Packt

用Python写网络爬虫

（第2版）

Python
Web Scraping
Second Edition

[德] 凯瑟琳·雅姆尔（Katharine Jarmul）
[澳] 理查德·劳森（Richard Lawson） 著
李斌 译

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

用Python写网络爬虫

（第2版）

Python
Web Scraping
Second Edition

[德] 凯瑟琳·雅姆尔（Katharine Jarmul）
[澳] 理查德·劳森（Richard Lawson） 著
李斌 译

人民邮电出版社
北京

图书在版编目（CIP）数据

用Python写网络爬虫 : 第2版 / (德) 凯瑟琳·雅姆尔 (Katharine Jarmul) , (澳) 理查德·劳森 (Richard Lawson) 著 ; 李斌译. -- 北京 : 人民邮电出版社, 2018.8
ISBN 978-7-115-47967-9

Ⅰ. ①用… Ⅱ. ①凯… ②理… ③李… 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043962号

版权声明

◆ 著　[德] 凯瑟琳·雅姆尔（Katharine Jarmul）
[澳] 理查德·劳森（Richard Lawson）
译　李　斌
责任编辑　傅道坤
责任印制　焦志炜

◆ 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
三河市祥达印刷包装有限公司印刷

◆ 开本：800×1000　1/16
印张：13.25
字数：183 千字　2018 年 8 月第 1 版
印数：1－3 000 册　2018 年 8 月河北第 1 次印刷

著作权合同登记号　图字：01-2017-8623 号

定价：49.00 元

读者服务热线：(010) 81055410　印装质量热线：(010) 81055316
反盗版热线：(010) 81055315
广告经营许可证：京东工商广登字 20170147 号

内容提要

本书讲解了如何使用 Python 来编写网络爬虫程序，内容包括网络爬虫简介，从页面中抓取数据的 3 种方法，提取缓存中的数据，使用多个线程和进程进行并发抓取，抓取动态页面中的内容，与表单进行交互，处理页面中的验证码问题，以及使用 Scarpy 和 Portia 进行数据抓取，并在最后介绍了使用本书讲解的数据抓取技术对几个真实的网站进行抓取的实例，旨在帮助读者活学活用书中介绍的技术。

本书适合有一定 Python 编程经验而且对爬虫技术感兴趣的读者阅读。

关于作者

Katharine Jarmul 是德国柏林的一位数据科学家和 Python 支持者。她经营了一家数据科学咨询公司——Kjamistan，为不同规模的企业提供诸如数据抽取、采集以及建模的服务。她从 2008 年开始使用 Python 进行编程，从 2010 年开始使用 Python 抓取网站，并且在使用网络爬虫进行数据分析和机器学习的不同规模的初创企业中工作过。读者可以通过 Twitter（`@kjam`）关注她的想法以及动态。

Richard Lawson 来自澳大利亚，毕业于墨尔本大学计算机科学专业。毕业后，他创办了一家专注于网络爬虫的公司，为超过 50 个国家的业务提供远程工作。他精通世界语，可以使用汉语和韩语对话，并且积极投身于开源软件事业。他目前正在牛津大学攻读研究生学位，并利用业余时间研发自主无人机。

关于审稿人

Dimitrios Kouzis-Loukas 在为大小型组织提供软件系统方面拥有超过 15 年的经验。他近期的项目通常是具有超低延迟及高可用性要求的分布式系统。他是语言无关论者，不过对 C++和 Python 略有偏好。他对开源有着坚定的信念，他希望他的贡献能够造福于各个社区和全人类。

Lazar Telebak 是一位自由的 Web 开发人员，专注于使用 Python 库/框架进行网络抓取、爬取和网页索引的工作。

他主要从事于处理自动化和网站抓取、爬取以及导出数据到不同格式（包括 CSV、JSON、XML 和 TXT）和数据库（如 MongoDB、SQLAlchemy 和 Postgres）的项目。

Lazer 还拥有前端技术和语言的经验，包括 HTML、CSS、JavaScript 和 jQuery。

前言

互联网包含了迄今为止最有用的数据集，并且大部分可以免费公开访问。但是，这些数据难以复用。它们被嵌入在网站的结构和样式当中，需要抽取出来才能使用。从网页中抽取数据的过程又称为网络爬虫，随着越来越多的信息被发布到网络上，网络爬虫也变得越来越有用。

本书使用的所有代码均已使用 Python 3.4+测试通过，并且可以在异步社区下载到。

本书内容

第 1 章，网络爬虫简介，介绍了什么是网络爬虫，以及如何爬取网站。

第 2 章，数据抓取，展示了如何使用几种库从网页中抽取数据。

第 3 章，下载缓存，介绍了如何通过缓存结果避免重复下载的问题。

第 4 章，并发下载，教你如何通过并行下载网站加速数据抓取。

第 5 章，动态内容，介绍了如何通过几种方式从动态网站中抽取数据。

第 6 章，表单交互，展示了如何使用输入及导航等表单进行搜索和登录。

第 7 章，验证码处理，阐述了如何访问被验证码图像保护的数据。

第 8 章，Scrapy，介绍了如何使用 Scrapy 进行快速并行的抓取，以及使用 Portia 的 Web 界面构建网络爬虫。

第 9 章，综合应用，对你在本书中学到的网络爬虫技术进行总结。

阅读本书的前提

为了有助于阐明爬取示例，我们创建了一个示例网站，其网址为http://example.python-scraping.com。用于生成该网站的源代码可以从异步社区获取到，其中包含了如何自行搭建该网站的说明。如果你愿意的话，也可以自己搭建它。

我们决定为本书示例搭建一个定制网站，而不是抓取活跃的网站，这样我们就对环境拥有了完全控制。这种方式提供了稳定性，因为活跃的网站要比书中的定制网站更新更加频繁，当你尝试运行爬虫示例时，代码可能已经无法工作。另外，定制网站允许我们自定义示例，便于阐释特定技巧并避免其他干扰。最后，活跃的网站可能并不欢迎我们使用它作为学习网络爬虫的对象，并且可能会封禁我们的爬虫。使用我们自己定制的网站可以规避这些风险，不过在这些例子中学到的技巧确实也可以应用到这些活跃的网站当中。

本书读者

本书假设你已经拥有一定的编程经验，并且本书很可能不适合零基础的初学者阅读。本书中的网络爬虫示例需要你具有 Python 语言以及使用 pip 安装模块的能力。如果你想复习一下这些知识，有一本非常好的免费在线书籍可以使用，其书名为 Dive Into Python，作者为 Mark Pilgrim，读者可在网上搜索并阅读。这本书也是我初学 Python 时所使用的资源。

此外，这些例子还假设你已经了解网页是如何使用 HTML 进行构建并通过 JavaScript 进行更新的知识。关于 HTTP、CSS、AJAX、WebKit 以及 Redis 的既有知识也很有用，不过它们不是必需的，这些技术会在需要使用时进行介绍。

资源与支持

本书由异步社区出品，社区（https://www.epubit.com/）为您提供相关资源和后续服务。

配套资源

本书提供如下资源：

- 本书源代码；
- 构建本书实例网站的源码。

要获得以上配套资源，请在异步社区本书页面中点击 配套资源，跳转到下载界面，按提示进行操作即可。注意：为保证购书读者的权益，该操作会给出相关提示，要求输入提取码进行验证。

提交勘误

作者和编辑尽最大努力来确保书中内容的准确性，但难免会存在疏漏。欢迎您将发现的问题反馈给我们，帮助我们提升图书的质量。

当您发现错误时，请登录异步社区，按书名搜索，进入本书页面，点击“提交勘误”，输入勘误信息，点击“提交”按钮即可。本书的作者和编辑会对您提交的勘误进行审核，确认并接受后，您将获赠异步社区的 100 积分。积分可用于在异步社区兑换优惠券、样书或奖品。

扫码关注本书

扫描下方二维码，您将会在异步社区微信服务号中看到本书信息及相关的服务提示。

与我们联系

我们的联系邮箱是 contact@epubit.com.cn。

如果您对本书有任何疑问或建议，请您发邮件给我们，并请在邮件标题中注明本书书名，以便我们更高效地做出反馈。

如果您有兴趣出版图书、录制教学视频，或者参与图书翻译、技术审校等工作，可以发邮件给我们；有意出版图书的作者也可以到异步社区在线提交投稿（直接访问 www.epubit.com/selfpublish/submission 即可）。

如果您是学校、培训机构或企业，想批量购买本书或异步社区出版的其他图书，也可以发邮件给我们。

如果您在网上发现有针对异步社区出品图书的各种形式的盗版行为，包括对图书全部或部分内容的非授权传播，请您将怀疑有侵权行为的链接发邮件给我们。您的这一举动是对作者权益的保护，也是我们持续为您提供有价值的内容的动力之源。

关于异步社区和异步图书

“异步社区”是人民邮电出版社旗下 IT 专业图书社区，致力于出版精品 IT 技术图书和相关学习产品，为作译者提供优质出版服务。异步社区创办于 2015 年 8 月，提供大量精品 IT 技术图书和电子书，以及高品质技术文章和视频课程。更多详情请访问异步社区官网 https://www.epubit.com。

“异步图书”是由异步社区编辑团队策划出版的精品 IT 专业图书的品牌，依托于人民邮电出版社近 30 年的计算机图书出版积累和专业编辑团队，相关图书在封面上印有异步图书的 LOGO。异步图书的出版领域包括软件开发、大数据、AI、测试、前端、网络技术等。

异步社区

微信服务号

目录

第 1 章　网络爬虫简介　1

1.1　网络爬虫何时有用 …… 1
1.2　网络爬虫是否合法 …… 2
1.3　Python 3 …… 3
1.4　背景调研 …… 4
1.4.1　检查 robots.txt …… 4
1.4.2　检查网站地图 …… 5
1.4.3　估算网站大小 …… 6
1.4.4　识别网站所用技术 …… 7
1.4.5　寻找网站所有者 …… 9
1.5　编写第一个网络爬虫 …… 11
1.5.1　抓取与爬取的对比 …… 11
1.5.2　下载网页 …… 12
1.5.3　网站地图爬虫 …… 15
1.5.4　ID 遍历爬虫 …… 17
1.5.5　链接爬虫 …… 19
1.5.6　使用 requests 库 …… 28
1.6　本章小结 …… 30

第 2 章　数据抓取 31

2.1　分析网页 32
2.2　3 种网页抓取方法 34
2.2.1　正则表达式 35
2.2.2　Beautiful Soup 37
2.2.3　Lxml 39
2.3　CSS 选择器和浏览器控制台 41
2.4　XPath 选择器 43
2.5　LXML 和家族树 46
2.6　性能对比 47
2.7　抓取结果 49
2.7.1　抓取总结 50
2.7.2　为链接爬虫添加抓取回调 51
2.8　本章小结 55

第 3 章　下载缓存 56

3.1　何时使用缓存 57
3.2　为链接爬虫添加缓存支持 57
3.3　磁盘缓存 60
3.3.1　实现磁盘缓存 62
3.3.2　缓存测试 64
3.3.3　节省磁盘空间 65
3.3.4　清理过期数据 66
3.3.5　磁盘缓存缺点 68
3.4　键值对存储缓存 69
3.4.1　键值对存储是什么 69
3.4.2　安装 Redis 70

3.4.3 Redis 概述 ····· 71
3.4.4 Redis 缓存实现 ····· 72
3.4.5 压缩 ····· 74
3.4.6 测试缓存 ····· 75
3.4.7 探索 requests-cache ····· 76
3.5 本章小结 ····· 78

第 4 章 并发下载 79

4.1 100 万个网页 ····· 79
4.2 串行爬虫 ····· 82
4.3 多线程爬虫 ····· 83
4.4 线程和进程如何工作 ····· 83
4.4.1 实现多线程爬虫 ····· 84
4.4.2 多进程爬虫 ····· 87
4.5 性能 ····· 91
4.6 本章小结 ····· 94

第 5 章 动态内容 95

5.1 动态网页示例 ····· 95
5.2 对动态网页进行逆向工程 ····· 98
5.3 渲染动态网页 ····· 104
5.3.1 PyQt 还是 PySide ····· 105
5.3.2 执行 JavaScript ····· 106
5.3.3 使用 WebKit 与网站交互 ····· 108
5.4 渲染类 ····· 111
5.5 本章小结 ····· 117

第 6 章　表单交互 119

6.1　登录表单 120
6.2　支持内容更新的登录脚本扩展 128
6.3　使用 Selenium 实现自动化表单处理 132
6.4　本章小结 135

第 7 章　验证码处理 136

7.1　注册账号 137
7.2　光学字符识别 140
7.3　处理复杂验证码 144
7.4　使用验证码处理服务 144
7.4.1　9kw 入门 145
7.4.2　报告错误 150
7.4.3　与注册功能集成 151
7.5　验证码与机器学习 153
7.6　本章小结 153

第 8 章　Scrapy 154

8.1　安装 Scrapy 154
8.2　启动项目 155
8.2.1　定义模型 156
8.2.2　创建爬虫 157
8.3　不同的爬虫类型 162
8.4　使用 shell 命令抓取 163
8.4.1　检查结果 165
8.4.2　中断与恢复爬虫 167

8.5 使用 Portia 编写可视化爬虫 ····· 170
8.5.1 安装 ····· 170
8.5.2 标注 ····· 172
8.5.3 运行爬虫 ····· 176
8.5.4 检查结果 ····· 176
8.6 使用 Scrapely 实现自动化抓取 ····· 177
8.7 本章小结 ····· 178

第 9 章 综合应用 179

9.1 Google 搜索引擎 ····· 179
9.2 Facebook ····· 184
9.2.1 网站 ····· 184
9.2.2 Facebook API ····· 186
9.3 Gap ····· 188
9.4 宝马 ····· 192
9.5 本章小结 ····· 196

第 1 章
网络爬虫简介

欢迎来到网络爬虫的广阔天地！网络爬虫被用于许多领域，收集不太容易以其他格式获取的数据。你可能是正在撰写新报道的记者，也可能是正在抽取新数据集的数据科学家。即使你只是临时的开发人员，网络爬虫也是非常有用的工具，比如当你需要检查大学网站上最新的家庭作业并且希望通过邮件发送给你时。无论你的动机是什么，我们都希望你已经准备好开始学习了！

在本章中，我们将介绍如下主题：

- 网络爬虫领域简介；
- 解释合法性质疑；
- 介绍 Python 3 安装；
- 对目标网站进行背景调研；
- 逐步完善一个高级网络爬虫；
- 使用非标准库协助抓取网站。

1.1 网络爬虫何时有用

假设我有一个鞋店，并且想要及时了解竞争对手的价格。我可以每天访问他们的网站，与我店铺中鞋子的价格进行对比。但是，如果我店铺中的鞋类

品种繁多，或是希望能够更加频繁地查看价格变化的话，就需要花费大量的时间，甚至难以实现。再举一个例子，我看中了一双鞋，想等到它促销时再购买。我可能需要每天访问这家鞋店的网站来查看这双鞋是否降价，也许需要等待几个月的时间，我才能如愿盼到这双鞋促销。上述这两个重复性的手工流程，都可以利用本书介绍的网络爬虫技术实现自动化处理。

在理想状态下，网络爬虫并不是必需品，每个网站都应该提供 API，以结构化的格式共享它们的数据。然而在现实情况中，虽然一些网站已经提供了这种 API，但是它们通常会限制可以抓取的数据，以及访问这些数据的频率。另外，网站开发人员可能会变更、移除或限制其后端 API。总之，我们不能仅仅依赖于 API 去访问我们所需的在线数据，而是应该学习一些网络爬虫技术的相关知识。

1.2　网络爬虫是否合法

尽管在过去 20 年间已经做出了诸多相关裁决，不过网络爬虫及其使用时法律所允许的内容仍然处于建设当中。如果被抓取的数据用于个人用途，且在合理使用版权法的情况下，通常没有问题。但是，如果这些数据会被重新发布，并且抓取行为的攻击性过强导致网站宕机，或者其内容受版权保护，抓取行为违反了其服务条款的话，那么则有一些法律判例可以提及。

在 Feist Publications, Inc. 起诉 Rural Telephone Service Co. 的案件中，美国联邦最高法院裁定抓取并转载真实数据（比如，电话清单）是允许的。在澳大利亚，Telstra Corporation Limited 起诉 Phone Directories Company Pty Ltd 这一类似案件中，则裁定只有拥有明确作者的数据，才可以受到版权的保护。而在另一起发生于美国的美联社起诉融文集团的内容抓取案件中，则裁定对美联社新闻重新聚合为新产品的行为是侵犯版权的。此外，在欧盟的 ofir.dk 起诉 home.dk 一案中，最终裁定定期抓取和深度链接是允许的。

还有一些案件中，原告控告一些公司抓取强度过大，尝试通过法律手段停