大数据及人工智能产教融合系列丛书

Python

网络爬虫从入门到实践

庄培杰◎编 著

前沿
全书基于Python 3.7.0讲解爬虫

全面
涵盖抓包、请求模拟、数据解析、反爬虫应对等众多知识点

实战
20多个经典实战案例详解，随学随用

超值
附赠22个Python爬虫项目源码，帮助新手轻松入门

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

大数据及人工智能产教融合系列丛书

Python 网络爬虫从入门到实践

庄培杰　编著

Publishing House of Electronics Industry
北京 • BEIJING

内 容 简 介

本书讲解了如何使用 Python 编写网络爬虫，涵盖爬虫的概念、Web 基础、Chrome、Charles 和 Packet Capture 抓包、urllib、Requests 请求库、lxml、Beautiful Soup、正则表达式解析数据、CSV、Excel、MySQL、Redis、MongoDB 保存数据、反爬虫策略应对、爬虫框架 Scrapy 的使用与部署，以及应用案例。

本书结构清晰、内容精练，代码示例典型实用，附带实践过程中遇到问题的解决方案，非常适合 Python 初学者和进阶读者阅读。

图书在版编目（CIP）数据

Python 网络爬虫从入门到实践 / 庄培杰编著. —北京：电子工业出版社，2019.8
（大数据及人工智能产教融合系列丛书）
ISBN 978-7-121-37105-9

Ⅰ. ①P… Ⅱ. ①庄… 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（2019）第 144186 号

责任编辑：李 冰　　特约编辑：王 纲
印　　刷：三河市鑫金马印装有限公司
装　　订：三河市鑫金马印装有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编 100036
开　　本：787×1 092　1/16　印张：19.5　字数：499 千字
版　　次：2019 年 8 月第 1 版
印　　次：2019 年 8 月第 1 次印刷
定　　价：79.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：libing@phei.com.cn。

前　言

笔者是一名 Android 开发工程师，在接触 Python 之前，每天的工作流程基本上都是接到新版本的需求→写新页面→修改接口和业务逻辑，非常乏味。

持续性的重复劳动，让笔者意识到一个问题：如果只会 Android 开发，能做的事情非常有限！例如，自己写一个 App，如果没有可供调用的 API，那么只能得到一个单机的 App。因为自己对后台相关的技术一窍不通，平时根本不用去了解这方面的知识，只要给后台发出请求，然后解析数据，显示到页面上就好。

笔者开始琢磨花点时间去学习后台开发的知识，候选方案有 Java、Kotlin、Python、PHP 和 Go。因为之前接触过 Python，加上笔者只是想写一个给自己的 App 调用的 API，所以最后选择了 Python 这门对初学者非常友好的编程语言。

笔者花了一周多的时间把 Python 的基本语法研究了一遍，发现 Python 语法简单、代码简洁。正当笔者准备去看 Flask 这个轻量级 Python Web 框架的文档时，问题来了：没有数据源。数据源都没有，写什么 API？于是，笔者把学习重心转移到 Python 爬虫编写上。

这仿佛打开了新的大门，简单的几行代码就完成了站点模拟请求、页面解析，还把图片下载到了本地。接下来，笔者又开始研究如何提高爬取效率、解析效率，以及采用多样化的数据存储形式。后来发现有些站点有反爬虫的策略，得不到数据，于是笔者又对此进行研究。爬虫学得差不多了，笔者又开始研究数据分析，分析爬取的结果，得出一些有用的结论，如分析招聘网站上某个工作岗位的行情等。

笔者相信掌握 Python 爬虫，会为你的工作、生活带来一些意想不到的便利。因笔者能力和成书时间所限，书中难免会有纰漏，相关代码也不是尽善尽美，希望读者批评指正，笔者将感激不尽。

本书特色

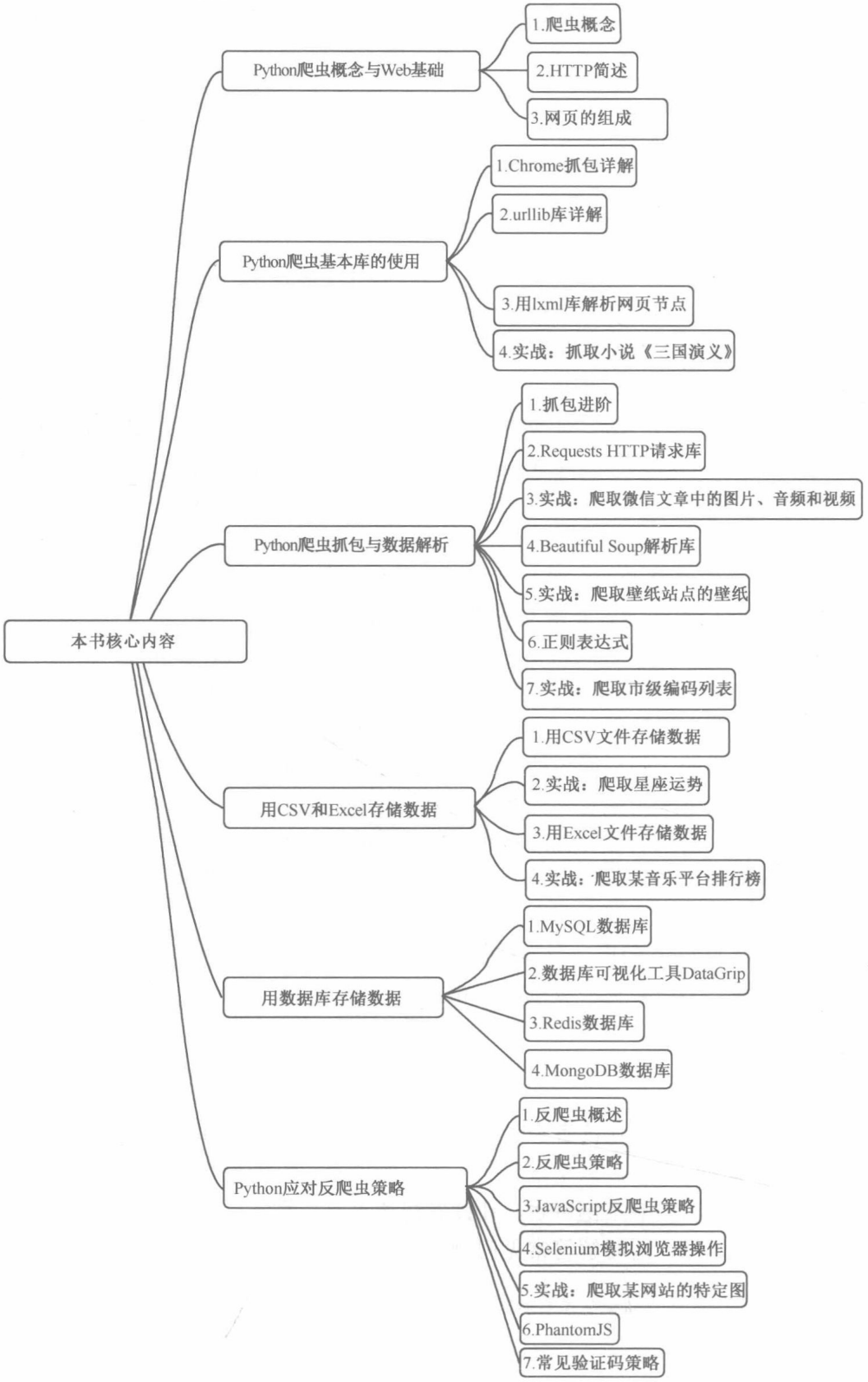
（1）清晰的学习路线，让读者少走弯路。

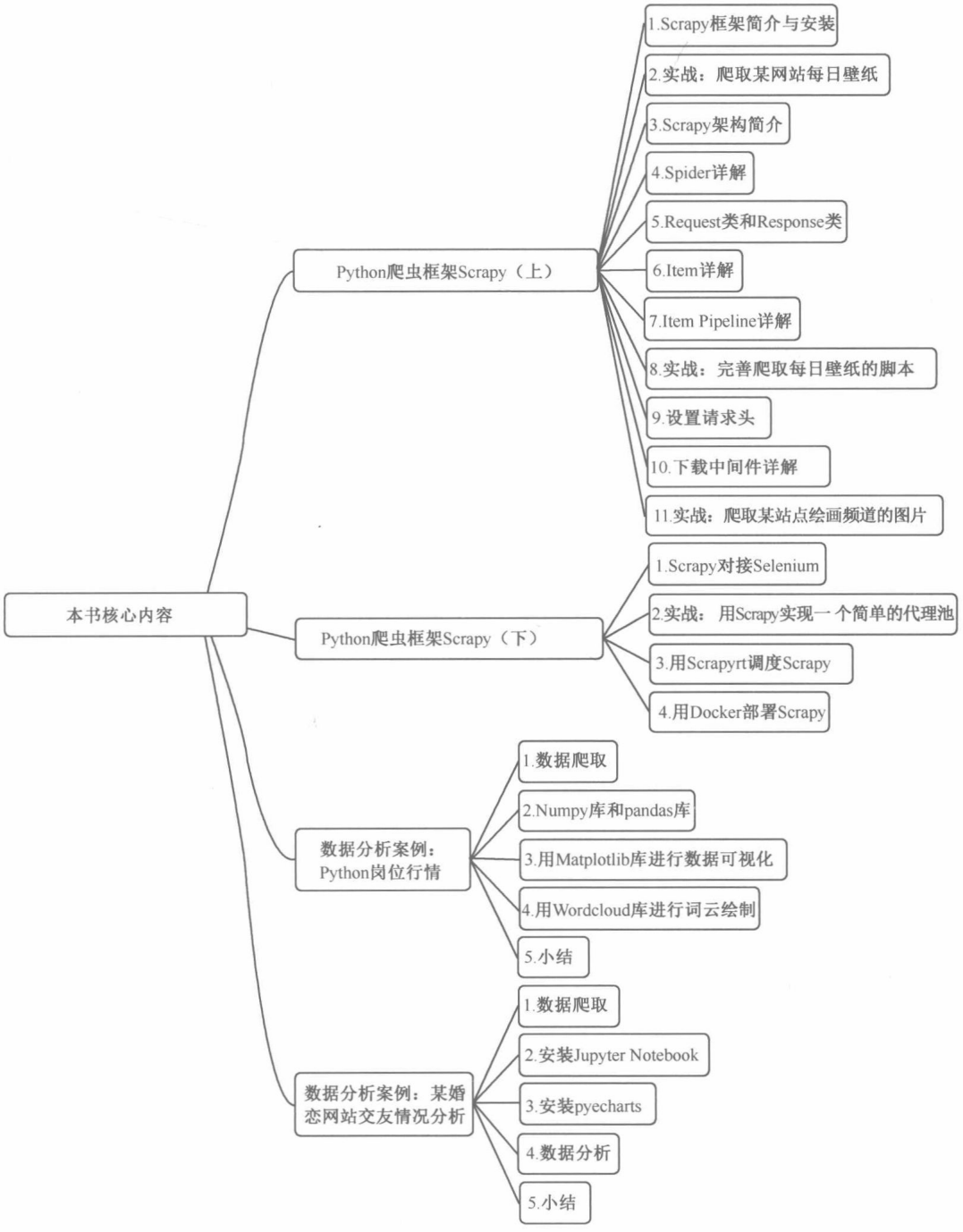
（2）精练的知识点讲解，只学重要的、核心的内容，不拖泥带水。

（3）简单而经典的实战案例讲解，让读者快速上手，举一反三。

（4）代码注释详尽，适时解答常见问题。

本书内容及体系结构


本书核心内容
Python爬虫概念与Web基础
1.爬虫概念
2.HTTP简述
3.网页的组成
Python爬虫基本库的使用
1.Chrome抓包详解
2.urllib库详解
3.用lxml库解析网页节点
4.实战：抓取小说《三国演义》
Python爬虫抓包与数据解析
1.抓包进阶
2.Requests HTTP请求库
3.实战：爬取微信文章中的图片、音频和视频
4.Beautiful Soup解析库
5.实战：爬取壁纸站点的壁纸
6.正则表达式
7.实战：爬取市级编码列表
用CSV和Excel存储数据
1.用CSV文件存储数据
2.实战：爬取星座运势
3.用Excel文件存储数据
4.实战：爬取某音乐平台排行榜
用数据库存储数据
1.MySQL数据库
2.数据库可视化工具DataGrip
3.Redis数据库
4.MongoDB数据库
Python应对反爬虫策略
1.反爬虫概述
2.反爬虫策略
3.JavaScript反爬虫策略
4.Selenium模拟浏览器操作
5.实战：爬取某网站的特定图
6.PhantomJS
7.常见验证码策略



本书读者对象

- 想要学习 Python 编程的人员。
- 有 Python 编程基础、想进阶的人员。
- 各计算机、软件专业的在校学生。
- 其他对 Python 爬虫感兴趣的读者。

目　录

第1章 Python 爬虫概念与 Web 基础

在开始学习如何编写 Python 爬虫之前，我们先学习一些和爬虫相关的概念和 Web 基础知识。

本章主要学习内容：

- 爬虫概念。
- HTTP 简述。
- 网页的组成（HTML，CSS，JavaScript）。

1.1 爬虫概念

本节将详细介绍与爬虫相关的一些概念，这将有助于我们后面的学习。

1.1.1 什么是爬虫

爬虫，即网络爬虫，又称网络蜘蛛（Web Spider），是一种按照一定规则，用来自动浏览或抓取万维网数据的程序。可以把爬虫程序看成一个机器人，它的功能就是模拟人的行为去访问各种站点，或者带回一些与站点相关的信息。它可以 24 小时不间断地做一些重复性的工作，还可以自动提取一些数据。

有一点要注意，本章标题虽然写明 Python 爬虫，但并不是只有 Python 才能编写爬虫程序。其他的编程语言也可以用来编写爬虫程序，如 PHP 有 phpspider 爬虫框架，Java 有 WebMagic 爬虫框架，C#有 DotnetSpider 爬虫框架等。

1.1.2 爬虫使用场景的引入

爬虫到底能干些什么呢？我们通过下面几个场景引入。

场景一：

大部分读者应该都有看网络小说的习惯，而去正版站点看小说一般是需要付费的，所以衍生了很多盗版小说站点。这种盗版小说站点一般通过挂广告的形式来盈利，在浏览器底部会嵌入各种类型的广告，用户单击这些广告会打开其对应的网址，盗版小说站点以此获利。

这些盗版小说站点的广告一般都是误导读者打开它。比如，广告上有个 X 按钮，读者通常以为单击 X 按钮可以关闭这个广告，殊不知是打开了广告，而且有些广告中含有不堪入目的内容。想一想，在拥挤的地铁里，旁边的人看到你打开了这样的网页，会有多尴尬。有没有办法既能看到小说又不用看这些烦人的广告呢？

答：通过爬虫可以解决这个问题，让爬虫只解析小说部分的内容并显示出来，甚至可以把整本小说的内容解析完，保存到本地，以便离线阅读。

场景二：

有些读者有一种类似于强迫症的行为，比如，快递预计今天会到，每隔一段时间你就会不由自主地输入单号看看快递到了没。有没有办法能摆脱这种频繁而又枯燥的工作呢？

答：对于这种轮询（每隔一段时间查询一次）的任务，我们可以编写一个定时爬虫，每隔一段时间自动去请求相应的站点，然后处理结果，判断其是否符合我们的预期，与人工刷新相比，爬虫的频率更快、效率更高。

场景三：

有一些站点，会通过签到或以做每日任务的形式来提高用户的活跃度，当坚持到一定天数后会发放一些小奖励。比如，坚持十五天就能获得一个小礼物。因为忙碌或其他原因，你中断了签到，使得之前的努力就都白费了。而且有些每日任务非常枯燥，但每天要为此花上好几分钟。有没有办法把这种任务交给程序来做，然后坐享其成呢？

答：可以把这些任务交给爬虫。对于签到这种操作，通过抓包获取签到所需的接口、参数和请求规则，接着让脚本每天定时执行即可。每日任务一般通过模拟单击的方式完成，通过 Selenium 自动化框架来模拟。对于 App 签到，则可以通过 Appium 移动端自动化测试框架来完成。

场景四：

如果你有志于从事 Python 开发相关的工作，想了解这个行业的一些情况，比如薪资、年限和相关要求等，而你身边又没有从事相关工作的朋友，怎样才能获取到这些信息呢？

答：可以编写爬虫爬取一些招聘网站，比如拉勾、前程无忧这类站点，把 Python 岗位相关的信息都抓取下来，然后通过数据分析基础三件套（NumPy、pandas、Matplotlib）进行基本的数据分析，以此获取和这个行业相关的一些信息。

1.1.3　爬虫的组成部分

相信在看完爬虫应用的四个场景后，读者对爬虫能做什么有了大致了解，接下来我们来了解爬虫由哪几部分组成。爬虫的三个组成部分如图 1.1 所示。

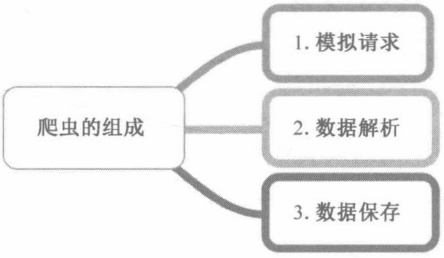


图 1.1　爬虫的三个组成部分

1.1.4　模拟请求

模拟请求就是如何把我们的爬虫伪装得像一个人一样去访问互联网站点。

- 最简单的站点，什么都不处理，只要发送请求，就会给出相应的结果。
- 稍微复杂一点的站点，会判断请求头中的 User-Agent 是否为浏览器请求，Host 字段是否为正确的服务器域名，以及 Referer 字段的地址是否合法。
- 再复杂一点的站点，需要登录后才能访问。登录后会持有一个 Cookie 或 Session 会话，你需要带着这个东西才能执行一些请求，否则都会跳到登录页。
- 更复杂一点的站点，登录很复杂，需要五花八门的验证码、最简单的数字图片加噪点、滑动验证码、点触验证码。除此之外，还有一些其他特立独行的验证方式，如最经典的微博宫格验证码、极验验证码的行为验证等。
- 还有更复杂的站点，其链接和请求参数都是加密的，需要研究、破解加密规则才能够模拟访问。
- 除此之外，还有一些反爬虫套路，如限制 IP 访问频次、JavaScript 动态加载数据等。

在模拟请求之前，先要了解请求规则，一般通过抓包工具来完成。

（1）对最简单的浏览器请求（以 Chrome 谷歌浏览器为例），在网页空白处右键单击并选择检查，或者依次单击如图 1.2 所示的“更多工具”→“开发者工具”。在 Windows 中打开开发者工具的快捷键 F12。

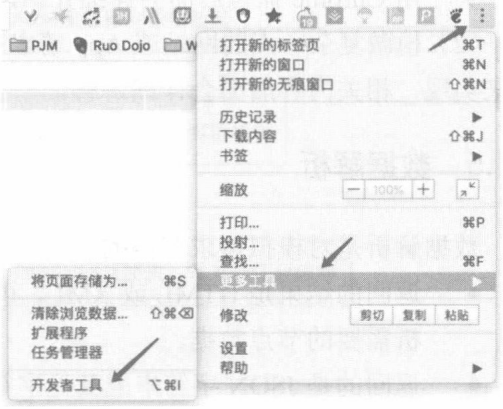


图 1.2　打开 Chrome 开发者工具