal -> <assign> END                    [set]

assign -> 'set' <variable> <expr>       [set]

expr -> <factor> <expr-tail>            [number, variable, ( ]

expr-tail -> ^                          [END, ) ]
expr-tail -> '+' <factor> <expr-tail>   [+]
expr-tail -> '-' <factor> <expr-tail>   [-]

factor -> <term> <factor-tail>          [number, variable, ( ]

factor-tail -> ^                        [+, -, END, ) ]
factor-tail -> '*' <term> <factor-tail> [*]
factor-tail -> '/' <term> <factor-tail> [/]

term -> <number>                        [number]
term -> <variable>                      [variable]
term -> '(' <expr> ')'                  [(]

tokens: (, ), num, var, -, +, /, *, set, end
```

Это довольно типичная грамматика простого языка выражений, допускающая произвольную вложенность выражений (некоторые примеры выражений можно найти в конце листинга модуля `testparser`, представленного в примере 19.15). Анализируемые строки представляют собой выражение или присваивание значения переменной (`set`). В выражениях участвуют числа, переменные и операторы `+`, `-`, `*` и `/`. Поскольку в грамматике `factor` вложен в `expr`, то `*` и `/` имеют более высокий приоритет (то есть связывают сильнее), чем `+` и `-`. Выражения могут заключаться в круглые скобки, чтобы переопределять старшинство операций, и все операторы являются левоассоциативными (например, выражение `1-2-3` интерпретируется как `(1-2)-3`).

Лексемы представляют собой наиболее элементарные компоненты языка выражений. За каждым грамматическим правилом, приведенным выше, в квадратных скобках следует список лексем, по которым оно выбирается. При анализе методом рекурсивного спуска мы определяем набор лексем, которые могут начинать подстроку правила, и с помощью этой информации заранее предсказываем правило, которое будет работать. Для повторяющихся правил (правил `-tail`) используется на-

бор возможных последующих лексем, чтобы знать, когда остановиться. Обычно лексемы распознаются обработчиком строк (*лексическим анализатором*, или «сканером»), а обработчик более высокого уровня (*синтаксический анализатор*, или «парсер») использует поток лексем для предсказания и прохода грамматических правил и подстрок.

Реализация парсера

Система организована в виде двух модулей, содержащих два класса:

- Сканер производит посимвольный анализ нижнего уровня.
- Парсер содержит в себе сканер и производит грамматический анализ высокого уровня.

Парсер отвечает также за вычисление значения выражения и тестирование системы. В данной версии парсер вычисляет выражение во время его грамматического разбора. Чтобы использовать систему, следует создать парсер, передать ему входную строку и вызвать метод `parse`. Впоследствии можно снова вызвать `parse` с новой строкой выражения.

Здесь намеренно произведено разделение труда. Сканер извлекает из строки лексемы, но ему ничего не известно о грамматике. Парсер обрабатывает грамматику, но находится в неведении относительно самой строки. Благодаря такой модульной организации программный код сохраняет относительную простоту. И это еще один пример отношения композиции объектно-ориентированного программирования в действии: парсеры содержат сканеры и делегируют им выполнение некоторых операций.

Модуль в примере 19.13 реализует задачу лексического анализа – обнаружения в выражении базовых лексем путем сканирования строки текста слева направо. Обратите внимание на простоту используемой здесь логики; такой анализ иногда можно выполнить с помощью регулярных выражений (описанных раньше), но, на мой взгляд, шаблон, необходимый для обнаружения и извлечения лексем в этом примере, окажется слишком сложным и хрупким. Если вам так не кажется, попробуйте переписать этот модуль с применением модуля `re`.

Пример 19.13. PP4E\Lang\Parser\scanner.py

```
"""
##############################################################################
сканнер (лексический анализатор)
##############################################################################
"""


import string
class SyntaxError(Exception): pass          # локальные ошибки
class LexicalError(Exception): pass         # бывшие строки

class Scanner:
```

```
    def __init__(self, text):
        self.next = 0
        self.text = text + '\0'

    def newtext(self, text):
        Scanner.__init__(self, text)

    def showerror(self):
        print('=> ', self.text)
        print('=> ', (' ' * self.start) + '^')

    def match(self, token):
        if self.token != token:
            raise SyntaxError(token)
        else:
            value = self.value
            if self.token != '\0':
                self.scan()               # очередная лексема/значение
            return value                  # вернуть предыдущее значение

    def scan(self):
        self.value = None
        ix = self.next
        while self.text[ix] in string.whitespace:
            ix += 1
        self.start = ix

        if self.text[ix] in ['(', ')', '-', '+', '/', '*', '\0']:
            self.token = self.text[ix]
            ix += 1

        elif self.text[ix] in string.digits:
            str = ''
            while self.text[ix] in string.digits:
                str += self.text[ix]
                ix += 1
            if self.text[ix] == '.':
                str += '.'
                ix += 1
                while self.text[ix] in string.digits:
                    str += self.text[ix]
                    ix += 1
                self.token = 'num'
                self.value = float(str)
            else:
                self.token = 'num'
                self.value = int(str) # соответствует long() в 3.x

        elif self.text[ix] in string.ascii_letters:
            str = ''
            while self.text[ix] in (string.digits + string.ascii_letters):
```

```
                str += self.text[ix]
                ix += 1
            if str.lower() == 'set':
                self.token = 'set'
            else:
                self.token = 'var'
                self.value = str

        else:
            raise LexicalError()
        self.next = ix
```

Класс модуля парсера создает и встраивает сканер для выполнения задач лексического анализа, занимается интерпретацией грамматических правил и вычислением результата выражения, как показано в примере 19.14.

Пример 19.14. PP4E\Lang\Parser\parser1.py

```
"""
##############################################################################
парсер (синтаксический анализатор, вычисляет выражение в процессе анализа)
##############################################################################
"""

class UndefinedError(Exception): pass
from scanner import Scanner, LexicalError, SyntaxError

class Parser:
    def __init__(self, text=''):
        self.lex = Scanner(text)           # встроить сканер
        self.vars = {'pi': 3.14159}        # добавить переменную

    def parse(self, *text):
        if text:                            # главная точка входа
            self.lex.newtext(text[0])       # использовать парсер повторно?
        try:
            self.lex.scan()                 # получить первую лексему
            self.Goal()                     # разобрать предложение
        except SyntaxError:
            print('Syntax Error at column:', self.lex.start)
            self.lex.showerror()
        except LexicalError:
            print('Lexical Error at column:', self.lex.start)
            self.lex.showerror()
        except UndefinedError as E:
            name = E.args[0]
            print("'%s' is undefined at column:" % name, self.lex.start)
            self.lex.showerror()

    def Goal(self):
        if self.lex.token in ['num', 'var', '(']:
```

```
            val = self.Expr()
            self.lex.match('\0')          # выражение?
            print(val)
        elif self.lex.token == 'set':     # команда set?
            self.Assign()
            self.lex.match('\0')
        else:
            raise SyntaxError()

    def Assign(self):
        self.lex.match('set')
        var = self.lex.match('var')
        val = self.Expr()
        self.vars[var] = val              # присвоить имени в словаре

    def Expr(self):
        left = self.Factor()
        while True:
            if self.lex.token in ['\0', ')']:
                return left
            elif self.lex.token == '+':
                self.lex.scan()
                left = left + self.Factor()
            elif self.lex.token == '-':
                self.lex.scan()
                left = left - self.Factor()
            else:
                raise SyntaxError()

    def Factor(self):
        left = self.Term()
        while True:
            if self.lex.token in ['+', '-', '\0', ')']:
                return left
            elif self.lex.token == '*':
                self.lex.scan()
                left = left * self.Term()
            elif self.lex.token == '/':
                self.lex.scan()
                left = left / self.Term()
            else:
                raise SyntaxError()

    def Term(self):
        if self.lex.token == 'num':
            val = self.lex.match('num')    # числа
            return val
        elif self.lex.token == 'var':
            if self.lex.value in self.vars.keys(): # keys(): EIBTI[1]!
```

[1] Explicit Is Better Than Implicit – «явное лучше неявного», один из принципов Python.– *Прим. ред.*

```
                val = self.vars[self.lex.value]    # найти значение имени
                self.lex.scan()
                return val
            else:
                raise UndefinedError(self.lex.value)
        elif self.lex.token == '(':
            self.lex.scan()
            val = self.Expr()            # подвыражение
            self.lex.match(')')
            return val
        else:
            raise SyntaxError()

if __name__ == '__main__':
    import testparser                    # программный код самотестирования
    testparser.test(Parser, 'parser1')   # тест локального объекта Parser
```

Если внимательно изучить этот пример, можно заметить, что парсер ведет словарь (`self.vars`) имен переменных: они сохраняются в словаре командой `set` и выбираются из него при появлении в выражении. Лексемы представляются строками с возможными ассоциированными значениями (числовое значение для чисел и строки для имен переменных).

Для обработки правил `expr-tail` и `factor-tail` в этом парсере используются итерации (циклы `while`) вместо рекурсии. За исключением этой оптимизации правила грамматики непосредственно отображаются в методы парсера: лексемы становятся обращениями к сканеру, а ссылки на вложенные правила становятся обращениями к другим методам.

При выполнении файла *parser1.py* как программы верхнего уровня выполняется его программный код самотестирования, который, в свою очередь, просто выполняет готовый тест, показанный в примере 19.15. Обратите внимание, что строки преобразуются в числа с помощью функции `int`, это гарантирует поддержку целых чисел неограниченной точности в целочисленных вычислениях, потому что сканер использует целые числа Python, которые всегда обеспечивают дополнительную точность по мере необходимости (отдельный синтаксис `long`, имевшийся в Python 2.X, использовать более не требуется).

Заметьте также, что смешанные операции над целыми числами и числами с плавающей точкой приводятся к операциям над числами с плавающей точкой, так как для фактических вычислений используются операторы Python. Кроме того, оператор деления / также унаследовал модель истинного деления от Python 3.X, которая сохраняет остаток от деления и возвращает результат с плавающей точкой, независимо от типов операндов. Мы могли бы просто использовать // в логике вычислений, чтобы сохранить прежнее поведение (или включить в грамматику оба оператора, / и //), но здесь мы будем следовать основным тенденциям Python 3.X.

Пример 19.15. PP4E\Lang\Parser\testparser.py

```
"""
##############################################################################
контрольный пример парсера
##############################################################################
"""

def test(ParserClass, msg):
    print(msg, ParserClass)
    x = ParserClass('4 / 2 + 3')          # допускаются различные парсеры
    x.parse()

    x.parse('3 + 4 / 2')                  # аналог eval('3 + 4 / 2')...
    x.parse('(3 + 4) / 2')                # 3.X: / - истинное деление
    x.parse('4 / (2 + 3)')                # // не поддерживается (пока)
    x.parse('4.0 / (2 + 3)')
    x.parse('4 / (2.0 + 3)')
    x.parse('4.0 / 2 * 3')
    x.parse('(4.0 / 2) * 3')
    x.parse('4.0 / (2 * 3)')
    x.parse('(((3))) + 1')

    y = ParserClass()
    y.parse('set a 4 / 2 + 1')
    y.parse('a * 3')
    y.parse('set b 12 / a')
    y.parse('b')

    z = ParserClass()
    z.parse('set a 99')
    z.parse('set a a + 1')
    z.parse('a')

    z = ParserClass()
    z.parse('pi')
    z.parse('2 * pi')
    z.parse('1.234 + 2.1')

def interact(ParserClass):                # ввод из командной строки
    print(ParserClass)
    x = ParserClass()
    while True:
        cmd = input('Enter=> ')
        if cmd == 'stop':
            break
        x.parse(cmd)
```

Сопоставьте следующие результаты с вызовами функции `print` в модуле самотестирования:

```
C:\...\PP4E\Lang\Parser> python parser1.py
parser1 <class '__main__.Parser'>
```

```
5.0
5.0
3.5
0.8
0.8
0.8
6.0
6.0
0.666666666667
4
9.0
4.0
100
3.14159
6.28318
3.334
```

Как обычно, парсер можно тестировать также в интерактивной оболочке, чтобы опробовать все его возможности:

```
C:\...\PP4E\Lang\Parser> python
>>> import parser1
>>> x = parser1.Parser()
>>> x.parse('1 + 2')
3
```

Ошибки перехватываются и описываются достаточно дружественным способом (предполагается, что счет начинается с нуля):

```
>>> x.parse('1 + a')
'a' is undefined at column: 4
=> 1 + a
=> ^
>>> x.parse('1+a+2')
'a' is undefined at column: 2
=> 1+a+2
=> ^
>>> x.parse('1 * 2 $')
Lexical Error at column: 6
=> 1 * 2 $
=> ^
>>> x.parse('1 * - 1')
Syntax Error at column: 4
=> 1 * - 1
=> ^
>>> x.parse('1 * (9')
Syntax Error at column: 6
=> 1 * (9
=> ^

>>> x.parse('1 + 2 / 3')      # в 3.X поведение операции деления изменилось
1.66666666667
>>> x.parse('1 + 2 // 3')
```

```
Syntax Error at column: 7
=> 1 + 2 // 3
=> ^
```

Чрезвычайно большие числа также с успехом обрабатываются парсером, потому что в процессе используются встроенные объекты и операторы Python:

```
>>> x.parse('888888888888888888888888888888888888888888888.9999999')
8.88888888889e+44
>>> x.parse('999999999999999999999999999999999999999999 + 2')
1000000000000000000000000000000000000000001
>>> x.parse('999999999999999999999999999999.88888888888 + 1.1')
1e+30
```

Кроме того, в модуле `testparser` реализована также возможность интерактивного вычисления выражений, если вам вздумается использовать парсер как простой калькулятор командной строки (или если вам надоест вводить вызовы методов). Передайте класс `Parser`, чтобы модуль `testparser` мог создать собственный экземпляр парсера:

```
>>> import testparser
>>> testparser.interact(parser1.Parser)
<class 'parser1.Parser'>
Enter=> 4 * 3 + 5
17
Enter=> 5 + 4 * 3
17
Enter=> (5 + 4) * 3
27
Enter=> set a 99
Enter=> set b 66
Enter=> a + b
165
Enter=> # + 1
Lexical Error at column: 0
=> # + 1
=> ^
Enter=> a * b + c
'c' is undefined at column: 8
=> a * b + c
=> ^
Enter=> a * b * + c
Syntax Error at column: 8
=> a * b * + c
=> ^
Enter=> a
99
Enter=> a * a * a
970299
Enter=> stop
>>>
```

Урок 3: разделяй и властвуй

Как демонстрирует система синтаксического анализа, модульная конструкция программы почти всегда приносит значительный выигрыш. С помощью средств структурирования программ на языке Python (функций, модулей, классов и так далее) большие задачи могут быть разбиты на маленькие контролируемые части, которые можно создавать и тестировать независимо друг от друга.

Например, сканер можно протестировать без парсера, создав его экземпляр с помощью входной строки и вызывая его методы `scan` или `match`. Можно даже проверить его в интерактивном режиме, из командной строки Python. Когда используется разделение программ на логические составляющие, становится проще понять их и модифицировать. Представьте себе, как бы выглядел парсер, если бы логика сканера была вложена в него, а не вызывалась.

Добавление интерпретатора дерева синтаксического анализа

Одна из слабостей программы `parser1` состоит в том, что она встраивает логику вычисления выражений в логику синтаксического анализа: результат вычисляется во время грамматического разбора строки. Это ускоряет вычисление, но может затруднить модификацию программного кода, особенно при росте системы. Проще говоря, можно изменить организацию программы с тем, чтобы отделить синтаксический анализ от вычисления. Вместо вычисления строки парсер может построить промежуточное представление, которое будет вычислено позднее. Дополнительным стимулом к созданию такого отдельного представления является возможность подвергнуть его анализу другими средствами (например, оптимизаторами, инструментами просмотра и так далее) – они могут запускаться в отдельных проходах по дереву.

В примере 19.16 показана версия `parser1`, реализующая эту идею. Парсер анализирует строку и строит *дерево синтаксического анализа*, то есть дерево экземпляров классов, которое представляет выражение и может быть вычислено на отдельной стадии. Дерево синтаксического анализа строится из классов, которые «умеют» вычислять себя: чтобы вычислить выражение, мы просто предложим дереву вычислить себя. Корневые узлы дерева просят своих потомков вычислить себя, а затем объединяют результаты, применяя единственный оператор. Фактически вычисление в этой версии является рекурсивным обходом дерева вложенных экземпляров классов, построенного парсером.

Пример 19.16. PP4E\Lang\Parser\parser2.py

```
"""
Синтаксический анализ, выполняемый отдельно от вычисления, конструирующий
дерево синтаксического анализа
"""

TraceDefault = False
class UndefinedError(Exception): pass

if __name__ == '__main__':
    from scanner import Scanner, SyntaxError, LexicalError # запускается здесь
else:
    from .scanner import Scanner, SyntaxError, LexicalError # из PyTree

##############################################################################
# интерпретатор (дерево умных объектов)
##############################################################################

class TreeNode:
    def validate(self, dict):        # проверка ошибок по умолчанию
        pass
    def apply(self, dict):           # механизм вычисления по умолчанию
        pass
    def trace(self, level):          # механизм анализа дерева по умолчанию
        print('.' * level + '<empty>')

# КЛАССЫ КОРНЕВЫХ ОБЪЕКТОВ

class BinaryNode(TreeNode):
    def __init__(self, left, right):        # наследуемые методы
        self.left, self.right = left, right # левая/правая ветви
    def validate(self, dict):
        self.left.validate(dict)            # ветви рекурсивного спуска
        self.right.validate(dict)
    def trace(self, level):
        print('.' * level + '[' + self.label + ']')
        self.left.trace(level+3)
        self.right.trace(level+3)

class TimesNode(BinaryNode):
    label = '*'
    def apply(self, dict):
        return self.left.apply(dict) * self.right.apply(dict)

class DivideNode(BinaryNode):
    label = '/'
    def apply(self, dict):
        return self.left.apply(dict) / self.right.apply(dict)

class PlusNode(BinaryNode):
    label = '+'
```

```
    def apply(self, dict):
        return self.left.apply(dict) + self.right.apply(dict)

class MinusNode(BinaryNode):
    label = '-'
    def apply(self, dict):
        return self.left.apply(dict) - self.right.apply(dict)

# КЛАССЫ ОБЪЕКТОВ-ЛИСТЬЕВ

class NumNode(TreeNode):
    def __init__(self, num):
        self.num = num                      # уже число
    def apply(self, dict):                  # проверка по умолчанию
        return self.num
    def trace(self, level):
        print('.' * level + repr(self.num)) # как прогр. код, было 'self.num'

class VarNode(TreeNode):
    def __init__(self, text, start):
        self.name = text                    # имя переменной
        self.column = start                 # номер позиции для ошибок
    def validate(self, dict):
        if not self.name in dict.keys():
            raise UndefinedError(self.name, self.column)
    def apply(self, dict):
        return dict[self.name]              # сначала проверить
    def assign(self, value, dict):
        dict[self.name] = value             # локальное расширение
    def trace(self, level):
        print('.' * level + self.name)

# КЛАССЫ КОМПОЗИТНЫХ ОБЪЕКТОВ

class AssignNode(TreeNode):
    def __init__(self, var, val):
        self.var, self.val = var, val
    def validate(self, dict):
        self.val.validate(dict)             # не проверять переменные
    def apply(self, dict):
        self.var.assign( self.val.apply(dict), dict )
    def trace(self, level):
        print('.' * level + 'set ')
        self.var.trace(level + 3)
        self.val.trace(level + 3)

##############################################################################
# парсер (синтаксический анализатор, построитель дерева)
##############################################################################

class Parser:
```

```
    def __init__(self, text=''):
        self.lex = Scanner(text)            # создать сканер
        self.vars = {'pi':3.14159}          # добавить константы
        self.traceme = TraceDefault

    def parse(self, *text):                 # внешний интерфейс
        if text:
            self.lex.newtext(text[0])       # использовать с новым текстом
        tree = self.analyse()               # разобрать строку
        if tree:
            if self.traceme:                # вывести дерево анализа?
                print(); tree.trace(0)
            if self.errorCheck(tree):       # проверка имен
                self.interpret(tree)        # вычислить дерево

    def analyse(self):
        try:
            self.lex.scan()                 # получить первую лексему
            return self.Goal()              # построить дерево анализа
        except SyntaxError:
            print('Syntax Error at column:', self.lex.start)
            self.lex.showerror()
        except LexicalError:
            print('Lexical Error at column:', self.lex.start)
            self.lex.showerror()

    def errorCheck(self, tree):
        try:
            tree.validate(self.vars)        # проверка ошибок
            return 'ok'
        except UndefinedError as instance:  # args - кортеж
            varinfo = instance.args
            print("'%s' is undefined at column: %d" % varinfo)
            self.lex.start = varinfo[1]
            self.lex.showerror()            # вернет None

    def interpret(self, tree):
        result = tree.apply(self.vars)      # дерево вычисляет себя само
        if result != None:                  # игнорировать результат 'set'
            print(result)                   # игнорировать ошибки

    def Goal(self):
        if self.lex.token in ['num', 'var', '(']:
            tree = self.Expr()
            self.lex.match('\0')
            return tree
        elif self.lex.token == 'set':
            tree = self.Assign()
            self.lex.match('\0')
            return tree
        else:
```

```
        raise SyntaxError()

    def Assign(self):
        self.lex.match('set')
        vartree = VarNode(self.lex.value, self.lex.start)
        self.lex.match('var')
        valtree = self.Expr()
        return AssignNode(vartree, valtree) # два поддерева

    def Expr(self):
        left = self.Factor()                # левое поддерево
        while True:
            if self.lex.token in ['\0', ')']:
                return left
            elif self.lex.token == '+':
                self.lex.scan()
                left = PlusNode(left, self.Factor()) # добавить корневой узел
            elif self.lex.token == '-':
                self.lex.scan()
                left = MinusNode(left, self.Factor()) # растет вверх/вправо
            else:
                raise SyntaxError()

    def Factor(self):
        left = self.Term()
        while True:
            if self.lex.token in ['+', '-', '\0', ')']:
                return left
            elif self.lex.token == '*':
                self.lex.scan()
                left = TimesNode(left, self.Term())
            elif self.lex.token == '/':
                self.lex.scan()
                left = DivideNode(left, self.Term())
            else:
                raise SyntaxError()

    def Term(self):
        if self.lex.token == 'num':
            leaf = NumNode(self.lex.match('num'))
            return leaf
        elif self.lex.token == 'var':
            leaf = VarNode(self.lex.value, self.lex.start)
            self.lex.scan()
            return leaf
        elif self.lex.token == '(':
            self.lex.scan()
            tree = self.Expr()
            self.lex.match(')')
            return tree
        else:
```

```
            raise SyntaxError()

################################################################################
# программный код самотестирования: использует свой парсер, тест для parser1
################################################################################

if __name__ == '__main__':
    import testparser
    testparser.test(Parser, 'parser2') # выполнить с классом Parser
```

Обратите внимание на способ обработки исключения обнаружения неопределенного имени в errorCheck. Когда классы исключений наследуют встроенный класс Exception, их экземпляры автоматически возвращают аргументы, переданные в вызов конструктора исключения в виде кортежа в своем атрибуте args, что удобно использовать для форматирования строк.

Отметьте также, что новый парсер повторно использует тот же самый модуль сканера. Чтобы перехватывать ошибки, возбуждаемые сканером, он также импортирует классы, идентифицирующие исключения сканера. И сканер, и парсер могут возбуждать исключения при ошибках (лексические ошибки, синтаксические ошибки и ошибки неопределенных имен). Они перехватываются на верхнем уровне парсера и завершают текущий разбор. При этом нет надобности устанавливать и проверять флаги состояния, чтобы завершить рекурсию. Поскольку математические действия выполняются с использованием целых чисел, чисел с плавающей точкой и операторов Python, обычно не требуется перехватывать ошибки переполнения или потери значимости. Но в данной реализации парсер не обрабатывает такие ошибки, как деление на ноль; они приводят к системному выходу из парсера с выводом содержимого стека Python и сообщения об ошибке. Выяснение причины и исправление этого оставляются в качестве упражнения.

Когда модуль parser2 запускается как программа верхнего уровня, мы получаем тот же результат, что и при запуске модуля parser1. Фактически они совместно используют один и тот же программный код тестирования – оба парсера передают свои объекты классов Parser функции testparser.test. А поскольку классы также являются объектами, мы точно так же можем передать эту версию парсера в интерактивный цикл модуля testparser: testparser.interact(parser2.Parser).

Внешне новый парсер ведет себя точно так же, как и оригинал, поэтому я не буду приводить его вывод здесь (запустите пример у себя, чтобы самим увидеть результаты). Следует, однако, отметить, что этот парсер поддерживает возможность использования его как в виде самостоятельного сценария, так и в виде пакета, доступного для импортирования из других программ, таких как PyTree, которой мы воспользуемся чуть ниже. Однако Python 3.X больше не включает собственный каталог модуля в путь поиска, поэтому в последнем случае мы вынуждены использовать синтаксис импорта по относительному пути в пакете и им-

портировать из другого каталога при тестировании в интерактивном режиме:

```
C:\...\PP4E\Lang\Parser> parser2.py
parser2 <class '__main__.Parser'>
5.0
```
...остальная часть вывода такая же, как в parser1...

```
C:\...PP4E\Lang\Parser> python
>>> import parser2
    from .scanner import Scanner, SyntaxError, LexicalError # из PyTree
ValueError: Attempted relative import in non-package
```
(ValueError: Попытка импорта по относительному пути отсутствующего пакета)

```
C:\...\PP4E\Lang\Parser> cd ..
C:\...\PP4E\Lang> Parser\parser2.py
parser2 <class '__main__.Parser'>
5.0
```
...остальная часть вывода такая же, как в parser1...

```
C:\...\PP4E\Lang> python
>>> from Parser import parser2
>>> x = parser2.Parser()
>>> x.parse('1 + 2 * 3 + 4')
11
>>> import Parser.testparser
>>> Parser.testparser.interact(parser2.Parser)
<class 'Parser.parser2.Parser'>
Enter=> 4 * 3 + 5
17
Enter=> stop
>>>
```

Использования в parser2 синтаксиса импорта пакета по полному пути вместо импорта по относительному пути или неквалифицированного импорта:

```
from PP4E.Lang.Parser import scanner
```

должно быть достаточно для всех трех случаев использования – как сценария и как импортируемого модуля из того же самого или из другого каталога – но для этого требуется указывать корректный путь, и такой подход выглядит излишеством при импортировании файла из того же каталога, где находится импортирующий его сценарий.

Структура дерева синтаксического анализа

В действительности, единственное существенное отличие этой последней версии парсера состоит в том, что он строит и использует деревья для вычисления выражения, вместо того чтобы вычислять его в ходе

анализа. Промежуточное представление выражения является деревом экземпляров классов, форма которого отражает порядок выполнения операторов. В этом парсере имеется также логика, с помощью которой выводится листинг созданного дерева с отступами, если атрибут traceme установлен в значение True (или 1). Отступы указывают на вложенность поддеревьев, а для двухместных операторов сначала выводятся левые поддеревья. Например:

```
C:\...\PP4E\Lang> python
>>> from Parser import parser2
>>> p = parser2.Parser()
>>> p.traceme = True
>>> p.parse('5 + 4 * 2')
[+]
...5
...[*]
......4
......2
13
```

При вычислении этого дерева метод apply рекурсивно вычисляет поддеревья и применяет корневые операторы к результатам. Здесь * вычисляется раньше +, так как находится ниже в дереве. Метод Factor поглощает подстроку * перед возвращением правого поддерева в Expr. Следующее дерево принимает иную форму:

```
>>> p.parse('5 * 4 - 2')

[-]
...[*]
......5
......4
...2
18
```

В этом примере * вычисляется прежде -. Метод Factor выполняет обход подстроки выражений * и / перед возвращением результирующего левого поддерева в Expr. Следующий пример немного сложнее, но он следует тем же правилам:

```
>>> p.parse('1 + 3 * (2 * 3 + 4)')
[+]
...1
...[*]
......3
......[+]
.........[*]
............2
............3
.........4
31
```

Деревья состоят из вложенных экземпляров классов. С точки зрения ООП, это еще один способ применения композиции. Так как узлы дерева являются просто экземплярами классов, это дерево можно создать и вычислить вручную:

```
PlusNode( NumNode(1),
          TimesNode( NumNode(3),
                     PlusNode( TimesNode(NumNode(2), NumNode(3)),
                               NumNode(4) ))).apply({})
```

Но точно так же можно позволить парсеру построить его (Python не настолько похож на Lisp, как вы, возможно, слышали).

Исследование деревьев синтаксического анализа с помощью PyTree

Но погодите – есть более удобный способ исследования структур деревьев синтаксического анализа. На рис. 19.1 показано дерево синтаксического анализа, созданное для строки `1 + 3 * (2 * 3 + 4)`, которое изображается в окне PyTree, программы с графическим интерфейсом для визуализации деревьев, представленной в конце главы 18. Это возможно только потому, что модуль `parser2` явно строит дерево синтаксического анализа (модуль `parser1` производит вычисления во время анализа), и благодаря общности и возможности повторного использования программного кода PyTree.

Если вы читали предыдущую главу, то вспомните, что программа PyTree способна нарисовать структуру данных почти любого дерева, но изначально настроена для работы с двоичными деревьями поиска и деревьями синтаксического анализа, которые изучаются в этой главе. Если в отображаемом дереве щелкнуть на изображении узла, будет вычислено значение поддерева с корнем в этом узле.

Программа PyTree облегчает процесс изучения и экспериментирования с парсером. Для определения формы дерева, получающегося для заданного выражения, запустите PyTree, щелкните на переключателе Parser (Парсер), введите выражение в поле ввода внизу и щелкните на кнопке input (ввести) (или нажмите клавишу Enter). Для генерации дерева по введенным данным будет запущен класс парсера, и графический интерфейс отобразит результат. В зависимости от операторов, используемых в выражении, некоторые деревья, весьма существенно отличающиеся по форме, возвращают один и тот же результат.

Попробуйте запустить PyTree на своем компьютере, чтобы получить более полное представление о процессе синтаксического анализа. (Мне хотелось бы показать большее число примеров деревьев, но я уже исчерпал количество страниц, отведенное для книги.)

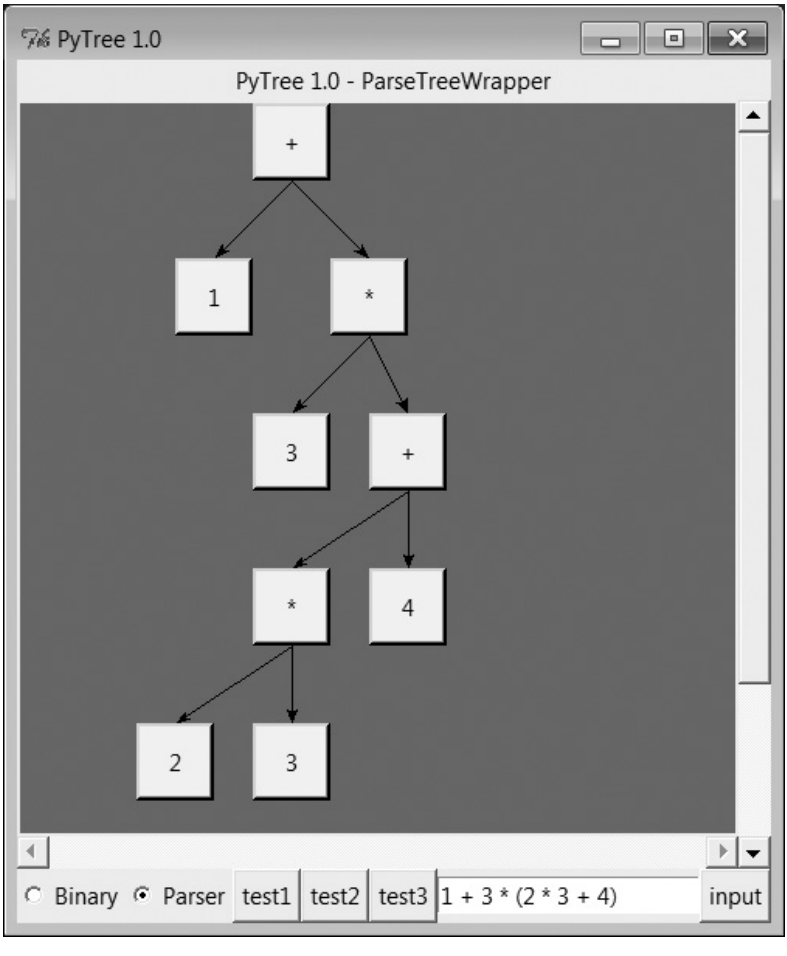


Рис. 19.1. *Дерево синтаксического анализа для строки «1 + 3 * (2 * 3 + 4)»*

Парсеры и возможности Python

Показанные выше самостоятельные программы парсеров демонстрируют некоторые интересные понятия и подчеркивают мощь языка Python как универсального инструмента программирования. В вашей работе написание парсеров может оказаться одной из типичных задач, которые вы решаете полностью с помощью обычных языков программирования, таких как C. Парсеры являются важным компонентом в широком круге приложений, но в некоторых случаях они не столь необходимы, как может показаться. Позвольте объяснить почему.

К данному моменту мы реализовали механизм анализа выражений и затем добавили интерпретатор дерева синтаксического анализа, что-

бы легче было модифицировать программный код. В настоящем виде парсер работает, но, возможно, медленнее, чем реализация на языке C. Если парсер используется часто, можно ускорить его работу, переместив выполнение некоторых операций в модули расширений C. Например, сначала можно перевести на C сканер, потому что он часто вызывается из парсера. В конце концов, можно добавить в грамматику компоненты, позволяющие выражениям обращаться к специфическим для приложения переменным и функциям.

Все это – путь к хорошему проектированию. Но для некоторых приложений такой подход может оказаться не самым лучшим при программировании на Python. Простейшим способом вычисления вводимого выражения в Python часто оказывается вызов встроенной функции `eval`. На самом деле обычно можно заменить всю программу вычисления выражения одним вызовом функции. В следующем разделе демонстрируется, как можно использовать этот прием, чтобы упростить системы синтаксического анализа в целом.

Важна центральная идея, лежащая в основе языка и подчеркиваемая в следующем разделе: если у вас уже имеется расширяемая, встраиваемая система языка высокого уровня, зачем изобретать еще одну? Python сам часто в состоянии удовлетворить потребности основанных на языке компонентов.

PyCalc: программа/объект калькулятора

В завершение этой главы я хочу показать практическое применение некоторых технологий синтаксического анализа, с которыми мы познакомились в предыдущем разделе. Этот раздел представляет PyCalc – программу-калькулятор на языке Python с графическим интерфейсом, аналогичным программам калькуляторов, имеющимся в большинстве оконных систем. Но, как и большинство примеров графических интерфейсов в этой книге, PyCalc имеет ряд преимуществ перед существующими калькуляторами. Так как программа PyCalc написана на языке Python, ее легко настраивать и переносить между оконными платформами. А так как она реализована с помощью классов, то является самостоятельной программой и библиотекой повторно используемых объектов.

Графический интерфейс простого калькулятора

Однако прежде чем показывать, как написать полноценный калькулятор, начнем с простого модуля, представленного в примере 19.17. В нем реализован графический интерфейс калькулятора с ограниченными возможностями, кнопки которого просто добавляют текст в поле ввода наверху, образуя строку выражения Python. При получении и запуске строки немедленно получается результат. На рис. 19.2 изображено окно, создаваемое этим модулем, если запустить его как сценарий верхнего уровня.

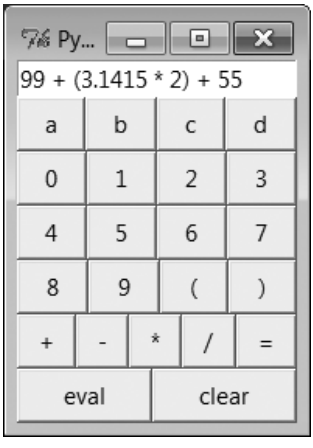


Рис. 19.2. *Сценарий calc0 в действии, в Windows 7 (результат=160,283)*

Пример 19.17. PP4E\Lang\Calculator\calc0.py

```
"""
упрощенный графический интерфейс калькулятора: выражение вычисляется
с помощью eval/exec
"""


from tkinter import *
from PP4E.Gui.Tools.widgets import frame, button, entry

class CalcGui(Frame):
    def __init__(self, parent=None):     # расширенный фрейм
        Frame.__init__(self, parent)     # по умолчанию - верхнего уровня
        self.pack(expand=YES, fill=BOTH) # все части растягиваются
        self.master.title('Python Calculator 0.1') # 6 фреймов плюс поле ввода
        self.master.iconname("pcalc1")

        self.names = {}                  # пространство имен для переменных
        text = StringVar()
        entry(self, TOP, text)

        rows = ["abcd", "0123", "4567", "89()"]
        for row in rows:
            frm = frame(self, TOP)
            for char in row:
                button(frm, LEFT, char,
                            lambda char=char: text.set(text.get() + char))

        frm = frame(self, TOP)
        for char in "+-*/=":
            button(frm, LEFT, char,
                        lambda char=char: text.set(text.get()+' '+char+' '))
```

```
        frm = frame(self, BOTTOM)
        button(frm, LEFT, 'eval', lambda: self.eval(text))
        button(frm, LEFT, 'clear', lambda: text.set(''))

    def eval(self, text):
        try:
            text.set(str(eval(text.get(), self.names, self.names))) # был 'x'
        except SyntaxError:
            try:
                exec(text.get(), self.names, self.names)
            except:
                text.set("ERROR") # не годится и как инструкция?
            else:
                text.set('')      # действует, как инструкция
        except:
            text.set("ERROR")     # другие ошибки вычисления выражения

if __name__ == '__main__': CalcGui().mainloop()
```

Создание графического интерфейса

Вряд ли может быть калькулятор проще, но он вполне пригоден для демонстрации основ. В этом окне имеются кнопки для ввода чисел, имен переменных и операторов. Оно строится путем вставки кнопок во фреймы: каждый ряд кнопок представляет собой вложенный экземпляр класса `Frame`, и сам графический интерфейс также является подклассом `Frame`, в который вставлены поле `Entry` и шесть вложенных фреймов рядов (точно так же можно было бы использовать компоновку по сетке). Фрейм калькулятора, поле ввода и кнопки сделаны растягиваемыми в импортированном вспомогательном модуле `widgets`, реализованном нами в примере 10.1.

Этот калькулятор конструирует строку, которая целиком должна быть передана интерпретатору Python при нажатии кнопки eval (вычислить). Поскольку в поле ввода можно ввести любое выражение или инструкцию языка Python, кнопки существуют только для удобства. На самом деле поле ввода не слишком отличается от командной строки. Попробуйте ввести `import sys`, а затем `dir(sys)`, чтобы вывести атрибуты модуля `sys` вверху в поле ввода – обычно калькулятор так не используют, однако эта особенность достаточно показательна.[1]

[1] И еще раз я должен предостеречь вас от выполнения таких строк, если вы не можете быть уверены в том, что они не нанесут вреда. Будьте осторожны, если такие строки могут вводиться пользователями, не пользующимися доверием, – они смогут получить доступ ко всему, что имеется на компьютере и что доступно интерпретатору Python. Подробнее о проблемах безопасности программного кода, реализующего графические интерфейсы, веб-страницы и в других контекстах, рассказывается в главах 9 и 15.

В конструкторе `CalcGui` кнопки реализованы в виде списков строк; каждая строка представляет ряд, а каждый символ в строке – кнопку. С целью сохранения дополнительных данных, необходимых функциям обратного вызова для каждой кнопки, используются `lambda`-выражения. Функции обратного вызова сохраняют символ кнопки и связанную переменную с текстовым значением, чтобы обеспечить добавление символа в конец текущей строки элемента ввода при нажатии кнопки.

Урок 4: встраивание лучше, чем парсеры

Вместо анализа и вычисления выражений вручную калькулятор использует функции `eval` и `exec` для вызова парсера и интерпретатора Python во время исполнения. По существу, калькулятор выполняет встроенный программный код Python из программы Python. Это возможно благодаря тому, что среда разработки Python (парсер и компилятор байт-кода) всегда являются частью систем, использующих Python. Поскольку различия между средой разработки и средой выполнения отсутствуют, программы на языке Python имеют возможность использовать парсер Python.

В результате всё вычисление выражения заменяется одним вызовом `eval` или `exec`. В более широком смысле это мощный прием, о котором следует помнить: сам язык Python может заменить много небольших специальных языков. Помимо того что сокращается время разработки, клиентам приходится учить лишь один язык, который может быть достаточно прост, чтобы конечный пользователь мог программировать на нем.

Кроме того, Python может приобретать особенности любого приложения. Если интерфейс языка требует специфических для приложения расширений, просто добавьте классы Python или экспортируйте API для использования во встроенном программном коде Python в качестве расширения на языке C. В результате интерпретации программного кода Python со специфическими для приложения расширениями необходимость в индивидуальных парсерах почти полностью отпадает.

Есть также важное дополнительное преимущество такого подхода: у встроенного программного кода Python есть доступ ко всем средствам и возможностям мощного развитого языка программирования. Он может пользоваться списками, функциями, классами, внешними модулями и даже такими крупными инструментами Python, как библиотека tkinter, хранилища shelve, потоки выполнения, сетевые сокеты и средства получения веб-страниц. Вам могут понадобиться годы, чтобы обеспечить такие функции в специальном парсере языка. Можете поинтересоваться у Гвидо.

Обратите внимание, что внутри циклов параметры передаются lambda-выражениям как *аргументы по умолчанию*. Вспомните, как в главе 7 говорилось, что ссылки внутри lambda-выражений (или внутри вложенных инструкций def) на имена в объемлющей области видимости разыменовываются в момент вызова вложенной функции, а не в момент ее создания. Когда позднее сгенерированная функция будет вызвана, ссылки внутри lambda-выражений будут отражать последние значения, полученные во внешней области видимости, которые могут не совпадать со значениями, имевшимися на момент выполнения lambda-выражений. Значения по умолчанию, напротив, вычисляются в момент создания функций и поэтому позволяют сохранять текущие значения переменных цикла. Без аргументов по умолчанию все кнопки отражали бы значение, полученное в последней итерации цикла.

Выполнение строк программного кода

Этот модуль реализует калькулятор с графическим интерфейсом в 45 строках программного кода (считая комментарии и пустые строки). Но если честно, то он несколько «мошенничает»: вычисление выражений делегируется интерпретатору Python. На самом деле большую часть работы здесь делают встроенные функции eval и exec:

eval

Анализирует, вычисляет и возвращает результат выражения, представленного в виде строки.

exec

Выполняет произвольную инструкцию Python, представленную в виде строки, но не возвращает значение.

Обе принимают дополнительные словари, которые могут использоваться как глобальные и локальные пространства имен для присваивания и вычисления имен, используемых в строках программного кода. Словарь self.names в калькуляторе становится таблицей символов для выполнения выражений калькулятора. Родственная функция Python compile может использоваться для предварительного компилирования строк кода до передачи их в eval и exec (используйте ее, если нужно выполнять одну и ту же строку многократно).

По умолчанию пространством имен строки программного кода является пространство имен вызывающей программы. Если здесь не передавать словари, то строки будут выполняться в пространстве имен метода eval. Поскольку локальное пространство имен метода исчезает после возврата из него, не было бы возможности сохранить имена, присваиваемые в строке. Обратите внимание на использование встроенных обработчиков исключений в методе eval:

1. Сначала предполагается, что строка является выражением, и вызывается встроенная функция eval.

2. В случае неудачи из-за синтаксической ошибки строка вычисляется как инструкция, с помощью функции exec.
3. Если и эта попытка оказывается безуспешной, сообщается об ошибке в строке (синтаксическая ошибка, неопределенное имя и так далее).

Инструкции и недопустимые выражения могут разбираться дважды, но издержки при этом не важны, и нельзя узнать, является ли строка выражением или инструкцией, не разобрав ее вручную. Обратите внимание, что кнопка eval (вычислить) вычисляет выражения, а кнопка = устанавливает переменные Python, выполняя оператор присваивания. Имена переменных являются комбинациями клавиш букв «abcd» (или любым именем, введенным непосредственно). Они присваиваются и вычисляются в словаре, используемом для представления пространства имен калькулятора.

Расширение и прикрепление

Клиенты, использующие реализацию этого калькулятора, получаются такие же простые, как сам калькулятор. Как и большинство графических интерфейсов на основе классов из библиотеки tkinter, данный интерфейс может быть расширен в подклассах – класс в примере 19.18 переопределяет конструктор простого калькулятора для вставки новых графических элементов.

Пример 19.18. PP4E\Lang\Calculator\calc0ext.py

```
from tkinter import *
from calc0 import CalcGui

class Inner(CalcGui):          # расширенный графический интерфейс
    def __init__(self):
        CalcGui.__init__(self)
        Label(self, text='Calc Subclass').pack() # добавить после
        Button(self, text='Quit', command=self.quit).pack() # подразумевается
                                                            # позиция top
Inner().mainloop()
```

Возможно встраивание калькулятора в класс контейнера – в примере 19.19 к общему родителю прикрепляются пакет виджетов простого калькулятора и дополнительные элементы.

Пример 19.19. PP4E\Lang\Calculator\calc0emb.py

```
from tkinter import *
from calc0 import CalcGui  # добавить родителя, без вызовов владельца

class Outer:
    def __init__(self, parent):                # встроить интерфейс
        Label(parent, text='Calc Attachment').pack() # side=top
        CalcGui(parent)                        # добавить фрейм калькулятора
```

```
        Button(parent, text='Quit', command=parent.quit).pack()

root = Tk()
Outer(root)
root.mainloop()
```

На рис. 19.3 показан результат запуска обоих сценариев, представленных выше, из различных командных строк. Оба имеют отдельное поле ввода вверху. Все работает, но чтобы увидеть более практическое применение такой техники повторного использования, нужно также сделать более реальным лежащий в основе калькулятор.

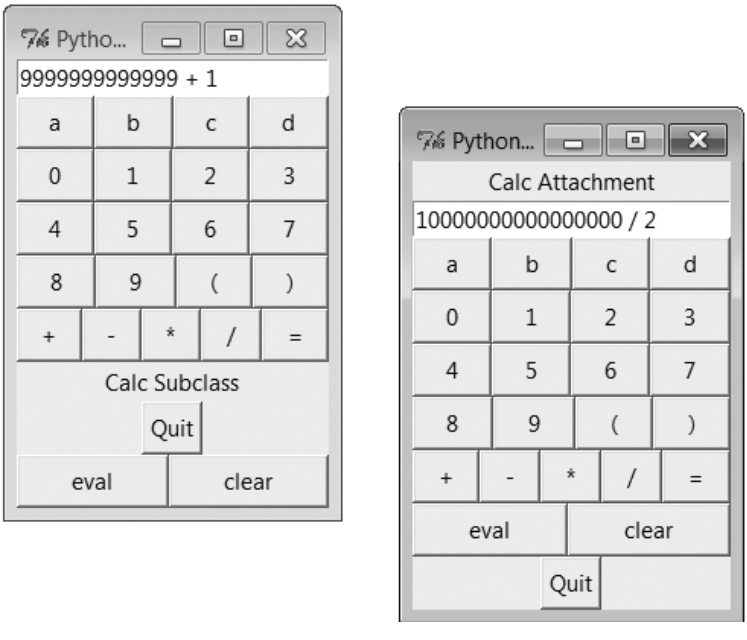


Рис. 19.3. *Объект из сценария calc0, прикрепленный и расширенный*

PyCalc – графический интерфейс «настоящего» калькулятора

Конечно, настоящие калькуляторы не конструируют строки выражений и не вычисляют их целиком. Такой подход мало чем отличается от прославленной командной строки Python. Обычно выражения вычисляются по частям по мере ввода, и временные результаты отображаются сразу, как будут вычислены. Реализовать такое поведение немного труднее: выражения должны вычисляться вручную вместо однократного вызова функции `eval`. Но конечный результат становится значительно полезнее и интуитивнее.

В этом разделе представлена реализация PyCalc – программы Python/tkinter, реализующей графический интерфейс такого традиционного калькулятора. Она имеет двоякое отношение к теме обработки текста и синтаксического анализа: анализирует и вычисляет выражения и реализует собственный язык, опирающийся на использование стека при вычислениях. Хотя логика ее вычислений сложнее, чем у простого калькулятора, представленного выше, тем не менее, она демонстрирует более развитую технологию программирования и служит эффектным финалом этой главы.

Урок 5: повторное использование – это сила

Несмотря на простоту, приемы встраивания и расширения графического интерфейса калькулятора, результаты которых показаны на рис. 19.3, иллюстрируют мощь Python как средства написания повторно используемого программного обеспечения. Благодаря возможности писать программы с применением модулей и классов отдельно написанные компоненты почти автоматически становятся универсальными инструментами. Особенности организации программ Python способствуют написанию повторно используемого программного кода.

Повторное использование программного кода является одним из главных достоинств языка Python и было одной из главных тем на всем протяжении этой книги. Для создания удачной объектно-ориентированной архитектуры нужны некоторый опыт и предусмотрительность, а выгоды, получаемые от повторного использования программного кода, могут стать очевидными не сразу. И иногда нам нужно быстро изменить реализацию, а не думать об использовании кода в будущем.

Но если писать программный код, подумывая иногда о его повторном использовании, то в долгосрочной перспективе можно сократить время разработки. Например, парсеры собственной разработки совместно использовали сканер, графический интерфейс калькулятора использует модуль `widgets` из главы 10, а в следующем разделе будет повторно использован класс `GuiMixin` из главы 10. Иногда можно выполнить часть работы, даже не начиная ее.

Работа с PyCalc

Как обычно, сначала рассмотрим графический интерфейс, а потом уже программный код. Запустить программу PyCalc можно из панелей запуска PyGadgets и PyDemos, находящихся в корневом каталоге дерева примеров, либо непосредственно, как файл *calculator.py*, листинг кото-

рого приведен ниже (например, щелкнуть на нем в менеджере файлов). На рис. 19.4 показано главное окно программы PyCalc. По умолчанию кнопки операндов окрашены в голубой цвет и имеют надписи черного цвета (а кнопки операторов – наоборот), но при этом параметры шрифта и цвета можно передать в метод конструктора класса графического интерфейса. Конечно, на черно-белых рисунках в книге этого не видно, поэтому просто запустите PyCalc у себя, чтобы понять, что я имею в виду.

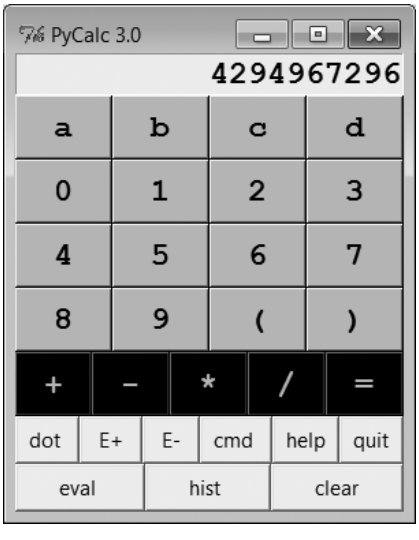


Рис. 19.4. *Калькулятор PyCalc в Windows 7*

Если запустить программу PyCalc, можно заметить, что она реализует обычную модель калькулятора – выражения вычисляются во время ввода, а не полностью, в конце. То есть части выражений вычисляются и отображаются, как только это становится возможным с учетом старшинства операторов и расставленных вручную скобок. Например, на рис. 19.4 изображен результат, полученный нажатием кнопки 2 и последующих нескольких нажатий кнопки *, с целью получить последовательность степеней двойки. Как действует это вычисление, я объясню через минуту.

Класс `CalcGui` в программе PyCalc конструирует графический интерфейс из фреймов с кнопками, подобно простому калькулятору из предыдущего раздела, но в PyCalc добавляется множество новых функций. Среди них еще один ряд командных кнопок, унаследованные от `GuiMixin` методы (представленные в главе 10), новая кнопка `cmd` (команда), выводящая немодальные диалоги для ввода произвольного программного кода на языке Python, и всплывающее окно с историей предыдущих вычислений. На рис. 19.5 показаны некоторые всплывающие окна PyCalc.

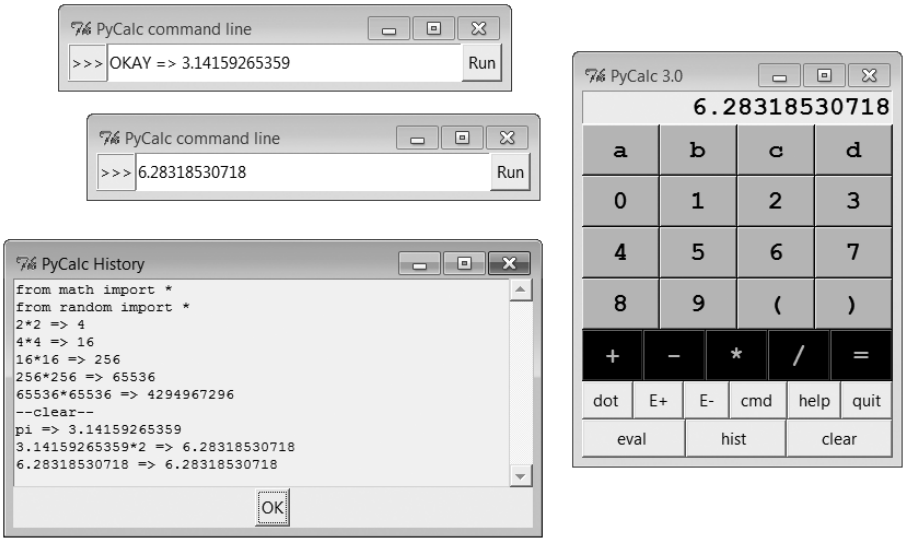


Рис. 19.5. *Калькулятор PyCalc с несколькими всплывающими окнами*

Вводить выражения в PyCalc можно щелчками на кнопках в графическом интерфейсе, вводя выражения целиком во всплывающих окнах командной строки или нажатием клавиш на клавиатуре. Программа PyCalc перехватывает события нажатий клавиш и интерпретирует их так же, как нажатия соответствующих кнопок; ввод символа «+» аналогичен нажатию кнопки +, клавиша пробела соответствует кнопке clear (очистить), клавиша Enter соответствует кнопке eval (вычислить), клавиша забоя (backspace) удаляет символ, а ввод символа «?» аналогичен нажатию кнопки help (справка).

Диалоги для ввода произвольного программного кода являются немодальными (их можно вывести в любом количестве). Они принимают любой программный код на языке Python – чтобы выполнить программный код в поле ввода, необходимо щелкнуть на кнопке Run (Запустить) или нажать клавишу Enter. Результат выполнения этого программного кода в пространстве имен калькулятора, определяемом словарем, помещается в главное окно и может использоваться в более крупных выражениях. Этим механизмом можно пользоваться, чтобы избежать необходимости применения внешних инструментов в расчетах. Например, в этих диалогах можно импортировать и использовать функции, написанные на Python или C. Текущее значение в главном окне калькулятора записывается также в поля ввода вновь открываемых диалогов для использования во вводимых выражениях.

Программа PyCalc поддерживает целые числа (неограниченной точности), отрицательные числа и числа с плавающей точкой просто потому, что они поддерживаются языком Python. Отдельные операнды и выра-

жения по-прежнему вычисляются вызовом встроенной функции `eval`, которая вызывает анализатор/интерпретатор Python на этапе исполнения. Имеется возможность присваивать значения переменным и обращаться к ним в главном окне с помощью кнопок с буквами, = и eval (вычислить). Присвоение выполняется в пространстве имен калькулятора (более сложные имена переменных можно вводить в диалогах ввода программного кода). Посмотрите в окне с историей вычислений, как используется переменная `pi`: программа PyCalc заранее импортирует имена из модулей `math` и `random` в пространство имен, в котором вычисляются выражения.

Вычисление выражений с применением стеков

Теперь, когда вы получили общее представление, чем занимается программа PyCalc, я должен немного рассказать о том, как она это делает. Большинство изменений в этой версии касается организации отображения выражений и вычисления выражений. Реализация PyCalc организована в виде двух классов:

Класс `CalcGui`

Управляет собственно графическим интерфейсом. Он контролирует события ввода и отвечает за поле отображения вверху главного окна. Однако он не вычисляет выражения – для этого он пересылает операторы и операнды, введенные в графическом интерфейсе, встроенному экземпляру класса `Evaluator`.

Класс `Evaluator`

Управляет двумя стеками. В одном стеке записываются текущие *операторы* (например, +), а в другом – текущие *операнды* (например, `3.141`). Временные результаты вычисляются по мере отправки новых операторов из `CalcGui` и помещаются на стек операндов.

Из этого можно сделать вывод, что механизм вычисления выражений сводится к жонглированию стеками операторов и операндов. В некотором смысле, для вычисления выражений калькулятор реализует простой *язык программирования* на основе стеков. При сканировании строк выражений слева направо во время ввода операнды попутно помещаются на стек; при этом операторы служат разделителями операндов и могут вызывать вычисление промежуточных результатов перед помещением на стек. Поскольку этот механизм выполняет запись информации о состоянии и производит переходы, для описания реализации языка калькулятора можно было бы использовать термин *конечный автомат*.

Ниже описывается общий сценарий:

1. Когда обнаруживается новый оператор (то есть когда нажимается кнопка или клавиша, соответствующая оператору), предыдущий операнд в поле ввода помещается на стек операндов.
2. Затем в стек операторов добавляется введенный оператор, но только после того, как все незавершенные операторы с более высоким стар-

шинством вытолкнуты со стека и применены к ожидающим обработки операндам (например, нажатие + вызывает срабатывание всех невыполненных операторов * в стеке).

3. При нажатии eval (вычислить) все оставшиеся операторы выталкиваются со стека и применяются к оставшимся операндам, и результатом является последнее значение, оставшееся в стеке операндов.

В завершение последнее значение в стеке операндов отображается в поле ввода калькулятора в готовности для использования в другой операции. Этот алгоритм вычислений лучше всего, вероятно, описать на примерах. Введем несколько выражений и посмотрим, как заполняются стеки вычислений.

Трассировка стека калькулятора PyCalc включается с помощью флага `debugme` в модуле – если он имеет значение `True`, стеки операторов и операндов выводятся в `stdout` каждый раз, когда класс `Evaluator` собирается применить оператор и *вытолкнуть* элементы со стеков. Чтобы увидеть эту информацию, запустите PyCalc в окне консоли. При каждом выталкивании элементов из стека выводится кортеж, содержащий списки стеков (*операторы*, *операнды*) – вершины стеков находятся в концах списков. Например, ниже приводится вывод в консоли после ввода и вычисления простой строки:

```
1) Нажатые клавиши: "5 * 3 + 4 <eval>"   [результат = 19]

([*], [5, 3])     [по нажатию '+': выведет "15"]
([+], [15, 4])    [по нажатию 'eval': выведет "19"]
```

Обратите внимание, что ожидающее обработки (находящееся в стеке) подвыражение * вычисляется при нажатии +: операторы * имеют более высокий приоритет, чем +, поэтому вычисления выполняются сразу, перед проталкиванием оператора +. При нажатии кнопки + поле ввода содержит число 3 – мы помещаем 3 на стек операндов, вычисляем подвыражение (5*3), помещаем результат на стек операндов, помещаем оператор + на стек операторов и продолжаем сканировать ввод пользователя. Когда в конце нажимается кнопка eval (вычислить), на стек операндов помещается число 4 и к оставшимся операндам на стеке применяется оператор +.

Поле ввода, находящееся в верхней части главного окна, также играет определенную роль в этом алгоритме. Поле ввода и стеки выражения объединяются классом калькулятора. Вообще говоря, поле ввода всегда содержит предыдущий операнд, когда нажимается кнопка оператора (например, при вводе последовательности `5 *`). Значение текстового поля должно быть помещено на стек операндов до того, как будет опознан оператор. Поэтому мы должны вытолкнуть результаты, прежде чем отобразить их после нажатия кнопки eval (вычислить) или ввода символа «)» (в противном случае результаты будут помещены на стек дважды) – они находились бы на стеке и отображались в поле ввода, от-

куда они немедленно будут помещены на стек еще раз, при вводе следующего оператора.

Для удобства и обеспечения точности вычислений после ввода оператора и перед вводом очередного операнда необходимо также стереть содержимое поля ввода (например, перед вводом 3 и 4 в выражении 5 * 3 + 4). Подобное стирание прежних значений выполняется также при нажатии кнопки eval (вычислить) или вводе символа «)» исходя из предположения, что следующее нажатие кнопки или клавиши изменит предыдущий результат – чтобы дать возможность ввести новое выражение после нажатия кнопки eval (вычислить) и новый операнд, следующий за оператором после ввода символа «)». Например, чтобы стереть 12, результат подвыражения в круглых скобках, при вводе 2 в выражении 5 + (3 * 4) * 2. Без этого стирания новые операнды просто добавлялись бы в конец текущего отображаемого значения. Данная модель также позволяет заменять промежуточный результат на операнд после ввода символа «)» вводом операнда вместо оператора.

Стеки выражений также задерживают выполнение операций с более низким приоритетом в процессе сканирования ввода. В следующем примере ожидающий оператор + не выполняется, пока не будет нажата кнопка *: поскольку оператор * имеет более высокий приоритет, нам необходимо отложить выполнение оператора +, пока не будет выполнен оператор *. Оператор * не выталкивается со стека, пока не будет получен его правый операнд 4. По нажатию кнопки eval (вычислить) со стека выталкиваются и применяются к элементам стека операндов два оператора – оператор *, находящийся на вершине стека операторов, применяется к операндам 3 и 4, находящимся на вершине стека операндов, а затем оператор + применяется к числу 5 и к числу 12, помещенному на стек оператором *:

```
2) Нажатые клавиши: "5 + 3 * 4 <eval>"   [результат = 17]

(['+', '*'], ['5', '3', '4'])  [по нажатию клавиши 'eval']
(['+'], ['5', '12'])           [выведет "17"]
```

Для выражений, содержащих операторы с одинаковым приоритетом, как показано ниже, выталкивание и вычисление выполняется сразу при сканировании слева направо без откладывания вычисления. Это приводит к линейному вычислению слева направо, если отсутствуют скобки: выражение 5+3+4 вычисляется как ((5+3)+4). Для операций + и * порядок не имеет значения:

```
3) Нажатые клавиши: "5 + 3 + 4 <eval>" [результат = 12]

(['+'], ['5', '3'])  [по нажатию второго оператора '+']
(['+'], ['8', '4'])  [по нажатию клавиши 'eval']
```

Ниже приводится более сложный пример. В данном случае все операторы и операнды помещаются на стек (откладываются), пока не будет вве-

ден символ «)» в конце. Чтобы обеспечить такую работу со скобками, открывающей скобке «(» присвоен наивысший приоритет и она помещается на стек операторов, чтобы отсрочить выполнение операторов, находящихся на стеке ниже, пока не будет введен символ «)». При вводе символа «)» подвыражение, заключенное в круглые скобки, выталкивается со стека и вычисляется, и в поле ввода выводится число 13. При нажатии кнопки eval (вычислить) вычисляется оставшаяся часть выражения ((3 * 13), (1 + 39)) и отображается окончательный результат (40). Этот результат в поле ввода превращается в левый операнд другого оператора.

```
4) Нажатые клавиши: "1 + 3 * ( 1 + 3 * 4 ) <eval>"   [результат = 40]
(['+', '*', '(', '+', '*'], ['1', '3', '1', '3', '4'])  [по нажатию
                                                         клавиши ')']
(['+', '*', '(', '+'], ['1', '3', '1', '12'])           [выведет "13"]
(['+', '*'], ['1', '3', '13'])                          [по нажатию
                                                         клавиши 'eval']
(['+'], ['1', '39'])
```

В действительности можно снова использовать любой временный результат: если снова нажать кнопку оператора, не вводя новых операндов, он применяется к результату предыдущего нажатия – значение в поле ввода дважды помещается на стек и выполняется оператор. Нажмите кнопку * множество раз после ввода 2, чтобы увидеть, как это действует (например, `2***`). Первое нажатие *, поместит 2 на стек операндов и `*` на стек операторов. Следующее нажатие * опять поместит 2 из поля ввода на стек операндов, вытолкнет и вычислит выражение на стеке (2 * 2), поместит результат обратно на стек и отобразит его. Каждое последующее нажатие * повторно будет помещать на стек текущее отображаемое значение и выполнять операцию, последовательно вычисляя квадраты чисел.

На рис. 19.6 показано, как выглядят оба стека в самой верхней точке во время сканирования выражения при трассировке предыдущего примера. Верхний оператор применяется к двум верхним операндам, и результат возвращается обратно для следующего оператора. Поскольку в работе участвуют два стека, результат напоминает преобразование выражения в строку вида `+1*3(+1*34` и вычисление ее справа налево. Однако в других случаях производится вычисление частей выражений и отображение промежуточных результатов, поэтому данный процесс не является простым преобразованием строки.

Наконец, следующий пример возбуждает ошибку. Программа PyCalc довольно небрежно относится к обработке ошибок. Многие ошибки оказываются невозможными благодаря самому алгоритму, но на таких ошибках, как непарные скобки, реализация алгоритма вычисления все же спотыкается. Вместо того чтобы пытаться явно обнаруживать все возможные случаи ошибок, в методе `reduce` используется общая инструкция `try`, которая перехватывает все ошибки: ошибки выражений, число-

вые ошибки, ошибки неопределенных имен, синтаксические ошибки и так далее.

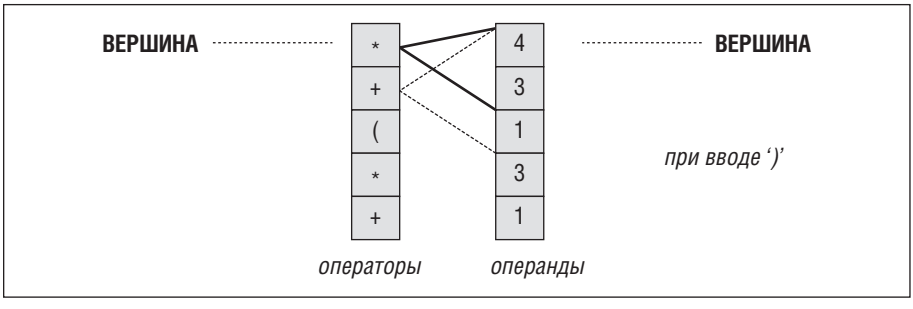


Рис. 19.6. *Стеки вычисления: 1 + 3 * (1 + 3 * 4)*

Операнды и временные результаты всегда помещаются на стек в виде строк, а все операторы применяются путем вызова функции `eval`. Если ошибка происходит внутри выражения, на стек операндов проталкивается результат `*ERROR*`, в результате чего все оставшиеся операторы также не смогут выполниться в `eval`. Вследствие этого операнд `*ERROR*` распространяется до вершины выражения и в конце оказывается последним операндом, который выводится в поле ввода, извещая о допущенной ошибке:

```
5) Нажатые клавиши: "1 + 3 * ( 1 + 3 * 4 <eval>" [результат = *ERROR*]
(['+', '*', '(', '+', '*'], ['1', '3', '1', '3', '4']) [по нажатию eval]
(['+', '*', '(', '+'], ['1', '3', '1', '12'])
(['+', '*', '('], ['1', '3', '13'])
(['+', '*'], ['1', '*ERROR*'])
(['+'], ['*ERROR*'])
(['+'], ['*ERROR*', '*ERROR*'])
```

Проследите в программном коде калькулятора, как выполняется этот и другие примеры, чтобы получить представление, как производятся вычисления с применением стеков. Как только вы поймете принцип действия механизма работы со стеками, вычисление выражений станет для вас простой задачей.

Исходный программный код PyCalc

Пример 19.20 содержит модуль с исходным программным кодом PyCalc, в котором эти идеи применены на практике в контексте графического интерфейса. Эта реализация состоит из одного файла (не считая импортированные и повторно использованные вспомогательные средства). Внимательно изучите этот программный код. Как обычно, ничто не может заменить попытки самостоятельной работы с программой, чтобы полнее понять ее функциональные возможности.

Обращайте особое внимание на комментарии, начинающиеся со слов «что сделать», где приводятся предложения по дальнейшему усовершенствованию. Подобно всем программным системам этот калькулятор может продолжать развиваться с течением времени (и фактически так и происходит с выходом каждого нового издания этой книги). Поскольку он написан на языке Python, подобные улучшения легко можно реализовать в будущем.

Пример 19.20. PP4E\Lang\Calculator\calculator.py

```
#!/usr/local/bin/python
"""
##############################################################################
PyCalc 3.0+: программа калькулятора и компонент графического интерфейса
на Python/tkinter.

Вычисляет выражения по мере ввода, перехватывает нажатия клавиш
на клавиатуре для ввода выражений; в версии 2.0 были добавлены диалоги
для ввода произвольного программного кода, отображение истории вычислений,
настройка шрифтов и цветов, вывод справочной информации о программе,
предварительный импорт констант из модулей math/random и многое другое;

3.0+ (PP4E, номер версии сохранен):
- адаптирована для работы под управлением Python 3.X (только)
- убрана обработка клавиши 'L' (тип long теперь отсутствует в языке)

3.0, изменения (PP3E):
- теперь для поля ввода вместо состояния 'disabled' используется состояние
'readonly', иначе оно окрашивается в серый цвет (исправлено в соответствии
с изменениями в версии 2.3 библиотеки Tkinter);
- исключено расширенное отображение точности для чисел с плавающей точкой
за счет использования str(), вместо `x`/repr() (исправлено в соответствии
с изменениями в Python);
- настраивается шрифт в поле ввода, чтобы текст в нем выглядел крупнее;
- используется флаг justify=right для поля ввода, чтобы обеспечить
выравнивание по правому краю, а не по левому;
- добавлены кнопки 'E+' и 'E-' (и обработка клавиши 'E') для ввода чисел
в экспоненциальной форме; вслед за нажатием клавиши 'E' вообще должен
следовать ввод цифр, а не знака + или -;
- убрана кнопка 'L' (но нажатие клавиши 'L' все еще обрабатывается): теперь
излишне, потому что Python автоматически преобразует числа, если они
оказываются слишком большими (в прошлом кнопка 'L' выполняла эту операцию
принудительно);
- повсюду используются шрифты меньшего размера;
- автоматическая прокрутка в конец окна с историей вычислений

что сделать: добавить режим включения запятых (смотрите str.format и пример
в "Изучаем Python"); добавить поддержку оператора '**'; разрешить ввод '+'
и 'J' для комплексных чисел; использовать новый тип Decimal для вещественных
чисел с фиксированной точностью; сейчас для ввода и обработки комплексных
```

```
чисел можно использовать диалог 'cmd', но такая возможность отсутствует
в главном окне; предупреждение: точность представления чисел и некоторые
особенности поведения PyCalc в настоящее время обусловлены особенностями
работы функции str();
##############################################################################
"""

from tkinter import *                           # виджеты, константы
from PP4E.Gui.Tools.guimixin import GuiMixin                  # метод quit
from PP4E.Gui.Tools.widgets import label, entry, button, frame # конструкторы
                                                              # виджетов
Fg, Bg, Font = 'black', 'skyblue', ('courier', 14, 'bold') # настр. по умолч.

debugme = True
def trace(*args):
    if debugme: print(args)

##############################################################################
# Основной класс - работает с интерфейсом пользователя; расширенный Frame
# в новом Toplevel или встроенный в другой элемент-контейнер
##############################################################################

class CalcGui(GuiMixin, Frame):
    Operators = "+-*/="                         # списки кнопок
    Operands = ["abcd", "0123", "4567", "89()"] # настраиваемые

    def __init__(self, parent=None, fg=Fg, bg=Bg, font=Font):
        Frame.__init__(self, parent)
        self.pack(expand=YES, fill=BOTH)      # все элементы растягиваются
        self.eval = Evaluator()               # встроить обработчик стека
        self.text = StringVar()               # создать связанную перемен.
        self.text.set("0")
        self.erase = 1                        # затем убрать текст "0"
        self.makeWidgets(fg, bg, font)        # построить граф. интерфейс
        if not parent or not isinstance(parent, Frame):
            self.master.title('PyCalc 3.0')   # заголов., если владеет окном
            self.master.iconname("PyCalc")       # то же для привязки клавиш
            self.master.bind('<KeyPress>', self.onKeyboard)
            self.entry.config(state='readonly')  # 3.0: не 'disabled'=серый
        else:
            self.entry.config(state='normal')
            self.entry.focus()

    def makeWidgets(self, fg, bg, font):         # 7 фреймов плюс поле ввода
        self.entry = entry(self, TOP, self.text) # шрифт, цвет настраиваемые
        self.entry.config(font=font)             # 3.0: make display larger
        self.entry.config(justify=RIGHT)         # 3.0: справа, не слева
        for row in self.Operands:
            frm = frame(self, TOP)
            for char in row:
```

```
            button(frm, LEFT, char,
                        lambda op=char: self.onOperand(op),
                        fg=fg, bg=bg, font=font)

        frm = frame(self, TOP)
        for char in self.Operators:
            button(frm, LEFT, char,
                        lambda op=char: self.onOperator(op),
                        fg=bg, bg=fg, font=font)

        frm = frame(self, TOP)
        button(frm, LEFT, 'dot ', lambda: self.onOperand('.'))
        button(frm, LEFT, ' E+ ', lambda: self.text.set(self.text.get()+'E+'))
        button(frm, LEFT, ' E- ', lambda: self.text.set(self.text.get()+'E-'))
        button(frm, LEFT, 'cmd ', self.onMakeCmdline)
        button(frm, LEFT, 'help', self.help)
        button(frm, LEFT, 'quit', self.quit)    # из guimixin

        frm = frame(self, BOTTOM)
        button(frm, LEFT, 'eval ', self.onEval)
        button(frm, LEFT, 'hist ', self.onHist)
        button(frm, LEFT, 'clear', self.onClear)

    def onClear(self):
        self.eval.clear()
        self.text.set('0')
        self.erase = 1

    def onEval(self):
        self.eval.shiftOpnd(self.text.get())  # посл. или единств. операнд
        self.eval.closeall()              # применить все оставш. операторы
        self.text.set(self.eval.popOpnd())    # вытолкнуть: след. оператор?
        self.erase = 1

    def onOperand(self, char):
        if char == '(':
            self.eval.open()
            self.text.set('(')                # очистить текст далее
            self.erase = 1
        elif char == ')':
            self.eval.shiftOpnd(self.text.get())  # послед. или единств.
                                                  # вложенный операнд
            self.eval.close()                 # вытолкнуть: след. оператор?
            self.text.set(self.eval.popOpnd())
            self.erase = 1
        else:
            if self.erase:
                self.text.set(char)           # очистить последнее значение
            else:
                self.text.set(self.text.get() + char) # иначе добавить
                                                      # в операнд
```

```
        self.erase = 0

    def onOperator(self, char):
        self.eval.shiftOpnd(self.text.get())  # втолкнуть лев. операнд
        self.eval.shiftOptr(char)             # вычислить выраж. слева?
        self.text.set(self.eval.topOpnd())    # втолкнуть оператор, показать
                                              # операнд|результат
        self.erase = 1                        # стереть при вводе след.
                                              # операнда или '('
    def onMakeCmdline(self):
        new = Toplevel()                      # новое окно верхнего уровня
        new.title('PyCalc command line')      # произвольный код Python
        frm = frame(new, TOP)                 # расширяется только Entry
        label(frm, LEFT, '>>>').pack(expand=NO)
        var = StringVar()
        ent = entry(frm, LEFT, var, width=40)
        onButton = (lambda: self.onCmdline(var, ent))
        onReturn = (lambda event: self.onCmdline(var, ent))
        button(frm, RIGHT, 'Run', onButton).pack(expand=NO)
        ent.bind('<Return>', onReturn)
        var.set(self.text.get())

    def onCmdline(self, var, ent):          # выполняет команду в окне
        try:
            value = self.eval.runstring(var.get())
            var.set('OKAY')                 # выполняет в eval
            if value != None:               # с пространством имен в словаре
                self.text.set(value)        # выражение или инструкция
                self.erase = 1
                var.set('OKAY => '+ value)
        except:                             # результат - в поле ввода
            var.set('ERROR')                # состояние в поле ввода окна
        ent.icursor(END)                    # позиция вставки после текста
        ent.select_range(0, END)            # выделить сообщ., чтобы след.
                                            # нажатие клавиши удалило его
    def onKeyboard(self, event):
        pressed = event.char                # обраб. событий клавиатуры
        if pressed != '':                   # как если бы нажата клавиша
            if pressed in self.Operators:
                self.onOperator(pressed)
            else:
                for row in self.Operands:
                    if pressed in row:
                        self.onOperand(pressed)
                        break
                else:                           # 4E: убрана клавиша 'Ll'
                    if pressed == '.':
                        self.onOperand(pressed) # может быть
                                                # началом операнда
                    if pressed in 'Ee':         # 2e10, без +/-
                       self.text.set(self.text.get()+pressed) # нет: не удал.
```

```
                    elif pressed == '\r':
                        self.onEval()           # Enter=eval
                    elif pressed == ' ':
                        self.onClear()          # пробел=очистить
                    elif pressed == '\b':
                        self.text.set(self.text.get()[:-1]) # забой
                    elif pressed == '?':
                        self.help()

    def onHist(self):
        # выводит окно с историей вычислений
        from tkinter.scrolledtext import ScrolledText # или PP4E.Gui.Tour
        new = Toplevel()                              #создать новое окно
        ok = Button(new, text="OK", command=new.destroy)
        ok.pack(pady=1, side=BOTTOM)      # добавл. первым - усекается посл.
        text = ScrolledText(new, bg='beige') # добавить Text + полосу прокрут.
        text.insert('0.0', self.eval.getHist()) # получить текст Evaluator
        text.see(END)                           # 3.0: прокрутить в конец
        text.pack(expand=YES, fill=BOTH)

        # новое окно закрывается нажатием кнопки ok или клавиши Enter
        new.title("PyCalc History")
        new.bind("<Return>", (lambda event: new.destroy()))
        ok.focus_set()          # сделать новое окно модальным:
        new.grab_set()          # получить фокус ввода, захватить приложение
        new.wait_window()       # не вернется до вызова new.destroy

    def help(self):
        self.infobox('PyCalc', 'PyCalc 3.0+\n'
                     'A Python/tkinter calculator\n'
                     'Programming Python 4E\n'
                     'May, 2010\n'
                     '(3.0 2005, 2.0 1999, 1.0 1996)\n\n'
                     'Use mouse or keyboard to\n'
                     'input numbers and operators,\n'
                     'or type code in cmd popup')

##############################################################################
# класс вычисления выражений встраивается в экземпляр CalcGui
# и используется им для вычисления выражений
##############################################################################

class Evaluator:
    def __init__(self):
        self.names = {}                         # простр. имен для переменных
        self.opnd, self.optr = [], []           # два пустых стека
        self.hist = []                          # журнал предыдущ. вычислений
        self.runstring("from math import *")    # предварит. импорт модулей
        self.runstring("from random import *") # в простр. имен калькулятора
```

```
    def clear(self):
        self.opnd, self.optr = [], []          # оставить имена нетронутыми
        if len(self.hist) > 64:                # ограничить размер истории
            self.hist = ['clear']
        else:
            self.hist.append('--clear--')

    def popOpnd(self):
        value = self.opnd[-1]        # вытолк./вернуть верх.|послед. операнд
        self.opnd[-1:] = []             # для отображ. и использования
        return value                    # или x.pop(), или del x[-1]

    def topOpnd(self):
        return self.opnd[-1]            # верхн. операнд (конец списка)

    def open(self):
        self.optr.append('(')           # трактовать '(' как оператор

    def close(self):                    # по ')' вытолкнуть до первой '('
        self.shiftOptr(')')             # ok если пусто: остается пустым
        self.optr[-2:] = []             # вытолкнуть, иначе снова будет
                                        # добавлен оператором
    def closeall(self):
        while self.optr:                # остальное по 'eval'
            self.reduce()               # послед. может быть именем перем.
        try:
            self.opnd[0] = self.runstring(self.opnd[0])
        except:
            self.opnd[0] = '*ERROR*'    # вытолкнуть, иначе снова добавится:

    afterMe = {'*': ['+', '-', '(', '='], # член класса
               '/': ['+', '-', '(', '='], # не выталкивать операторы для клав.
               '+': ['(', '='],         # если это предыдущ. оператор: push
               '-': ['(', '='],         # иначе: pop/eval предыдущ. оператор
               ')': ['(', '='],         # все левоассоциативные
               '=': ['('] }

    def shiftOpnd(self, newopnd):       # втолкнуть операнд для оператора,
        self.opnd.append(newopnd)       # ')', eval

    def shiftOptr(self, newoptr):       # применить операторы с приорит. <=
        while (self.optr and
               self.optr[-1] not in self.afterMe[newoptr]):
            self.reduce()
        self.optr.append(newoptr) # втолкнуть этот оператор над результатом
                            # операторы предполаг. стирание след. операндом
    def reduce(self):
        trace(self.optr, self.opnd)
        try:                                # свернуть верхнее выражение
            operator = self.optr[-1]        # вытолк. верх. оператор (в конце)
```

```
            [left, right] = self.opnd[-2:]  # вытолк. 2 верх.
                                            # операнда (в конце)
            self.optr[-1:] = []             # удалить срез на месте
            self.opnd[-2:] = []
            result = self.runstring(left + operator + right)
            if result == None:
                result = left           # присваивание? клавиша имени перем.
            self.opnd.append(result)    # втолкнуть строку результ. обратно
        except:
            self.opnd.append('*ERROR*') # ошибка стека/числа/имени

    def runstring(self, code):
        try: # 3.0: not `x`/repr
            result = str(eval(code, self.names, self.names)) # вычислить
            self.hist.append(code + ' => ' + result)    # добавить в журнал
        except:
            exec(code, self.names, self.names)          # инструкция: None
            self.hist.append(code)
            result = None
        return result

    def getHist(self):
        return '\n'.join(self.hist)

    def getCalcArgs():
        from sys import argv        # получить арг. команд. строки в словаре
        config = {}                 # пример: -bg black -fg red
        for arg in argv[1:]:        # шрифт пока не поддерживается
            if arg in ['-bg', '-fg']: # -bg red' -> {'bg':'red'}
                try:
                    config[arg[1:]] = argv[argv.index(arg) + 1]
                except:
                    pass
        return config

if __name__ == '__main__':
    CalcGui(**getCalcArgs()).mainloop() # по умолчанию окно верхнего уровня
```

Использование PyCalc как компонента

Я использую PyCalc как самостоятельную программу на моем компьютере, но этот калькулятор можно также использовать и в контексте других графических интерфейсов. Как и большинство классов реализации графического интерфейса в этой книге, PyCalc можно настраивать, наследуя его в подклассах, или встраивать в более крупные графические интерфейсы. Модуль, представленный в примере 19.21, демонстрирует один из способов повторного использования класса `CalcGui` путем расширения и встраивания подобно тому, как это делалось выше для простого калькулятора.

Пример 19.21. PP4E\Lang\Calculator\calculator_test.py

```
"""
проверка калькулятора: используется как расширяемый и встраиваемый компонент
графического интерфейса
"""

from tkinter import *
from calculator import CalcGui

def calcContainer(parent=None):
    frm = Frame(parent)
    frm.pack(expand=YES, fill=BOTH)
    Label(frm, text='Calc Container').pack(side=TOP)
    CalcGui(frm)
    Label(frm, text='Calc Container').pack(side=BOTTOM)
    return frm

class calcSubclass(CalcGui):
    def makeWidgets(self, fg, bg, font):
        Label(self, text='Calc Subclass').pack(side=TOP)
        Label(self, text='Calc Subclass').pack(side=BOTTOM)
        CalcGui.makeWidgets(self, fg, bg, font)
        #Label(self, text='Calc Subclass').pack(side=BOTTOM)

if __name__ == '__main__':
    import sys
    if len(sys.argv) == 1:        # % calculator_test.py
        root = Tk()               # запуск 3 калькуляторов в том же процессе
        CalcGui(Toplevel())       # каждый в новом окне верхнего уровня
        calcContainer(Toplevel())
        calcSubclass(Toplevel())
        Button(root, text='quit', command=root.quit).pack()
        root.mainloop()
    if len(sys.argv) == 2:         # % calculator_testl.py -
        CalcGui().mainloop()       # как самостоятельное окно (корневое)
    elif len(sys.argv) == 3:       # % calculator_test.py - -
        calcContainer().mainloop() # как встраиваемый компонент
    elif len(sys.argv) == 4:       # % calculator_test.py - - -
        calcSubclass().mainloop()  # как настраиваемый суперкласс
```

На рис. 19.7 показан результат выполнения этого сценария без аргументов командной строки. Мы получили экземпляры первоначального класса калькулятора плюс класс контейнера и подкласс, определенные в этом сценарии, которые все прикреплены к новым окнам верхнего уровня.

Два окна справа повторно используют базовый программный код PyCalc, выполняющийся в левом окне. Все эти окна действуют в рамках одного процесса (например, закрытие одного окна вызовет закрытие остальных), но все они действуют как независимые окна. Обратите внимание,

что при таком выполнении трех калькуляторов в одном процессе у каждого из них имеется собственное пространство имен для вычисления выражений, потому что это атрибут экземпляра класса, а не глобальная переменная уровня модуля. Поэтому установка переменных в одном калькуляторе касается только этого калькулятора и не изменяет значений, присваиваемых в других окнах. Аналогично у каждого калькулятора имеется собственный объект, управляющий стеками вычислений, благодаря чему вычисления в одном окне не отображаются в других окнах и не влияют на них.

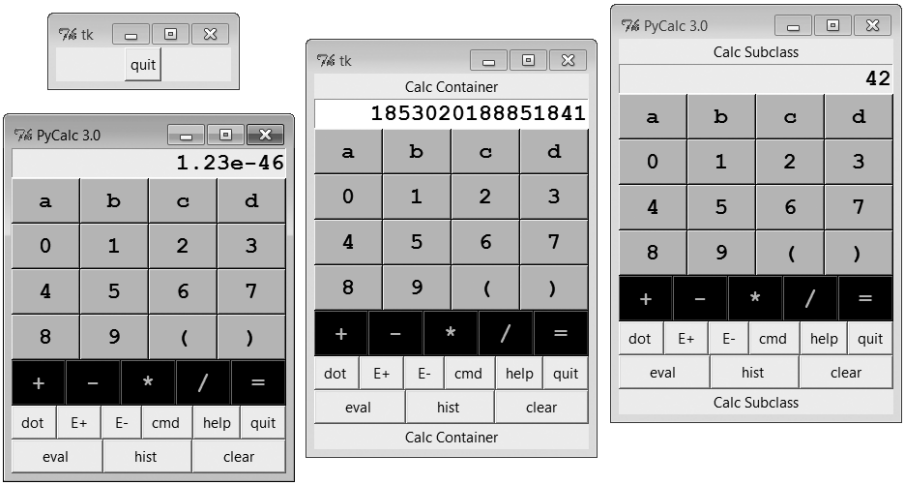


Рис. 19.7. *Сценарий calculator_test: прикрепление и расширение*

Два расширения в этом сценарии, конечно, искусственные – они просто добавляют метки вверху и внизу окна – но сама концепция имеет широкий круг применения. Класс калькулятора можно использовать повторно, прикрепляя его к любому графическому интерфейсу, где требуется калькулятор, и произвольным образом настраивать его с помощью подклассов. Это повторно используемый виджет.

Добавление новых кнопок в новые компоненты

Один очевидный способ повторного использования калькулятора заключается в добавлении новых функциональных кнопок выражений – квадратных корней, обратных значений, кубов и так далее. Такие операции можно вводить в диалогах ввода произвольного программного кода, но кнопки несколько более удобны. Такие функции можно добавить и в саму основную реализацию калькулятора, но поскольку набор таких полезных функций может различаться в зависимости от пользователей и приложений, более правильным подходом явится их добавление в отдельные расширения. Например, класс в примере 19.22 добав-

ляет в PyCalc несколько новых кнопок путем встраивания его (то есть прикрепления) в контейнер.

Пример 19.22. PP4E\Lang\Calculator\calculator_plus_emb.py

```
"""
###############################################################################
контейнер с дополнительным рядом кнопок, выполняющих типичные операции;
более практичное усовершенствование: добавляет кнопки для дополнительных
операций (sqrt, 1/x и так далее), используя прием встраивания/композиции
без создания подкласса; новые кнопки добавляются после всего фрейма CalGui,
из-за особенностей операции компоновки;
###############################################################################
"""

from tkinter import *
from calculator import CalcGui, getCalcArgs
from PP4E.Gui.Tools.widgets import frame, button, label

class CalcGuiPlus(Toplevel):
    def __init__(self, **args):
        Toplevel.__init__(self)
        label(self, TOP, 'PyCalc Plus - Container')
        self.calc = CalcGui(self, **args)
        frm = frame(self, BOTTOM)
        extras = [('sqrt', 'sqrt(%s)'),
                  ('x^2 ', '(%s)**2'),
                  ('x^3 ', '(%s)**3'),
                  ('1/x ', '1.0/(%s)')]
        for (lab, expr) in extras:
            button(frm, LEFT, lab, (lambda expr=expr: self.onExtra(expr)))
        button(frm, LEFT, ' pi ', self.onPi)

    def onExtra(self, expr):
        text = self.calc.text
        eval = self.calc.eval
        try:
            text.set(eval.runstring(expr % text.get()))
        except:
            text.set('ERROR')

    def onPi(self):
        self.calc.text.set(self.calc.eval.runstring('pi'))

if __name__ == '__main__':
    root = Tk()
    button(root, TOP, 'Quit', root.quit)
    CalcGuiPlus(**getCalcArgs()).mainloop()        # -bg,-fg в calcgui
```

Поскольку PyCalc реализован как класс Python, аналогичного эффекта всегда можно добиться, расширив PyCalc в новом подклассе, вместо его встраивания, как показано в примере 19.23.

Пример 19.23. PP4E\Lang\Calculator\calculator_plus_ext.py

```
"""
###############################################################################
усовершенствование, добавляющее дополнительный ряд кнопок, выполняющих
типичные операции; более практичное усовершенствование: добавляет кнопки
для дополнительных операций (sqrt, 1/x и так далее) в подклассе путем
наследования оригинального класса, а не встраивания; новые кнопки
отображаются перед фреймом, присоединенным к низу класса calcgui;
###############################################################################
"""

from tkinter import *
from calculator import CalcGui, getCalcArgs
from PP4E.Gui.Tools.widgets import label, frame, button

class CalcGuiPlus(CalcGui):
    def makeWidgets(self, *args):
        label(self, TOP, 'PyCalc Plus - Subclass')
        CalcGui.makeWidgets(self, *args)
        frm = frame(self, BOTTOM)
        extras = [('sqrt', 'sqrt(%s)'),
                  ('x^2 ', '(%s)**2'),
                  ('x^3 ', '(%s)**3'),
                  ('1/x ', '1.0/(%s)')]
        for (lab, expr) in extras:
            button(frm, LEFT, lab, (lambda expr=expr: self.onExtra(expr)))
        button(frm, LEFT, ' pi ', self.onPi)

    def onExtra(self, expr):
        try:
            self.text.set(self.eval.runstring(expr % self.text.get()))
        except:
            self.text.set('ERROR')

    def onPi(self):
        self.text.set(self.eval.runstring('pi'))

if __name__ == '__main__':
    CalcGuiPlus(**getCalcArgs()).mainloop() # передает -bg, -fg
```

Обратите внимание, что функции обратного вызова этих кнопок принудительно используют вещественное деление при обращении величин, потому что так действует оператор / в Python 3.X (оператор //, выполняющий деление целых чисел, отсекает остатки). Кнопки также заключают значения поля ввода в скобки, чтобы обойти проблемы старшинства операций. Вместо этого можно было бы преобразовывать текст поля ввода в число и выполнять операции с действительными числами, но Python автоматически проделывает всю работу, когда строки выражений выполняются в необработанном виде.

Отметьте также, что добавляемые в этом сценарии кнопки выполняют операции над текущим значением поля ввода непосредственно. Это не совсем то же самое, что операторы выражений, применяемые со стековым механизмом вычислений (чтобы сделать их настоящими операторами, нужны дополнительные переделки). Тем не менее эти кнопки подтверждают то, что должны продемонстрировать эти сценарии. – они используют PyCalc как компонент, как снаружи, так и изнутри.

Наконец, для тестирования расширенных классов калькулятора и параметров настройки PyCalc в примере 19.24 представлен сценарий, который отображает четыре отдельных окна калькулятора (этот сценарий запускается программой PyDemos).

Пример 19.24. PP4E\Lang\Calculator\calculator_plusplus.py

```
#!/usr/local/bin/python
"""
демонстрация всех 3 разновидностей калькулятора
каждая из них является отдельным объектом калькулятора и окном
"""


from tkinter import Tk, Button, Toplevel
import calculator, calculator_plus_ext, calculator_plus_emb

root=Tk()
calculator.CalcGui(Toplevel())
calculator.CalcGui(Toplevel(), fg='white', bg='purple')
calculator_plus_ext.CalcGuiPlus(Toplevel(), fg='gold', bg='black')
calculator_plus_emb.CalcGuiPlus(fg='black', bg='red')
Button(root, text='Quit Calcs', command=root.quit).pack()
root.mainloop()
```

На рис. 19.8 изображен результат – четыре независимых калькулятора в окнах верхнего уровня, открытых одним и тем же процессом. Два окна справа представляют специализированное повторное использование PyCalc как компонента и диалог справки справа внизу. Хотя в данной книге это может быть не видно, во всех четырех используются разные цветовые схемы – классы калькулятора принимают параметры настройки цвета и шрифта и при необходимости передают их по цепочке вызовов.

Как мы узнали ранее, эти калькуляторы могут также выполняться в виде независимых процессов и запускаться с помощью модуля `launchmodes`, с которым мы познакомились в главе 5. В действительности панели запуска PyGadgets и PyDemos так и запускают калькуляторы, поэтому дополнительные подробности смотрите в их программном коде. И, как обычно, читайте программный код и экспериментируйте с ним самостоятельно, чтобы получить более полное представление о том, как он действует. В конце концов – это Python.

Эта глава завершает обсуждение темы синтаксического анализа в этой книге. Следующая и последняя техническая глава познакомит нас с приемами интеграции Python с программами, написанными на компилируемых языках программирования, таких как C и C++. Далеко не всем требуется знать, как реализуется такая интеграция, поэтому некоторые читатели могут пропустить ее и перейти к заключительной главе 21. Однако, так как большинство программистов на Python используют библиотеки C, обернутые в модули Python (даже если такое обертывание они выполняют не сами), я рекомендую бегло просмотреть следующую главу, прежде чем вы отложите эту книгу.

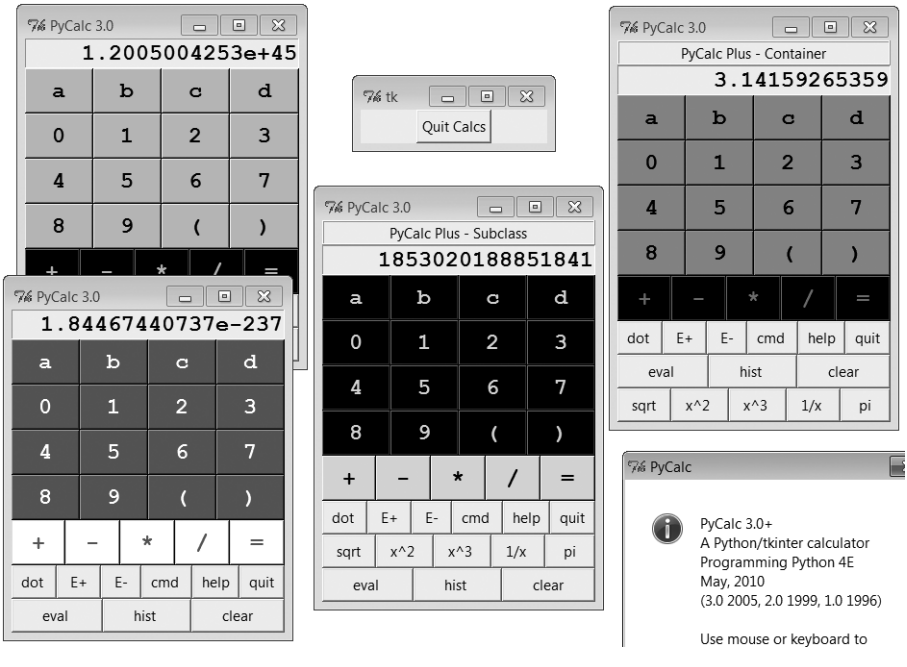


Рис. 19.8. *Сценарий calculator_plusplus: расширение, встраивание и настройка!*

Урок 6: получайте удовольствие

В заключение об одном менее ощутимом, но важном аспекте программирования на Python. Пользователи-новички часто отмечают, что на Python легко «выразить, что тебе нужно», не увязнув при этом в сложном синтаксисе или неясных правилах. Это язык, дружественный программисту. Действительно, совсем не редкость, что программы на языке Python работают с первой попытки.

Как мы видели в этой книге, такая отличительная особенность обусловлена рядом факторов – отсутствием объявлений и этапа компиляции, простым синтаксисом, практичными встроенными объектами и так далее. Python специально спроектирован для ускорения разработки (мысль, которую мы подробнее обсудим в главе 21). Многие пользователи в результате получают весьма выразительный и отзывчивый язык, пользоваться которым одно удовольствие.

Например, приведенные выше варианты программ калькуляторов были первоначально набросаны за полдня, в ходе осмысления не до конца сформированных целей. Не было аналитического этапа, формального проектирования и формальной стадии программирования. Я ввел с клавиатуры некоторые идеи, и они заработали. Более того, интерактивная природа Python позволила мне экспериментировать с новыми идеями и немедленно получать результат. После первоначальной разработки калькулятор был отшлифован и дополнен, но основная реализация сохранилась неизменной.

Естественно, такой расслабленный режим программирования годится не для каждого проекта. В каких-то случаях требуется уделить предварительному проектированию больше внимания. Для более сложных задач в Python есть модульные конструкции и поддерживаются системы, которые могут расширяться на Python или C. А графический интерфейс простого калькулятора многие, возможно, не сочтут «серьезной» программной разработкой. Но такие программы тоже нужны.

20

Интеграция Python/C

«Я заблудился в C»

На протяжении всей книги мы работали с программами, написанными на языке Python. Мы пользовались интерфейсами к внешним службам и писали многократно используемые инструменты на языке Python, но мы использовали только язык Python. Независимо от размера и практичности наших программ они до самого последнего символа были написаны на языке Python.

Для многих программистов и разработчиков сценариев это и есть конечная цель. Такое автономное программирование является одним из основных способов применения Python. Как мы видели, Python поставляется *в полном комплекте* – с интерфейсами к системным инструментам, протоколам Интернета, графическим интерфейсам, хранилищам данных и многим другим имеющимся средствам. Кроме того, для большого круга задач, с которыми мы столкнемся на практике, уже имеются готовые решения в мире открытого программного обеспечения. Система PIL, например, предоставляет возможность обрабатывать изображения в графических интерфейсах на основе библиотеки tkinter после простого запуска мастера установки.

Но для некоторых систем важнейшей характеристикой языка является возможность интеграции (или совместимость) с языком программирования C. На самом деле роль Python как языка расширений и интерфейсов в больших системах является одной из причин его популярности и того, что его часто называют «управляющим» языком. Его архитектура поддерживает *гибридные* системы, в которых смешаны компоненты, написанные на разных языках программирования. Так как у каждого языка есть свои сильные стороны, возможность отбирать и набирать

компонент за компонентом является мощным методом. Python можно включать во все системы, где требуется простой и гибкий языковый инструмент, – полагаясь на сохранение скорости выполнения в случаях, когда это имеет значение.

Компилируемые языки, такие как C и C++, оптимизированы по скорости *выполнения*, но программировать на них сложно и разработчикам, и, особенно, конечным пользователям, если им приходится адаптировать программы. Python оптимизирован по скорости *разработки*, поэтому при использовании сценариев Python для управления или индивидуальной подгонки программных компонентов, написанных на C или C++, получаются быстрые и гибкие системы и резко ускоряется скорость разработки. Например, перенос отдельных компонентов программ с языка Python на язык C может увеличить скорость выполнения программ. Кроме того, системы, спроектированные так, что их подгонка возлагается на сценарии Python, не требуют поставки полных исходных текстов и изучения конечными пользователями сложных или специализированных языков.

В этой последней технической главе мы коротко познакомимся с инструментами, предназначенными для организации взаимодействий с программными компонентами, написанными на языке C. Кроме того, мы обсудим возможность использования Python в качестве встроенного языкового инструмента в других системах и поговорим об интерфейсах для расширения сценариев Python новыми модулями и типами, реализованными на языках программирования, совместимых с C. Помимо этого мы кратко рассмотрим другие технологии интеграции, менее связанные с языком C, такие как Jython.

Обратите внимание на слово «кратко» в предыдущем абзаце. Из-за того, что далеко не всем программистам на Python требуется владение этой темой; из-за того, что для этого требуется изучать язык C и порядок создания файлов с правилами сборки для утилиты make; и из-за того, что это заключительная глава и без того весьма объемной книги, в этой главе будут опущены подробности, которые можно найти и в комплекте стандартных руководств по языку Python, и непосредственно в исходных текстах самого языка Python. Вместо этого мы рассмотрим ряд простых примеров, которые помогут вам начать двигаться в этом направлении и продемонстрируют некоторые возможности для систем Python.

Расширение и встраивание

Прежде чем перейти к программному коду, я хотел бы начать с определения, что здесь подразумевается под словом «интеграция». Несмотря на то, что этот термин имеет почти такое же широкое толкование, как и слово «объект», наше внимание в этой главе будет сосредоточено на тесной интеграции – когда передача управления между языками осуществляется простыми, прямыми и быстрыми вызовами функций. Ра-

нее мы рассматривали возможность организации менее тесной связи между компонентами приложения, используя механизмы взаимодействия между процессами и сетевые инструменты, такие как сокеты и каналы, но в этой части книги мы будем знакомиться с более непосредственными и эффективными приемами.

При объединении Python с компонентами, написанными на языке С (или на других компилируемых языках), на «верхнем уровне» может оказаться либо Python, либо С. По этой причине существуют две различные модели интеграции и два различных прикладных интерфейса:

Интерфейс расширения

Для выполнения скомпилированного программного кода библиотек на языке С из программ Python.

Интерфейс встраивания

Для выполнения программного кода Python из скомпилированных программ С.

Расширение играет три основные роли: оптимизация программ – перенос частей программы на язык С порой является единственной возможностью повысить общую производительность; использование существующих библиотек – библиотеки, открытые для использования в Python, получают более широкое распространение; поддержка возможностей в программах Python, не поддерживаемых самим языком непосредственно, – программный код на языке Python обычно не может обращаться к устройствам в абсолютных адресах памяти, например, но может вызывать функции на языке С, способные обеспечить это. Например, ярким представителем расширения в действии является пакет NumPy для Python: обеспечивая интеграцию с оптимизированными библиотеками реализаций численных алгоритмов, он превращает Python в гибкую и эффективную систему программирования вычислительных задач, сопоставимую с Matlab.

Встраивание обычно играет роль подгонки – выполняя программный код Python, подготовленный пользователем, система позволяет модифицировать себя без необходимости полностью пересобирать ее исходный программный код. Например, некоторые программы реализуют промежуточный уровень настройки на языке Python, который может использоваться для подгонки программы по месту за счет изменения программного кода Python. Кроме того, встраивание иногда используется для передачи событий обработчикам на языке Python. Инструменты конструирования графических интерфейсов на языке Python, например, обычно используют прием встраивания для передачи пользовательских событий.

На рис. 20.1 приводится схема этой традиционной двойной модели интеграции. В модели расширения управление передается из Python в программный код на языке С через промежуточный связующий уровень. В модели встраивания программный код на языке С обрабатывает объ-

екты Python и выполняет программный код Python, вызывая функции Python C API. Поскольку в модели расширения Python находится «сверху», он устанавливает определенный порядок интеграции, который можно автоматизировать с помощью таких инструментов, как SWIG – генератор программного кода, с которым мы познакомимся далее в этой главе, воспроизводящий программный код связующего уровня, необходимого для обертывания библиотек C и C++. Поскольку в модели внедрения Python занимает подчиненное положение, при ее использовании именно он предоставляет комплект инструментов API, которые могут использоваться программами C.

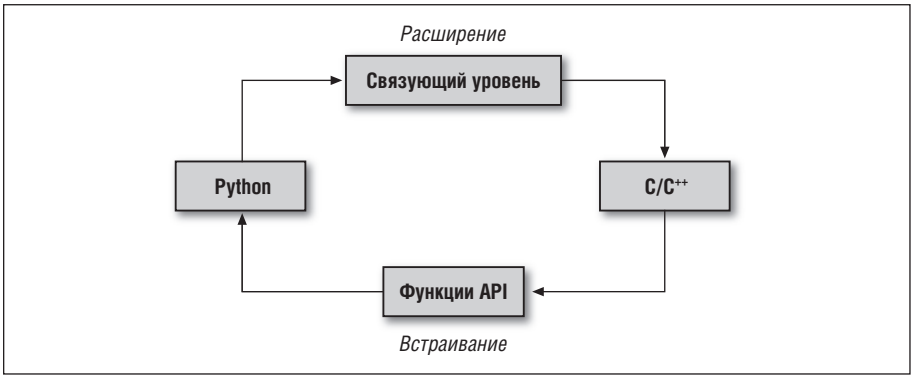


Рис. 20.1. *Традиционная модель интеграции*

В некоторых случаях разделение не такое четкое. Например, используя модуль `ctypes`, обсуждаемый ниже, сценарии Python могут вызывать библиотечные функции без применения связующего программного кода C. В таких системах, как Cython (и ее предшественнице Pyrex), различия становятся еще более существенными – библиотеки C создаются из комбинации программного кода на языках Python и C. В Jython и IronPython используется похожая модель, только язык C в них замещается языками Java и C#, а сама интеграция в значительной мере автоматизирована. Мы познакомимся с этими альтернативами далее в этой главе. А пока сосредоточимся на традиционных моделях интеграции Python/C.

В этой главе мы сначала познакомимся с моделью расширения, а затем перейдем к исследованию основ встраивания. Хотя мы будем рассматривать эти модели в отдельности, следует иметь в виду, что во многих системах обе эти модели комбинируются. Например, встроенный программный код Python, выполняемый из C, может также использовать присоединенные расширения C для взаимодействия с содержащим его приложением. А в системах, основанных на обратных вызовах, библиотеки C, доступные через интерфейсы расширения, в дальнейшем могут

использовать технику встраивания для выполнения обработчиков обратного вызова Python.

Например, при создании кнопок с помощью библиотеки tkinter в Python, представленной ранее в этой книге, мы вызывали функции в библиотеке C через прикладной интерфейс расширения. Когда позднее пользователь нажимал эти кнопки, библиотека C реализации инструментов графического интерфейса перехватывала события и передавала их функциям на языке Python, используя модель встраивания. Хотя большая часть деталей скрыта для программного кода Python, тем не менее, в таких системах управление часто свободно передается между языками. Архитектура Python открыта и реентерабельна, что позволяет соединять языки произвольным образом.

Дополнительные примеры интеграции Python/C, не вошедшие в книгу, вы найдете в исходных текстах самого языка Python – его каталоги *Modules* и *Objects* являются богатыми источниками информации. Большинство встроенных элементов Python, использованных в этой книге, – от простых, таких как целые числа и строки, до более сложных, вроде файлов, системных вызовов, tkinter и DBM, – используют ту же, представленную выше структуру интеграции. Приемы интеграции, используемые в них, можно изучать по дистрибутиву исходного кода Python и применять при реализации собственных расширений.

Кроме того, руководства «Extending and Embedding» и «Python/C API» содержат достаточно полную информацию и предоставляют сведения, дополняющие эту главу. Если вы планируете использовать приемы интеграции, в качестве следующего шага вам обязательно следует ознакомиться с этими руководствами. Например, в руководствах приводится дополнительная информация о типах расширений на языке C, о применении расширений C в программах с поддержкой многопоточной модели выполнения и о запуске нескольких интерпретаторов в программах, использующих модель встраивания, которую мы в значительной мере опустим здесь.

Расширения на C: обзор

Поскольку сам Python сегодня написан на C, компилированные расширения Python могут быть написаны на любых языках, совместимых с C в отношении стеков вызова и компоновки. В их число входят C, а также C++ с надлежащими объявлениями «extern C» (автоматически помещаемыми в файлы заголовков Python). Независимо от языка, используемого для реализации расширений, существует два вида расширений Python на компилируемых языках:

Модули на C

Библиотеки инструментов, воспринимаемые клиентами как файлы модулей Python.

Типы C

Множественные экземпляры объектов, ведущих себя, как встроенные типы и классы.

Обычно модули расширений на C используются для реализации простых библиотек функций и в программном коде Python они появляются в виде импортируемых модулей (отсюда и их название). Типы C используются для программирования объектов, генерирующих множественные экземпляры, каждый из которых имеет собственную информацию о состоянии и способен поддерживать операторы выражений подобно многим классам Python. Типы C обладают теми же возможностями, что и встроенные типы и классы Python. Они обладают методами, поддерживают операции сложения, индексирования, извлечения срезов и так далее.

Чтобы обеспечить возможность взаимодействия, как модули, так и типы C должны предоставить программный код «связующего слоя», который будет выполнять трансляцию вызовов и передавать данные между двумя языками. Этот слой регистрирует свои операции в интерпретаторе Python в виде указателей на функции C. Связующий слой C отвечает за преобразование аргументов, передаваемых из Python в C, и преобразование результатов, передаваемых из C в Python. Сценарии Python просто экспортируют расширения C и пользуются ими, как если бы они в действительности были написаны на Python. Поскольку все действия по трансляции осуществляет программный код C, взаимодействие со стороны сценариев Python выглядит простым и прозрачным.

Модули и типы на C ответственны также за передачу сообщений об ошибках обратно в Python, при обнаружении ошибок, сгенерированных вызовами Python API, и за управление счетчиками ссылок сборщика мусора для объектов, сохраняемых слоем C неограниченно долго – объекты Python, хранимые в программном коде C, не будут утилизированы, пока вы обеспечите возможность уменьшения счетчиков ссылок до нуля. Модули и типы C могут компоноваться с Python статически (на этапе сборки) или динамически (при первом импортировании). При соблюдении всех этих условий расширения на C способны стать еще одним инструментом, доступным для использования в сценариях Python.

Простой модуль расширения на C

Такова история вкратце. Чтобы создать модуль на C, необходимо писать программный код на языке C, а чтобы создать тип на C, требуется рассказать о многом таком, что невозможно уместить в этой главе. Эта книга не ставит перед собой целью дать вам знания о программировании на языке C, если у вас их еще нет, тем не менее, для придания конкретности вышесказанному нам придется обратиться к программному коду. Поскольку модули на C проще, а типы C обычно экспортируют мо-

дуль C с функцией конструктора экземпляра, начнем изучение основ программирования модулей C с короткого примера.

Как уже упоминалось, при добавлении новых или существующих компонентов C в Python необходимо написать на C слой интерфейсной («связующей») логики, обеспечивающий межъязыковую диспетчеризацию и трансляцию данных. Файл с исходным программным кодом на C, представленный в примере 20.1, демонстрирует, как написать такой слой вручную. В нем реализован простой модуль расширения на C с именем `hello` для использования в сценариях Python, содержащий функцию с именем `message`, которая просто возвращает передаваемую ей в качестве аргумента строку вместе с добавленным в ее начало некоторым текстом. Сценарии Python смогут вызывать эту функцию, как обычно, но она написана на C, а не на Python.

Пример 20.1. PP4E\Integrate\Extend\Hello\hello.c

```
/*****************************************************************
* Простой модуль расширения на C для Python с именем "hello";
* скомпилируйте его в файл ".so" где-нибудь в пути поиска python,
* импортируйте и вызовите функцию hello.message;
*****************************************************************/
#include <Python.h>
#include <string.h>

/* функции модуля */
static PyObject *                           /* возвращаемый объект */
message(PyObject *self, PyObject *args)     /* self не использ. в модулях */
{                                           /* args - из Python */
    char *fromPython, result[1024];
    if (! PyArg_Parse(args, "(s)", &fromPython)) /* преобраз. Python -> C */
        return NULL;                             /* null=исключение */
    else {
        strcpy(result, "Hello, ");           /* создать строку C */
        strcat(result, fromPython);          /* добавить строку из Python */
        return Py_BuildValue("s", result);   /* преобразовать C -> Python */
    }
}

/* таблица регистрации */
static PyMethodDef hello_methods[] = {
    {"message", message, METH_VARARGS, "func doc"}, /* имя, адрес функции */
                                                    /* формат, описание */
    {NULL, NULL, 0, NULL}                    /* признак конца таблицы */
};

/* структура определения модуля */
static struct PyModuleDef hellomodule = {
    PyModuleDef_HEAD_INIT,
    "hello",    /* имя модуля */
    "mod doc",  /* описание модуля, может быть NULL */
```

```
    -1,       /* размер структуры для каждого экземпляра, -1=глоб. перем. */
    hello_methods         /* ссылка на таблицу методов */
};

/* инициализация модуля */
PyMODINIT_FUNC
PyInit_hello()            /* вызывается первой инструкцией импорта */
{                         /* имя имеет значение при динамической загрузке */
    return PyModule_Create(&hellomodule);
}
```

Этот модуль на C имеет стандартную структуру, состоящую из 4 частей, описанных в комментариях, которой следуют все модули на C и которая заметно изменилась в Python 3.X. В конечном итоге, программный код Python вызовет функцию `message` из этого файла C, передав объект строки и получив обратно новый объект строки. Сначала, однако, нужно как-то скомпоновать файл с интерпретатором Python. Для использования этого файла C в сценарии Python скомпилируйте его в динамически загружаемый объектный файл (например, *hello.so* в Linux или *hello.dll* в Cygwin для Windows) с помощью make-файла, такого как в примере 20.2, и поместите полученный объектный файл в каталог, присутствующий в пути поиска модулей, так же, как это делается для файлов *.py* или *.pyc*.

Пример 20.2. PP4E\Integrate\Extend\Hello\makefile.hello

```
################################################################
# Компилирует файл hello.c в разделяемый объектный файл в Cygwin,
# динамически загружаемый при первом импорте в Python.
################################################################

PYLIB = /usr/local/bin
PYINC = /usr/local/include/python3.1

hello.dll: hello.c
        gcc hello.c -g -I$(PYINC) -shared -L$(PYLIB) -lpython3.1 -o hello.dll
clean:
        rm -f hello.dll core
```

Это make-файл для Cygwin, использующий `gcc` для компиляции программного кода на C в Windows. На других платформах make-файлы аналогичны, но могут отличаться в деталях. Как рассказывалось в главе 5, Cygwin – это Unix-подобная оболочка и комплект библиотек для Windows. Для работы с примерами в этой главе вам необходимо либо установить Cygwin в свою систему Windows, либо изменить make-файлы в соответствии с требованиями вашего компилятора и платформы. Не забудьте включить путь к каталогу установки Python с помощью флага `-I`, чтобы обеспечить доступ к заголовочным файлам Python, а также путь к двоичным файлам библиотеки Python с помощью флага `-L`, если это необходимо. Я использовал имена каталогов для моей уста-

новки Python 3.1 на моем ноутбуке после сборки из исходных текстов. Обратите также внимание, что для оформления отступов в make-файлах следует использовать символы табуляции, если вы собираетесь копировать их содержимое из электронной версии книги, где они заменяются пробелами.

Теперь, чтобы задействовать make-файл из примера 20.2 для сборки модуля расширения из примера 20.1, просто введите в оболочке стандартную команду `make` (в этом примере используется оболочка Cygwin и я перенес одну строку для большей ясности):

```
.../PP4E/Integrate/Extend/Hello$ make -f makefile.hello
gcc hello.c -g -I/usr/local/include/python3.1 -shared
               -L/usr/local/bin -lpython3.1 -o hello.dll
```

Эта команда создаст разделяемый объектный файл – библиотеку *.dll* в Cygwin для Windows. При компиляции таким способом Python автоматически будет загружать и компоновать модуль C при первом импорте сценарием Python. К моменту импортирования двоичный файл библиотеки *.dll* должен находиться в каталоге, расположенном в пути поиска модулей Python, подобно обычным файлам *.py*. Поскольку при импорте интерпретатор Python всегда начинает поиск с текущего рабочего каталога, примеры из этой главы будут работать, если запускать их из того же каталога, где выполнялась компиляция (.), без необходимости копировать или перемещать файлы. В крупных системах обычно требуется помещать скомпилированные расширения в каталог, перечисленный в переменной окружения `PYTHONPATH` или в файлах *.pth*, или использовать утилиту *distutils* Python для установки их в подкаталог *site-packages* в стандартной библиотеке.

Наконец, чтобы вызвать функцию C из программы Python, просто импортируйте модуль `hello` и вызовите его функцию `hello.message` со строкой – обратно вы получите обычную строку Python:

```
.../PP4E/Integrate/Extend/Hello$ python
>>> import hello                        # импортировать модуль C
>>> hello.message('world')              # вызвать функцию C
'Hello, world'
>>> hello.message('extending')
'Hello, extending'
```

Вот и все – вы только что вызвали из Python функцию интегрированного модуля C. Самое важное, на что нужно обратить внимание, – это то, что функция C выглядит точно так, как если бы она была написана на Python. Вызывающая программа на языке Python отправляет и получает при вызове обычные строковые объекты – интерпретатор Python осуществляет маршрутизацию вызова функции C, а функция C сама справляется с преобразованием данных Python/C.

На самом деле мало что выдает в `hello` модуль расширения на C, за исключением имени файла. Программный код на языке Python импорти-

рует модуль и загружает его атрибуты, как если бы он был написан на Python. Модули расширений на C даже отвечают на вызовы функции `dir`, как обычно, и обладают стандартными атрибутами модуля и имени файла, хотя имя файла в этом случае не оканчивается на *.py* или *.pyc* – единственный очевидный признак, позволяющий сказать, что это библиотека на C:

```
>>> dir(hello)                          # атрибуты модуля на C
['__doc__', '__file__', '__name__', '__package__', 'message']

>>> hello.__name__, hello.__file__
('hello', 'hello.dll')

>>> hello.message                       # объект функции на C
<built-in function message>
>>> hello                               # объект модуля на C
<module 'hello' from 'hello.dll'>

>>> hello.__doc__                       # строка документирования
'mod doc'                               # в программном коде C
>>> hello.message.__doc__
'func doc'

>>> hello.message()                     # ошибки тоже обрабатываются
TypeError: argument must be sequence of length 1, not 0
(TypeError: аргумент должен быть последовательностью
с длиной 1, а не 0)
```

Подобно любому модулю Python к расширению на C можно обращаться из файла сценария. Сценарий на языке Python, представленный в примере 20.3, импортирует и использует модуль расширения на C из примера 20.1.

Пример 20.3. PP4E\Integrate\Extend\Hello\hellouse.py

```
"импортирует и использует модуль расширения на C"

import hello
print(hello.message('C'))
print(hello.message('module ' + hello.__file__))

for i in range(3):
    reply = hello.message(str(i))
    print(reply)
```

Этот сценарий можно запускать, как всякий другой, – при первом импорте модуля `hello` интерпретатор Python автоматически найдет объектный файл *.dll* модуля в каталоге, входящем в путь поиска, и динамически включит его в процесс. Весь вывод этого сценария представляет строки, возвращаемые функцией C из файла *hello.c*:

```
.../PP4E/Integrate/Extend/Hello$ python hellouse.py
Hello, C
Hello, module /cygdrive/c/.../PP4E/Integrate/Extend/Hello/hello.dll
Hello, 0
Hello, 1
Hello, 2
```

Дополнительные сведения о программировании модулей на языке C, а также советы по компиляции и компоновке вы найдете в стандартных руководствах по языку Python. В качестве альтернативы использованию make-файлов обратите также внимание на файлы *disthello.py* и *disthello-alt.py* в пакете с примерами. Ниже приводится краткое описание исходного программного кода первого из них:

```
# для сборки: python disthello.py build
# получившаяся библиотека dll появится в подкаталоге build

from distutils.core import setup, Extension
setup(ext_modules=[Extension('hello', ['hello.c'])])
```

Это сценарий на языке Python, который выполняет компиляцию расширения на C с помощью пакета инструментов `distutils` – стандартного компонента Python, используемого для сборки, установки и распространения расширений Python, написанных на языке Python или C. Основная задача пакета `distutil` состоит в том, чтобы автоматизировать сборку и установку пакетов дистрибутивов переносимым способом, но он также знает, как компилировать расширения на C. Обычно в состав дистрибутивов включается файл *setup.py*, который выполняет установку в подкаталог *site-packages* стандартной библиотеки. К сожалению, пакет `distutils` слишком крупный, чтобы его можно было использовать в этой главе – дополнительные подробности о нем смотрите в двух его руководствах в наборе руководств по языку Python.

Генератор интегрирующего программного кода SWIG

Как вы уже наверняка поняли, программирование расширений на C вручную может превратиться в довольно сложную задачу (это практически неизбежно при программировании на языке C). Я показал основы расширений C в этой главе, чтобы вам стала понятна лежащая в основе структура. Но сегодня расширения C обычно лучше и проще реализуются с помощью инструмента, генерирующего весь необходимый связующий интегрирующий программный код автоматически. В мире Python существует множество таких инструментов, включая SIP, SWIG и Boost.Python. Мы рассмотрим эти альтернативы в конце данной главы. Входящая в перечень система SWIG широко используется разработчиками Python.

SWIG – Simplified Wrapper and Interface Generator (упрощенный генератор оболочек и интерфейсов) – является системой с открытыми исходными текстами. Первоначально она была создана Дэйвом Бизли (Dave Beazley), а теперь разрабатывается сообществом, как и сам Python. Она использует объявления типов C и C++ для генерации законченных модулей расширения C, интегрирующих существующие библиотеки для использования в сценариях Python. Генерируемые модули расширения на C (и C++) являются законченными: они автоматически осуществляют преобразование данных, поддерживают протоколы ошибок, управляют счетчиками ссылок и прочее.

То есть SWIG автоматически генерирует весь связующий программный код, необходимый для подключения компонентов C и C++ к программам Python, – просто запустите SWIG, скомпилируйте полученный от него результат и вы получите готовое к работе расширение. Вам остается только разобраться с деталями компиляции и компоновки, а остальную часть работы, связанной с созданием расширения на C, выполнит SWIG.

Простой пример SWIG

Чтобы не писать программный код на C вручную, как в предыдущем разделе, а воспользоваться инструментом SWIG, достаточно написать функцию на языке C, которую нужно использовать из Python, без всякой логики интеграции с Python, как если бы функцию требовалось вызывать только из C. В примере 20.4 демонстрируется переделанная версия функции из примера 20.1.

Пример 20.4. PP4E\Integrate\Extend\HelloLib\hellolib.c

```
/*******************************************************************
* простой файл библиотеки на C с единственной функцией "message",
* которая предназначена для использования в программах Python.
* Здесь нет ничего, что говорило бы о Python, - эта функция на языке C
* может вызываться из программ на C, а также из Python
* (с помощью связующего программного кода).
*******************************************************************/

#include <string.h>
#include <hellolib.h>

static char result[1024];      /* не экспортируется */

char *
message(char *label)           /* экспортируется */
{
    strcpy(result, "Hello, "); /* создать строку C */
    strcat(result, label);     /* добавить label */
    return result;             /* вернуть результат */
}
```

Теперь определим обычный заголовочный файл C и объявим эту функцию внешней, как показано в примере 20.5. Это можно счесть излишеством для такого маленького примера, но это подчеркивает суть.

Пример 20.5. PP4E\Integrate\Extend\HelloLib\hellolib.h

```
/*************************************************************************
* Определяет имена, экспортируемые файлом hellolib.c в пространство имен C,
* а не программ Python - экспортируемые имена для последних определяются
* в таблице регистрации методов в программном коде модуля расширения
* Python, а не в этом файле .h;
*************************************************************************/

extern char *message(char *label);
```

Теперь вместо всего связующего программного кода расширения Python, показанного в предыдущем разделе, напишем входной файл объявлений типов SWIG, как в примере 20.6.

Пример 20.6. PP4E\Integrate\Extend\Swig\hellolib.i

```
/*********************************************************************
* Файл описания модуля Swig для файла библиотеки на C.
* Создание расширения выполняется командой "swig -python hellolib.i".
*********************************************************************/

%module hellowrap

%{
#include <hellolib.h>
%}

extern char *message(char*); /* или: %include "../HelloLib/hellolib.h" */
                        /* или: %include hellolib.h, и испол. флаг -I arg */
```

Этот файл описывает сигнатуру функции C. Вообще говоря, SWIG сканирует файлы, содержащие объявления ANSI C и C++. Его входной файл может иметь вид описания интерфейса (обычно с расширением *.i*) или быть заголовочным файлом или файлом с исходным программным кодом C/C++. Формат файлов интерфейсов, подобных этому, чаще всего используется для ввода – они могут содержать комментарии в формате C или C++, объявления типов, как в стандартных заголовочных файлах, и директивы SWIG, начинающиеся с `%`. Например:

`%module`

Определяет имя модуля, которое будет использоваться при импорте в Python.

`%{...%}`

Содержит программный код, добавляемый в создаваемый файл оболочки дословно.

`extern`

Объявляет экспортируемые имена в обычном синтаксисе ANSI C/C++.

`%include`

Заставляет SWIG сканировать другой файл (флаги `-I` определяют пути поиска).

В этом примере можно было заставить SWIG прочесть файл заголовков *hellolib.h*, представленный в примере 20.5, непосредственно. Но одним из преимуществ специальных входных файлов SWIG, таких как *hellolib.i*, является то, что с их помощью можно отобрать функции, которые будут заключены в оболочку и экспортированы в Python, и это дает более полный контроль над процессом генерации.

SWIG – это вспомогательная программа, которая запускается из сценариев компиляции, а не язык программирования, поэтому здесь особенно нечего больше показывать. Просто добавьте в свой make-файл запуск SWIG и скомпилируйте его вывод для компоновки с Python. В примере 20.7 демонстрируется способ осуществления этого в Cygwin.

Пример 20.7. PP4E\Integrate\Extend\Swig\makefile.hellolib-swig

```
##########################################################################
# Использование SWIG для интеграции hellolib.c с программами Python
# в Cygwin. В текущей версии SWIG (>1.3.13) библиотека DLL должна иметь
# ведущий "_" в имени, потому что также создается файл .py без "_" в имени.
##########################################################################

PYLIB = /usr/local/bin
PYINC = /usr/local/include/python3.1
CLIB  = ../HelloLib
SWIG  = /cygdrive/c/temp/swigwin-2.0.0/swig

# библиотека плюс ее обертка
_hellowrap.dll: hellolib_wrap.o $(CLIB)/hellolib.o
        gcc -shared hellolib_wrap.o $(CLIB)/hellolib.o \
                        -L$(PYLIB) -lpython3.1 -o $@

# генерирует модуль-обертку
hellolib_wrap.o: hellolib_wrap.c $(CLIB)/hellolib.h
        gcc hellolib_wrap.c -g -I$(CLIB) -I$(PYINC) -c -o $@

hellolib_wrap.c: hellolib.i
        $(SWIG) -python -I$(CLIB) hellolib.i

# программный код библиотеки C (в другом каталоге)
$(CLIB)/hellolib.o: $(CLIB)/hellolib.c $(CLIB)/hellolib.h
        gcc $(CLIB)/hellolib.c -g -I$(CLIB) -c -o $(CLIB)/hellolib.o

clean:
        rm -f *.dll *.o *.pyc core
```

```
force:
        rm -f *.dll *.o *.pyc core hellolib_wrap.c hellowrap.py
```

При запуске входного файла *hellolob.i* в этом make-файле SWIG создаст два файла:

hellolib_wrap.c

Связующий программный код модуля расширения на C.

hellolibwrap.py

Модуль Python, импортирующий сгенерированный модуль расширения на C.

Первый получает имя по имени входного файла, а второй – из директивы `%module`. В действительности, в настоящее время SWIG генерирует два модуля: для достижения интеграции он использует комбинацию программного кода на языках Python и C code. В конечном итоге сценарии импортируют сгенерированный файл модуля на языке Python, который импортирует сгенерированный и скомпилированный модуль на C. Если вы готовы пробиваться до конца, то сгенерированный программный код можно найти в пакете с примерами для книги; следует учесть, что он может измениться со временем и является слишком обобщенным, чтобы быть простым.

Для создания модуля C этот make-файл просто запускает SWIG, чтобы получить связующий программный код, компилирует полученный результат, компилирует оригинальный программный код библиотеки на C и затем объединяет результаты с компилированной оберткой, создавая *_hellowrap.dll*, библиотеки DLL, которую будет загружать *hellowrap.py* при импортировании в сценариях Python:

```
.../PP4E/Integrate/Extend/Swig$ dir
hellolib.i makefile.hellolib-swig

.../PP4E/Integrate/Extend/Swig$ make -f makefile.hellolib-swig
/cygdrive/c/temp/swigwin-2.0.0/swig -python -I../HelloLib hellolib.i
gcc hellolib_wrap.c -g -I../HelloLib -I/usr/local/include/python3.1
                 -c -o hellolib_wrap.o
gcc ../HelloLib/hellolib.c -g -I../HelloLib -c -o ../HelloLib/hellolib.o
gcc -shared hellolib_wrap.o ../HelloLib/hellolib.o \
                -L/usr/local/bin -lpython3.1 -o _hellowrap.dll

.../PP4E/Integrate/Extend/Swig$ dir
_hellowrap.dll hellolib_wrap.c hellowrap.py
hellolib.i     hellolib_wrap.o makefile.hellolib-swig
```

В результате получается файл динамически загружаемого модуля расширения на C, готовый к импортированию программным кодом на языке Python. Подобно всем модулям файл *_hellowrap.dll*, вместе с *hellowrap.py*, должен быть помещен в каталог, находящийся в пути поиска модулей Python (их можно оставить в каталоге, где выполнялась компиляция, если импортирующий сценарий запускается из того же ката-

лога). Обратите внимание, что имя файла библиотеки *.dll* должно начинаться с символа подчеркивания. Это является обязательным требованием SWIG, потому что этот инструмент дополнительно создает файл *.py* с тем же именем без подчеркивания – если имена будут совпадать, то импортироваться сможет только какой-то один модуль, а нам необходимы оба (сценарии должны импортировать модуль *.py*, который в свою очередь импортирует библиотеку *.dll*).

Как обычно в программировании на языке C, вам, возможно, придется повозиться с make-файлом, чтобы заставить его работать в вашей системе. Однако как только вам удастся запустить make-файл, работу можно считать законченной. Сгенерированный модуль на C используется точно так же, как созданная вручную версия, показанная выше, за исключением того, что SWIG позаботился об автоматизации наиболее сложных этапов. Передача вызовов функций из Python в программный код на C, представленный в примере 20.4, и возврат результатов осуществляется через сгенерированный инструментом SWIG слой – благодаря SWIG все это «просто работает»:

```
.../PP4E/Integrate/Extend/Swig$ python
>>> import hellowrap       # импортировать связующий слой + файл библиотеки
>>> hellowrap.message('swig world') # при импорте всегда
'Hello, swig world'                 # первым просматривается cwd

>>> hellowrap.__file__
'hellowrap.py'
>>> dir(hellowrap)
['__builtins__', '__doc__', '__file__', '__name__', '_hellowrap', ...
'message']

>>> hellowrap._hellowrap
<module '_hellowrap' from '_hellowrap.dll'>
```

Иными словами, научившись пользоваться SWIG, можно в целом забыть детали интеграции, о которых рассказывалось в этой главе. На самом деле SWIG настолько сведущ в генерации связующего программного кода Python, что обычно значительно проще и менее чревато ошибками писать расширения на C для Python сначала как чистые библиотеки C или C++, а затем добавлять их в Python, пропуская файлы заголовков через SWIG, как продемонстрировано здесь.

Мы лишь слегка коснулись возможностей SWIG. Гораздо больше информации об этом инструменте можно почерпнуть из его руководства по Python, доступного на сайте *http://www.swig.org*. Несмотря на простоту примеров в этой главе, SWIG обладает достаточно широкими возможностями интеграции библиотек, таких же сложных, как расширения Windows и широко используемые графические библиотеки, например OpenGL. Мы еще вернемся к этому инструменту далее в этой главе и исследуем его модель «теневых классов», предназначенную для обер-

тывания классов C++. А теперь перейдем к изучению примера более полезного расширения.

Создание оберток для функций окружения C

Следующий наш пример является модулем расширения C, который интегрирует функции getenv и putenv из стандартной библиотеки C, предназначенные для работы с переменными окружения, для использования в сценариях Python. В примере 20.8 представлен файл на языке C, достигающий поставленной цели за счет создания связующего слоя вручную.

Пример 20.8. PP4E\Integrate\Extend\Cenviron\cenviron.c

```
/*************************************************************************
* Модуль расширения на C для Python с именем "cenviron". Обертывает
* библиотечные функции getenv/putenv для использования в программах Python.
*************************************************************************/

#include <Python.h>
#include <stdlib.h>
#include <string.h>

/***********************/
/* 1) функции модуля   */
/***********************/

static PyObject *                              /* возвращаемый объект */
wrap_getenv(PyObject *self, PyObject *args)    /* self не используется */
{                                              /* args - из python */
    char *varName, *varValue;
    PyObject *returnObj = NULL;                /* null=исключение */

    if (PyArg_Parse(args, "(s)", &varName)) {  /* Python -> C */
        varValue = getenv(varName);            /* вызов getenv из библ. C */
        if (varValue != NULL)
            returnObj = Py_BuildValue("s", varValue); /* C -> Python */
        else
            PyErr_SetString(PyExc_SystemError, "Error calling getenv");
    }
    return returnObj;
}

static PyObject *
wrap_putenv(PyObject *self, PyObject *args)
{
    char *varName, *varValue, *varAssign;
    PyObject *returnObj = NULL;

    if (PyArg_Parse(args, "(ss)", &varName, &varValue))
```

```
    {
        varAssign = malloc(strlen(varName) + strlen(varValue) + 2);
        sprintf(varAssign, "%s=%s", varName, varValue);
        if (putenv(varAssign) == 0) {
            Py_INCREF(Py_None);             /* успешный вызов C */
            returnObj = Py_None;            /* ссылка на None */
        }
        else
            PyErr_SetString(PyExc_SystemError, "Error calling putenv");
    }
    return returnObj;
}

/***************************/
/* 2) таблица регистрации */
/***************************/

static PyMethodDef cenviron_methods[] = {
    {"getenv", wrap_getenv, METH_VARARGS, "getenv doc"}, /* имя, адрес,... */
    {"putenv", wrap_putenv, METH_VARARGS, "putenv doc"}, /* имя, адрес,... */
    {NULL, NULL, 0, NULL}                                /* признак конца */
};

/*************************/
/* 3) определение модуля */
/*************************/

static struct PyModuleDef cenvironmodule = {
    PyModuleDef_HEAD_INIT,
    "cenviron",                /* имя модуля */
    "cenviron doc",            /* описание модуля, может быть NULL */
    -1, /* размер структуры для каждого экземпляра, -1=глоб. перем. */
    cenviron_methods           /* ссылка на таблицу методов */
};

/****************************/
/* 4) инициализация модуля */
/****************************/

PyMODINIT_FUNC
PyInit_cenviron()  /* вызывается первой инструкцией импорта */
{                  /* имя имеет значение при динамической загрузке */
    return PyModule_Create(&cenvironmodule);
}
```

Хотя этот пример достаточно представителен, тем не менее, сейчас он менее полезен, чем когда он был включен в первое издание этой книги, – как мы узнали во второй части, можно не только получать значения переменных окружения из таблицы `os.environ`, но и автоматически вызывать функцию C `putenv`, присваивая значения ключам в этой таблице, чтобы экспортировать новые значения в слой программного кода C.

То есть обращение os.environ['key'] загрузит значение переменной окружения 'key', а операция os.environ['key']=value присвоит значение переменной окружения как в Python, так и в C.

Второе действие – присвоение в C – было добавлено в Python после выхода первого издания книги. Кроме дополнительного показа приемов программирования расширений этот пример все же служит практической цели: даже сегодня изменения переменных окружения в программном коде на C, связанном с процессом Python, не видны при обращении к os.environ в программном коде Python. То есть после запуска программы таблица os.environ отражает только последующие изменения, сделанные в программном коде Python.

Кроме того, несмотря на то, что в настоящее время в модуле os Python имеются обе функции, putenv и getenv, их интеграция выглядит неполной. Изменения в os.environ приводят к вызову os.putenv, но прямой вызов os.putenv не изменяет содержимое os.environ, поэтому значения в os.environ могут не соответствовать действительности. А функция os.getenv в настоящее время просто обращается к os.environ и поэтому может не замечать изменения в окружении, выполненные уже после запуска процесса, за пределами программного кода Python. Такое положение вещей может изредка приводить к проблемам, но все равно этот модуль расширения на C нельзя назвать полностью бесполезным. Для правильного взаимодействия переменных окружения со встроенным программным кодом C необходимо вызывать функции или библиотеки языка C непосредственно (по крайней мере, пока Python опять не изменит эту модель!).

Файл *cenviron.c*, представленный в примере 20.8, создает модуль Python с именем cenviron, который идет немного дальше, чем предыдущий пример, – он экспортирует две функции, явно определяет описания некоторых исключений и обращается к счетчику ссылок объекта Python None (он не создается заново, поэтому необходимо добавить ссылку перед отправкой его в Python). Как и раньше, чтобы добавить этот модуль в Python, следует откомпилировать и скомпоновать этот программный код в объектный файл. В примере 20.9 представлен make-файл для Cygwin, который компилирует исходный программный код на языке C в модуль, готовый для динамического связывания при импортировании.

Пример 20.9. PP4E\Integrate\Extend\Cenviron\makefile.cenviron

```
#####################################################################
# Компилирует cenviron.c в cenviron.dll - разделяемый объектный файл
# в Cygwin, динамически загружаемый первой инструкцией импорта.
#####################################################################

PYLIB = /usr/local/bin
PYINC = /usr/local/include/python3.1

cenviron.dll: cenviron.c
```

```
        gcc cenviron.c -g -I$(PYINC) -shared -L$(PYLIB) -lpython3.1 -o $@
clean:
        rm -f *.pyc cenviron.dll
```

Для сборки введите в оболочке команду `make -f makefile.cenviron`. Перед запуском сценария убедитесь, что файл *.dll* находится в каталоге, включенном в путь поиска Python (текущий рабочий каталог тоже годится):

```
.../PP4E/Integrate/Extend/Cenviron$ python
>>> import cenviron
>>> cenviron.getenv('USER') # подобно os.environ[key], но загружает заново
'mark'
>>> cenviron.putenv('USER', 'gilligan') # подобно os.environ[key]=value
>>> cenviron.getenv('USER') # программный код на C тоже видит изменения
'gilligan'
```

Как и прежде, `cenviron` является настоящим объектом модуля Python после импорта со всей информацией, обычно прикрепляемой к модулям, и в случае ошибок корректно возбуждает исключения и передает описание ошибок:

```
>>> dir(cenviron)
['__doc__', '__file__', '__name__', '__packge__', 'getenv', 'putenv']
>>> cenviron.__file__
'cenviron.dll'
>>> cenviron.__name__
'cenviron'

>>> cenviron.getenv
<built-in function getenv>
>>> cenviron
<module 'cenviron' from 'cenviron.dll'>

>>> cenviron.getenv('HOME')
'/home/mark'
>>> cenviron.getenv('NONESUCH')
SystemError: Error calling getenv
```

Ниже приводится пример задачи, которая решается с помощью этого модуля (но необходимо, чтобы вызовы `getenv` осуществлялись присоединенным кодом C, а не Python; перед запуском этого сеанса я изменил значение переменной `USER` в командной оболочке с помощью команды `export`):

```
.../PP4E/Integrate/Extend/Cenviron$ python
>>> import os
>>> os.environ['USER']                  # инициализируется из оболочки
'skipper'
>>> from cenviron import getenv, putenv # прямые вызовы функций C
>>> getenv('USER')
'skipper'
>>> putenv('USER', 'gilligan')          # изменится для C, но не для Python
>>> getenv('USER')
```

```
'gilligan'
>>> os.environ['USER']            # ой! - значения не загружаются заново
'skipper'
>>> os.getenv('USER')             # то же самое
'skipper'
```

Добавление классов-оберток в простые библиотеки

В настоящей версии модуль расширения на C экспортирует интерфейс, основанный на функциях, но его функции можно обернуть программным кодом Python, который придаст интерфейсу любой желаемый вид. Например, в примере 20.10 осуществляется доступ к функциям через словарь и реализуется интеграция с объектом os.environ – это гарантирует сохранение объекта в актуальном состоянии даже после внесения изменений вызовом функций нашего расширения на C.

Пример 20.10. PP4E\Integrate\Extend\Cenviron\envmap.py

```
import os
from cenviron import getenv, putenv    # получить методы модуля на C

class EnvMapping:                      # обернуть классом Python
    def __setitem__(self, key, value):
        os.environ[key] = value        # при записи: Env[key]=value
        putenv(key, value)             # записать в os.environ

    def __getitem__(self, key):
        value = getenv(key)            # при чтении: Env[key]
        os.environ[key] = value        # проверка целостности
        return value

Env = EnvMapping()                     # создать один экземпляр
```

Чтобы задействовать этот модуль, клиенты могут импортировать его объект Env и использовать синтаксис Env['var'] обращения к словарям для обращения к переменным окружения. Класс в примере 20.11 делает еще один шаг вперед и экспортирует функции как квалифицированные имена атрибутов вместо вызовов или ключей – он позволяет обращаться к переменным окружения, используя синтаксис Env.var доступа к атрибутам.

Пример 20.11. PP4E\Integrate\Extend\Cenviron\envattr.py

```
import os
from cenviron import getenv, putenv    # получить методы модуля на C

class EnvWrapper:                      # обернуть классом Python
    def __setattr__(self, name, value):
        os.environ[name] = value       # при записи: Env.name=value
        putenv(name, value)            # записать в os.environ

    def __getattr__(self, name):
```

```
        value = getenv(name)          # при чтении: Env.name
        os.environ[name] = value      # проверка целостности
        return value

Env = EnvWrapper()                    # создать один экземпляр
```

Ниже показано, как действуют наши обертки, опирающиеся на использование функций доступа к переменным окружения из нашего модуля расширения на C. Самое важное, на что следует обратить внимание, это то, что функции расширений можно обернуть самыми разными интерфейсами, реализовав обертки на языке Python в дополнение к расширениям на C:

```
>>> from envmap import Env
>>> Env['USER']
'skipper'
>>> Env['USER'] = 'professor'
>>> Env['USER']
'professor'
>>>
>>> from envattr import Env
>>> Env.USER
'professor'
>>> Env.USER = 'gilligan'
>>> Env.USER
'gilligan'
```

Обертывание функций окружения C с помощью SWIG

Модули расширения можно писать вручную, как мы это только что сделали, но это совершенно необязательно. Поскольку этот пример в действительности просто создает обертки для имеющихся функций в стандартных библиотеках C, весь файл *cenviron.c* с программным кодом C, представленным в примере 20.8, можно заменить простым входным файлом SWIG, который выглядит, как показано в примере 20.12:

Пример 20.12. PP4E\Integrate\Extend\Swig\Environ\environ.i

```
/*************************************************************************
* Файл Swig описания модуля, чтобы сгенерировать весь программный код обертки
* Python для функций getenv/putenv из библиотеки C: "swig -python environ.i".
*************************************************************************/

%module environ

extern char * getenv(const char *varname);
extern int putenv(char *assignment);
```

И все готово. Ну, почти – все же нужно еще пропустить этот файл через SWIG и скомпилировать полученный вывод. Как и прежде, просто добавьте в свой make-файл строки для вызова SWIG, скомпилируйте результат в разделяемый объект для динамического связывания и все бу-

дет готово. В примере 20.13 представлен make-файл для Cygwin, который решает эту задачу.

Пример 20.13. PP4E\Integrate\Extend\Swig\Environ\makefile.environ-swig

```
# компилирует расширение environ из программного кода, сгенерированного SWIG

PYLIB = /usr/local/bin
PYINC = /usr/local/include/python3.1
SWIG  = /cygdrive/c/temp/swigwin-2.0.0/swig

_environ.dll: environ_wrap.c
       gcc environ_wrap.c -g -I$(PYINC) -L$(PYLIB) -lpython3.1 -shared -o $@

environ_wrap.c: environ.i
        $(SWIG) -python environ.i

clean:
        rm -f *.o *.dll *.pyc core environ_wrap.c environ.py
```

При запуске с файлом *environ.i* утилита SWIG создаст два файла и два модуля – *environ.py* (интерфейсный модуль Python) и *environ_wrap.c* (файл модуля со связующим программным кодом, который будет скомпилирован в файл *_environ.dll* и будет импортироваться файлом *.py*). Так как функции, для которых здесь создаются обертки, находятся в стандартных присоединяемых библиотеках C, объединять с генерируемым программным кодом нечего – этот make-файл просто запускает SWIG и компилирует файл обертки в модуль расширения C, готовый к импортированию:

```
.../PP4E/Integrate/Extend/Swig/Environ$ make -f makefile.environ-swig
/cygdrive/c/temp/swigwin-2.0.0/swig -python environ.i
gcc environ_wrap.c -g -I/usr/local/include/python3.1 -L/usr/local/bin
-lpython3.1
    -shared -o _environ.dll
```

И вот теперь действительно все. Полученный модуль расширения на C связывается при импортировании и используется как прежде (за исключением того, что все сложности были решены с помощью SWIG):

```
.../PP4E/Integrate/Extend/Swig/Environ$ ls
_environ.dll environ.i environ.py environ_wrap.c makefile.environ-swig

.../PP4E/Integrate/Extend/Swig/Environ$ python
>>> import environ
>>> environ.getenv('USER')
'gilligan'
>>> environ.__name__, environ.__file__, environ
('environ', 'environ.py', <module 'environ' from 'environ.py'>)
>>> dir(environ)
[ ... '_environ', 'getenv', 'putenv' ... ]
```

Если внимательно рассмотреть последний листинг сеанса, можно заметить, что на этот раз я не вызывал функцию `putenv`. Как оказывается, тому есть веская причина: функция `putenv` в библиотеке C ожидает получить строку вида «USER=Gilligan», которая станет частью окружения. Для программного кода C это означает, что мы должны выделить новый блок памяти для передачи функции – для удовлетворения этого требования в примере 20.8 мы использовали функцию `malloc`. Однако нет простого и непосредственного способа обеспечить выделение памяти на стороне Python. В предыдущей версии Python было достаточно сохранить строку, передаваемую функции `putenv`, во временной переменной Python, но в Python 3.X и/или SWIG 2.0 этот прием больше не работает. Для исправления ситуации можно либо написать собственную функцию на C, либо использовать средства SWIG отображения типов, которые позволяют настраивать обработку передачи данных. В интересах экономии места мы оставим решение этой проблемы в качестве самостоятельного упражнения – подробности смотрите в документации SWIG.

Обертывание классов C++ с помощью SWIG

До сих пор в этой главе мы имели дело с модулями расширений на C – простыми библиотеками функций. Чтобы реализовать на C поддержку объектов, позволяющих создавать множество экземпляров, необходимо написать расширение типа, а не модуль. Подобно классам в языке Python типы C генерируют объекты, позволяющие создавать множество экземпляров, и могут переопределять (то есть предоставлять свою реализацию) операторы выражений Python и операции над типом. Типы C также могут поддерживать возможность создания подклассов, подобно классам Python, в значительной степени благодаря тому, что различия между типами и классами в Python 3.X были практически полностью стерты.

Увидеть, как выглядят типы C, вы можете в собственной библиотеке с исходными текстами Python – загляните в каталог *Objects*. Программный код с определением типа на C может оказаться очень длинным – в нем необходимо определить алгоритм создания экземпляров, именованные методы, реализацию операторов, тип итератора и так далее и связать все это воедино с помощью таблиц – но это в основном типовой программный код, который имеет похожую структуру для большинства типов.

Создавать новые типы на C можно вручную, и в некоторых случаях в этом есть определенный смысл. Но в целом в этом нет необходимости – благодаря тому, что SWIG умеет генерировать связующий программный код для *классов C++*, имеется возможность *автоматически* генерировать весь необходимый программный код расширения на C и класса обертки для интеграции таких объектов, просто обрабатывая соответствующее объявление класса с помощью SWIG. Прием оберты-

вания классов C++ обеспечивает возможность создания типов данных подобно приему создания типов на C, но он может оказаться существенно проще в реализации благодаря тому, что SWIG берет на себя решение всех проблем интеграции языков.

Вот как это происходит: на основе имеющегося определения класса C++ и специальных параметров командной строки SWIG генерирует следующие компоненты:

- Написанный на C++ модуль расширения Python с функциями доступа, взаимодействующими с методами и членами класса C++.
- Написанный на Python модуль с классом-оберткой (на языке SWIG он называется «теневым» классом), взаимодействующий с модулем функций доступа к классу C++.

Как и прежде, чтобы задействовать SWIG, нужно написать и отладить класс, как если бы он использовался исключительно в программном коде C++. Затем просто запустить SWIG в make-файле для сканирования объявления класса C++ и скомпилировать вывод. В итоге, импортируя теневые классы, в сценариях Python можно использовать классы C++, как если бы они были написаны на языке Python. Программы Python не только могут создавать и использовать экземпляры классов C++, но и выполнять их индивидуальную подгонку путем создания подклассов сгенерированного теневого класса.

Простое расширение с классом C++

Чтобы увидеть, как все это работает, нам нужен класс C++. Для иллюстрации напишем простой класс, который будет использоваться в сценариях Python. Чтобы понимать, о чем рассказывается в этом разделе, вам необходимо знать язык C++. SWIG поддерживает расширенные возможности языка C++ (включая шаблоны и перегрузку функций и операторов), но для иллюстрации я постарался сохранить этот пример максимально простым. В следующих файлах C++ определен класс `Number` с четырьмя методами (`add`, `sub`, `square` и `display`), переменной класса (`data`), конструктором и деструктором. В примере 20.14 представлен заголовочный файл.

Пример 20.14. PP4E\Integrate\Extend\Swig\Shadow\number.h

```
class Number
{
public:
    Number(int start);   // конструктор
    ~Number();           // деструктор
    void add(int value); // изменяет член data
    void sub(int value);
    int square();        // возвращает значение
    void display();      // выводит член data
    int data;
};
```

В примере 20.15 представлена реализация класса C++. При вызове большинства методов выводится сообщение трассировки операций класса. Обратите внимание, что для этого используется функция printf вместо типичного для C++ объекта cout. Когда-то таким способом решалась проблема смешивания вывода через cout в C++ и стандартный поток вывода в Python 2.X в Cygwin. Возможно в настоящее время в этом нет необходимости – из-за того, что система вывода в Python 3.X с включенной буферизацией может произвольно смешиваться с выводом из C++, в C++ обычно желательно принудительно выталкивать содержимое потока вывода (вызовом fflush(stdout) или cout<<flush), чтобы при выводе он не смешивался с текстом, в конце которого отсутствует символ перевода строки. Довольно странная, но вполне типичная особенность, когда смешиваются несопоставимые языковые системы.

Пример 20.15. PP4E\Integrate\Extend\Swig\Shadow\number.cxx

```
////////////////////////////////////////////////////////////////////
// реализация класса C++, который, возможно, будет использоваться
// из программного кода Python;
// предупреждение: обычно оба приема вывода, через cout и print,
// работают, но я столкнулся с проблемой при смешивании
// вывода C++/Python в Cygwin, которая была решена с помощью printf
////////////////////////////////////////////////////////////////////

#include "number.h"
#include "stdio.h"                // вместо #include "iostream.h"

Number::Number(int start) {
    data = start;                 // вывод из python поступает в stdout
    printf("Number: %d\n", data);  // или: cout << "Number: " << data << endl;
}

Number::~Number() {
    printf("~Number: %d\n", data);
}

void Number::add(int value) {
    data += value;
    printf("add %d\n", value);
}

void Number::sub(int value) {
    data -= value;
    printf("sub %d\n", value);
}

int Number::square() {
    return data * data;    // после вывода, fflush(stdout)
                           // или cout << flush
}
```

```
void Number::display() {
    printf("Number=%d\n", data);
}
```

Чтобы вы могли сравнить языки, ниже показано, как этот класс используется в программе C++. Программа в примере 20.16 создает объект Number, вызывает его методы, напрямую извлекает и устанавливает его атрибут данных (в C++ делается различие между «членами» и «методами», в то время как в Python те и другие обычно называют «атрибутами»).

Пример 20.16. PP4E\Integrate\Extend\Swig\Shadow\main.cxx

```
#include "iostream.h"
#include "number.h"

main()
{
    Number *num;
    int res, val;

    num = new Number(1);   // создать экземпляр класса C++
    num->add(4);           // вызвать его методы
    num->display();
    num->sub(2);
    num->display();

    res = num->square();   // метод возвращает значение
    cout << "square: " << res << endl;

    num->data = 99;        // установить значение члена data C++
    val = num->data;       // получить значение члена data C++
    cout << "data: " << val << endl;
    cout << "data+1: " << val + 1 << endl;

    num->display();
    cout << num << endl;   // вывести значение указателя на экземпляр
    delete num;            // вызвать деструктор
}
```

Чтобы скомпилировать этот программный код в Cygwin (или в Linux) в выполняемый файл, можно использовать компилятор C++ командной строки g++. Если вы используете иную систему, вам придется импровизировать – между разными компиляторами C++ слишком много различий, чтобы их здесь перечислять. Введите команду компиляции непосредственно или укажите цель cxxtest в представленном выше make-файле, находящемся в каталоге с этим примером, и затем запустите вновь созданную программу C++:

```
.../PP4E/Integrate/Extend/Swig/Shadow$ make -f makefile.number-swig cxxtest
g++ main.cxx number.cxx -Wno-deprecated
```

```
.../PP4E/Integrate/Extend/Swig/Shadow$ ./a.exe
Number: 1
add 4
Number=5
sub 2
Number=3
square: 9
data:   99
data+1: 100
Number=99
0xe502c0
~Number: 99
```

Обертывание классов C++ с помощью SWIG

Но достаточно о C++: вернемся к Python. Чтобы использовать класс C++ `Number` из предыдущего раздела в сценариях Python, нужно написать или сгенерировать слой связующей логики между двумя языками, как в предыдущих примерах расширений на C. Чтобы автоматически сгенерировать этот слой, напишем входной файл SWIG, как показано в примере 20.17.

Пример 20.17. PP4E\Integrate\Extend\Swig\Shadow\number.i

```
/*************************************************************
* Файл Swig с описанием модуля для обертывания класса C++.
* Генерация выполняется командой "swig -c++ -python number.i".
* Сгенерированный модуль C++ сохраняется в файле number_wrap.cxx;
* модуль 'number' ссылается на теневой класс в модуле number.py.
*************************************************************/

%module number

%{
#include "number.h"
%}

%include number.h
```

Этот интерфейс просто указывает SWIG, что он должен прочесть данные с сигнатурой типа класса C++ из включаемого заголовочного файла *number.h*. SWIG снова создает по объявлению класса два разных модуля Python:

`number_wrap.cxx`

Модуль расширения C++ с функциями доступа к классу.

`number.py`

Модуль с теневым классом Python, служащим оберткой для функций доступа.

Первый из них должен быть скомпилирован в двоичный файл библиотеки. Второй предназначен для импортирования и использования скомпилированного представления первого файла, и в конечном итоге он сам должен импортироваться в сценариях Python. Как и в примерах с простыми функциями, SWIG решает проблему интеграции за счет комбинирования программного кода на языках Python и C++.

После запуска SWIG make-файл для Cygwin, представленный в примере 20.18, объединит сгенерированный модуль *number_wrap.cxx* с программным кодом обертки C++ и реализацию класса и создаст файл *_number.dll* – динамически загружаемый модуль расширения, который к моменту импортирования должен находиться в пути поиска модулей Python, а также файл *number.py* (все файлы в данном случае находятся в текущем рабочем каталоге).

Как и прежде, при использовании современной версии SWIG имя модуля скомпилированного расширения на C должно начинаться с символа подчеркивания: *_number.dll*, следуя соглашениям Python, а не в других форматах, использовавшихся в более ранних версиях. Модуль *number.py* теневого класса импортирует *_number.dll*. Не забывайте использовать ключ `-c++` командной строки SWIG; ранее использовавшийся ключ `-shadow` более не требуется указывать, чтобы создать обертку для класса в дополнение к низкоуровневому интерфейсу функций модуля, так как он подразумевается по умолчанию.

Пример 20.18. PP4E\Integrate\Extend\Swig\Shadow\makefile.number-swig

```
##########################################################################
# Использование SWIG для интеграции класса C++ number.h с программами Python.
# Обновление: имя "_number.dll" имеет важное значение, потому что теневой
# класс импортирует модуль _number.
# Обновление: ключ "-shadow" командной строки swig более не требуется
# (подразумевается по умолчанию).
# Обновление: swig более не создает файл .doc и его не нужно удалять здесь
# (это довольно старая история).
##########################################################################

PYLIB = /usr/local/bin
PYINC = /usr/local/include/python3.1
SWIG  = /cygdrive/c/temp/swigwin-2.0.0/swig

all: _number.dll number.py

# обертка + действительный класс
_number.dll: number_wrap.o number.o
        g++ -shared number_wrap.o number.o -L$(PYLIB) -lpython3.1 -o $@

# генерирует модуль обертки класса
number_wrap.o: number_wrap.cxx number.h
        g++ number_wrap.cxx -c -g -I$(PYINC)
```

```
number_wrap.cxx: number.i
        $(SWIG) -c++ -python number.i

number.py: number.i
        $(SWIG) -c++ -python number.i

# программный код обертки класса C++
number.o: number.cxx number.h
        g++ number.cxx -c -g -Wno-deprecated

# тестовая программа не на Python
cxxtest:
        g++ main.cxx number.cxx -Wno-deprecated

clean:
        rm -f *.pyc *.o *.dll core a.exe
force:
        rm -f *.pyc *.o *.dll core a.exe number_wrap.cxx number.py
```

Как обычно, воспользуйтесь этим make-файлом, чтобы сгенерировать и откомпилировать необходимый связующий код в виде модуля расширения, который может импортироваться программами Python:

```
.../PP4E/Integrate/Extend/Swig/Shadow$ make -f makefile.number-swig
/cygdrive/c/temp/swigwin-2.0.0/swig -c++ -python number.i
g++ number_wrap.cxx -c -g -I/usr/local/include/python3.1
g++ number.cxx -c -g -Wno-deprecated
g++ -shared number_wrap.o number.o -L/usr/local/bin -lpython3.1 -o _number.dll

.../PP4E/Integrate/Extend/Swig/Shadow$ ls
_number.dll  makefile.number-swig  number.i   number_wrap.cxx
a.exe        number.cxx            number.o   number_wrap.o
main.cxx     number.h              number.py
```

Использование класса C++ в Python

После того как связующий программный код будет сгенерирован и откомпилирован, сценарии Python смогут обращаться к классу C++, как если бы он был написан на Python. Фактически импортируемый теневой класс в модуле *number.py*, который выполняется поверх модуля расширения, является сгенерированным программным кодом Python. В примере 20.19 повторяются тесты классов из файла *main.cxx*. Однако здесь класс C++ используется из языка программирования Python – это может показаться удивительным, но со стороны Python программный код выглядит вполне естественным.

Пример 20.19. PP4E\Integrate\Extend\Swig\Shadow\main.py

```
"""
пример использования класса C++ в Python (модуль c++ + теневой класс py)
этот сценарий выполняет те же самые тесты, что и файл main.cxx C++
"""
```

```
from number import Number # импортировать теневой класс для модуля C++

num = Number(1)         # создать объект класса C++ в Python
num.add(4)              # вызвать его методы из Python
num.display()           # num хранит указатель 'this' языка C++
num.sub(2)
num.display()

res = num.square()      # преобразует возвращаемое из C++ значение int
print('square: ', res)

num.data = 99        # установить член data C++, сгенерированный __setattr__
val = num.data       # получить член data C++, сгенерированный __getattr__
print('data: ', val) # вернет обычный объект Python целого числа
print('data+1: ', val + 1)

num.display()
print(num)              # вызовет repr в теневом классе
del num                 # автоматически вызовет деструктор C++
```

Так как класс C++ и его обертки автоматически загружаются при импорте модуля *number.py* теневого класса, этот сценарий выполняется, как всякий другой:

```
.../PP4E/Integrate/Extend/Swig/Shadow$ python main.py
Number: 1
add 4
Number=5
sub 2
Number=3
square: 9
data:   99
data+1: 100
Number=99
<number.Number; proxy of <Swig Object of type 'Number *' at 0x7ff4bb48> >
~Number: 99
```

Этот вывод производится в основном методами класса C++ и в значительной степени совпадает с результатами работы программы *main.cxx*, представленной в примере 20.16 (кроме вывода информации об экземпляре – теперь это экземпляр теневого класса Python).

Использование низкоуровневого модуля расширения

SWIG реализует интеграцию как комбинацию программного кода C++/Python, но при желании всегда можно использовать сгенерированные функции доступа в модуле, как показано в примере 20.20. Эта версия напрямую использует модуль расширения на C++ без применения теневого класса и демонстрирует, как теневой класс отображает вызовы программного кода C++.

Пример 20.20. PP4E\Integrate\Extend\Swig\Shadow\main_low.py

```
"""
выполняет те же тесты, что и main.cxx и main.py,
но использует низкоуровневый интерфейс функций доступа на C
"""

from _number import *  # модуль-обертка расширения c++

num = new_Number(1)
Number_add(num, 4)     # явная передача указателя 'this'
Number_display(num)    # использовать функции доступа в модуле на C
Number_sub(num, 2)
Number_display(num)
print(Number_square(num))

Number_data_set(num, 99)
print(Number_data_get(num))
Number_display(num)
print(num)
delete_Number(num)
```

Этот сценарий порождает по сути такой же вывод, как *main.py*; но так как сценарий был несколько упрощен, то экземпляр класса C++ в данном случае является более низкоуровневым объектом, чем теневой класс.

```
.../PP4E/Integrate/Extend/Swig/Shadow$ python main_low.py
Number: 1
add 4
Number=5
sub 2
Number=3
9
99
Number=99
_6025aa00_p_Number
~Number: 99
```

Наследование классов C++ в подклассах Python

Как видите, модули расширений вполне возможно использовать напрямую, но здесь не видно преимуществ перехода от использования теневого класса к использованию функций. При использовании теневого класса вы одновременно получаете основанный на объектах интерфейс к C++ и доступный для наследования объект Python. Так, модуль Python, представленный в примере 20.21, расширяет класс C++, добавляя в метод `add` вызов функции `print` и определяя новый метод `mul`. Так как теневой класс – это класс Python, все это действует естественным образом.

Пример 20.21. PP4E\Integrate\Extend\Swig\Shadow\main_subclass.py

```
"подкласс класса C++ в Python (подкласс сгенерированного теневого класса)"

from number import Number  # импортировать теневой класс

class MyNumber(Number):
    def add(self, other):  # расширить метод
        print('in Python add...')
        Number.add(self, other)
    def mul(self, other):  # добавить новый метод
        print('in Python mul...')
        self.data = self.data * other

num = MyNumber(1)          # те же тесты, что и в main.cxx, main.py
num.add(4)                 # использовать подкласс Python теневого класса
num.display()              # add() - специализированная в Python версия
num.sub(2)
num.display()
print(num.square())

num.data = 99
print(num.data)
num.display()

num.mul(2)                 # mul() - реализован на языке Python
num.display()
print(num)                 # repr из теневого суперкласса
del num
```

Теперь метод `add` выводит дополнительные сообщения, а метод `mul` автоматически изменяет член `data` класса C++ при присваивании `self.data` – программный код на языке Python расширяет программный код на языке C++:

```
.../PP4E/Integrate/Extend/Swig/Shadow$ python main_subclass.py
Number: 1
in Python add...
add 4
Number=5
sub 2
Number=3
9
99
Number=99
in Python mul...
Number=198
<__main__.MyNumber; proxy of <Swig Object of type 'Number *' at 0x7ff4baa0> >
~Number: 198
```

Иными словами, SWIG упрощает использование библиотек классов C++ в качестве базовых классов в сценариях Python. Среди всего прочего, эта возможность позволяет использовать существующие библиотеки классов C++ в сценариях Python и оптимизировать их, программируя отдельные фрагменты иерархии классов на C++, когда это необходимо. Практически то же самое можно делать, создавая типы на C, поскольку в настоящее время типы являются классами (и наоборот), но процедура обертывания классов C++ с помощью SWIG часто оказывается существенно проще.

Исследование оберток в интерактивном сеансе

Как обычно, класс C++ можно импортировать в интерактивной оболочке и немного поэкспериментировать с ним – этот прием позволяет не только продемонстрировать здесь некоторые важные особенности, но и протестировать обернутые классы C++ в интерактивном сеансе Python:

```
.../PP4E/Integrate/Extend/Swig/Shadow$ python
>>> import _number
>>> _number.__file__   # класс C++ плюс сгенерированный связующий модуль
'_number.dll'
>>> import number      # сгенерированный модуль Python теневого класса
>>> number.__file__
'number.py'

>>> x = number.Number(2)   # создать экземпляр класса C++ в Python
Number: 2
>>> y = number.Number(4)   # создать еще один объект C++
Number: 4
>>> x, y
(<number.Number; proxy of <Swig Object of type 'Number *' at 0x7ff4bcf8> >,
<number.Number; proxy of <Swig Object of type 'Number *' at 0x7ff4b998> >)

>>> x.display()      # вызвать метод C++ (как в C++: x->display())
Number=2
>>> x.add(y.data)    # извлечь значение члена data в C++, вызвать метод C++
add 4
>>> x.display()
Number=6

>>> y.data = x.data + y.data + 32  # присвоить значение члену data в C++
>>> y.display()                    # y хранит указатель C++ 'this'
Number=42

>>> y.square()                     # метод с возвращаемым значением
1764
>>> t = y.square()
>>> t, type(t)                     # в Python 3.X типы - это классы
(1764, <class 'int'>)
```

Естественно, в этом примере используется очень маленький класс C++, только чтобы подчеркнуть основные моменты, но даже на этом уровне прозрачность интеграции Python/C++, которую обеспечивает SWIG, восхищает. Программный код Python использует члены и методы класса C++, как если бы они были реализованы на языке Python. Кроме того, такая же прозрачность интеграции сохраняется и при переходе к использованию более реалистичных библиотек классов C++.

Так чего же мы добились? На данный момент ничего особенного, но если всерьез начать пользоваться SWIG, то самым большим недостатком может оказаться отсутствие на сегодняшний день в SWIG поддержки *всех* особенностей C++. Если в классах используются довольно сложные особенности C++ (а их достаточно много), то может потребоваться создать для SWIG упрощенные описания типов классов, а не просто запускать SWIG с исходными заголовочными файлами классов, поэтому вам обязательно следует ознакомиться с руководством по SWIG и с информацией на сайте проекта, чтобы получить больше подробностей по этой и другим темам.

Однако в обмен на такие компромиссы SWIG может полностью исключить необходимость вручную писать связующие слои для доступа к библиотекам C и C++ из сценариев Python. Если вам доводилось когда-либо писать такой программный код вручную, вам должно быть ясно, что это *очень* большой выигрыш.

Если же вы пойдете путем работы вручную, обратитесь к стандартным руководствам Python по расширению за дополнительными сведениями по вызовам API, используемым в этой главе, а также к дополнительным инструментам создания расширений, о которых мы не имеем возможности рассказывать в этой книге. Расширения на C могут лежать в широком диапазоне от коротких входных файлов SWIG до программного кода, крепко связанного с внутренним устройством интерпретатора Python. Практика показывает, что первые из них переносят разрушительное действие времени значительно лучше, чем вторые.

Другие инструменты создания расширений

В заключение темы создания расширений я должен упомянуть о существовании других инструментов, кроме SWIG, многие из которых имеют своих сторонников. В этом разделе будут представлены некоторые наиболее популярные инструменты в этой области. Как обычно, дополнительную информацию по этой и другим темам ищите в Сети. Подобно SWIG все перечисленные ниже инструменты начинали свою жизнь как сторонние программные компоненты, устанавливаемые отдельно. Однако в версии Python 2.5 расширение *ctypes* было включено в состав стандартной библиотеки.

SIP

SIP в сравнении со SWIG – как маленький глоточек в сравнении с большим глотком[1], представляет собой более легковесную альтернативу (фактически и свое название этот инструмент получил по этой же причине). Согласно информации на домашней странице этого инструмента SIP упрощает создание связующих модулей Python к библиотекам C и C++. Первоначально этот инструмент разрабатывался для создания PyQt – набора связующих модулей к библиотеке инструментов Qt, но его с успехом можно использовать для создания модулей к любым другим библиотекам C или C++. В состав SIP входят генератор программного кода и модуль поддержки Python.

Во многом подобно SWIG генератор программного кода обрабатывает комплект файлов спецификаций и генерирует программный код на языке C или C++, из которого потом путем компиляции создаются модули расширений. Модуль поддержки SIP для Python предоставляет функции, позволяющие автоматически генерировать программный код. В отличие от SWIG, система SIP изначально проектировалась исключительно для интеграции языков Python и C/C++. Система SWIG, в свою очередь, способна генерировать обертки для множества других языков сценариев и потому может рассматриваться, как более сложный проект.

ctypes

Система ctypes – это модуль интерфейса доступа к внешним функциям (Foreign Function Interface, FFI) для Python. Он позволяет сценариям Python обращаться к компилированным функциям в двоичных файлах библиотек непосредственно и динамически, для чего не требуется генерировать или писать интегрирующую обертку на языке C, с которыми мы познакомились в этой главе, а достаточно написать соответствующий программный код на самом языке Python. То есть связующий программный код для доступа к библиотеке пишется исключительно на языке Python, а не на C. Главное преимущество такого подхода состоит в том, что вам не требуется программировать на языке C или использовать систему сборки программ на языке C, чтобы получить доступ к функциям C из сценария Python. Один из недостатков заключается в возможной потере скорости на этапе маршрутизации вызовов, что, впрочем, зависит от применяемых способов измерений.

Согласно документации система ctypes позволяет сценариям Python вызывать функции, экспортируемые динамическими (DLL) и разделяемыми библиотеками, и содержит инструменты, дающие возможность создавать, получать и изменять данные сложных типов C на языке Python. Кроме того, предусматривается возможность создания

[1] Здесь игра слов. SWIG можно перевести как «большой глоток», а SIP – «как маленький глоток». – *Прим. перев.*

на языке Python функций обратного вызова, которые будут вызываться из программного кода C, а экспериментальный генератор программного кода в `ctypes` позволяет автоматически создавать обертки для библиотек из заголовочных файлов C. Система `ctypes` работает в Windows, Mac OS X, Linux, Solaris, FreeBSD и OpenBSD. Она может использоваться и в других системах, поддерживающих пакет `libffi`. Версия `ctypes` для Windows включает пакет `ctypes.com`, который дает возможность программному коду Python вызывать и создавать собственные интерфейсы COM. Более подробную информацию о функциональных возможностях системы ctypes, включенной в состав стандартной библиотеки, ищите в руководствах по библиотеке Python.

Boost.Python

Система Boost.Python – это библиотека C++, которая обеспечивает прозрачность взаимодействий между языками программирования C++ и Python за счет использования IDL-подобной модели. Используя эту систему, разработчики обычно пишут небольшой объем обертывающего программного кода C++ для создания разделяемой библиотеки, используемой в сценариях Python. Система Boost.Python обслуживает ссылки, обратные вызовы, преобразование типов и решает задачи управления памятью. Поскольку она предназначена для обертывания интерфейсов C++, отпадает необходимость изменять обертываемый программный код C++. Как и другие инструменты, эту систему удобно использовать для обертывания существующих библиотек, а также для разработки совершенно новых расширений.

Реализация интерфейсов для больших библиотек может оказаться более сложным делом, чем использование генераторов программного кода SWIG и SIP, но это намного проще, чем писать обертки вручную, и такой подход способен обеспечить более полный контроль, чем полностью автоматизированные инструменты. Кроме того, Py++ и более старая система Pyste предоставляют в распоряжение Boost.Python генераторы программного кода, с помощью которых пользователи могут определять, какие классы и функции должны экспортироваться, используя простой файл с описанием интерфейса. Для анализа заголовочных файлов и извлечения информации, необходимой для создания программного кода C++, обе они используют парсер GCC-XML.

Cython (и Pyrex)

Cython, дальнейшее развитие системы Pyrex, является языком, специально предназначенным для создания модулей расширений Python. Он позволяет писать файлы, в которых допускается произвольно смешивать программный код Python и типы данных C и компилировать их в расширения C для Python. В принципе, при использовании этого языка разработчикам вообще не приходится иметь дело

с Python/C API, потому что Cython автоматически решает такие задачи, как проверка ошибок и подсчет ссылок.

С технической точки зрения, Cython – это самостоятельный язык программирования, *похожий* на Python, позволяющий использовать объявления типов данных и вызовы функций языка C. При этом практически любой программный код на языке Python также является допустимым программным кодом на языке Cython. Компилятор Cython преобразует программный код Python в программный код C, который производит вызовы Python/C API. В этом отношении Cython напоминает более старый проект преобразования Python2C. Позволяя комбинировать программный код Python и C, Cython предлагает иной подход, чем другие системы создания интегрирующего программного кода.

CXX, weave и другие

Систему *CXX* грубо можно назвать C++-версией обычного C API языка Python, которая решает задачи подсчета ссылок, передачи исключений, проверок типов и управления памятью, необходимых в расширениях C++. Кроме того, система CXX позволяет сосредоточиться на прикладных аспектах программного кода. Дополнительно система CXX экспортирует части стандартной библиотеки шаблонов контейнеров C++, совместимых с последовательностями Python.

Пакет *weave* позволяет включать программный код C/C++ в программный код Python. Он является частью пакета SciPy (*http://www.scipy.org*), но его можно устанавливать отдельно, как самостоятельную систему. На странице *http://www.python.org* вы найдете ссылки на дополнительные проекты в этой области, для упоминания которых здесь у нас недостаточно места.

Другие языки: Java, C#, FORTRAN, Objective-C и прочие

Несмотря на то, что в этой главе мы сосредоточились на языках C и C++, тем не менее, в мире открытых исходных текстов можно найти прямую поддержку возможности смешивания Python с другими языками программирования. Сюда входят языки программирования, компилируемые в двоичное представление, такие как C, а также другие, некомпилируемые языки.

Например, предоставляя полноценные компиляторы в байт-код, системы Jython и IronPython позволяют писать программы Python, практически прозрачно взаимодействующие с компонентами на языках Java и C#/.NET. Альтернативные проекты, JPype и Python для .NET, также поддерживают интеграцию Java и C#/.NET с обычным программным кодом на CPython (стандартная реализация Python на языке C) без привлечения альтернативных компиляторов в байт-код.

Кроме того, системы f2py и PyFort предоставляют интеграцию с программным кодом на языке FORTRAN. Имеются также другие инструменты, обеспечивающие интеграцию с такими языками, как Del-

phi и Objective-C. В их число входит проект PyObjC, цель которого – обеспечить интеграцию Python и Objective-C. Он поддерживает возможность разработки на языке Python приложений на основе технологии Cocoa GUI в Mac OS X.

Информацию об инструментах интеграции с другими языками программирования ищите в Сети. Посмотрите также страницу wiki на сайте *http://www.python.org*, где перечислено большое количество других языков программирования, для которых имеется поддержка интеграции, включая Prolog, Lisp, TCL и другие.

Поскольку многие из этих систем поддерживают двунаправленную передачу управления, необходимую для реализации обеих моделей, расширения и встраивания, мы еще вернемся к ним в конце этой главы, когда будем обсуждать вопросы интеграции в целом. Однако прежде нам необходимо развернуться на 180 градусов и исследовать другую сторону интеграции Python/C: *встраивание*.

Соединение Python и C++

Стандартная реализация Python в данное время написана на языке C, поэтому к интерпретатору Python относятся все обычные правила соединения программ C с программами C++. Поэтому не требуется реализация никаких специальных механизмов для Python, но нужно сделать несколько замечаний.

При *встраивании* Python в программу C++ не нужно придерживаться каких-то отдельных правил. Просто скомпонуйте библиотеку Python и вызывайте ее функции из C++. Заголовочные файлы Python автоматически заключаются в объявления `extern "C" {...}`, чтобы подавить корректировку имен C++. Поэтому библиотека на языке Python выглядит для C++, как любой другой компонент C – нет необходимости перекомпиляции самого Python компилятором C++.

При *расширении* Python с помощью компонентов C++ заголовочные файлы Python остаются дружественными к C++, поэтому вызовы Python API в расширениях на C++ действуют, как любые другие вызовы из C++ в C. Но необходимо следить, чтобы в программном коде расширения, видимом Python, использовались объявления `extern "C"`, благодаря чему их сможет вызывать программный код C, на котором написан Python. Например, для создания оболочки класса C++ SWIG генерирует модуль расширения C++, в котором функция инициализации объявлена именно таким способом.

Встраивание Python в C: обзор

До сих пор в этой главе мы рассматривали лишь одну половину интеграции Python/C: вызов функций C из Python. Этот способ интеграции является, пожалуй, наиболее широко используемым – он позволяет программистам ускорять выполнение программ, переводя их на C, и использовать внешние библиотеки, создавая для них обертки в виде модулей расширений и типов на C. Но столь же полезным может оказаться и обратное: вызов Python из C. Путем реализации отдельных компонентов приложения на встраиваемом программном коде Python мы даем возможность изменять их на месте, не поставляя заказчику весь программный код системы.

В данном разделе рассказывается об этой второй стороне интеграции Python/C. Здесь говорится об интерфейсах C к Python, позволяющих программам, написанным на C-совместимых языках, выполнять программный код на языке Python. В этом режиме Python выступает в качестве встроенного управляющего языка (или, как иногда говорят, «макроязыка»). Хотя встраивание представляется здесь, по большей части, изолированно, нужно помнить, что лучше всего рассматривать поддержку интеграции в Python как единое целое. Структура системы обычно определяет соответствующий подход к интеграции: расширения C, вызовы встроенных функций или то и другое вместе. В завершение этой главы мы обсудим нескольких крупных платформ интеграции, таких как Jython и IronPython, которые предоставляют более широкие возможности интеграции компонентов.

Обзор API встраивания в C

Первое, что следует отметить в API встроенных вызовов Python, это меньшую его структурированность, чем у интерфейсов расширения. Для встраивания Python в C может потребоваться более творческий подход, чем при расширении: программист должен реализовать интеграцию с Python, выбирая из всей совокупности средств C, а не писать программный код, имеющий типовую структуру. Положительной стороной такой свободной структуры является возможность объединять в программах встроенные вызовы и стратегии, создавая произвольные архитектуры интеграции.

Отсутствие строгой модели встраивания в значительной мере является результатом менее четко обозначенных целей. При *расширении* Python есть четкое разделение ответственности между Python и C и ясная структура интеграции. Модули и типы C должны соответствовать модели модулей/типов Python путем соблюдения стандартных структур расширений. В результате интеграция оказывается незаметной для клиентов Python: расширения C выглядят, как объекты Python, и выполняют большую часть работы. При этом имеются дополнительные инструменты, такие как SWIG, обеспечивающие автоматизацию интеграции.

Но при *встраивании* Python структура не так очевидна – так как внешним уровнем является C, не совсем ясно, какой модели должен придерживаться встроенный код Python. В C может потребоваться выполнять загружаемые из модулей объекты, загружаемые из файлов или выделенные в документах строки и так далее. Вместо того чтобы решать, чего можно и чего нельзя делать в C, Python предоставляет набор общих инструментов интерфейса встраивания, применяемых и организуемых согласно целям встраивания.

Большинство этих инструментов соответствует средствам, доступным программам Python. В табл. 20.1 перечислены некоторые наиболее часто встречающиеся вызовы API, используемые для встраивания, и их эквиваленты в Python. В целом, если можно установить, как решить задачи встраивания с помощью чистого программного кода Python, то, вероятно, найдутся средства C API, позволяющие достичь таких же результатов.

Таблица 20.1. Часто используемые функции API

Вызов C API	**Эквивалент Python**
PyImport_ImportModule	import module, __import__
PyImport_GetModuleDict	sys.modules
PyModule_GetDict	module.__dict__
PyDict_GetItemString	dict[key]
PyDict_SetItemString	dict[key]=val
PyDict_New	dict = {}
PyObject_GetAttrString	getattr(obj, attr)
PyObject_SetAttrString	setattr(obj, attr, val)
PyObject_CallObject	funcobj(*argstuple)
PyEval_CallObject	funcobj(*argstuple)
PyRun_String	eval(exprstr), exec(stmtstr)
PyRun_File	exec(open(filename().read())

Так как встраивание основывается на выборе вызова API, знакомство с Python C API совершенно необходимо для решения задач встраивания. В этой главе представлен ряд характерных примеров встраивания и обсуждаются стандартные вызовы API, но нет полного списка всех имеющихся в нем инструментов. Разобравшись с приведенными примерами, возможно, вы обратитесь к руководствам Python по интеграции за дополнительными сведениями о том, какие вызовы есть в этой области. Как уже упоминалось выше, в Python есть два стандартных руководства для программистов, занимающихся интеграцией с C/C++:

«Extending and Embedding», учебник по интеграции, и «Python/C API», справочник по библиотеке времени выполнения Python.

Самые свежие версии этих руководств можно найти на сайте *http://www.python.org*, и, возможно, они уже были установлены на вашем компьютере вместе с самим Python. Помимо данной главы эти два руководства послужат лучшим источником свежей и полной информации по средствам Python API.

Что представляет собой встроенный код?

Прежде чем переходить к деталям, разберемся с базовыми понятиями встраивания. Когда в этой книге говорится о «встроенном» программном коде Python, имеется в виду любая программная структура на языке Python, которая может быть выполнена из C посредством прямого вызова функции. Вообще говоря, встроенный программный код Python может быть нескольких видов:

Строки программного кода

Программы на C могут представлять программы Python в виде символьных строк и выполнять их как выражения или инструкции (подобно встроенным функциям `eval` и `exec` языка Python).

Вызываемые объекты

Программы на C могут загружать или обращаться к вызываемым объектам Python, таким как функции, методы и классы, и вызывать их со списками аргументов (например, используя синтаксис Python `func(*pargs, *kargs)`).

Файлы с программным кодом

Программы на C могут выполнять целые программные файлы Python, импортируя модули и выполняя файлы сценариев через API или обобщенные системные вызовы (например, `popen`).

Физически в программу C обычно встраивается двоичная библиотека Python. Программный код Python, выполняемый из C, может поступать из разнообразных источников:

- Строки программного кода могут загружаться из файлов, могут быть получены в результате ввода пользователем, из баз данных и файлов-хранилищ, выделены из файлов HTML или XML, могут читаться через сокеты, строиться или находиться непосредственно в программах C, передаваться функциям расширения C из программного кода регистрации Python и так далее.
- Вызываемые объекты могут загружаться из модулей Python, возвращаться другими вызовами Python API, передаваться функциям расширения C из программного кода регистрации Python и так далее.
- Файлы с программным кодом просто существуют в виде файлов, модулей и выполняемых сценариев.

Регистрация является приемом, часто используемым при организации обратных вызовов, который будет более подробно изучен далее в этой главе. Но что касается строк программного кода, возможных источников существует столько, сколько их есть для строк символов C. Например, программы на C могут динамически строить произвольный программный код на языке Python, создавая и выполняя строки.

Наконец, после получения некоторого программного кода Python, который должен быть выполнен, необходим какой-то способ связи с ним: программный код Python может потребовать передачи входных данных из слоя C и может создать вывод для передачи результатов обратно в C. На самом деле встраивание в целом представляет интерес, когда у встроенного программного кода имеется доступ к содержащему его слою C. Обычно средство связи определяется видом встроенного программного кода:

- Строки программного кода, являющиеся выражениями Python, возвращают значение выражения в виде выходных данных. Кроме того, как входные, так и выходные данные могут иметь вид глобальных переменных в том пространстве имен, в котором выполняется строка программного кода – C может устанавливать значения переменных, служащих входными данными, выполнять программный код Python и получать переменные с результатами его выполнения. Входные и выходные данные можно также передавать с помощью *вызовов функций*, экспортируемых расширениями на C, – программный код Python может с помощью модулей или типов C получать или устанавливать переменные в охватывающем слое C. Схемы связи часто являются комбинированными. Например, программный код C может заранее назначать глобальные имена объектам, экспортирующим информацию о состоянии и интерфейсные функции во встроенный программный код Python.[1]
- Вызываемые объекты могут получать входные данные в виде аргументов функции и возвращать результаты в виде возвращаемых значений. Переданные изменяемые аргументы (например, списки, словари, экземпляры классов) можно использовать во встроенном коде одновременно для ввода и вывода – сделанные в Python изменения

[1] Если нужен конкретный пример, вернитесь к обсуждению языков шаблонов в части, посвященной разработке сценариев для Интернета. В таких системах программный код Python обычно встраивается в файл HTML веб-страницы, в пространстве имен глобальным переменным присваиваются объекты, предоставляющие доступ к окружению веб-броузера, и программный код Python выполняется в пространстве имен, где осуществлено назначение объектам. Мне приходилось работать над проектом, в котором делалось нечто схожее, но программный код Python был встроен в документы XML, а объекты, присваиваемые глобальным переменным в пространстве имен программного кода, представляли собой виджеты графического интерфейса. По сути это был обычный программный код Python, встроенный и выполняемый программным кодом C.

сохраняются в объектах, которыми владеет C. Для связи с C объекты могут также использовать технику глобальных переменных и интерфейса расширений C, описанную для строк.

- Файлы программного кода по большей части могут использовать для связи такую же технику, как строки кода. При выполнении в качестве отдельных программ файлы могут также использовать приемы взаимодействий между процессами (IPC).

Естественно, все виды встроенного программного кода могут обмениваться данными с C, используя общие средства системного уровня: файлы, сокеты, каналы и другие. Однако обычно эти способы действуют медленнее и менее непосредственно. В данном случае нас по-прежнему интересует интеграция, основанная на вызовах функций.

Основные приемы встраивания

Как можно заключить из предшествующего обзора, встраивание предоставляет большую гибкость. Чтобы проиллюстрировать стандартные приемы встраивания в действии, в этом разделе представлен ряд коротких программ на языке C, которые тем или иным способом выполняют программный код Python. Большинство этих примеров используют простой файл модуля Python, представленный в примере 20.22.

Пример 20.22. PP4E\Integrate\Embed\Basics\usermod.py

```
"""
######################################################################
Программы на C будут выполнять программный код в этом модуле Python
в режиме встраивания. Этот файл может изменяться без необходимости
изменять слой C. Это обычный, стандартный программный код Python
(преобразования выполняются программой на C). Должен находиться в пути
поиска модулей Python. Программы на C могут также выполнять программный
код модулей из стандартной библиотеки, таких как string.
######################################################################
"""

message = 'The meaning of life...'

def transform(input):
    input = input.replace('life', 'Python')
    return input.upper()
```

Если вы хотя бы минимально знакомы с языком Python, то заметите, что этот файл определяет строку и функцию. Функция возвращает переданную ей строку после выполнения замены в строке и перевода всех символов в ней в верхний регистр. Из Python пользоваться модулем просто:

```
.../PP4E/Integrate/Embed/Basics$ python
>>> import usermod                    # импортировать модуль
```

```
>>> usermod.message                    # извлечь строку
'The meaning of life...'
>>> usermod.transform(usermod.message) # вызвать функцию
'THE MEANING OF PYTHON...'
```

При надлежащем использовании API ненамного сложнее использовать этот модуль в C.

Выполнение простых строк программного кода

Проще всего, пожалуй, запустить программный код Python из C можно с помощью функции API `PyRun_SimpleString`. Обращаясь к ней, программы C могут выполнять программы Python, представленные в виде массивов символьных строк C. Эта функция весьма ограничена: весь программный код выполняется в одном и том же пространстве имен (модуль `__main__`), строки программного кода должны быть инструкциями Python (не выражениями), отсутствует простой способ обмена входными и выходными данными с выполняемым программным кодом Python.

Тем не менее, это простой способ для начала. Кроме того, с добавлением импортируемого модуля расширения C, который может задействоваться встраиваемым программным кодом Python для обеспечения связи с объемлющим слоем C, такой прием может с успехом обеспечивать множественные цели встраивания. Для демонстрации основ в примере 20.23 приводится программа C, которая выполняет программный код Python и получает те же результаты, что были получены в интерактивном сеансе, приведенном в предыдущем разделе.

Пример 20.23. PP4E\Integrate\Embed\Basics\embed-simple.c

```
/*********************************************************************
* простые строки программного кода: C действует как интерактивная
* оболочка, код выполняется в модуле __main__, вывод не посылается в C;
*********************************************************************/

#include <Python.h>            /* определение стандартного API */

main() {
    printf("embed-simple\n");
    Py_Initialize();
    PyRun_SimpleString("import usermod");         /* загрузить файл .py */
    PyRun_SimpleString("print(usermod.message)"); /* в пути поиска Python */
    PyRun_SimpleString("x = usermod.message"); /* компилировать и выполнить */
    PyRun_SimpleString("print(usermod.transform(x))");
    Py_Finalize();
}
```

Во-первых, нужно отметить, что при встраивании Python программы C всегда вызывают функцию `Py_Initialize`, чтобы инициализировать подключаемые библиотеки Python, прежде чем использовать какие-либо

другие функции API, и обычно вызывают функцию `Py_Finalize` для завершения интерпретатора.

Остальная часть программного кода проста – C передает интерпретатору Python готовые строки, примерно совпадающие с теми, что вводились в интерактивной оболочке. В действительности, можно было бы объединить все строки с программным кодом Python в одну строку через символ `\n` и передать интерпретатору в виде одной большой строки. Внутренне `PyRun_SimpleString` вызывает компилятор и интерпретатор Python для выполнения переданных из C строк. Как обычно, компилятор Python всегда есть в системах, где установлен Python.

Компиляция и выполнение

Чтобы создать самостоятельно выполняемую программу из этого файла исходного программного кода C, необходимо скомпоновать результат его компиляции с файлом библиотеки Python. В данной главе под «библиотекой» обычно подразумевается двоичный файл библиотеки, создаваемый при компиляции Python, а не стандартная библиотека Python.

На сегодняшний день все, что касается Python и может потребоваться в C, компилируется в один библиотечный файл при сборке интерпретатора (например, *libpython3.1.dll* в Cygwin). Функция `main` программы поступает из программного кода C, а в зависимости от расширений, установленных для Python, может потребоваться компоновка внешних библиотек, к которым обращается библиотека Python.

В предположении, что никаких дополнительных библиотек расширений не требуется, в примере 20.24 представлен минимальный make-файл для Cygwin в Windows, с помощью которого собирается программа на C из примера 20.23. Как всегда, конкретные особенности make-файла зависят от платформы, поэтому ищите дополнительные указания в руководствах по Python. Данный make-файл использует путь подключаемых файлов Python при поиске *Python.h* на этапе компиляции и добавляет файл библиотеки Python на конечном этапе компоновки, чтобы сделать возможными вызовы API в программе на C.

Пример 20.24. PP4E\Integrate\Embed\Basics\makefile.1

```
# make-файл для Cygwin, создающий выполняемую программу на C
# со встроенным Python, в предположении, что не требуется
# компоновка с библиотеками внешних модулей;
# использует заголовочные файлы Python, подключает файл библиотеки Python;
# те и другие могут находиться в других каталогах (например, /usr);

PYLIB = /usr/local/bin
PYINC = /usr/local/include/python3.1

embed-simple: embed-simple.o
        gcc embed-simple.o -L$(PYLIB) -lpython3.1 -g -o embed-simple
```

```
embed-simple.o: embed-simple.c
        gcc embed-simple.c -c -g -I$(PYINC)
```

Чтобы собрать программу с помощью этого файла, запустите утилиту make с этим файлом, как обычно (как и прежде, не забывайте, что отступы в make-файлах должны оформляться с помощью символов табуляции):

```
.../PP4E/Integrate/Embed/Basics$ make -f makefile.1
gcc embed-simple.c -c -g -I/usr/local/include/python3.1
gcc embed-simple.o -L/usr/local/bin -lpython3.1 -g -o embed-simple
```

На практике все может оказаться не так просто, и понадобится терпение, чтобы добиться результата. В действительности, для сборки всех программ из этого раздела в Cygwin я использовал make-файл из примера 20.25.

Пример 20.25. PP4E\Integrate\Embed\Basics\makefile.basics

```
# make-файл для сборки сразу всех 5 примеров встраивания в Cygwin

PYLIB = /usr/local/bin
PYINC = /usr/local/include/python3.1

BASICS = embed-simple.exe \
         embed-string.exe \
         embed-object.exe \
         embed-dict.exe \
         embed-bytecode.exe

all: $(BASICS)

embed%.exe: embed%.o
        gcc embed$*.o -L$(PYLIB) -lpython3.1 -g -o $@

embed%.o: embed%.c
        gcc embed$*.c -c -g -I$(PYINC)

clean:
        rm -f *.o *.pyc $(BASICS) core
```

На некоторых платформах может потребоваться подключить другие библиотеки, если файл используемой библиотеки Python был собран с внешними зависимостями. Фактически ваша библиотека Python может потребовать подключения значительно большего числа внешних библиотек, и, честно говоря, проследить все зависимости компоновщика может оказаться не так-то просто. Необходимые библиотеки могут зависеть от платформы и установки Python, поэтому я мало чем могу помочь для облегчения этого процесса (в конце концов, это C). Здесь действуют стандартные правила разработки на языке C.

Одно замечание: если вы будете много заниматься встраиванием и столкнетесь с проблемами внешних зависимостей, на некоторых платфор-

мах может потребоваться скомпилировать Python на своей машине из исходных текстов, *отключив* все ненужные расширения в файлах с настройками (подробности смотрите в пакете с исходными текстами Python). В результате будет создана библиотека Python с минимумом внешних зависимостей, компоновать программы с которой значительно легче.

Получив рабочий make-файл, запускайте его для сборки программ C с подключенными библиотеками Python:

```
.../PP4E/Integrate/Embed/Basics$ make -f makefile.basics clean
rm -f *.o *.pyc embed-simple.exe embed-string.exe embed-object.exe
embed-dict.exe embed-bytecode.exe core

.../PP4E/Integrate/Embed/Basics$ make -f makefile.basics
gcc embed-simple.c -c -g -I/usr/local/include/python3.1
gcc embed-simple.o -L/usr/local/bin -lpython3.1 -g -o embed-simple.exe
gcc embed-string.c -c -g -I/usr/local/include/python3.1
gcc embed-string.o -L/usr/local/bin -lpython3.1 -g -o embed-string.exe
gcc embed-object.c -c -g -I/usr/local/include/python3.1
gcc embed-object.o -L/usr/local/bin -lpython3.1 -g -o embed-object.exe
gcc embed-dict.c -c -g -I/usr/local/include/python3.1
gcc embed-dict.o -L/usr/local/bin -lpython3.1 -g -o embed-dict.exe
gcc embed-bytecode.c -c -g -I/usr/local/include/python3.1
gcc embed-bytecode.o -L/usr/local/bin -lpython3.1 -g -o embed-bytecode.exe
rm embed-dict.o embed-object.o embed-simple.o embed-bytecode.o embed-string.o
```

После сборки с любым make-файлом вы можете запустить получившуюся программу на C, как обычно:

```
.../PP4E/Integrate/Embed/Basics$ ./embed-simple
embed-simple
The meaning of life...
THE MEANING OF PYTHON...
```

Этот вывод производят в основном инструкциями, вызывающими функцию Python `print`, которые C посылает присоединенной библиотеке Python. Это напоминает использование программистом интерактивной оболочки Python.

Разумеется, строки с программным кодом на языке Python, выполняемые программой на C, вероятно, не стоит жестко определять в программе C, как показано в этом примере. Вместо этого их можно загружать из текстового файла или получать из графического интерфейса, извлекать из файлов HTML или XML, получать из базы данных или через сокеты и так далее. При использовании таких внешних источников строки с программным кодом Python, выполняемые из C, можно произвольно изменять без необходимости перекомпилировать выполняющую их программу на C. Они могут изменяться на сайте или конечными пользователями системы. Чтобы использовать строки программного кода с максимальной пользой, мы должны перейти к более гибким инструментам API.

Практические особенности: В Python 3.1 и Cygwin для Windows мне пришлось сначала включить в переменную окружения `PYTHONPATH` текущий рабочий каталог, чтобы обеспечить работоспособность примеров встраивания, с помощью команды `export PYTHONPATH=.`. Кроме того, мне потребовалось использовать команду `./embed-simple` для запуска программы, потому что каталог `.` также изначально отсутствовал в переменной оболочки `PATH` и он не записывается в нее при установке Cygwin.

В вашем случае ситуация может быть иной, но если вы столкнетесь с проблемами, попробуйте выполнить встроенные команды Python `import sys` и `print sys.path` из C, чтобы посмотреть, как в действительности выглядит путь поиска модулей Python, а за дополнительной информацией о настройке пути поиска для встраиваемых приложений обращайтесь к руководству «Python/C API».

Выполнение строк программного кода с использованием результатов и пространств имен

В примере 20.26 используются следующие вызовы API для выполнения строк программного кода, возвращающих значения выражений в C:

`Py_Initialize`

Инициализирует присоединенные библиотеки Python, как и раньше.

`PyImport_ImportModule`

Импортирует модуль Python и возвращает указатель на него.

`PyModule_GetDict`

Загружает объект словаря атрибутов модуля.

`PyRun_String`

Выполняет строку программного кода в явно указанных пространствах имен.

`PyObject_SetAttrString`

Присваивает значение атрибуту объекта, имя которого определяется аргументом `namestring`.

`PyArg_Parse`

Преобразует объект возвращаемого значения Python в формат C.

С помощью функции, выполняющей импорт, загружается пространство имен модуля `usermod`, представленного в примере 20.22, чтобы строки программного кода в нем могли выполняться непосредственно и обращаться к именам, определенным в этом модуле, не квалифицируя их. Функция `PyImport_ImportModule` напоминает инструкцию Python `import`, но объект импортированного модуля возвращается в C, а не присваивается переменной Python. Поэтому она скорее похожа на встроенную функцию Python `__import__`.

Однако фактическим выполнением программного кода здесь занимается функция PyRun_String. Она принимает строку программного кода, флаг режима парсера и указатели на объекты словарей, которые будут играть роль глобального и локального пространств имен при выполнении строки программного кода. Флаг режима может иметь значение Py_eval_input для выполнения выражения или Py_file_input для выполнения инструкции. При выполнении выражения эта функция возвращает результат вычисления выражения (в виде указателя на объект PyObject*). Два аргумента с указателями на словари пространств имен позволяют различать глобальную и локальную области видимости, но обычно они задают один и тот же словарь, благодаря чему программный код выполняется в одном пространстве имен.

Пример 20.26. PP4E\Integrate\Embed\Basics\embed-string.c

```
/* строки программного кода с результатами и пространствами имен */

#include <Python.h>

main() {
    char *cstr;
    PyObject *pstr, *pmod, *pdict;
    printf("embed-string\n");
    Py_Initialize();

    /* получить usermod.message */
    pmod = PyImport_ImportModule("usermod");
    pdict = PyModule_GetDict(pmod);
    pstr = PyRun_String("message", Py_eval_input, pdict, pdict);

    /* преобразовать в C */
    PyArg_Parse(pstr, "s", &cstr);
    printf("%s\n", cstr);

    /* присвоить usermod.X */
    PyObject_SetAttrString(pmod, "X", pstr);

    /* вывести usermod.transform(X) */
     (void) PyRun_String("print(transform(X))", Py_file_input, pdict, pdict);
    Py_DECREF(pmod);
    Py_DECREF(pstr);
    Py_Finalize();
}
```

После компиляции и запуска этот файл выводит те же результаты, что и его предшественник:

```
.../PP4E/Integrate/Embed/Basics$ ./embed-string
embed-string
The meaning of life...
THE MEANING OF PYTHON...
```

Но здесь выполняются совершенно иные действия. На этот раз C загружает, преобразует и выводит значение атрибута `message` модуля Python путем непосредственного выполнения выражения и присваивает значение глобальной переменной (`X`) в пространстве имен модуля, которая играет роль входных данных для функции Python `print`.

Благодаря тому, что функция выполнения строки в этой версии позволяет задавать пространства имен, можно лучше расчленить встроенный программный код, выполняемый системой, – для разных групп можно определять отдельные пространства имен, чтобы избежать изменения переменных в других группах. А так как эта функция возвращает результат, облегчается связь со встроенным программным кодом; результаты выражений являются выходными данными, а присваивания глобальным переменным в пространстве имен, в котором выполняется программный код, позволяют передавать входные данные.

Прежде чем двинуться дальше, я должен остановиться на трех проблемах, возникающих здесь. Во-первых, эта программа также уменьшает счетчики ссылок для объектов, передаваемых ей из Python, с помощью функции `Py_DECREF`, описанной в руководствах по Python/C API. Строго говоря, вызывать эти функции здесь необязательно (память объектов все равно освобождается при выходе из программ), но они демонстрируют, как интерфейсы встраивания должны управлять счетчиками ссылок, когда Python передает владение ими в C. Если бы, скажем, это была функция, вызываемая из более крупной системы, обычно потребовалось бы уменьшить счетчик, чтобы позволить Python освободить память объектов.

Во-вторых, в настоящей программе следует сразу проверять значения, возвращаемые *всеми* функциями API, чтобы обнаруживать ошибки (например, неудачу при импорте). В данном примере проверка ошибок была опущена для упрощения программного кода, но ее обязательно следует выполнять, чтобы повысить надежность программ.

И, в-третьих, имеется родственная функция, позволяющая выполнять сразу весь файл с программным кодом, но она не демонстрируется в этой главе: `PyRun_File`. Поскольку всегда имеется возможность загрузить текст из файла и выполнить его как единственную строку с помощью `PyRun_String`, основное преимущество функции `PyRun_File` сводится к отсутствию необходимости выделять память для содержимого файла. В таких строках, содержащих многострочный текст, для разделения строк как обычно используется символ `\n`, а блоки программного кода выделяются отступами.

Вызов объектов Python

В двух последних разделах мы занимались выполнением строк программного кода, но программы на C могут также легко работать с объектами Python. Программа в примере 20.27 выполняет ту же задачу,

что и примеры 20.23 и 20.26, но использует другие инструменты API для прямого взаимодействия с объектами в модуле Python:

`PyImport_ImportModule`

Импортирует модуль в C, как и прежде.

`PyObject_GetAttrString`

Загружает значение атрибута объекта по имени.

`PyEval_CallObject`

Вызывает функцию Python (или класс, или метод).

`PyArg_Parse`

Преобразует объекты Python в значения C.

`Py_BuildValue`

Преобразует значения C в объекты Python.

С обеими функциями преобразования мы познакомились ранее в этой главе. Вызов функции `PyEval_CallObject` в этой версии является ключевым: он выполняет импортированную функцию, передавая ей кортеж аргументов подобно синтаксической конструкции `func(*args)` в языке Python. Значение, возвращаемое функцией Python, поступает в C в виде `PyObject*`, обобщенного указателя на объект Python.

Пример 20.27. PP4E\Integrate\Embed\Basics\embed-object.c

```
/* получает и вызывает объекты из модулей */

#include <Python.h>

main() {
    char *cstr;
    PyObject *pstr, *pmod, *pfunc, *pargs;
    printf("embed-object\n");
    Py_Initialize();

    /* получить usermod.message */
    mpmod = PyImport_ImportModule("usermod");
    pstr = PyObject_GetAttrString(pmod, "message");

    /* преобразовать строку в C */
    PyArg_Parse(pstr, "s", &cstr);
    printf("%s\n", cstr);
    Py_DECREF(pstr);

    /* вызвать usermod.transform(usermod.message) */
    pfunc = PyObject_GetAttrString(pmod, "transform");
    pargs = Py_BuildValue("(s)", cstr);
    pstr = PyEval_CallObject(pfunc, pargs);
    PyArg_Parse(pstr, "s", &cstr);
    printf("%s\n", cstr);
```

```
    /* освободить объекты */
    Py_DECREF(pmod);
    Py_DECREF(pstr);
    Py_DECREF(pfunc);  /* в main() это делать необязательно */
    Py_DECREF(pargs);  /* поскольку вся память и так освобождается */
    Py_Finalize();
}
```

После компиляции и выполнения получается тот же результат:

```
.../PP4E/Integrate/Embed/Basics$ ./embed-object
embed-object
The meaning of life...
THE MEANING OF PYTHON...
```

Но на этот раз весь вывод создается средствами языка C – сначала путем получения значения атрибута `message` модуля Python, а затем путем прямого извлечения и вызова объекта функции `transform` модуля и вывода возвращаемого им значения, пересылаемого в C. Входными данными функции `transform` является аргумент функции, а не глобальная переменная, которой предварительно присвоено значение. Обратите внимание, что на этот раз `message` извлекается как атрибут модуля, а не в результате выполнения строки программного кода с именем переменной; часто есть несколько способов получить один и тот же результат, используя разные функции API.

Вызов функций в модулях, как показано в этом примере, дает простой способ организации встраивания. Программный код в файле модуля можно произвольно менять, не перекомпилируя выполняющую его программу на C. Обеспечивается также модель прямой связи: входные и выходные данные могут иметь вид аргументов функций и возвращаемых значений.

Выполнение строк в словарях

Когда выше мы использовали функцию `PyRun_String` для выполнения выражений и получения результатов, программный код выполнялся в пространстве имен существующего модуля Python. Однако иногда для выполнения строк программного кода удобнее создавать совершенно новое пространство имен, независимое от существующих файлов модулей. Программа на языке C, представленная в примере 20.28, демонстрирует, как это делается. Пространство имен создается как новый объект словаря Python, при этом в процессе участвует ряд новых для нас функций API:

`PyDict_New`

Создает новый пустой объект словаря.

`PyDict_SetItemString`

Выполняет присваивание по ключу словаря.

PyDict_GetItemString

Загружает значение из словаря по ключу.

PyRun_String

Выполняет строку программного кода в пространствах имен, как и ранее.

PyEval_GetBuiltins

Получает модуль с областью видимости встроенных элементов.

Главная хитрость здесь заключается в новом словаре. Входные и выходные данные для строк встроенного программного кода отображаются в этот словарь при его передаче в качестве словарей пространств имен кода в вызове функции PyRun_String. В итоге программа на C из примера 20.28 работает в точности, как следующий программный код Python:

```
>>> d = {}
>>> d['Y'] = 2
>>> exec('X = 99', d, d)
>>> exec('X = X + Y', d, d)
>>> print(d['X'])
101
```

Но здесь каждая операция Python заменяется вызовом C API.

Пример 20.28. PP4E\Integrate\Embed\Basics\embed-dict.c

```
/* создает новый словарь пространства имен для строки программного кода */

#include <Python.h>

main() {
    int cval;
    PyObject *pdict, *pval;
    printf("embed-dict\n");
    Py_Initialize();

    /* создать новое пространство имен */
    pdict = PyDict_New();
    PyDict_SetItemString(pdict, "__builtins__", PyEval_GetBuiltins());

    PyDict_SetItemString(pdict, "Y", PyLong_FromLong(2)); /* dict['Y'] = 2 */
    PyRun_String("X = 99", Py_file_input, pdict, pdict); /* вып. инструкц. */
    PyRun_String("X = X+Y", Py_file_input, pdict, pdict); /* то же X и Y */
    pval = PyDict_GetItemString(pdict, "X");      /* получить dict['X'] */

    PyArg_Parse(pval, "i", &cval);        /* преобразовать в C */
    printf("%d\n", cval);                 /* результат = 101 */
    Py_DECREF(pdict);
    Py_Finalize();
}
```

После компиляции и выполнения эта программа на C выведет следующее:

```
.../PP4E/Integrate/Embed/Basics$ ./embed-dict
embed-dict
101
```

На этот раз вывод получился иным: он отражает значение переменной Python X, присвоенное встроенными строками программного кода Python и полученное в C. В целом C может получать атрибуты модуля, либо вызывая функцию PyObject_GetAttrString с объектом модуля, либо обращаясь к словарю атрибутов модуля с помощью PyDict_GetItemString (также можно использовать строки выражений, но не напрямую). В данном случае модуля нет вообще, поэтому для обращения к пространству имен программного кода из C используется доступ к словарю по индексу.

Помимо возможности расчленения пространств имен строк с программным кодом, не зависимых от файлов модулей Python базовой системы, эта схема предоставляет естественный механизм связи. Значения, сохраняемые в новом словаре перед выполнением программного кода, служат входными данными, а имена, которым производится присвоение встроенным программным кодом, впоследствии могут извлекаться из словаря как выходные данные. Например, переменная Y во второй строке ссылается на имя, которому в C присваивается значение 2; значение переменной X присваивается программным кодом Python и извлекается позднее в программе C, как результат вывода.

Здесь есть один тонкий прием, требующий пояснения: словари, используемые в качестве пространств имен выполняемого программного кода, обычно должны содержать ссылку __builtins__ на пространство имен области видимости встроенных объектов, которая устанавливается программным кодом следующего вида:

```
PyDict_SetItemString(pdict, "__builtins__", PyEval_GetBuiltins());
```

Это неочевидное требование и обычно для модулей и встроенных компонентов, таких как функция exec, оно обрабатывается самим интерпретатором Python. Однако в случае использования собственных словарей пространств имен мы должны определять эту ссылку вручную, если выполняемый программный код должен иметь возможность ссылаться на встроенные имена. Это требование сохраняется в Python 3.X.

Предварительная компиляция строк в байт-код

Наконец, при вызове объектов функций Python из C вы фактически выполняете откомпилированный байт-код, связанный с объектом (например, телом функции), который обычно создается на этапе импортирования вмещающего модуля. При выполнении строк Python должен предварительно скомпилировать строку. Так как компиляция является медленным процессом, это может повлечь существенные потери време-

ни, если предполагается многократное выполнение строки. Чтобы избежать их, откомпилируйте строку в объект байт-кода, который будет выполняться позднее, с помощью функций API, представленных в примере 20.29.

`Py_CompileString`

Компилирует строку программного кода, возвращает объект байт-кода.

`PyEval_EvalCode`

Выполняет объект скомпилированного байт-кода.

Первая из этих функций принимает флаг режима, обычно передаваемый функции `PyRun_String`, и второй строковый аргумент, используемый только в сообщениях об ошибках. Вторая функция принимает два словаря пространств имен. Эти две функции API использованы в примере 20.29 для компиляции и выполнения трех строк программного кода Python.

Пример 20.29. PP4E\Integrate\Embed\Basics\embed-bytecode.c

```
/* предварительная компиляция строк программного кода в объекты байт-кода */

#include <Python.h>
#include <compile.h>
#include <eval.h>

main() {
    int i;
    char *cval;
    PyObject *pcode1, *pcode2, *pcode3, *presult, *pdict;
    char *codestr1, *codestr2, *codestr3;
    printf("embed-bytecode\n");

    Py_Initialize();
    codestr1 = "import usermod\nprint(usermod.message)"; /* инструкции */
    codestr2 = "usermod.transform(usermod.message)";     /* выражение */
    codestr3 = "print('%d:%d' % (X, X ** 2), end=' ')";  /* входное знач. X */

    /* создать новый словарь пространства имен */
    pdict = PyDict_New();
    if (pdict == NULL) return -1;
    PyDict_SetItemString(pdict, "__builtins__", PyEval_GetBuiltins());

    /* скомпилировать строки программного кода в объекты байт-кода */
    pcode1 = Py_CompileString(codestr1, "<embed>", Py_file_input);
    pcode2 = Py_CompileString(codestr2, "<embed>", Py_eval_input);
    pcode3 = Py_CompileString(codestr3, "<embed>", Py_file_input);

    /* выполнить скомпилированный байт-код в пространстве имен словаря */
    if (pcode1 && pcode2 && pcode3) {
        (void)    PyEval_EvalCode((PyCodeObject *)pcode1, pdict, pdict);
```

```
        presult = PyEval_EvalCode((PyCodeObject *)pcode2, pdict, pdict);
        PyArg_Parse(presult, "s", &cval);
        printf("%s\n", cval);
        Py_DECREF(presult);

        /* выполнить объект байт-кода несколько раз */
        for (i = 0; i <= 10; i++) {
            PyDict_SetItemString(pdict, "X", PyLong_FromLong(i));
            (void) PyEval_EvalCode((PyCodeObject *)pcode3, pdict, pdict);
        }
        printf("\n");
    }
    /* освободить использовавшиеся объекты */
    Py_XDECREF(pdict);
    Py_XDECREF(pcode1);
    Py_XDECREF(pcode2);
    Py_XDECREF(pcode3);
    Py_Finalize();
}
```

Эта программа объединяет в себе ряд приемов, которые мы уже видели. Например, пространством имен, в котором выполняются скомпилированные строки, является вновь создаваемый словарь (а не существующий объект модуля), а входные данные передаются строкам программного кода в виде предустановленных переменных в пространстве имен. После компиляции и выполнения первая часть вывода программы аналогична предыдущим примерам из этого раздела, но последняя строчка представляет 11-кратное выполнение одной и той же предварительно откомпилированной строки:

```
.../PP4E/Integrate/Embed/Basics$ embed-bytecode
embed-bytecode
The meaning of life...
THE MEANING OF PYTHON...

0:0 1:1 2:4 3:9 4:16 5:25 6:36 7:49 8:64 9:81 10:100
```

Если ваша система выполняет строки программного кода Python несколько раз, предварительное компилирование их в байт-код позволит получить значительное ускорение. Этот этап не является обязательным в других контекстах, где производится обращение к вызываемым объектам Python, включая типичный случай использования технологии встраивания, представленный в следующем разделе.

Регистрация объектов для обработки обратных вызовов

В примерах, приводившихся до сего момента, программный код на языке C выполнялся и вызывал программный код Python из обычной последовательности команд основной программы. Однако программы

не всегда работают таким образом. В некоторых случаях они создаются на основе архитектуры *управления событиями*, когда программный код выполняется только в ответ на те или иные события. Событием может являться щелчок конечным пользователем на кнопке в графическом интерфейсе, получение сигнала, переданного операционной системой, или просто выполнение программой действий, ассоциируемых с вводом данных в таблицу.

Так или иначе, программный код в такой архитектуре обычно организуется в виде обработчиков *обратных вызовов* – фрагментов программного кода, выполнение которых организовано логикой, обрабатывающей события. Для реализации обработчиков обратного вызова в такой системе довольно легко можно использовать встроенный программный код Python. В действительности для запуска обработчиков на языке Python слой обработки событий может просто использовать инструменты API вызова встроенного кода, с которыми мы познакомились в этой главе.

Единственный новый прием в этой модели состоит в том, чтобы сообщить слою C о том, какой программный код должен выполняться для каждого события. Обработчики каким-то образом должны быть зарегистрированы в C для связи с возможными событиями. В целом существует большое разнообразие способов добиться такой связи между программным кодом и событиями. Например, программы на языке C могут:

- Загружать и вызывать *функции* по имени события из одного или нескольких файлов *модулей*.
- Загружать и выполнять *строки* программного кода, ассоциированные с именами событий, из *базы данных*.
- Загружать и выполнять программный код, ассоциируемый с *тегами* событий в HTML или XML.
- Выполнять программный код Python, который обращается к C с запросом о том, что должно быть выполнено.

И так далее. В действительности механизмом регистрации обратного вызова может стать все, что каким-то образом ассоциирует объекты или строки с идентификаторами. У некоторых из таких приемов есть свои преимущества. Например, получение программного кода обратного вызова из файлов модулей поддерживает динамическую повторную загрузку (`imp.reload` действует в отношении модулей, но не обновляет объекты, хранящиеся непосредственно). При этом ни в одной из первых трех схем не требуется писать специальные программы Python, занимающиеся только регистрацией обработчиков для последующего выполнения.

Однако чаще, по-видимому, для регистрации обработчиков обратного вызова используется последний подход: регистрация обработчиков в C программным кодом Python при обратном вызове C через интерфейс расширений. Хотя в этой схеме есть свои минусы, она предоставляет ес-

тественную и прямую модель в ситуациях, где обратные вызовы ассоциируются с большим числом объектов.

Рассмотрим, например, графический интерфейс, построенный в виде дерева объектов виджетов в сценарии Python. Если с каждым объектом дерева может быть связан обработчик события, проще будет регистрировать обработчики при вызове методов виджетов, образующих дерево. При связывании обработчиков с объектами виджетов в отдельной структуре, такой как файл модуля или файл XML, потребуется дополнительная работа с перекрестными ссылками, чтобы поддерживать соответствие обработчиков дереву.

Фактически, если вам требуется более практичный пример обработчиков событий на Python, загляните в реализацию библиотеки tkinter, которую мы широко использовали в этой книге. В библиотеке tkinter используются обе технологии, расширения и встраивания. Интерфейс *расширения* в ней используется для регистрации обработчиков обратного вызова на Python, которые позднее запускаются в ответ на события в графическом интерфейсе, с применением интерфейса *встраивания*. Вы можете исследовать реализацию библиотеки tkinter, которая включена в пакет с исходными текстами Python, – логика взаимодействия с библиотекой Tk в ней довольно сложна для восприятия, но используемая в ней модель является достаточно прямолинейной.

Реализация регистрации

В файлах C и Python, которые будут представлены в этом разделе, демонстрируются основы приемов программирования, используемых при реализации явно регистрируемых обработчиков обратных вызовов. Сначала рассмотрим программу на языке C, представленную в примере 20.30, которая реализует интерфейсы регистрации обработчиков на языке Python, а также программный код для запуска этих обработчиков в ответ на события:

Маршрутизатор событий

Функция `Route_Event` реагирует на событие, вызывая объект функции Python, ранее переданный из Python в C.

Регистрация обработчика обратного вызова

Функция `Register_Handler` сохраняет переданный ей указатель на объект функции Python в глобальной переменной C. Python вызывает `Register_Handler` через простой модуль расширения C `cregister`, создаваемый этим файлом.

Возбуждение событий

Для моделирования действительных событий можно вызвать через созданный модуль C функцию Python `Trigger_Event`, которая сгенерирует событие.

Иными словами, в этом примере используются оба знакомых нам интерфейса – расширения и встраивания – для регистрации и вызова программного кода обработчика события на Python. Чтобы лучше узнать, как действует этот прием, внимательно изучите его реализацию в примере 20.30.

Пример 20.30. PP4E\Integrate\Embed\Regist\cregister.c

```
#include <Python.h>
#include <stdlib.h>

/*********************************************/
/* 1) передача событий объекту Python        */
/* вместо этого можно было бы выполнять строки */
/*********************************************/

static PyObject *Handler = NULL; /* содержит объект Python в C */

void Route_Event(char *label, int count)
{
    char *cres;
    PyObject *args, *pres;

    /* вызов обработчика Python */
    args = Py_BuildValue("(si)", label, count);  /* создать список
                                                    аргументов */
    pres = PyEval_CallObject(Handler, args);     /* применение: выполнить
                                                    вызов */
    Py_DECREF(args);             /* добавить контроль ошибок */
    if (pres != NULL) {
        /* использовать результат и уменьшить счетчик ссылок на него */
        PyArg_Parse(pres, "s", &cres);
        printf("%s\n", cres);
        Py_DECREF(pres);
    }
}

/*************************************************************/
/* 2) модуль расширения python для регистрации обработчиков  */
/* python импортирует этот модуль для установки обработчиков */
/*************************************************************/

static PyObject *
Register_Handler(PyObject *self, PyObject *args)
{
    /* сохранить вызываемый объект Python */
    Py_XDECREF(Handler);                 /* вызывался прежде? */
    PyArg_Parse(args, "(O)", &Handler);  /* один аргумент */
    Py_XINCREF(Handler);         /* добавить ссылку */
    Py_INCREF(Py_None);          /* вернуть 'None': успех */
```

```
    return Py_None;
}

static PyObject *
Trigger_Event(PyObject *self, PyObject *args)
{
    /* позволить Python имитировать событие, перехваченное C */
    static count = 0;
    Route_Event("spam", count++);
    Py_INCREF(Py_None);
    return Py_None;
}

static PyMethodDef cregister_methods[] = {
    {"setHandler", Register_Handler, METH_VARARGS, ""}, /* имя, адр.
                                                           функ.,...*/
    {"triggerEvent", Trigger_Event, METH_VARARGS, ""},
    {NULL, NULL, 0, NULL}                              /* конец таблицы */
};

static struct PyModuleDef cregistermodule = {
    PyModuleDef_HEAD_INIT,
    "cregister",       /* имя модуля */
    "cregister mod",   /* описание модуля, может быть NULL */
    -1, /* размер структуры для каждого экземпляра, -1=глоб. перем. */
    cregister_methods  /* ссылка на таблицу методов */
};

PyMODINIT_FUNC
PyInit_cregister()     /* вызывается первой инструкцией импортирования */
{
    return PyModule_Create(&cregistermodule);
}
```

Конечно, эта программа на C является модулем расширения Python, а не самостоятельной программой C со встроенным программным кодом Python (хотя программа на C могла бы быть верхним уровнем). Для компиляции ее в файл динамически загружаемого модуля выполните make-файл, представленный в примере 20.31 в Cygwin (или аналогичный ему make-файл на других платформах). Как мы узнали выше в этой главе, получающийся файл *cregister.dll* будет загружен при первом импорте в сценарии Python, если поместить его в каталог, находящийся в пути поиска модулей Python (например, . или содержащийся в PYTHONPATH).

Пример 20.31. PP4E\Integrate\Embed\Regist\makefile.regist

```
#####################################################################
# make-файл для Cygwin, выполняющий сборку cregister.dll, динамически
# загружаемого модуля расширения на C (разделяемого), который
# будет импортироваться сценарием register.py
#####################################################################
```

```
PYLIB = /usr/local/bin
PYINC = /usr/local/include/python3.1

CMODS = cregister.dll

all: $(CMODS)

cregister.dll: cregister.c
        gcc cregister.c -g -I$(PYINC) -shared -L$(PYLIB) -lpython3.1 -o $@

clean:
        rm -f *.pyc $(CMODS)
```

Теперь, когда у нас есть модуль расширения C для регистрации и вызова обработчиков Python, нам нужны лишь какие-нибудь обработчики Python. Модуль Python, представленный в примере 20.32, определяет две функции обработчиков обратного вызова, импортирует модуль расширения C для регистрации обработчиков и генерирует события.

Пример 20.32. PP4E\Integrate\Embed\Regist\register.py

```
"""
##########################################################################
в Python, регистрирует обработчики событий, которые будут вызываться из C;
скомпилируйте и скомпонуйте программный код на C и запустите этот сценарий
командой 'python register.py'
##########################################################################
"""


##############################################
# эти функции Python будут вызываться из C;
# обрабатывать события и возвращать результат
##############################################

def callback1(label, count):
    return 'callback1 => %s number %i' % (label, count)

def callback2(label, count):
    return 'callback2 => ' + label * count

##################################################
# Python вызывает модуль расширения C
# для регистрации обработчиков, возбуждает события
##################################################

import cregister

print('\nTest1:')
cregister.setHandler(callback1) # зарегистрировать функцию-обработчик
for i in range(3):
    cregister.triggerEvent()  # имитировать события, перехватываемые слоем C
```

```
print('\nTest2:')
cregister.setHandler(callback2)
for i in range(3):
    cregister.triggerEvent()  # передаст события функции callback2
```

Вот и все – интеграция обработчиков обратного вызова Python/C готова к работе. Чтобы запустить систему, выполните сценарий Python. Он зарегистрирует одну функцию, сгенерирует три события, затем изменит обработчик событий и повторит все снова:

```
.../PP4E/Integrate/Embed/Regist$ make -f makefile.regist
gcc cregister.c -g -I/usr/local/include/python3.1 -shared -L/usr/local/bin
-lpython3.1 -o cregister.dll

.../PP4E/Integrate/Embed/Regist$ python register.py

Test1:
callback1 => spam number 0
callback1 => spam number 1
callback1 => spam number 2

Test2:
callback2 => spamspamspam
callback2 => spamspamspamspam
callback2 => spamspamspamspamspam
```

Этот вывод производит функция в программе C маршрутизации событий, но в нем присутствуют значения, возвращаемые функциями обработчиков в модуле Python. В действительности здесь незаметно много чего происходит. Когда Python возбуждает событие, между двумя языками происходит такая передача управления:

1. Из Python в функцию маршрутизации событий C.
2. Из функции маршрутизации событий C в функцию-обработчик Python.
3. Обратно в функцию маршрутизации событий C (которая выводит полученные результаты).
4. Наконец, обратно в сценарий Python.

Таким образом, мы переходим из Python в C, оттуда в Python и еще раз обратно. Попутно управление передается через интерфейсы расширения и встраивания. Во время работы обработчика обратного вызова Python активными оказываются два уровня Python и один уровень C посередине. К счастью, это действует – API Python отличается реентерабельностью, поэтому не нужно беспокоиться по поводу того, что одновременно активны несколько уровней интерпретатора Python. Каждый уровень выполняет свой программный код и действует независимо.

Попробуйте по результатам работы и исходному программному коду этого примера проследить порядок выполнения операций, чтобы лучше понять, как он действует. А теперь перейдем к последнему короткому при-

меру, пока у нас еще есть время и место для исследований, – для симметрии попробуем использовать классы Python из программного кода на C.

Использование классов Python в программах C

Ранее в этой главе мы научились использовать классы C++ в Python, создавая для них обертки с помощью SWIG. А можно ли пойти в обратном направлении – использовать классы Python из других языков? Оказывается, для этого нужно только применить уже продемонстрированные интерфейсы.

Напомню, что сценарии Python создают объекты экземпляров классов путем *вызова* объектов классов, как если бы они являлись функциями. Чтобы сделать это из C (или C++), нужно просто сделать те же шаги: импортировать класс из модуля, создать кортеж аргументов и вызвать класс для создания экземпляра с помощью тех же инструментов C API, которыми вызываются функции Python. Создав экземпляр, можно получать его атрибуты и методы теми же средствами, с помощью которых извлекаются значения глобальных переменных из модуля. Доступ к вызываемым объектам и атрибутам везде выполняется одинаково.

Чтобы продемонстрировать это на практике, в примере 20.33 определен модуль с простым классом Python, которым можно пользоваться в C.

Пример 20.33. PP4E\Integrate\Embed\Pyclasss\module.py

```
# вызывать этот класс из C для создания объектов

class klass:
    def method(self, x, y):
        return "brave %s %s" % (x, y) # вызывать из C
```

Проще некуда, но достаточно для иллюстрации основ. Как обычно, этот модуль должен находиться в пути поиска модулей Python (например, в текущем каталоге или содержащемся в PYTHONPATH), иначе импортировать его при вызове из C не удастся, равно как и из сценария Python. Как вы уже наверняка знаете, раз уж дочитали книгу до этого места, к этому классу Python можно обратиться из программы на языке Python, как показано ниже:

```
...\PP4E\Integrate\Embed\Pyclass$ python
>>> import module                          # импортировать модуль
>>> object = module.klass()                # создать экземпляр класса
>>> result = object.method('sir', 'robin')  # вызвать метод класса
>>> print(result)
brave sir robin
```

В Python это весьма просто. Те же действия можно выполнить и в C, но для этого потребуется немного больше программного кода. Программа на C, представленная в примере 20.34, выполняет эти шаги, организуя вызовы соответствующих инструментов Python API.

Пример 20.34. PP4E\Integrate\Embed\Pyclass\objects.c

```
#include <Python.h>
#include <stdio.h>

main() {
    /* запускает объекты, используя низкоуровневые вызовы */
    char *arg1="sir", *arg2="robin", *cstr;
    PyObject *pmod, *pclass, *pargs, *pinst, *pmeth, *pres;

    /* экземпляр = module.klass() */
    Py_Initialize();
    pmod   = PyImport_ImportModule("module");       /* получить модуль */
    pclass = PyObject_GetAttrString(pmod, "klass"); /* получить класс */
    Py_DECREF(pmod);

    pargs = Py_BuildValue("()");
    pinst = PyEval_CallObject(pclass, pargs);       /* вызвать class() */
    Py_DECREF(pclass);
    Py_DECREF(pargs);

    /* результат = instance.method(x,y) */
    pmeth = PyObject_GetAttrString(pinst, "method"); /* связанный метод */
    Py_DECREF(pinst);
    pargs = Py_BuildValue("(ss)", arg1, arg2);  /* преобразовать в Python */
    pres  = PyEval_CallObject(pmeth, pargs);    /* вызвать method(x,y) */
    Py_DECREF(pmeth);
    Py_DECREF(pargs);

    PyArg_Parse(pres, "s", &cstr);              /* преобразовать в C */
    printf("%s\n", cstr);
    Py_DECREF(pres);
}
```

Просмотрите этот файл и разберитесь в деталях. Нужно просто понять, как задача решалась бы в Python, а затем вызывать эквивалентные функции C в Python API. Для компиляции этого исходного текста в выполняемую программу C запустите make-файл в каталоге, где находится этот файл (он аналогичен уже рассмотренным make-файлам). После компиляции запустите, как любую другую программу C:

```
.../PP4E/Integrate/Embed/Pyclass$ ./objects
brave sir robin
```

Этот вывод может разочаровать, но он действительно отражает значения, возвращаемые в C методом класса из файла *module.py*. Программе на C пришлось немало потрудиться, чтобы получить эту короткую строку: она импортировала модуль, загрузила класс, создала экземпляр, загрузила и вызвала метод экземпляра, попутно выполняя преобразование данных и управляя счетчиком ссылок. В награду за все труды программа C может применять показанный в этом файле прием для повторного использования *любых* классов Python.

Конечно, на практике этот пример был бы сложнее. Как уже говорилось, обычно нужно проверять значения, возвращаемые всеми вызовами Python API, чтобы убедиться в их успешном выполнении. Например, попытка импортировать модуль в этом программном коде C может закончиться неудачей, если модуль находится вне пути поиска. Если не перехватывать возврат указателя NULL, то программа почти наверняка закончится крахом при попытке использовать этот указатель (когда он понадобится). В примере 20.35 представлен программный код из примера 20.34, в который была добавлена проверка всех ошибок – программный код получился объемным, но надежным.

Пример 20.35. PP4E\Integrate\Embed\Pyclasss\objects-err.c

```
#include <Python.h>
#include <stdio.h>
#define error(msg) do { printf("%s\n", msg); exit(1); } while (1)

main() {
    /* запускает объекты, используя низкоуровневые вызовы
       и проверяя ошибки */
    char *arg1="sir", *arg2="robin", *cstr;
    PyObject *pmod, *pclass, *pargs, *pinst, *pmeth, *pres;

    /* экземпляр = module.klass() */
    Py_Initialize();
    pmod = PyImport_ImportModule("module");          /* получить модуль */
    if (pmod == NULL)
        error("Can't load module");

    pclass = PyObject_GetAttrString(pmod, "klass");  /* получить класс */
    Py_DECREF(pmod);
    if (pclass == NULL)
        error("Can't get module.klass");

    pargs = Py_BuildValue("()");
    if (pargs == NULL) {
        Py_DECREF(pclass);
        error("Can't build arguments list");
    }
    pinst = PyEval_CallObject(pclass, pargs);        /* вызвать class() */
    Py_DECREF(pclass);
    Py_DECREF(pargs);
    if (pinst == NULL)
        error("Error calling module.klass()");

    /* результат = instance.method(x,y) */
    pmeth = PyObject_GetAttrString(pinst, "method"); /* связанный метод */
    Py_DECREF(pinst);
    if (pmeth == NULL)
        error("Can't fetch klass.method");
```

```
    pargs = Py_BuildValue("(ss)", arg1, arg2); /* преобразовать в Python */
    if (pargs == NULL) {
        Py_DECREF(pmeth);
        error("Can't build arguments list");
    }
    pres = PyEval_CallObject(pmeth, pargs);    /* вызвать method(x,y) */
    Py_DECREF(pmeth);
    Py_DECREF(pargs);
    if (pres == NULL)
        error("Error calling klass.method");

    if (!PyArg_Parse(pres, "s", &cstr))        /* преобразовать в C */
        error("Can't convert klass.method result");
    printf("%s\n", cstr);
    Py_DECREF(pres);
}
```

Эти 53 строки программного кода на C (не считая make-файла) получают тот же результат, что и 4 строки в интерактивной оболочке Python, которые мы видели выше, – не самый лучший показатель с точки зрения производительности разработчика! Но как бы то ни было, использованная модель позволяет программам на C и C++ использовать Python так же, как программы на Python могут использовать C и C++. В заключении к этой книге, которое следует чуть ниже, я отмечу, что такие комбинации часто могут оказаться более мощными, чем отдельные части, составляющие их.

Другие темы интеграции

В данной книге термин *интеграция* в основном означал соединение Python с компонентами, написанными на C или C++ (либо другими языками, совместимыми с C), в режимах расширения или встраивания. Но с более широкой точки зрения интеграция также включает в себя любые другие технологии, позволяющие образовывать из компонентов Python крупные гетерогенные системы. В данном последнем разделе кратко рассматривается ряд технологий интеграции помимо средств C API, исследованных нами.

Jython: интеграция с Java

С Jython мы впервые столкнулись в главе 12 и еще раз – выше в этой главе, при обсуждении вопросов расширения. Однако в действительности Jython представляет собой более обширную платформу интеграции. Jython компилирует программный код Python в байт-код Java для выполнения под управлением JVM. Получающаяся при этом система на основе Java поддерживает два вида интеграции:

- *Расширение*: Jython использует API *отражения* (рефлексии) Java, позволяющий программам Python автоматически обращаться к библиотекам классов Java. API отражения Java предоставляет

информацию о типах Java на этапе выполнения и служит той же цели, что и связующий программный код, который мы генерировали в этой части книги для подключения библиотек C к Python. Однако в Jython эта информация о типах в момент выполнения позволяет в значительной мере автоматизировать вызов функций Java в сценариях Python – нет необходимости писать или генерировать связующий программный код.

- *Встраивание*: Jython также предоставляет API класса Java `PythonInterpreter`, позволяющий программам Java выполнять программный код Python в отдельном пространстве имен подобно инструментам C API, использовавшимся для выполнения строк программного кода Python из программ C. Кроме того, поскольку Jython реализует все объекты Python как экземпляры класса Java `PyObject`, слой Java, содержащий встроенный код Python, легко может обрабатывать объекты Python.

Иными словами, Jython позволяет создавать расширения для Python на Java и встраивать его в Java подобно стратегиям интеграции с C, рассмотренным в этой части книги. Добавляя простой язык сценариев в приложения Java, Jython играет практически ту же роль, что и инструменты интеграции C, которые мы только что изучили.

Однако развитие Jython обычно отстает от развития CPython, а его опора на библиотеки классов и среду выполнения Java влечет за собой зависимости Java, которые могут быть значимы в определенных сценариях разработки, ориентированных на Python. Но как бы то ни было, JPython предоставляет удивительно гладкую модель интеграции и служит идеальным языком сценариев для приложений Java. За более подробной информацией о Jython обращайтесь на сайт *http://www.jython.org* и к поисковым системам.

IronPython: интеграция с C#/.NET

IronPython, также упоминавшийся ранее, обеспечивает интеграцию с C#/.NET, аналогичную той, что обеспечивает Jython для Java (и фактически принадлежит тому же автору) – он обеспечивает гладкую интеграцию между программным кодом Python и программными компонентами, написанными для фреймворка .NET, а также его реализации Mono в Linux. Подобно Jython IronPython компилирует исходный программный код Python в байт-код .NET и выполняет программы под управлением системы времени выполнения. Таким образом интеграция с внешними компонентами получается также гладкой. В результате, как и в случае с Jython, Python превращается в простой в использовании язык сценариев для приложений на основе C#/.NET и в многоцелевой инструмент быстрой разработки, дополняющий C#. За более подробной информацией об IronPython обращайтесь на сайт *http://www.ironpython.org* или к поисковым системам.

Интеграция в Windows с помощью COM

Модель COM определяет стандартную и независимую от языка объектную модель, с помощью которой компоненты, написанные на разных языках, могут интегрироваться и поддерживать связь. Пакет расширения PyWin32 для Python позволяет программам Python реализовывать серверы и клиенты в модели интерфейса COM. По существу, модель COM предоставляет мощный способ интеграции программ Python с программами, написанными на других языках, поддерживающих COM, таких как Visual Basic. Сценарии Python могут также с помощью вызовов COM управлять такими популярными приложениями Microsoft, как Word и Excel, так как эти системы регистрируют собственные интерфейсы объектов COM. С другой стороны, в модели COM используется промежуточный механизм косвенной адресации и отсутствует возможность использовать ее на других платформах, как в случае использования других решений, перечисленных здесь. Дополнительную информацию о поддержке COM и о других расширениях для Windows можно найти в Интернете и в книге Марка Хаммонда (Mark Hammond) и Энди Робинсона (Andy Robinson) «Python Programming on Win32», выпущенной издательством O'Reilly.

Интеграция CORBA

Имеется также поддержка, в значительной степени распространяемая с исходными текстами, использования Python в контексте приложений, основанных на CORBA. *CORBA* (Common Object Request Broker – стандартная архитектура брокера объектных запросов) представляет собой независимый от языка способ построения распределенных систем из компонентов, связанных между собой архитектурой объектной модели. По существу, она представляет собой еще один способ интеграции компонентов Python в большие системы. Поддержка CORBA в Python включает в себя находящиеся в общественном владении системы, такие как OmniORB. Как и модель COM, CORBA является большой системой – слишком большой, чтобы можно было даже поверхностно рассмотреть ее в этой книге. Дополнительные сведения ищите в Интернете.

Другие языки

Как уже обсуждалось выше, вы также можете найти прямую поддержку интеграции Python с другими языками программирования, включая FORTRAN, Objective-C и др. В большинстве случаев обеспечивается поддержка расширения (вызова функций, написанных на других языках) и встраивания (обработка вызовов из других языков). Дополнительные подробности вы найдете в обсуждении выше или выполнив поиск в Интернете. Кроме того, некоторые наблюдатели относят к этой категории недавно появившуюся систему pyjamas –

компилируя программный код Python в программный код JavaScript, она позволяет программам на языке Python получить доступ к технологии AJAX и прикладным интерфейсам на основе веб-броузеров в контексте полнофункциональных интернет-приложений (Rich Internet Applications), обсуждавшихся ранее в этой книге; смотрите главы 7, 12 и 16.

Интеграция на основе сетевых протоколов

Наконец, в мире Python существует также поддержка интеграции на основе протоколов Интернета транспортировки данных, включая SOAP и XML-RPC. Отправляя вызовы по сети, такие системы обеспечивают поддержку распределенных архитектур и дают начало понятию веб-служб. Поддержка протокола XML-RPC присутствует в стандартной библиотеке Python, тем не менее, выполните поиск в Интернете, чтобы получить больше информации об этих протоколах.

Как видите, существует большое число возможностей в области интеграции. Лучший совет, который я могу дать на прощание, это учитывать, что разные инструменты предназначены для решения разных задач. Модули расширений и типы на C идеально подходят для оптимизации систем и интеграции с библиотеками, но фреймворки предлагают другие способы интеграции компонентов – Jython и IronPython используются при работе с Java и .NET, COM – для повторного использования и реализации объектов в Windows, XML-RPC – для создания распределенных служб и так далее. Как обычно, лучшими практически всегда будут инструменты, больше всего подходящие для ваших программ.

VI

Финал

Подобно первой части книги последнюю ее часть составляет единственная, на этот раз более короткая глава, в которой подводятся некоторые итоги:

Глава 21

В этой главе обсуждаются роли Python и область его применения. В ней исследуются некоторые более широкие идеи о круге обычного применения Python, а кроме того, представлены расширенные возможности в том объеме, который позволяет оставшееся пространство в книге. Большая часть этой главы представляет философский взгляд на вещи, но помимо этого в ней выделяются некоторые основные причины использования таких инструментов, как Python.

Обратите внимание на отсутствие приложений справочного характера. За дополнительными справочными материалами обращайтесь к стандартным руководствам по языку Python, доступным в Интернете, или к коммерческим печатным справочникам, таким как «Python Pocket Reference» издательства O'Reilly, и другим, которые вы сможете отыскать в привычных местах в Интернете.

Дополнительный материал по базовому языку Python можно найти в книге «Изучаем Python»[1]. Кроме всего прочего, в четвертом издании этой книги исследуются более сложные инструменты языка, такие как свойства, дескрипторы, декораторы и метаклассы, которые мы пропустили здесь, потому что они относятся к категории базовых особенностей. В книге «Изучаем Python» также более подробно, чем здесь, рассматриваются вопросы работы с текстом Юникода, так как это является отличительной особенностью Python 3.

---

[1] Марк Лутц «Изучаем Python», 4-е издание, СПб.: Символ-Плюс, 2010.

А за помощью по другим темам, относящимся к Python, обращайтесь к ресурсам, ссылки на которые вы найдете на официальном веб-сайте Python, *http://www.python.org*, или поищите в Интернете с помощью своей любимой поисковой системы.

21

Заключение: Python и цикл разработки

Предисловие к четвертому изданию

Эту заключительную главу я написал еще в 1995 году для первого издания данной книги. В то время роль Python в интеграции при использовании его в качестве управляющего языка верхнего уровня считалась для Python более важной, чем в настоящее время. С тех пор Python вошел в четверку или пятерку самых используемых языков программирования. С ростом популярности языка его ориентация на высокое качество и удобочитаемость программного кода и как следствие положительное влияние на производительность разработчиков представляются доминирующими факторами успеха Python.

На практике большинство программистов, использующих Python, пишут программный код исключительно на языке Python, даже не зная или не задумываясь об использовании внешних библиотек. Как правило, лишь небольшое количество разработчиков интегрирует внешние библиотеки для использования в своем программном коде Python. В настоящее время проблема интеграции по-прежнему имеет большое значение, тем не менее, для большинства основными преимуществами Python являются качество и производительность (смотрите врезку «“Тайное знание” разработчиков Python» в конце главы 1, где подробнее говорится о современной философии Python).

Это смещение интересов обусловило некоторое сужение набора тем для заключения. Вследствие этого я удалил все «комментарии», приводимые в предыдущих изданиях. Однако само заключение было оставлено в этом издании, отчасти из почтения к его исторической ценности, отчасти потому, что оно отражает идеалы Python, сразу выдвинувшие его под огни рампы, и отчасти потому, что оно остается значимым для пользователей Python, продолжающих создавать гибридные системы (ну, и еще из-за шуток).

В конце концов, многие из нас сейчас уже отплыли от пресловутого острова, о котором говорится в этой главе, благодаря развитию инструментов, таких как Python. Следует признать, что за последние 15 лет в языке Python было реализовано большинство заложенных в нем изначально идей, о которых было рассказано в этой книге. Выбор, который стоит перед нами в настоящее время, похоже, заключается в том, чтобы или не утяжелять лодку, или научиться плавать.

«Книга заканчивается, пора уже и о смысле жизни»

Или, во всяком случае, о смысле Python. Во введении к этой книге я обещал, что мы вернемся к вопросу о круге задач, решаемых на Python, после того как посмотрим на его применение на практике. Поэтому в завершение приведу некоторые совершенно субъективные замечания по поводу более широкой роли этого языка. В значительной степени эта глава не изменилась с момента выхода первого издания книги в 1995 году, но это потому, что неизменными остались факторы, которые вытолкнули Python в центр внимания разработчиков программного обеспечения.

Как говорилось во врезке в предисловии к книге, Python нацелен на *качество*, *производительность*, *переносимость* и *интеграцию*. Надеюсь, что эта книга продемонстрировала некоторые из преимуществ, определяемые такой направленностью, в действии. По пути мы видели возможность использования Python в таких областях, как системное программирование, разработка графических интерфейсов, создание сценариев для Интернета, работа с базами данных и обработка текста и многих других. И мы на практике убедились в возможности применения языка и его библиотек для создания действующего и масштабируемого программного обеспечения, выходящего за границы того, что часто относят к «написанию сценариев». Я надеюсь, что при этом вы получили некоторое удовольствие – это тоже является частью работы с Python.

В этом заключении я хотел бы, чтобы после долгой прогулки среди деревьев мы снова увидели лес в целом; рассмотрим более конкретно роли, которые играет Python. Например, использование Python в качестве инструмента создания прототипов может существенно влиять на весь цикл разработки.

«Как-то мы неправильно программируем компьютеры»

Это одна из наиболее затасканных фраз в нашей отрасли. Все же, если на некоторое время задуматься о картине в целом, большинство из нас, видимо, согласится, что что-то мы «недоделали». За несколько последних десятилетий индустрия программного обеспечения достигла значительного прогресса в упрощении процесса разработки (кто-нибудь помнит еще ввод данных с перфокарт?). В то же время стоимость разработки определенных потенциально полезных компьютерных программ часто остается настолько высокой, что они становятся невыгодными.

Кроме того, при создании систем с помощью современных инструментов и парадигм часто наблюдается значительное отставание от графика работ. Разработка программного обеспечения по-прежнему не поддается количественным методам, применяемым в других областях конструирования. В мире программирования не столь уж редко берут некоторую наилучшую оценку времени выполнения нового проекта и умножают на коэффициент два или три, чтобы учесть непредвиденные затраты при разработке. Очевидно, что такая ситуация не удовлетворяет менеджеров, разработчиков и конечных пользователей.

«Фактор Гиллигана»

Было предположено, в шутку, что если бы у разработчиков программного обеспечения был святой покровитель, этой чести не удостоился бы никто, кроме Гиллигана, персонажа весьма популярного американского телешоу 1960-х «Остров Гиллигана». Гиллиган – загадочный, обутый в тапочки первый помощник капитана, виноватый, по общему мнению людей, оказавшихся на острове, в происшедшем кораблекрушении.

В самом деле ситуация с Гиллиганом кажется до странности знакомой. Выброшенные на необитаемый остров и располагающие лишь самыми скромными из достижений современной технологии, Гиллиган и его товарищи вынуждены бороться за выживание, используя лишь то, что дает им природа. В эпизоде за эпизодом мы видим, как Профессор изобретает крайне замысловатые средства, чтобы улучшить их жизнь на этом заброшенном острове, но на стадии реализации вечно неловкий Гиллиган все портит.

Но ясно, что виноват в этом не несчастный Гиллиган. Можно ли рассчитывать, что с помощью элементарных технологий, доступных в таких условиях, будут реализованы проекты таких сложных устройств, как бытовая техника и устройства связи? У него просто не было необходимых средств. Не исключено, что инженерные способности Гиллигана были на высочайшем уровне, но что можно собрать из бананов и кокосов?

И неизменно, раз за разом, кончается тем, что Гиллиган неумышленно срывает лучшие планы Профессора, неправильно используя и в конечном счете уничтожая его изобретения. Если бы он крутил педали своего самодельного неподвижного велосипеда быстрее (как его убеждали), все было бы хорошо. Но в конце неизбежно кокосы летят в воздух, а обломки пальмовых ветвей падают ему на голову, и бедного Гиллигана в очередной раз ругают за провал технологии.

Как ни драматичен этот образ, некоторые наблюдатели считают его поразительной метафорой программной индустрии. Как и Гиллигана, нас, разработчиков программ, часто вынуждают решать задачи с помощью явно негодных средств. Как и у Гиллигана, наши намерения правильны, но сдерживает технология. И, как несчастный Гиллиган, мы неизбежно навлекаем на себя гнев руководства, когда наши проекты завершаются с опозданием. С бананами и кокосами задачу не решить...

Делать правильно

Конечно, фактор Гиллигана – преувеличение, внесенное для комического эффекта. Но мало кто станет утверждать, что узкое место между идеями и их воплощением в работающих системах полностью исчезло. Даже сегодня стоимость разработки программного обеспечения значительно превосходит стоимость аппаратной части. Обязательно ли программирование должно быть таким сложным?

Рассмотрим ситуацию внимательней. В общем и целом причина сложности разработки программного обеспечения не связана с той ролью, которую оно должно выполнять, – обычно это четко очерченный процесс реального мира. Скорее она связана с отображением задач реального мира на модели, которые могут исполняться компьютерами. А это отображение осуществляется в контексте языков и средств программирования.

Поэтому путь к устранению узкого места программного обеспечения должен лежать, по крайней мере частично, в оптимизации самого программирования путем применения инструментов, соответствующих задаче. В этой практической области можно сделать многое – есть ряд чисто искусственных издержек, вызываемых инструментами, которыми мы пользуемся в настоящее время.

Цикл разработки для статических языков

При использовании традиционных статических языков невозможно избежать издержек, связанных с переходом от программного кода к работающим системам: этапы компиляции и сборки вводят обязательную задержку в процесс разработки. В некоторых случаях обычным является, когда ожидание завершения цикла сборки приложения, написанного на статическом языке, доходит до нескольких часов в неделю. С учетом современной практики разработки, включающей итеративный процесс компиляции, тестирования и повторной компиляции, такие задержки оказываются дорогостоящими и деморализующими (а то и физически мучительными).

Конечно, в разных организациях это может происходить по-разному, а в некоторых областях приложений требования, предъявляемые к производительности, оправдывают задержки в цикле разработки. Но мне приходилось работать в таких условиях разработки на C++, где программисты шутили, что после запуска проектов на перекомпиляцию можно идти обедать. И это было не совсем шуткой.

Искусственные сложности

Многие традиционные средства программирования таковы, что за деревьями можно не увидеть леса: процедура программирования становится настолько сложной, что за ней плохо видна задача, выполняемая программой в реальном мире. Традиционные языки отвлекают драгоценное внимание на синтаксические проблемы и разработку программного кода, ведущего учет системных ресурсов. Очевидно, что сложность не является самоцелью – это нужно четко заявить. Тем не менее, некоторые современные инструменты настолько сложны, что сам язык усложняет задачу и затягивает процесс разработки.

Одним языком не угодишь всем

Многие традиционные языки косвенно поощряют однородные, одноязычные системы. Усложняя интеграцию, они препятствуют использованию многоязычных инструментов. Поэтому, не имея возможности выбрать инструмент, более подходящий для конкретной задачи, разработчики часто вынуждены использовать один и тот же язык для всех компонентов приложения. Так как нет языков, которые одинаково хороши во всем, такое ограничение неизбежно наносит ущерб – как продукту, так и производительности труда программиста.

Пока наши компьютеры не станут умны настолько, чтобы понимать указания не хуже нас (возможно, не самая разумная из целей), задача программирования останется. Но в данный момент можно достичь существенного прогресса, упростив механику этой задачи. На эту тему я и хочу сейчас поговорить.

И тут появляется Python

Если эта книга достигла целей, которые перед ней ставились, вы должны сейчас хорошо понимать, почему Python называют «управляющим языком нового поколения». В сравнении с аналогичными инструментами у него есть важные отличия, которые мы наконец можем суммировать:

Tcl

Подобно Tcl, Python может использоваться в качестве встроенного языка расширения. В отличие от Tcl, Python является также полнофункциональным языком программирования. Для многих пользователей средства Python для работы со структурами данных и поддержка программирования в целом открывают возможность применения его в самых разных областях. Язык Tcl продемонстрировал пользу интеграции интерпретируемых языков с модулями на языке C. Python предоставляет аналогичные возможности плюс мощный объектно-ориентированный язык: это не просто процессор командных строк.

Perl

Как и Perl, Python можно применять для создания инструментов командной оболочки, облегчая использование системных служб. В отличие от Perl, у Python простой, легко читаемый синтаксис и замечательно ясная архитектура. Благодаря этому некоторым программистам проще пользоваться Python – он лучше подходит для программ, которые должны повторно использоваться или сопровождаться другими программистами. Бесспорно, Perl является мощным инструментом системного администрирования. Но привлекательность Python возрастает с выходом за рамки обработки текста и файлов.

Scheme/Lisp

Подобно Scheme (и Lisp), Python поддерживает динамический контроль типов, поэтапную разработку и метапрограммирование; в нем открыт доступ к информации о состоянии интерпретатора и поддерживается создание программ на этапе исполнения. В отличие от Lisp, в Python используется процедурный синтаксис, знакомый пользователям таких основных языков, как C и Pascal. Когда требуется, чтобы конечные пользователи писали расширения, это становится важным преимуществом.

Smalltalk

Как и Smalltalk, Python поддерживает объектно-ориентированное программирование (ООП) в контексте динамического языка. В отличие от Smalltalk, Python не включает в систему объектов базовые конструкции управляющей логики программы. Чтобы работать с Python, пользователям не придется осваивать понятие оператора `if` как объекта, получающего сообщения, – Python более традиционен.

Icon

Как и Icon, Python поддерживает множество типов данных и операций высокого уровня, таких как списки, словари и срезы. Недавно появившиеся в Python инструменты, такие как итераторы и генераторы, похожи на механизм поиска с возвратом (backtracking) в языке Icon. В отличие от Icon, Python по существу прост. Программистам (и конечным пользователям) не нужно овладевать эзотерическими понятиями, такими как перебор с возвратом, чтобы начать работу.

BASIC

Как и у современных структурированных диалектов BASIC, у Python интерпретируемая/интерактивная природа. В отличие от большинства диалектов языка BASIC, в Python имеется стандартная поддержка развитых функций программирования, таких как классы, модули, исключения, типы данных высокого уровня и обобщенная интеграция с C. И подобно Visual Basic Python представляет собой кроссплатформенное решение, которое не контролируется коммерчески ориентированными фирмами.

Java

Подобно Java, Python является многоцелевым языком, поддерживающим ООП, исключения и модульную организацию программ и компилирующим программы в переносимый байт-код. В отличие от Java, Python имеет простой синтаксис, а встроенные типы данных обеспечивают более высокую скорость разработки. Программы на языке Python обычно в три-пять раз короче эквивалентных программ на языке Java.

C/C++

Подобно C и C++, Python является многоцелевым языком и может использоваться для решения долгосрочных стратегических задач системного программирования. В отличие от компилирующих языков в целом, Python прекрасно подходит также для решения тактических задач, как язык быстрой разработки. Программы на языке Python получаются короче, проще и гибче, чем те же программы, написанные на компилирующих языках. Например, благодаря отсутствию в программном коде Python объявлений типов данных или размеров он получается не только более кратким, но и может использоваться для решения более широкого круга задач.

Все эти и другие языки имеют свои достоинства и уникальные сильные стороны – на самом деле Python позаимствовал из таких языков большинство своих характеристик. Цель Python не состоит в том, чтобы заменить все другие языки – разные задачи требуют разных инструментов, и одной из главных идей Python является разработка в смешанной языковой среде. Но соединившиеся в Python передовые конструкции программирования и средства интеграции делают его выбор естественным в областях, о которых рассказывалось в этой книге, и многих других.

А как насчет того узкого места?

Вернемся к первоначальному вопросу: как можно облегчить разработку программ? В какой-то мере Python действительно является «всего лишь еще одним компьютерным языком». Несомненно, что с теоретической точки зрения язык Python не представляет множества радикальных нововведений. Так почему нас должен интересовать Python, когда уже есть столько языков?

Чем интересен Python и что может быть его крупным вкладом в дело разработки программ, так это не особенности синтаксиса или семантики, а взгляд на мир: сочетание инструментов в Python делает быструю разработку приложений вполне достижимой целью. Вкратце, Python способствует быстрой разработке благодаря наличию таких характеристик:

- Быстрый цикл разработки.
- Объектно-ориентированный язык очень высокого уровня.
- Средства интеграции, способствующие многоязыковым разработкам.

Более точно, Python борется с узким местом разработки программного обеспечения по четырем направлениям, рассмотренным в следующих разделах.

Python обеспечивает цикл разработки без промежуточных стадий

Цикл разработки на языке Python значительно короче, чем при использовании традиционных инструментов. В Python отсутствуют этапы компиляции и компоновки – программы на языке Python просто импортируют модули на этапе выполнения и пользуются содержащимися в них объектами. Благодаря этому программы можно запускать сразу после внесения в них изменений. А в тех случаях, когда можно осуществлять динамическую перезагрузку модулей, оказываются возможными изменение и перезагрузка частей выполняющейся программы без ее остановки. На рис. 21.1 показано воздействие Python на цикл разработки.

Поскольку программа на языке Python интерпретируется, продолжение разработки после внесения изменений в программе происходит без задержки. А так как синтаксический анализатор Python в основанных на Python системах является встроенным, можно легко модифицировать программы во время выполнения. Например, мы уже видели, как разработанные на Python программы с графическим интерфейсом позволяют разработчикам изменять программный код, обрабатывающий нажатие кнопки, при остающемся активным графическом интерфейсе – результат изменений можно видеть немедленно, если нажать кнопку снова. Не требуется останавливать программу и компилировать ее заново.

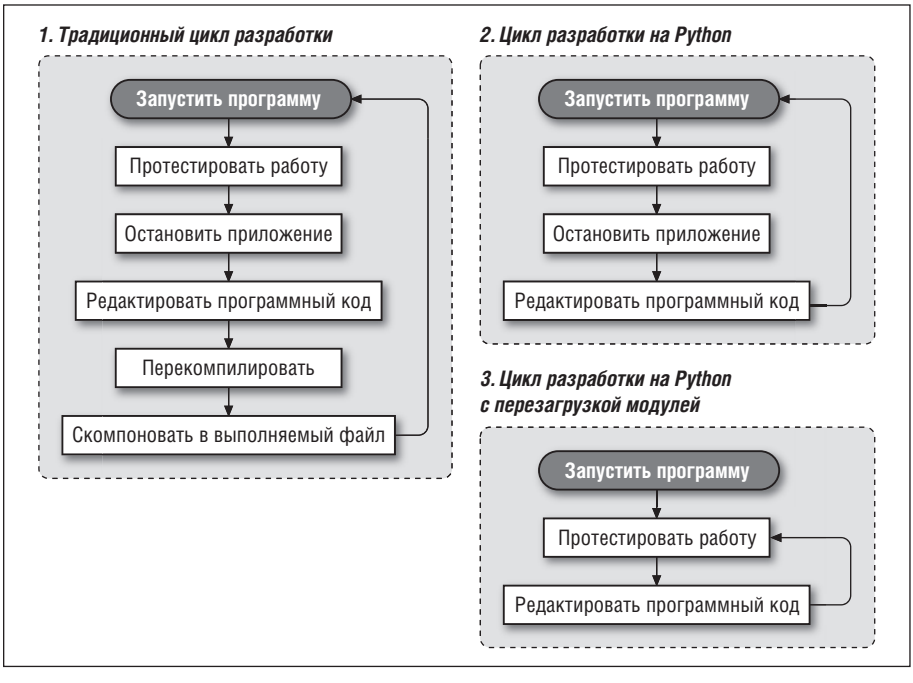


Рис. 21.1. *Циклы разработки*

В широком смысле весь процесс разработки на языке Python является практикой быстрого прототипирования. Python предоставляет возможность экспериментальной, интерактивной разработки программ и поощряет поэтапную разработку систем путем независимого тестирования компонентов и последующей их сборки. Мы видели на практике, что можно переключаться между тестированием компонентов (испытаниями программных единиц) и тестированием систем в целом (комплексными испытаниями) произвольным образом, как показано на рис. 21.2.

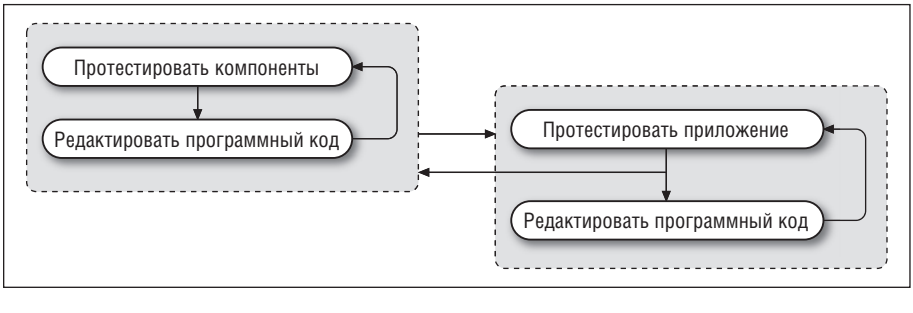


Рис. 21.2. *Поэтапная разработка*

Python является «выполняемым псевдокодом»

То, что Python является языком очень высокого уровня, означает, что остается меньше операций, которые нужно программировать и контролировать. Отсутствия этапов компиляции и сборки в действительности недостаточно для разрешения собственно проблемы узкого места цикла разработки. Например, наличие интерпретатора C или C++ могло бы обеспечить ускоренный цикл разработки, но все же оказалось бы почти бесполезным для быстрой разработки: сам язык слишком сложен и имеет низкий уровень.

Но так как Python к тому же простой язык, значительно ускоряется сам процесс разработки. Например, динамический контроль типов, встроенные объекты и уборка мусора в значительной мере устраняют необходимость вручную писать программный код для учета ресурсов, в отличие от таких языков низкого уровня, как C и C++. Ввиду отсутствия необходимости объявления типов, управления памятью и реализации стандартных структур данных программы Python обычно занимают лишь малую долю объема своих эквивалентов на C или C++. Писать и читать приходится меньше, а потому меньше возможности сделать ошибку в программе.

Благодаря отсутствию большей части программного кода учета ресурсов программы на языке Python легче понимать и они лучше отражают действительную задачу, которую призваны решать. А высокий уровень языка Python не только позволяет быстрее реализовывать алгоритмы, но и облегчает изучение языка.

Python – это правильное ООП

Чтобы объектно-ориентированное программирование (ООП) приносило пользу, оно должно быть простым в применении. Python увеличивает гибкость ООП, привнося его в динамический язык. Еще важнее, что механизм классов в нем представляет собой упрощенное подмножество C++, и именно такое упрощение обеспечивает полезность ООП в контексте инструмента быстрой разработки. Например, при обсуждении классов структур данных в этой книге, мы видели, что динамический контроль типов в Python позволяет применять один класс ко множеству типов объектов: нам не пришлось писать версии для каждого поддерживаемого типа. В обмен на отсутствие контроля типов Python получает гибкость и подвижность.

На самом деле в Python так просто применять ООП, что нет причин не использовать его в приложении почти всюду. Модель классов Python обладает достаточно мощными возможностями, позволяющими применять ее в сложных программах, но поскольку доступ к ним прост, это не помеха в решении нашей задачи.

Python способствует созданию гибридных приложений

Как было показано ранее, поддержка расширения и встраивания в Python дает возможность использовать его в многоязыковых системах. Без наличия хороших средств интеграции даже самый лучший язык быстрой разработки приложений оказывается «изолированным боксом», не слишком полезным в современных условиях ведения разработки. Но средства интеграции Python позволяют применять его в гибридных, многокомпонентных приложениях. Одним из последствий является то, что системы могут одновременно использовать преимущества быстрой разработки на Python и высокой скорости выполнения традиционных языков, таких как C.

Несмотря на то, что Python можно использовать как самостоятельный инструмент, тем не менее, он не навязывает такой режим. Напротив, Python поощряет интегрированный подход к разработке приложений. Поддерживая произвольное смешивание Python с традиционными языками, Python стимулирует целый ряд парадигм разработки в диапазоне от чистого прототипирования до чистой эффективности. В абстрактном виде это показано на рис. 21.3.

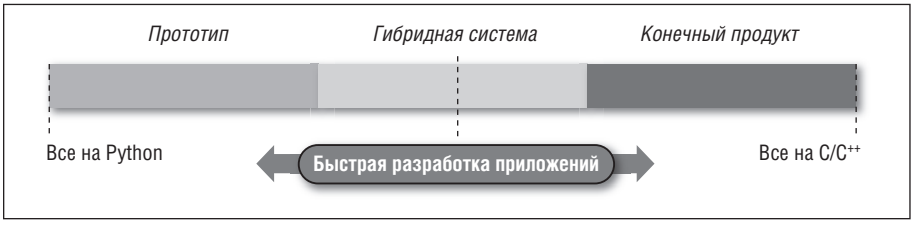


Рис. 21.3. *«Ползунок» режима разработки*

На левом краю спектра мы оптимизируем скорость разработки. При перемещении к правому краю оптимизируется скорость выполнения. Для каждого проекта оптимальная пропорция лежит где-то посередине. Python позволяет не только выбрать правильную пропорцию для проекта, но и изменять ее позднее произвольным образом при изменении потребностей:

Идем вправо

Разработку проектов можно начинать на Python с левого конца шкалы и постепенно перемещаться вправо, модуль за модулем, по мере необходимости оптимизируя конечный продукт.

Идем влево

Аналогично можно переносить важные части имеющихся приложений C или C++ в левый конец шкалы, перекладывая их на язык Python, чтобы обеспечить поддержку программирования и настройки продукта конечным пользователем.

Такая гибкость режима разработки важна в практических условиях. Python оптимизирован по скорости разработки, но одного этого недостаточно. Сами по себе ни C, ни Python не решают проблемы узкого места разработки; объединившись, они могут достичь значительно большего. Как показано на рис. 21.4, помимо самостоятельного использования одно из наиболее частых применений Python заключается в разбиении систем на клиентские составляющие, получающие преимущества простоты использования Python, и серверные модули, требующие производительности статических языков типа C, C++ или FORTRAN.

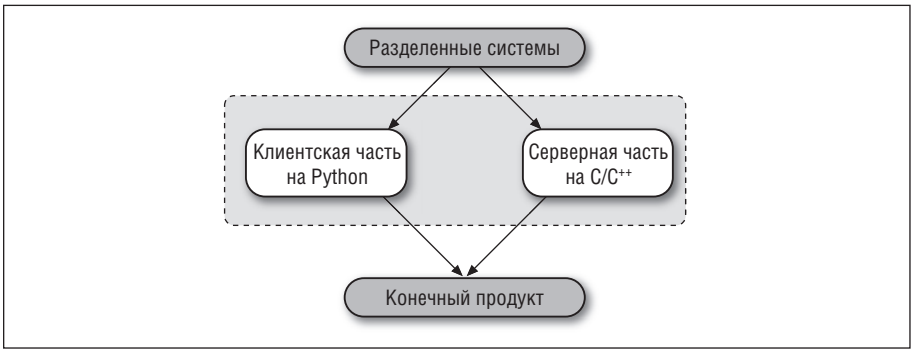


Рис. 21.4. *Гибридная архитектура*

Как при добавлении клиентских интерфейсов в существующие системы, так и при разработке их с начала такое разделение труда может открыть пользователям доступ к системе, не раскрывая ее внутреннего устройства.

При разработке новых систем также можно сначала целиком написать их на языке Python, а затем при необходимости оптимизировать конечный продукт, перенося критические для производительности компоненты на компилируемые языки. А так как модули Python и C выглядят для клиентов одинаково, переход на компилированные расширения оказывается прозрачным.

Не во всех ситуациях имеет смысл создавать прототипы. Иногда лучше заранее разделить систему на клиентскую часть на Python и серверную на C/C++. От прототипирования также мало пользы при доработке имеющихся систем. Но там, где его можно применять, прототипирование на раннем этапе может оказаться очень полезным. Путем предварительного создания прототипа на Python можно быстрее показать результаты. Вероятно, еще важнее то, что конечные пользователи могут активнее участвовать в ранних стадиях процесса разработки, как показано на рис. 21.5. В результате получаются системы, точнее соответствующие исходным требованиям.

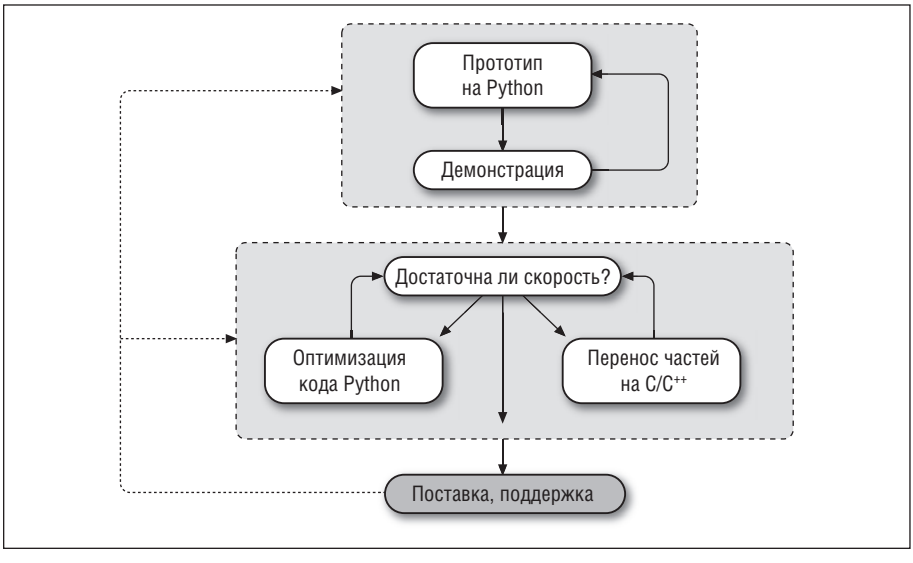


Рис. 21.5. *Создание прототипов на Python*

По поводу потопления «Титаника»

Проще говоря, Python в действительности представляет собой больше, чем язык; он предполагает определенную философию разработки. Понятия создания прототипов, быстрой разработки и гибридных приложений, разумеется, не новы. Но хотя преимущества таких способов разработки широко признаны, существовал недостаток инструментов для применения этих принципов на практике без потери мощности программирования. Это один из главных пробелов, которые заполняет архитектура Python: *Python предоставляет простой, но мощный язык для быстрой разработки приложений наряду со средствами интеграции, необходимыми для применения его в практических условиях разработки*.

Такое сочетание делает Python уникальным среди аналогичных инструментов. Например, Tcl служит хорошим инструментом интеграции, но не является развитым языком; Perl является мощным языком системного администрирования, но слабым инструментом интеграции. А тесное соединение в Python мощного динамического языка и интеграции открывает дорогу к более быстрым способам разработки. Благодаря Python больше не нужно выбирать между быстрой разработкой и быстрым выполнением.

К настоящему времени должно быть ясно, что один язык программирования не может удовлетворять всем потребностям разработки. На самом деле потребности иногда оказываются противоречивыми: эффективность и гибкость, по-видимому, всегда будут вступать в конфликт.

С учетом высокой стоимости создания программного обеспечения важно уметь выбирать между скоростью разработки и скоростью выполнения. Хотя машинное время дешевле времени программистов, нельзя полностью игнорировать эффективность программ.

Но имея такой инструмент, как Python, вообще не приходится выбирать между двумя целями. Так же как плотник не станет забивать гвоздь с помощью цепной пилы, так и разработчики программного обеспечения теперь достаточно вооружены, чтобы использовать нужный инструмент для решаемой в данный момент задачи: Python – если важна скорость разработки, компилируемые языки – если доминирует эффективность, и сочетание того и другого, когда цели не определены так однозначно.

Более того, нам не требуется жертвовать повторным использованием программного кода или полностью переписывать продукт перед поставкой, когда используется быстрая разработка с помощью Python. Быстрая разработка не отменяет преимуществ при эксплуатации:

Повторное использование

Так как Python является объектно-ориентированным языком высокого уровня, он способствует написанию повторно используемого программного обеспечения и правильно спроектированных систем.

Готовность продукта к поставке

Так как Python предназначен для использования в гибридных системах, нет необходимости переносить на более эффективные языки сразу весь программный код.

При типичной разработке на языке Python можно написать на нем интерфейсную часть и инфраструктуру системы, чтобы облегчить разработку и модификацию, а ядро, в целях эффективности, писать на C или C++. В таких системах Python называют верхушкой айсберга – это та часть, которая видна конечным пользователям, как на рис. 21.6.

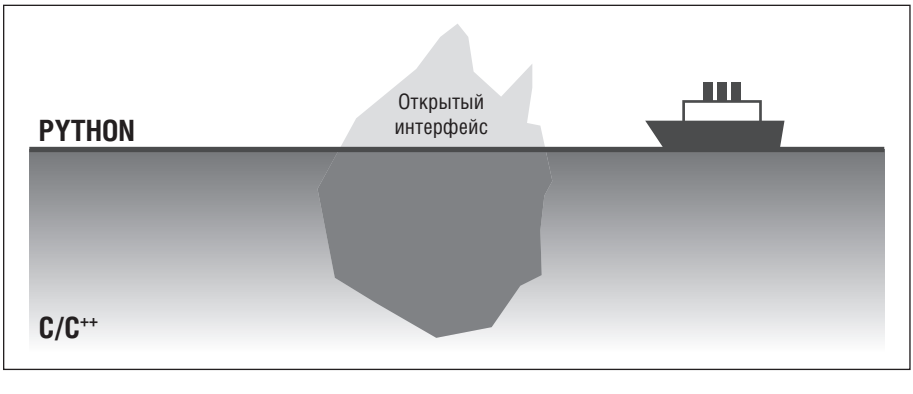


Рис. 21.6. *«Потопление Титаника» многоязыковыми системами*

В такой архитектуре используется лучшее от каждой из частей: ее можно расширить с помощью дополнительного программного кода Python или модулей расширения C, в зависимости от требований, предъявляемых к производительности. Вот один из многих возможных сценариев разработки с использованием нескольких языков:

Системные интерфейсы

Оформление библиотек в виде модулей расширения Python облегчает доступ к ним.

Настройка конечным пользователем

Передача программной логики встроенному программному коду Python позволяет вносить изменения на месте.

Чистое прототипирование

Прототипы на языке Python можно переводить на C все сразу или по частям.

Перенос существующего кода

Перенос существующего программного кода с C на Python делает его более простым и гибким.

Самостоятельное использование

Конечно, использование Python в самостоятельном виде усиливает имеющуюся библиотеку его инструментов.

Архитектура Python позволяет применять его любым способом, наиболее осмысленным для каждого конкретного проекта.

Так что же такое Python: продолжение

Как было показано в этой книге, Python является многогранным инструментом, применимым в самых разных областях. Что можно сказать о Python, подводя итоги? Рассматривая его лучшие качества, можно сказать, что язык Python:

- Универсальный
- Объектно-ориентированный
- Интерпретируемый
- Очень высокого уровня
- Открыто разрабатываемый
- Широко переносимый
- Доступный бесплатно
- Актуальный и согласованный

Язык Python полезен как для разработки самостоятельных приложений, так и для создания расширений, и оптимизирован для повышения *производительности разработчиков* по многим направлениям. Но дей-

ствительное значение Python должен определить сам читатель. Поскольку Python является универсальным инструментом, то, чем он «является», зависит от того, как вы станете его использовать.

Заключительный анализ...

Я надеюсь, что эта книга что-то рассказала вам о Python как о языке и о ролях, которые он играет. Однако ничто не заменит опыта практического программирования на Python. Обязательно возьмите какие-нибудь справочные материалы, которые помогут вам на этом пути.

Задача программирования компьютеров, вероятно, всегда будет сложной. Может быть и хорошо, что сохранится потребность в умных разработчиках программного обеспечения, искусных в переводе задач реального мира в форму, доступную для выполнения компьютерами, по крайней мере, в обозримом будущем. В конце концов, будь это просто, нам бы не платили денег. Никакой инструмент или язык не сможет полностью отменить программирование для реальных проектов.

Существующие в настоящее время практика разработки и инструменты добавляют сложности в наши задачи: многие препятствия, стоящие перед разработчиками, являются чисто искусственными. Мы многого достигли в повышении скорости работы компьютеров, пора сосредоточить внимание на скорости разработки. В эпоху постоянно уплотняющихся графиков производительность труда должна иметь первостепенное значение.

Python как реализация парадигмы совместного использования нескольких языков обладает потенциалом стимулировать применение методов разработки, при которых нарастает совокупный выигрыш – от быстрой разработки и от использования традиционных языков. Python не решит всех проблем индустрии разработки программного обеспечения, но он дает надежду на то, что программирование станет проще, быстрее и, по крайней мере, немного приятнее.

Возможно, он не гарантирует нам отплытие с того острова, но даст нам инструменты, отличные от бананов и кокосов.

Нравоучительная история о Perl и Python

(Следующий текст был размещен в телеконференции *rec.humor.funny* Ларри Хастингсом (Larry Hastings) и воспроизводится здесь с разрешения автора оригинала.)

Это прочно засевшая в моем сознании сцена из фильма «Империя наносит ответный удар», перефразированная, чтобы служить моральным уроком для честолюбивых программистов.

МЕСТО ДЕЙСТВИЯ: ДАГОБА – ДЕНЬ

Люк, с пристегнутым к его спине Йодой, карабкается вверх по толстой лозе, растущей на болоте, пока не добирается до лаборатории статистики планеты Дагоба. Тяжело дыша, он продолжает свои занятия – поиск, установку новых пакетов, регистрацию в качестве суперпользователя и замену командных сценариев двухлетней давности, написанных на Python.

ЙОДА: Код! Да. Сила программиста исходит из сопровождаемости кода его. Но берегись Perl. Сжатый синтаксис... больше одного способа сделать это... переменные по умолчанию. Темная сторона сопровождаемости кода это. Легко они текут, быстро приходят, когда ты пишешь код. Однажды по темному пути пойдя, навсегда судьбу свою свяжешь с ним, поглотит тебя он.

ЛЮК: А Perl лучше, чем Python?

ЙОДА: Нет... нет... нет. Быстрее, легче, соблазнительней.

ЛЮК: Но как я узнаю, почему Python лучше, чем Perl?

ЙОДА: Ты узнаешь. Когда пройдет полгода и код свой ты попробуешь прочесть.

Алфавитный указатель

A

Apache, веб-сервер, 29, 467
asyncore.py, модуль, 86

B

base64, модуль, 37
BASIC, язык программирования, 967
BeautifulSoup, стороннее расширение, 433, 451
binascii, модуль, 37
binhex, модуль, 37
Boost.Python, система, 925

C

CalcGui, класс, 868, 870
cenviron, модуль, 907
cgi, модуль
  escape, функция, 534, 551, 618, 628
  FieldStorage, класс, 509, 535, 618
  возможности, 37, 465
  и программа PyMailCGI, 592
  класс FieldStorage, 493
CGI-сценарии, 25, 460
  Hello World, программа, 530
  возможности, 462
  добавление взаимодействий с пользователем, 489
  добавление картинок, 486
  добавление таблиц, 486
  запуск, 466
  изменение размещения элементов формы ввода, 510
  использование cookies, 524
  и функции, 496
  определение, 460, 462
  отладка, 503
  первая веб-страница, 472
  передача параметров
    в адресах URL, 513, 520
    в скрытых полях форм, 516, 521
  преобразование строк, 503
  примеры, 479
  проблема декодирования текста, 214
  проверка данных, 538
  расширение модели, 528
  рефакторинг программного кода, 540
  соглашения по именованию, 482
  соглашения по экранированию, 551
  создание, 464
  сохранение информации о состоянии, 518
  табличная верстка формы, 499
  установка, 481
chmod, команда, 479
COM (Component Object Model – объектная модель компонентов), 29, 957
cookies
  вопросы безопасности, 650
  использование в сценариях CGI, 524
  обработка с помощью модуля urllib.request, 526
  определение, 522
  получение, 524
  создание, 523
CORBA, протокол
  возможности постоянного хранения, 676
  вопросы интеграции, 957
  поддержка ORB, 28
cregister, модуль, 947
ctypes, модуль
  возможности, 924
  вопросы интеграции, 892
  и двоичные данные, 118
CXX, система, 926
Cygwin, оболочка
  ветвление серверов, 60
Cygwin, система
  выполнение строк программного кода, 934
  простой класс C++, 915
  простой модуль на C, 896
Cython, система, 892, 925

C, язык программирования
getenv, функция, 905
putenv, функция, 905, 912
в сравнении с языком Python, 967
и инструмент SWIG, 899, 910
и классы, 952
интерфейс встраивания, 891, 928
интерфейс расширения, 891, 893
поддержка Python, 889
поиск совпадений с шаблонами в файлах заголовков, 824
создание оберток для функций, 905
стандартный API встраивания, 928
C++, язык программирования, 912, 927, 967

D

dbm, модуль, 678, 681
DBM, файлы
и ограничения модуля shelve, 699
и Юникод, 692
определение, 676, 677
особенности использования, 678
стандартные операции, 680
delete, команда (SQL), 719
distutils, пакет, 899
Django, фреймворк, 26, 513
DOM, парсеры, 826, 828, 831
Durus, система, 702

E

Earley, алгоритм, 838
ElementTree, пакет, 827
ElementTree, парсеры, 827, 828, 832
email, пакет, 177
Message, объекты, 205
анализ и составление электронных писем, 203
базовые интерфейсы, 208
возможности, 37, 178, 203
заголовки сообщений с адресами, 222
интернационализированные заголовки сообщений, 220
и программа PyMailCGI, 592, 624
и программа PyMailGUI, 361
кодировки для текстового содержимого, 213, 215
необходимость декодирования сообщений перед анализом, 212
обзор ограничений, 211
проблемы создания текста сообщений, 226, 231
Evaluator, класс, 870
eval, функция
поддержка в программе PyCalc, 864
поддержка синтаксического анализа, 838
преобразование строк, 503
эквивалент в C API, 930
exec, функция
поддержка в программе PyCalc, 864
поддержка синтаксического анализа, 838
преобразование строк, 503
эквивалент в C API, 930
Expat, парсер, 827

F

f2py, система, 926
FastCGI, расширение, 528
FieldStorage, класс, 493, 509, 535, 618
FORTRAN, язык программирования, 926
FTP (File Transfer Protocol), протокол
get и put, утилиты, 127
возможности, 121
FTP, объекты
cwd, метод, 170
delete, метод, 155
mkd, метод, 170
nlst, метод, 148, 155
ftplib, модуль
возможности, 36, 37
добавление пользовательского интерфейса, 136
передача деревьев каталогов, 168
передача каталогов, 144
передача файлов, 121

G

gcc, команда, 896
getaddresses, функция, 223
getenv, функция, 907
getfile, модуль, 135
на основе FTP, 129
примеры на стороне сервера, 472
примеры сценариев на стороне сервера, 571
getfile, сценарий
на основе сокетов, 105
getpass.getpass, метод, 124
Google App Engine, фреймворк, 26
Grail, броузер, 24
grid, менеджер компоновки, 111
система загрузки файлов, 111

H

Hello World, программа, 530
HList, виджет, 447
holmes, оболочка экспертной системы
  возможности, 807
  строки правил, 805
HTML
  извлечение текста в программе PyMailGUI, 369
  и интернет-приложения, 668
  конструирование веб-страниц, 472
  конфликты с URL, 557
  ограничение прав доступа к файлам, 479
  основы, 473
  поддержка синтаксического анализа, 28
  скрытые поля форм, 626
  соглашения по экранированию в сценариях CGI, 551
  теги таблиц, 489
  теги форм, 491
  экранирование текста сообщения и паролей, 628
html.entities, модуль, 836
HTMLgen, инструмент, 29
html.parser, модуль, 827, 834
  возможности, 37
  инструмент синтаксического анализа разметки HTML, 28
  проблемы извлечения данных, 298
HTTP, протокол
  доступ к веб-сайтам, 296
  поддержка cookies, 522
http.client, модуль, 37, 296
http.cookiejar, модуль, 37, 523, 526
http.cookies, модуль, 37, 523
HTTP_COOKIE, переменная окружения, 524
http.server, модуль, 37, 296
HTTPS (защищенный HTTP), 646

I

IANA (Internet Assigned Numbers Authority – полномочный комитет по надзору за номерами, используемыми в Интернете), 38
Icon, язык программирования, 967
IETF (Internet Engineering Task Force – рабочая группа инженеров Интернета), 38
imaplib, модуль, 37, 177
insert, команда (SQL), 714
IPC (Inter-Process Communication – взаимодействие между процессами)
  сокеты, 30
IronPython
  возможности разработки, 28
  вопросы интеграции, 892, 956
  обзор, 926

J

Java, язык программирования, 967
Jython, компилятор
  возможности разработки, 28
  вопросы интеграции, 892, 955
  обзор, 307, 926

K

kill, команда оболочки, 67
kwParsing, система, 838

L

LALR, парсеры, 838
LAMP, аббревиатура, 22
languages2common, модуль, 542
languages2reply, модуль, 549
Lisp, язык программирования, 966
listdir, функция
  получение списка файлов на удаленном сервере, 148
  сохранение локальных файлов, 155
LIST, команда (FTP), 148

M

M2Crypto, сторонний пакет, 650
mailconfig, модуль, 609
  и программа PyMailCGI, 589
  и программа PyMailGUI, 377
MailFetcher, класс, 262
MailParser, класс, 274
MailSender, класс, 252
MailTool, класс, 252
mailtools, вспомогательный пакет
  MailFetcher, класс, 262
  MailParser, класс, 274
  MailSender, класс, 252
  MailTool, класс, 252
  selftest.py, модуль, 249, 283
  и клиент pymail, 286
  и программа PyMailCGI, 592, 624
  и программа PyMailGUI, 348, 367
  обзор, 249
  файл инициализации, 250
marshal, модуль, 689

Message, объекты
get_content_charset, метод, 218
get_payload, метод, 215
возможности, 205
сообщения, состоящие из нескольких частей, 209
составление сообщений, 208
mimetypes, модуль
guess_extension, функция, 207
guess_type, функция, 207
возможности, 37
выбор режима передачи, 149, 155
mod_python, модуль, 29, 467, 529
Monty Python, музыкальная тема, 134
multiprocessing, модуль
и проблемы переносимости серверов сокетов, 71
MVC (model-view-controller – модель-представление-контроллер), архитектура, 26
mysql-python, интерфейс, 677

N

NLTK, комплект программ и библиотек, 840
nntplib, модуль, 293
возможности, 37
NNTP (Network News Transfer Protocol), протокол, 293
NumPy, расширение, 891

O

ORM (механизмы объектно-реляционного отображения)
возможности, 738
модель, 26
os.path, модуль
samefile, функция, 569
split, функция, 580
os, модуль
kill, функция, 67
stat, функция, 569
отображение переменных окружения
Env, объект, 909
доступ к переменным окружения, 907

P

Perl, язык программирования, 966, 977
pickle, модуль
Pickler, класс, 684
Unpickler, класс, 684
возможности, 37, 683
дополнительная информация, 808
ограничения, 700
playfile, модуль, 135
Plone, конструктор веб-сайтов, 26
PLY, система, 838
poplib, модуль
возможности, 37, 178
и программа PyMailCGI, 589, 624
и сценарий pymail, 239
сценарий чтения почты, 183
popmail, сценарий, 184
POP (Post Office Protocol – почтовый протокол)
и программа PyMailCGI, 611
и программа PyMailGUI, 367
модуль настройки электронной почты, 180
обзор, 179
сценарий чтения почты, 183
print, функция
и сценарии CGI, 506
PSP (Python Server Pages – серверные страницы Python), 29
putenv, функция, 907
putfile, модуль, 136
PyArg_Parse, функция API, 937, 940
Py_BuildValue, функция API, 940
PyCalc, программа
CalcGui, класс, 868, 870
Evaluator, класс, 870
возможности, 866
добавление новых кнопок, 883
использование, 867
исходный программный код, 874
как компонент, 881
построение графического интерфейса, 862
расширение и прикрепление, 865
Py_CompileString, функция API, 944
PyCrypto, система, 646
PyDict_GetItemString, функция API, 929, 942
PyDict_New, функция API, 929, 941
PyDict_SetItemString, функция API, 929, 941
PyErrata, веб-сайт, 670
PyEval_CallObject, функция API, 929, 940
PyEval_EvalCode, функция API, 944
PyEval_GetBuiltins, функция API, 942
PyForm, пример, 740
PyFort, система, 926
PyImport_GetModuleDict, функция API, 929

PyImport_ImportModule, функция API, 929, 937, 940
Py_Initialize, функция API, 937
pymail, сценарий
почтовый клиент командной строки, 238
возможности, 314
обновление, 286
PyMailCGI, программа
вспомогательные модули
внешние компоненты, 643
настройки, 643
обзор, 642
общий вспомогательный модуль, 655
потоки ввода-вывода, 655
шифрование паролей, 645
и модуль cgi, 592
корневая страница, 598
настройка, 601
новое в версии для третьего издания, 593
новое в версии для четвертого издания, 590
обзорное представление, 595
обзор реализации, 586
обработка загруженной почты
обзор, 630
операция удаления и номера POP, 637
ответ и пересылка, 632
удаление, 632
общая информация, 585, 586
опробование примеров, 595
отправка почты по SMTP
единство внешнего вида, 608
обзор, 602
страница составления сообщений, 602
страницы с сообщениями об ошибках, 607
сценарий отправки, 603, 609
передача информации о состоянии, 587
чтение почты по протоколу POP
передача информации о состоянии, 626
протоколы защиты данных, 620
страница ввода пароля, 611
страница выбора почты из списка, 613
страница просмотра сообщений, 622
экранирование текста сообщений и паролей, 628
PyMailGUIHelp, модуль, 440
PyMailGUI, программа
автономная работа, 344
возможности, 330
загрузка почты, 338
запуск, 316, 331
идеи по усовершенствованию, 447
извлечение простого текста из разметки HTML, 837
интернационализация содержимого электронной почты, 326
интерфейс загрузки с сервера, 343
и программа PyMailCGI, 662
компоненты
altconfigs, каталог, 377, 444, 449
html2text, модуль, 430, 448
ListWindows, модуль, 387, 449
mailconfig, модуль, 338, 348, 377, 433, 449
messagecache, модуль, 421
poputil, модуль, 425
PyMailGUIHelp, модуль, 440
SharedNames, модуль, 385
textConfig, модуль, 440
ViewWindows, модуль, 409
wraplines, модуль, 427
главный модуль, 382
обзор реализации, 380
многооконный интерфейс, 377
многопоточная модель выполнения, 339
модули с исходными текстами и их объем, 310
обзор, 314
обработка содержимого электронной почты в формате HTML, 369
основные изменения, 318
особенности адресации, 359
ответ на сообщения и пересылка, 359
отправка почты и вложений, 347
поддержка интернационализации содержимого почты, 371
поддержка протокола POP, 367
политика поддержки Юникода, 326
просмотр почты и вложений, 351
синхронизация, 367
сообщения о состоянии, 377
стратегия представления, 317
удаление сообщений, 365
PyModule_GetDict, функция API, 929, 937
PyObject_CallObject, функция API, 929

PyObject_GetAttrString, функция API, 929, 940
PyObject_SetAttrString, функция API, 929, 937
PyParsing, система, 838
PyPI, веб-сайт, 29
Pyrex, система, 892, 925
PyRun_File, функция API, 929
PyRun_SimpleString, функция API, 933
PyRun_String, функция API, 929, 937, 941, 942
PySerial, интерфейс, 118
PythonInterpreter, класс API, 956
PYTHONUNBUFFERED, переменная окружения, 484
Python, язык программирования
  возможности разработки сценариев для Интернета, 25
  в сравнении с другими языками, 966
PyTree, программа, 793, 858
PyWin32, пакет
  возможности разработки, 29
PyXML SIG, заинтересованная группа, 833

Q

QUERY_STRING, переменная окружения, 551
quopri, модуль, 37

R

Register_Handler, функция, 947
RETR, строка (FTP), 128
re, модуль
  compile, функция, 811, 816
  escape, функция, 817
  findall, функция, 812, 814, 817
  finditer, функция, 817
  match, функция, 811, 814, 817
  search, функция, 812, 814, 817
  split, функция, 817
  subn, функция, 817
  sub, функция, 817
  возможности, 810
rfc822, модуль, 208
RFC822, спецификация, 176
RIA (Rich Internet Application – полнофункциональные интернет-приложения), 669
rotor, модуль, 646, 647
Route_Event, функция, 947

S

SAX, парсеры, 826, 828, 829
Scheme, язык программирования, 966
SciPy, пакет, 926
select, команда (SQL), 716
select, модуль
  мультиплексирование серверов, 79
sendmail, программа, 191
shelve, модуль, 690
  уникальность объектов, 698
Smalltalk, язык программирования, 966
smtplib, модуль
  возможности, 37, 178, 192
  и программа PyMailCGI, 589
  и программа PyMailGUI, 348, 456
  и сценарий pymail, 239
smtpmail, сценарий, 192
SMTP (Simple Mail Transfer Protocol), протокол
  и программа PyMailCGI, 602
  обзор, 190
  отправка почты из интерактивной оболочки, 202
  стандарт форматирования даты и времени, 201
  сценарий отправки почты, 192
SOAP, протокол
  возможности постоянного хранения, 676
  вопросы интеграции, 958
  поддержка веб-служб, 27
socketserver, модуль, 37, 76
socket, модуль
  возможности, 37
  программирование, 39
SPARK, система, 838
SQLAlchemy, система, 709
SQLite, система баз данных
  введение, 712
  выполнение запросов, 716
  выполнение обновлений, 718
  добавление записей, 714
  обзор, 712
  создание баз данных и таблиц, 713
SQLObject, система, 709, 738
SSL (Secure Sockets Layer – уровень защищенных сокетов), 620
ssl, модуль, 37, 620
stdin, стандартный поток ввода
  и сценарии CGI, 465
stdout, стандартный поток вывода
  и сценарии CGI, 465
STOR, строка (FTP), 128

string, модуль
ascii_uppercase, константа, 799
дополнительная информация, 808
строковые методы, 799
strop, модуль, 808
struct, модуль
возможности, 37
и последовательные порты, 118
str, тип объектов
обработка шаблонов с помощью операций замены и форматирования, 800
строковые методы, 798
SWIG, инструмент, 899, 910, 912
sys.stderr, стандартный поток ошибок
и программа PyMailCGI, 655
и сценарии CGI, 504
sys.stdout, стандартный поток вывода
и программа PyMailCGI, 655

T

Tcl, язык программирования, 966
telnetlib, модуль, 37
Telnet, служба, 144, 179
testparser, модуль, 855
Text, класс виджетов
и программа PyMailGUI, 369, 450
Tix, расширение
HList, виджет, 447
TkinterTreectrl, стороннее расширение, 447
traceback, модуль
и программа PyMailCGI, 607
и сценарии CGI, 505
Trigger_Event, функция, 947
Turbo Gears, коллекция инструментов, 26
Twisted, фреймворк, 29
возможности разработки серверов, 86

U

Unix, платформа
изменение прав доступа, 479
предотвращение появления зомби, 66
update, команда (SQL), 719
URL, адреса
и программа PyMailGUI, 448
конфликты с HTML, 557
минимальные, 477
передача параметров, 495, 513, 520, 617
синтаксис, 474
соглашения по экранированию в сценариях CGI, 551
чтение почты, 621
urllib, пакет
загрузка файлов, 125
и программа PyMailCGI, 589
создание клиентских сценариев, 300
urllib.parse, модуль
quote_plus, функция, 618
urlencode, функция, 618
возможности, 37
вызов программ, 304
и программа PyMailCGI, 618
и сценарии CGI, 465
экранирование адресов URL, 553
urllib.request, модуль
urlopen, функция, 498
urlretrieve, интерфейс, 302
возможности, 37, 300
загрузка файлов, 125
и анализ HTML, 834
и программа PyMailCGI, 609
и сценарии CGI, 467
обработка cookies, 526
поддержка cookies, 523, 526
тестирование без броузера, 497
uu, модуль, 37

W

weave, система, 926

X

xdrlib, модуль, 37
xml, пакет
возможности, 826
возможности разработки, 28
xml.etree, пакет, 832
xmlrpc, пакет, 307, 827
XML-RPC, протокол
возможности постоянного хранения, 676
вопросы интеграции, 958
поддержка веб-служб, 27

Y

YAPPS, генератор парсеров, 838

Z

zip, функция, суммирование, 803
ZODB, система
возможности, 702

особенности использования, 704
Zope, библиотека инструментов, 26, 513

А

адреса электронной почты, 222
алгебра отношений, 770
анализ
необходимость декодирования сообщений перед анализом, 212
содержимого электронных писем, 203
анализ HTML, 834
анализ XML, 827
с помощью парсера DOM, 831
с помощью парсера ElementTree, 832
с помощью парсера SAX, 829
с помощью регулярных выражений, 828
сторонние инструменты, 833

Б

базы данных, 707
на стороне сервера, 527
свободно распространяемые интерфейсы, 708
создание с помощью SQLite, 713
базы данных SQL
возможности, 707, 709
вспомогательные сценарии SQL, 729
дополнительные ресурсы, 737
загрузка таблиц базы данных из файлов, 725
обзор интерфейса, 709
создание словарей записей, 719
учебник по API на примере SQLite, 712
безопасность
и программа PyMailCGI, 601
и шифрование паролей, 645
библиотечные модули, 36
брокер объектных запросов (Object Request Broker, ORB), 28
броузеры
Grail, 24
и сценарий отправки почты в PyMailCGI, 609
поддержка cookies, 522
тестирование с помощью модуля urllib.request, 497
буферизация
и каналы, 102
построчная, 99
потоков вывода, 95

В

веб-сайты
вопросы проектирования, 513
доступ, 296
веб-серверы
запуск CGI-сценариев, 466
запуск локального сервера, 467
корневая страница с примерами, 470
веб-службы, 27
веб-страницы
конструирование в CGI-сценариях, 472
совместное использование объектов, 542
веб-сценарии
альтернативные решения, 667
и PyMailGUI, 662
и настольные приложения, 663
преимущества и недостатки, 661
веб-фреймворки, 25
ветвление процессов
и сокеты, 63
серверы, 58
виджеты
PyCalc, программа, 881
вложения
отправка с помощью PyMailCGI, 592
отправка с помощью PyMailGUI, 347
просмотр с помощью программы PyMailGUI, 351
распространение, 453
вложенные структуры
выгрузка локального дерева каталогов, 168
сериализация, 685
вспомогательные модули
интерфейс к протоколу POP, 644
вызываемые объекты
обзор, 939
определение, 930
регистрация, 945
выполнение запросов
с помощью SQLite, 716
выполнение программ
PyMailGUI, 316
pymail, почтовый клиент командной строки, 290
сценарии CGI, 482
сценарий командной строки pymail, 244

Г

Гвидо ван Россум, 24

генераторы списков, суммирование, 803
гиперссылки, 448, 474, 513
графические интерфейсы
построение интерфейса для PyCalc, 860, 862
графы
определение, 779
перевод графов на классы, 782
реализация, 780

Д

двоичные деревья, 774
встроенные возможности, 774
определение, 774
реализация, 775
двоичные дистрибутивы, 452
деревья каталогов
передача с помощью ftplib, 168
дочерние программы
пример с сокетами, 52
дочерние процессы
завершение, 63

З

заголовки сообщений
mailtools, вспомогательный пакет, 253, 274
интернационализированные, 220
передача в скрытых полях форм, 638
с адресами, 222
загрузка деревьев каталогов с сервера, 176
записи
добавление с помощью SQLite, 714
создание словарей, 719
запросы
автоматизация, 722
параметры, 476
запуск программ
серверные сценарии, 466
зарезервированные номера портов, 56

И

идентификаторы компьютеров, 31
изображения
добавление в сценариях CGI, 486
имена компьютеров, 31
инструмент чтения почты, 180, 183
интеграция встраиванием
API языка C, 928
определение, 891
основные приемы, 932
регистрация объектов для обработки обратных вызовов, 945
интеграция расширением
SWIG, инструмент, 899
другие инструменты создания расширений, 923
обзор, 893
определение, 891
простой класс C++, 913
простой модуль на C, 894
интернационализация
поддержка для содержимого электронной почты, 326, 371
интерфейсы к последовательным портам, 118
информация о состоянии
SAX, парсеры, 827
и интернет-приложения, 668
комбинирование приемов сохранения, 529
передача в программе PyMailCGI, 587, 626
сохранение в базах данных на стороне сервера, 527
сохранение в сценариях CGI, 518
источники документации
об анализе XML, 833

К

каналы
и буферизация, 102
и сокеты, 103
каталоги
передача деревьев каталогов с помощью ftplib, 168
передача каталогов с помощью ftplib, 144
соглашения по именованию, 482
классы
изменение классов хранимых объектов, 696
и множества, 765
и стеки, 749
и язык программирования C, 952
перевод графов, 782
реорганизация сценариев на основе классов, 163
клиент/сервер, архитектура
определение, 34
передача файлов, 561
клиентские сценарии
mailtools, вспомогательный пакет, 249

urllib, пакет, 300
анализ и составление электронных писем, 203
возможности разработки, 25
доступ к веб-сайтам, 296
доступ к телеконференциям, 293
и интернет-приложения, 667
и программа PyMailGUI, 309
и протоколы, 34
используемые методы сокетов, 45
обработка электронной почты, 176
обслуживание нескольких клиентов, 58
отправка электронной почты по SMTP, 190
параллельный запуск нескольких клиентов, 52
передача деревьев каталогов, 168
передача каталогов с помощью ftplib, 144
передача файлов с помощью ftplib, 121
передача файлов через Интернет, 121, 125, 127
поддержка в Python, 120
почтовый клиент командной строки, 238
прочие возможности создания, 306
чтение электронной почты по протоколу POP, 179
кнопки
добавление в программу PyCalc, 883
кодировки текстового содержимого, 213, 215, 274
колонки в файле, суммирование, 802
конец строки, символ
сценарии CGI, 483

Л

лексический анализ, 837

М

минимальные адреса URL, 477
множества
алгебра отношений, 770
встроенные возможности, 761
и классы, 765
и функции, 763
определение, 760
перевод на использование словарей, 766
поддерживаемые операции, 760
модуль настройки электронной почты, 180
мультиплексирование серверов, 79

Н

наследование
классов C++, 920
номера портов
зарезервированные, 56
определение, 31
правила протоколов, 33

О

обертывание
классов C++ с помощью SWIG, 912
облачные вычисления, 26
обработка без подключения к Интернету, программа PyMailGUI, 344
обработка текста и синтаксический анализ
PyCalc, программа, 860
XML и HTML, 826
парсеры, написанные вручную
грамматика выражений, 840
с помощью регулярных выражений, 809
стратегии, 797
строковые методы, 798
обработка текстов на естественных языках, 840
обработчики обратных вызовов
регистрация объектов, 945
обработчики сигналов, 66
объектно-ориентированные базы данных, 676
объектно-реляционные отображения (ORM)
возможности, 676
объект соответствия (модуль re), 811
объект файла
readlines, метод, 802
объект шаблона (модуль re), 811
объекты
вызываемые, 930, 939, 945
ограничения модуля shelve, 698
преобразование в строки, 682
сериализованные, 676, 682
совместное использование разными страницами, 542
ООП (объектно-ориентированное программирование)
вопросы программирования, 970
оптимизация

непосредственная модификация списка в памяти, 755
перевод множеств на использование словарей, 766
стеки в виде деревьев кортежей, 753
отладка сценариев CGI, 503

П

параллельная обработка
и проблемы переносимости серверов сокетов, 71
параллельный запуск нескольких клиентов, 52
параметры
запроса, 476
передача в адресах URL, 495, 513, 520
передача в скрытых полях форм, 516, 521
пароли
страница ввода пароля PyMailCGI, 611
шифрование, 645
экранирование в HTML, 628
передача в двоичном режиме, 156
передача в текстовом режиме, 155
передача деревьев каталогов, 168
передача каталогов, 144
передача файлов
между клиентами и серверами, 561
с помощью ftplib, 121
через Интернет, 121, 125, 127
переменные оболочки
создание оберток для функций, 905
подклассы
рекурсивная выгрузка, 169
полнофункциональные интернет-приложения (Rich Internet Application, RIA), 26, 669
последовательности
обращение и сортировка, 787
перестановки, 785
постоянное хранение данных
ZODB, система, 702
базы данных SQL, 707
возможности, 676
и файлы shelve, 690
механизмы объектно-реляционного отображения, 738
сериализованные объекты, 682
построчная буферизация, 99
потоки ввода-вывода
CGI-сценарии, 465
вспомогательный модуль перенаправления потоков ввода-вывода, 89
и программа PyMailCGI, 655
придание сокетам внешнего вида потоков ввода-вывода, 88
потоки выполнения
и программа PyMailGUI, 322, 339
и серверы, 73
права доступа
ограничение, 479
преобразование строк, 503
проверка орфографии, 452
программирование на языке Python
возможности разработки сценариев для Интернета, 25
программирование сокетов, 38
производительность
и программа PyMailCGI, 663
и программа PyMailGUI, 456, 663
и стеки, 752, 760
и строковые методы, 808
протоколы
возможности разработки, 25
и модуль pickle, 688
обзор стандартов, 38
определение, 32
правила нумерации портов, 33
структура, 35
процессы-зомби, 64
путь поиска, сценарии CGI, 483

Р

разрешения
и сценарии CGI, 568
распространение вложений, 453
регулярные выражения
re, модуль, 810, 816
и строковые операции, 813
ограничения, 837
определение, 809
поиск совпадений с шаблонами в файлах заголовков, 824
приемы сопоставления, 809
примеры шаблонов, 822
синтаксис шаблонов, 817
синтаксический анализ, 828
реорганизация программного кода
с применением классов, 163
с применением функций, 158
рефакторинг программного кода
в сценариях CGI, 540

С

связанные методы
поддержка в программе PyMailGUI, 341
серверные сценарии
ветвление серверов, 58
и протоколы, 34
используемые методы сокетов, 42
многопоточные серверы, 73
мультиплексирование серверов с помощью select, 79
простой файловый сервер на Python, 104
серверы
привязка сокетов к зарезервированным портам, 57
сериализация, 683
сериализованные объекты
определение, 676, 683
сетевые сценарии
возможности разработки, 25
и библиотечные модули, 36
и протоколы, 32
и сокеты, 30, 38
обслуживание нескольких клиентов, 58
придание сокетам внешнего вида файлов и потоков ввода-вывода, 88
простой файловый сервер на Python, 104
синтаксический анализ
HTML, 826
XML, 826
обзор, 28
дополнительные инструменты синтаксического анализа, 837
методом рекурсивного спуска, 840
определение, 796
парсеры, написанные вручную
грамматика выражений, 841
добавление интерпретатора дерева синтаксического анализа, 850
и возможности Python, 859
и графический интерфейс PyTree, 858
реализация, 842
структура дерева синтаксического анализа, 856
с помощью методов split и join, 801
с помощью регулярных выражений, 828
строк правил, 805
синхронизация
mailtools, вспомогательный пакет, 265
и программа PyMailGUI, 367
система загрузки файлов, 105
скрытые поля
передача параметров, 516, 521
скрытые поля форм
передача информации о состоянии, 626
передача текста заголовка, 638
словари
выполнение строк программного кода, 941
перевод множеств на использование, 766
реализация поиска на графе, 780
создание, 719
суммирование с помощью словарей, 804
события
возбуждение, 947
маршрутизация, 947
соглашения по именованию
сценарии CGI, 482
соединения, объект (FTP)
cwd, метод, 170
mkd, метод, 170
retrbinary, метод, 123, 129, 149
retrlines, метод, 129, 149
storbinary, метод, 133
storlines, метод, 133
создание
словарей записей, 719
сценариев CGI, 464
сокетов, объекты
bind, метод, 42
close, метод, 46
connect, метод, 46
listen, метод, 43
makefile, метод, 88
send, метод, 46
setblocking, метод, 85
создание, 42
сокеты
базовые возможности, 30, 40
возможности разработки, 25
запуск программ
на локальном компьютере, 47
на удаленном компьютере, 48
и ветвление процессов, 63
идентификаторы компьютеров, 31
и каналы, 103
определение, 19

параллельный запуск нескольких клиентов, 52
передача строк байтов и объектов, 44
подключение к зарезервированным портам, 56
практическое использование, 50
придание сокетам внешнего вида файлов и потоков ввода-вывода, 88
программирование, 38
спам, 197, 449
списки
как стеки, 745
непосредственная модификация в памяти, 755
стеки
Stack, класс, 749
stack, модуль, 747
в виде деревьев кортежей, 753
встроенные возможности, 745
вычисление выражений с применением, 870
как списки, 745
настройка и мониторинг производительности, 752
непосредственная модификация списка в памяти, 755
определение, 744
хронометраж усовершенствований, 757
строки, 809
ограничения модуля shelve, 698
преобразование, 503
строки правил, синтаксический анализ, 805
строки программного кода
выполнение, 933
выполнение в словарях, 941
выполнение с использованием результатов и пространств имен, 937
определение, 930
предварительная компиляция, 943
строковые методы
endswith, метод, 799
find, метод, 798
format, метод, 798
isdigit, метод, 799
isupper, метод, 799
join, метод, 798, 801
replace, метод, 798
rjust, метод, 799
rstrip, метод, 799
split, метод, 798, 801
startswith, метод, 799
strip, метод, 798
upper, метод, 799
и производительность, 808
и регулярные выражения, 813
структуры данных
в сравнении со встроенными типами, 791
графы, 779
двоичные деревья, 774
обращение и сортировка последовательностей, 787
перестановки последовательностей, 785
реализация множеств, 760
реализация стеков, 744
создание подклассов встроенных типов, 771
структуры протоколов, 35
сценарии
автоматизация выполнения запросов, 722
вспомогательные, SQL, 729
реорганизация с применением
классов, 163
функций, 158
сценарии для Интернета
возможности разработки, 25
и библиотечные модули, 36
и протоколы, 32
и сокеты, 30, 38
обслуживание нескольких клиентов, 58
придание сокетам внешнего вида файлов и потоков ввода-вывода, 88
простой файловый сервер на Python, 104
сценарии на стороне сервера, 460
запуск, 466
корневая страница с примерами, 470
обзор, 460
передача файлов между клиентами и серверами, 561
сценарий отправки почты, 192

Т

таблицы
верстка форм, 499
загрузка из файлов, 725
использование описаний, 720
поддержка ORM, 676
создание, 486
создание с помощью SQLite, 713
сценарий вывода таблицы, 731

сценарии загрузки таблиц, 730
теги
таблиц, 489
форм, 491
текстовые файлы
и буферизация потоков вывода, 95
телеконференции
доступ, 293
обработка сообщений, 455
типы объектов, хранимые в хранилищах shelve, 693

У

удаление деревьев каталогов на сервере, 172
удаленные сайты
выгрузка каталогов, 153
загрузка деревьев каталогов с сервера, 176
загрузка каталогов, 144
удаление деревьев каталогов на сервере, 172
удаленные серверы, 52, 485
уровень защищенных сокетов (Secure Sockets Layer, SSL), 620

Ф

файлы
возможности постоянного хранения, 676
загрузка в таблицы, 725
закрытые, 568
передача между клиентами и серверами, 561
передача файлов с помощью ftplib, 121
передача файлов через Интернет, 121, 125, 127
придание сокетам внешнего вида файлов, 88
соглашения по именованию, 482
с программным кодом, 930
суммирование по колонкам, 802
файлы shelve
writeback, аргумент, 692
ограничения, 698
определение, 676, 690
особенности использования, 691
сохранение объектов встроенных типов, 693
сохранение экземпляров классов, 694
стандартные операции, 692
файлы с байт-кодом
предварительная компиляция строк программного кода, 943
фактор Гиллигана, 963
фильтрация спама, 449
формы
action, параметр, 491
method, параметр, 491
изменение размещения элементов, 510
имитация введенных данных, 535
многократно используемая утилита имитации, 545
передача параметров в скрытых полях, 516, 521
поля ввода, 491
совместное использование объектов разными страницами, 542
табличная верстка, 499
теги HTML, 491
функции
и множества, 763
и сценарии CGI, 496
реорганизация сценариев на основе функций, 158

Х

хеширования, прием, 678

Ч

чтение с экрана, 498

Ш

шифрование паролей, 645

Э

экземпляры классов
и постоянное хранение, 676
сериализация, 687
сохранение в хранилищах shelve, 694
электронная почта
mailtools, вспомогательный пакет, 249
анализ и составление электронных писем, 203
загрузка, 338
и интернационализация, 326, 371
и поддержка Юникода, 177
обработка содержимого в формате HTML, 324, 369
обработка через Интернет, 176
особенности адресации, 359

ответ на сообщения и пересылка, 359
отправка вложений, 347, 592
отправка из интерактивной оболочки, 202
отправка по SMTP, 190, 602
ошибка рассинхронизации с почтовым ящиком, 637
почтовый клиент командной строки, 238
просмотр вложений, 351, 593
удаление, 365
чтение по протоколу POP, 179
чтение с использованием прямых адресов URL, 621

Ю

Юникод
mailtools, вспомогательный пакет, 253, 263, 274
заголовки сообщений с адресами, 222
интернационализированные заголовки сообщений, 220
и программа PyMailCGI, 606
и программа PyMailGUI, 326, 457
и электронная почта, 177
кодировки для текстового содержимого, 213
необходимость декодирования сообщений перед анализом, 212
обзор ограничений пакета email, 211
проблемы создания текста сообщений, 226, 231