有趣、好玩、易学、有深度的
Python
网络爬虫课程

LEARNING PYTHON
WEB CRAWLER

Python
网络爬虫实例教程

视频讲解版

齐文光 编著

有趣的 Python 爬虫课程，**通俗**的讲解，**有深度**的实战
Python 进阶的**好途径**

扫一扫书中**二维码**，跟着视频**轻松学**
视频时长超过 **400** 分钟

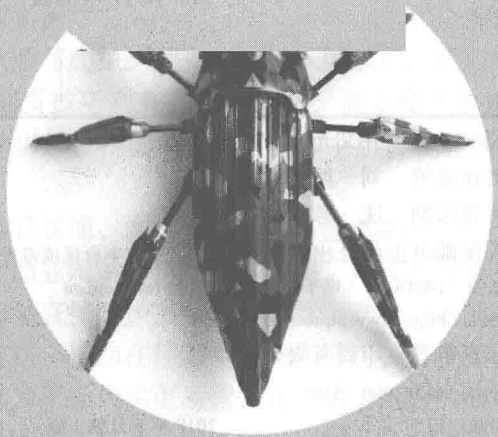
中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

Python
网络爬虫课程

LEARNING PYTHON
WEB CRAWLER

Python
网络爬虫实例教程

视频讲解版



人民邮电出版社
北京

图书在版编目（CIP）数据

Python网络爬虫实例教程 ：视频讲解版 / 齐文光编著. -- 北京 ：人民邮电出版社，2018.8
ISBN 978-7-115-48465-9

Ⅰ. ①P… Ⅱ. ①齐… Ⅲ. ①软件工具－程序设计
Ⅳ. ①TP311.561

中国版本图书馆CIP数据核字(2018)第101074号

内 容 提 要

本书共 16 章，详细介绍爬虫的基础知识、编写简单定向爬虫和使用 Scrapy 爬虫框架。第 1～3 章介绍爬虫的基础知识和网页解析基础；第 4～8 章用实例演示编写定向爬虫、模拟登录、应对反爬虫和爬取动态网页等；第 9 章介绍 Scrapy 基础知识；第 10、第 11 章讲解两个最常用的 Scrapy 爬虫类；第 12、第 13 章讲解 Scrapy 应对反爬虫、向网站提交数据和登录网站的方法；第 14 章用实例演示存储数据到数据库；第 15 章简单讲解爬虫去重、分布式爬虫编写和爬虫部署；第 16 章为综合实例，并且简单介绍爬取数据的分析。本书运用大量实例为读者演示编写爬虫的技巧，每一章都包含本章小结及要求，以帮助读者巩固所学内容。

本书面向对爬虫技术感兴趣的读者，介绍使用 Python 语言编写爬虫的各种技巧和方法。对希望深入学习 Python 编程的初学者，本书也很适合作为进阶读物。

◆ 编　　著　齐文光
责任编辑　刘　博
责任印制　沈　蓉　彭志环

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京市艺辉印刷有限公司印刷

◆ 开本：800×1000　1/16
印张：13.5　　2018 年 8 月第 1 版
字数：268 千字　　2018 年 8 月北京第 1 次印刷

定价：49.80 元

读者服务热线：(010)81055256　印装质量热线：(010)81055316
反盗版热线：(010)81055315

前言

爬虫技术是一门非常有趣、有用、易学、易令人产生成就感的技术。人们利用爬虫技术可以下载感兴趣的图片、小说，可以自动化地完成很多需要人工操作的事情，如定时抢购某件商品。对企业来讲，爬虫的作用显得更加重要，很多公司依赖于爬虫技术获取公开数据，为企业发展提供服务。在招聘网站上，爬虫工程师的薪酬非常高。

爬虫技术学起来容易上手，相信各位读者看完第 4 章的基础爬虫实例，就可以编程爬取很多网站，这对建立信心、激发学习兴趣非常关键，从这个角度看，爬虫技术也非常适合作为学习编程语言的进阶内容。

虽然爬虫技术易学、有用、有趣，但要真正系统地掌握爬虫技术，能够独立地解决数据获取过程中遇到的难题，还需要深入、系统地掌握爬虫知识。经常有读者觉得爬虫教学用例繁杂，技巧介绍不明确，学习起来很难掌握；或者内容比较片面，难以把学习的例子应用到其他网页爬取中。针对以上问题，本书在编写过程中，特别注重两点：一是简单易学，二是系统深入。本书为了简单明了地向读者介绍编写爬虫的技巧，着重选择那些既能体现编写技巧，页面又相对干净的例子；为了让读者能够比较爬虫框架与手写爬虫的不同，本书还多次使用两种方法爬取相同的网站，这些都非常有利于读者学习。

本书不仅精挑细选爬取实例，内容组织上也注重深入性和全面性，希望尽量为读者演示各种爬取技巧和方法。从手写爬虫到爬虫框架，从多层页面爬取到图片下载，从应对反爬虫到模拟登录，从各种翻页技巧到查找网页元素，甚至爬虫去重技术和分布式爬虫部署，书中都有详细的演示和讲解，相信读者在读完本书后，能够系统地掌握使用 Python 编写爬虫的技术。

为了使代码讲解内容易看易懂，本书直接提供了全部的代码，读者可以参考书中的代码编写爬虫，但是要注意，商业网站的更新速度很快，可能在你看到本书的时候，网站已经做了或大或小的改版，如果直接照抄书中代码，就会产生一些问题。因此，读者应该重点学习编写爬虫的技巧和方法，相信在仔细阅读完本书后，读者完全可以应对各种各样的网页改版问题。此外，本书为了让代码更易读，在代码中用到的如户型、楼层、小区等变量使用了拼音命名，这样处理的优点是可读性较好，但是在面试或公司生产环境中编写代码，还是应该尽量使用英文作为变量名称。

本书提供了配套讲解视频，读者扫描书中二维码即可免费观看，也可到网易云课堂搜索“Python 爬虫零基础入门到进阶实战”，观看本书配套视频。

由于编者水平有限，加上爬虫技术本身发展迅速，书中难免有不足和不当之处，恳请读者批评、指正，在此表示衷心感谢。

编者

2018 年 3 月

目录

第 1 章

网络爬虫概述

1.1 认识网络爬虫

1.1.1 网络爬虫的含义

在大数据时代，人类社会的数据正以前所未有的速度增长。数据蕴含着巨大的价值，无论是对个人工作、生活，还是对企业未来的发展和创新商业模式，都有着很大的帮助。充分挖掘数据潜在价值，能帮助人们找到更合适的合作对象、更便宜的生活用品，也能帮助企业找到更好的细分市场，有针对性地为企业日后的发展提供数据支撑。数据让人们更好地掌握市场动向，更好地应对市场，产生新的合理的决策。

数据背后所隐藏的巨大商业价值正开始被越来越多的人所重视，那么数据从何而来？可以从网上找数据，但是人工提取数据效率太低，从经济角度也不可行。购买数据是一个办法，但是目前公开交易的数据少之又少，很难与多样化的数据需求匹配。因此，对很多人和企业来说，如果想获取全面、有效、准确的数据，编写爬虫抓取数据是一种明智之选，这就用到了这本书的主题——网络爬虫。

网络爬虫是一种程序，编写网络爬虫的主要目的是将互联网上的网页下载到本地并提取出相关数据。网络爬虫可以自动化地浏览网络中的信息，然后根据制定的规则下载和提取信息。

如图 1-1 所示，如果把互联网比喻成一个蜘蛛网，那么网络爬虫就是在网上爬来爬去的蜘蛛。简单来讲，网络爬虫主要完成两个任务：一是下载目标网页，二是从目标网页中提取需要的数据。

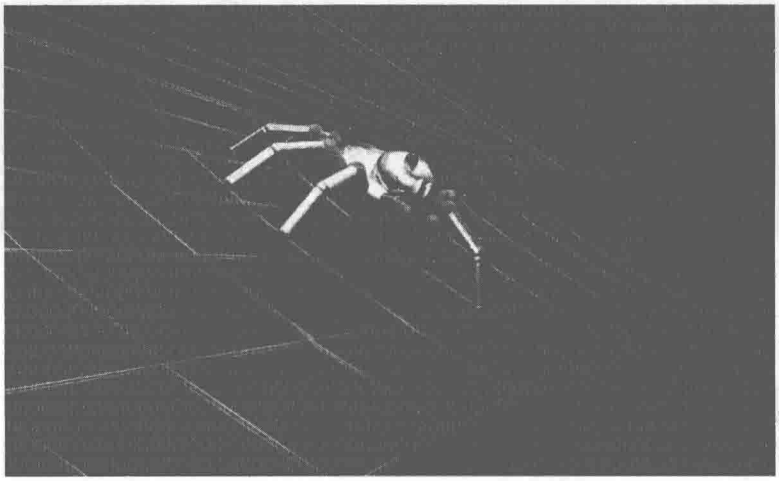


图 1-1　网络爬虫示意图

1.1.2　网络爬虫的主要类型

网络爬虫按照系统结构和实现技术，大致可以分为以下几种类型：通用网络爬虫、聚焦网络爬虫、增量式网络爬虫、深层页面爬虫。实际的网络爬虫系统通常是几种爬虫技术相结合实现的。

1. 通用网络爬虫

通用网络爬虫又称全网爬虫，爬行对象从一些种子 URL 扩充到整个 Web，主要为门户站点、搜索引擎和大型 Web 服务提供商采集数据。

2. 聚焦网络爬虫

聚焦网络爬虫是指选择性地爬行那些与预先定义好的主题相关页面的网络爬虫。与通用网络爬虫相比，聚焦网络爬虫只需要爬行与主题相关的页面，极大地节省了硬件和网络资源，保存的页面也因数量少而更新快，还可以很好地满足一些特定人群对特定领域信息的需求。聚焦网络爬虫是需要我们关注的重点爬虫类型。

3. 增量式网络爬虫

增量式网络爬虫是指对已下载网页采取增量式更新和只爬行新产生的或者已经发生变化的网页的爬虫，它能够在一定程度上保证所爬行的页面是尽可能新的页面。与周期性爬行和刷新页面的网络爬虫相比，增量式爬虫只会在需要的时候爬行新产生或发生更新的页面，并不重新下载没有发生变化的页面，可有效减少数据下载量，及时更新已爬行的网

页，减小时间和空间上的耗费，但是增加了爬行算法的复杂度和实现难度。后面的章节将对增量式网络爬虫和去重方法做简要介绍。

4. 深层页面爬虫

Web 页面按存在方式分为表层网页和深层网页。表层网页是传统搜索引擎可以索引的页面，是以超链接可以到达的静态网页为主构成的 Web 页面。深层网页是大部分内容不能通过静态链接获取的，隐藏在搜索表单后的，只有用户提交一些关键词才能获得的 Web 页面。例如那些用户注册后内容才可见的网页就属于深层页面。后面的章节将向读者介绍让爬虫登录一个网站、爬取深层页面的方法。

1.1.3 简单网络爬虫的架构

前面已经介绍网络爬虫的两个主要任务是下载目标网页和从网页中解析信息。为了完成这两个任务，一个简单的网络爬虫就要包含图 1-2 所示的 4 个部分。

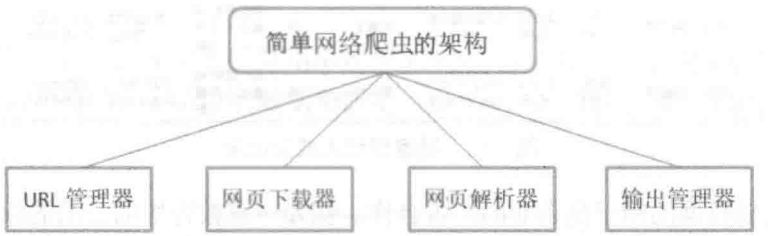


图 1-2 简单网络爬虫的架构

URL 管理器：管理将要爬取的 URL，防止重复抓取和循环抓取。

网页下载器：这是下载网页的组件，用来将互联网上 URL 对应的网页下载到本地，是爬虫的核心部分之一。

网页解析器：这是解析网页的组件，用来从网页中提取有价值的数据，是爬虫的另一个核心部分。

输出管理器：这是保存信息的组件，用来把解析出来的内容输出到文件或数据库中。

以上 4 个部分是一个简单的爬虫架构，这里通过介绍简单的爬虫架构，让读者对爬虫有一个直观的印象，后面的章节将详细讲解网络爬虫架构的实现。

1.1.4 网络爬虫的应用场景

网络爬虫的应用十分广泛，不仅应用在搜索引擎上，普通用户和企业在抓取数据、

分析数据的时候都需要借助于网络爬虫。这里用两个小例子来简单说明网络爬虫的应用场景。

假如现在有人想把北京的房子卖掉，需要委托给链家（或者我爱我家）的一位房产经纪人，就需要了解经纪人的业务能力，选择业务能力较强的经纪人，然而这两个房产中介网站并没有给出经纪人之间的成交对比。如果学习了爬虫技术，就可以写个爬取经纪人成交数据的爬虫，用爬虫爬取链家（或者我爱我家）网所有经纪人的成交记录，如图 1-3 所示。然后在经纪人的成交房产类型、成交数量、成交时间及成交价格之间做对比分析，从而找出成交能力最强的经纪人。

	cjtaoshu	mendian	cjzongjia	zhiwei	haoping	cjdanjia	cjxiaoqu	xingming	cjzhouqi	biaoqian	cjlouceng	cjshijian	congyenianxian	bankuai
0	37	红莲北里店	251.0	店经理	97% 141	43997 元/平	红莲北里 3室1厅 57平	郭海龙	36	房东信赖,销售达人,带看活跃	南 北/高楼层/6层	签约时间: 2015-05-24	4-5年	马连道
1	37	红莲北里店	159.0	店经理	97% 141	36969 元/平	红莲南里 1室1厅 43平	郭海龙	36	房东信赖,销售达人,带看活跃	南/高楼层/7层	签约时间: 2015-05-10	4-5年	马连道
2	37	红莲北里店	257.0	店经理	97% 141	39046 元/平	常青藤嘉园 1室1厅 65平	郭海龙	36	房东信赖,销售达人,带看活跃	北/低楼层/16层	签约时间: 2015-04-26	4-5年	马连道
3	37	红莲北里店	243.0	店经理	97% 141	41313 元/平	红莲北里 2室1厅 58平	郭海龙	36	房东信赖,销售达人,带看活跃	南 北/高楼层/6层	签约时间: 2015-04-04	4-5年	马连道
4	37	红莲北里店	372.5	店经理	97% 141	42053 元/平	广安门外大街 3室1厅 88平	郭海龙	36	房东信赖,销售达人,带看活跃	东 南 西 北/中楼层/18层	签约时间: 2015-04-01	4-5年	马连道

图 1-3 链家经纪人成交记录

另一个比较典型的例子是企业的广告合作。例如一家教育培训公司的领导，想要跟知乎上关注编程语言的意见领袖合作推广公司的培训课程，就需要了解在知乎的编程领域，哪位意见领袖的粉丝最多，哪位意见领袖的粉丝是公司的潜在培训对象。这时可以编写一个爬取知乎用户信息（包含从事领域、粉丝数量等内容）的爬虫，然后根据爬取下来的信息做一个简单的统计分析，从而找到可以寻求合作的优质对象。

以上两个例子在本书都有实现。

1.2 Python 网络爬虫技术概况

1.2.1 Python 中实现 HTTP 请求

本节主要介绍 Python 中都有哪些库和框架可以帮助我们实现网络爬虫。这里要特别说明一点的是，本书的代码和程序全部是在 Python 3.6.3 版本中实现的，也可以直接在

Python 3 的其他版本中运行。虽然大部分代码在 Python 2 中也可以运行，但并不推荐读者使用 Python 2，毕竟 Python 2 已经成为过去，Python 3 才是未来。

前面已经介绍，网页下载器是爬虫的核心部分之一，下载网页就需要实现 HTTP 请求，在 Python 中实现 HTTP 请求比较常用的主要有两个库。

一是 Urllib 库。Urllib 库是 Python 内置的 HTTP 请求库，可以直接调用。

二是 Requests 库。Requests 库是用 Python 语言编写的，基于 Urllib，采用 Apache2 Licensed 开源协议的 HTTP 库。它比 Urllib 更加方便，使用它可以节约我们大量的工作，完全满足 HTTP 的测试需求。Requests 是一个纯 Python 编写的、简单易用的 HTTP 库。

这两种实现 HTTP 请求的库中，Requests 库最简单，功能也最丰富，完全可以满足 HTTP 测试需求，是本书中手写简单爬虫的主力库，推荐读者学习和使用。至于 Urllib 库，后面的章节将做简单的介绍，让读者有所了解。

1.2.2 Python 中实现网页解析

所谓网页解析器，简单地说就是用来解析 HTML 网页的工具，它主要用于从 HTML 网页信息中提取需要的、有价值的数据和链接。在 Python 中解析网页主要用到图 1-4 所示的 3 种工具。

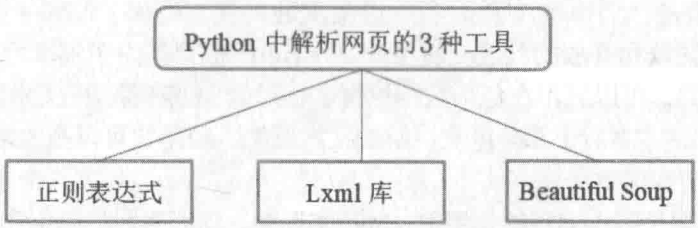


图 1-4 Python 解析网页常用的 3 种工具

一是正则表达式。正则表达式（regular expression）描述了一种字符串匹配的模式（pattern），可以用来检查一个串是否含有某种子串，将匹配的子串替换或者从某个串中取出符合某个条件的子串等。正则表达式的优点是基本能用正则表达式来提取想要的所有信息，效率比较高，但缺点也很明显——正则表达式不是很直观，写起来比较复杂。

二是 Lxml 库。这个库使用的是 XPath 语法，同样是效率比较高的解析库。XPath 是一门在 XML 文档中查找信息的语言。XPath 可用来在 XML 文档中对元素和属性进行

遍历。XPath 比较直观易懂，配合 Chrome 浏览器或 Firefox 浏览器，写起来非常简单，它的代码速度运行快且健壮，一般来说是解析数据的最佳选择，Lxml 是本书中解析网页的主力工具。

三是 Beautiful Soup。Beautiful Soup 是一个可以从 HTML 或 XML 文件中提取数据的 Python 库。它能够通过我们喜欢的转换器实现惯用的文档导航、查找。Beautiful Soup 编写效率高，能帮程序员节省数小时甚至数天的工作时间。Beautiful Soup 比较简单易学，但相比 Lxml 和正则表达式，解析速度慢很多。

总结起来，无论正则表达式、Beautiful Soup 库还是 Lxml 库，都能满足我们解析网页的需求，但 Lxml 使用的 XPath 语法简单易学、解析速度快，是本书推荐读者使用的网页解析工具。

1.2.3 Python 爬虫框架

前面介绍的 HTTP 请求库和网页解析技术都是一步步手写爬虫时使用的，Python 中还有很多帮助实现爬虫项目的半成品——爬虫框架。爬虫框架允许根据具体项目的情况，调用框架的接口，编写少量的代码实现一个爬虫。爬虫框架实现了爬虫要实现的常用功能，能够节省编程人员开发爬虫的时间，帮助编程人员高效地开发爬虫。

在 Python 中，爬虫框架很多，常见的 Python 爬虫框架主要有 Scrapy 框架、Pyspider 框架、Cola 框架等。

Scrapy 框架是 Python 中最著名、最受欢迎的爬虫框架。它是一个相对成熟的框架，有着丰富的文档和开放的社区交流空间。Scrapy 框架是为了爬取网站数据、提取结构性数据而编写的，可以应用在包括数据挖掘、信息处理或存储历史数据等一系列的程序中。Scrapy 框架是本书后半部分重点讲解的技术框架，利用它可以高效地爬取 Web 页面并提取有价值的结构化数据。

Pyspider 框架是国人编写的、用 Python 实现的、功能强大的网络爬虫系统，能在浏览器界面上进行脚本的编写、功能的调度和爬取结果的实时查看，后端使用常用的数据库进行爬取结果的存储，还能定时设置任务与任务优先级等。读者如果有兴趣，可以查看它的相关文档。

Cola 框架是一个分布式的爬虫框架，用户只需编写几个特定的函数，而无需关注分布式运行的细节，任务会被自动分配到多台机器上，整个过程对用户是透明的。

Python 还有很多其他的爬虫框架，它们各有特点，读者可以上网查阅相关材料。本书将深入讲解 Scrapy 框架的使用。

1.3 搭建开发环境

1.3.1 代码运行环境

本书所讲解的爬虫技术都是基于 Python 语言实现的，希望读者尽可能地了解 Python 语言的基础语法。为了方便在自己的计算机上实现本书的代码，读者可以尝试搭建与本书一致的开发环境。

为了照顾大多数入门 Python 爬虫的同学，本书中的代码都是在如下运行环境中编写的：Windows 10 操作系统；Python 3.6.3。

在 Windows 平台环境下，可以按照以下步骤搭建开发环境。

1. 下载 Python

从官网下载与操作系统、位数对应的 Python 版本。图 1-5 所示为 Python 官网页面，单击导航栏中的 Downloads 即可选择下载。如果计算机是 64 位的 Windows 系统，就可以选择从 Windows 版本下载页中下载 Windows x86-64 executable installer 这个可执行安装文件。

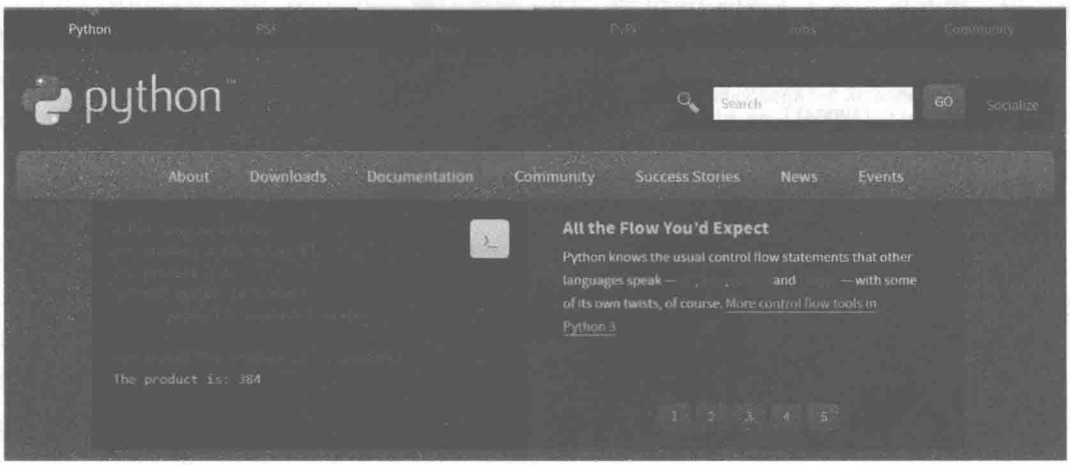


图 1-5 Python 官网页面

2. 安装 Python

单击运行下载下来的 python-3.6.5-amd64.exe 安装文件（这是最新版本的正式安装文