

精通Python网络爬虫：核心技术、框架与项目实战



[精通Python网络爬虫：核心技术、框架与项目实战 下载链接1](#)

著者:韦玮

出版者:机械工业出版社

出版时间:2017-3

装帧:平装

isbn:9787111562085

为什么写这本书

网络爬虫其实很早就出现了，最开始网络爬虫主要应用在各种搜索引擎中。在搜索引擎中，主要使用通用网络爬虫对网页进行爬取及存储。

随着大数据时代的到来，我们经常需要在海量数据的互联网环境中搜集一些特定的数据并对其进行分析，我们可以使用网络爬虫对这些特定的数据进行爬取，并对一些无关的数据进行过滤，将目标数据筛选出来。对特定的数据进行爬取的爬虫，我们将其称为聚焦网络爬虫。在大数据时代，聚焦网络爬虫的应用需求越来越大。

目前在国内Python网络爬虫的书籍基本上都是从国外引进翻译的，国内的本版书籍屈指可数，故而我跟华章的副总编杨福川策划了这本书。本书的撰写过程中各方面的参考资料非常少，因此完成本书所花费的精力相对来说是非常大的。

本书从系统化的视角，为那些想学习Python网络爬虫或者正在研究Python网络爬虫的朋友们提供了一个全面的参考，让读者可以系统地学习Python网络爬虫的方方面面，在理解并掌握了本书的实例之后，能够独立编写出自己的Python网络爬虫项目，并且能够胜任Python网络爬虫工程师相关岗位的工作。

同时，本书的另一个目的是，希望可以给大数据或者数据挖掘方向的从业者一定的参考，以帮助这些读者从海量的互联网信息中爬取需要的数据。所谓巧妇难为无米之炊，有了这些数据之后，从事大数据或者数据挖掘方向工作的读者就可以进行后续的分析处理了。

本书的主要内容和特色

本书是一本系统介绍Python网络爬虫的书籍，全书注重实战，涵盖网络爬虫原理、如何手写Python网络爬虫、如何使用Scrapy框架编写网络爬虫项目等关于Python网络爬虫的方方面面。

本书的主要特色如下：

系统讲解Python网络爬虫的编写方法，体系清晰。

结合实战，让读者能够从零开始掌握网络爬虫的基本原理，学会编写Python网络爬虫以及Scrapy爬虫项目，从而编写出通用爬虫及聚焦爬虫，并掌握常见网站的爬虫反屏蔽手段。

有配套免费视频，对于书中的难点，读者可以直接观看作者录制的对应视频，加深理解。

拥有多个爬虫项目编写案例，比如博客类爬虫项目案例、图片类爬虫项目案例、模拟登录爬虫项目等。除此之外，还有很多不同种类的爬虫案例，可以让大家在理解这些案例之后学会各种类型爬虫的编写方法。

总之，在理解本书内容并掌握书中实例之后，读者将能胜任Python网络爬虫工程师方向的工作并学会各种类型网络爬虫项目的编写。此外，本书对于大数据或数据挖掘方向的从业者也非常有帮助，比如可以利用Python网络爬虫轻松获取所需的数据信息等。

本书面向的读者

Python网络爬虫初学者

网络爬虫工程师

大数据及数据挖掘工程师

高校计算机专业的学生

其他对Python或网络爬虫感兴趣的人员

. 如何阅读本书

本书分为四篇，共计20章。

第一篇为理论基础篇（第1~2章），主要介绍了网络爬虫的基础知识，让大家从零开始对网络爬虫有一个比较清晰的认识。

第二篇为核心技术篇（第3~9章），详细介绍了网络爬虫实现的核心技术，包括网络爬虫的工作原理、如何用Urllib库编写网络爬虫、爬虫的异常处理、正则表达式、爬虫中Cookie的使用、手写糗事百科爬虫、手写链接爬虫、手写微信爬虫、手写多线程爬虫、浏览器伪装技术、Python网络爬虫的定向爬取技术及实例等。学完这一部分内容，读者就可以写出自己的爬虫了。这部分的爬虫编写采用的是一步步纯手写的方式进行的，没有采用框架。

第三篇为框架实现篇（第10~17章），主要详细介绍了如何用框架实现Python网络爬虫项目。使用框架实现Python网络爬虫项目相较于手写方式更加便捷，主要包括Python爬虫框架分类、Scrapy框架在各系统中的安装以及如何避免各种“坑”、如何用Scrapy框架编写爬虫项目、Scrapy框架架构详解、Scrapy的中文输出与存储、在Scrapy中如何使用for循环实现自动网页爬虫、如何通过CrawlSpider实现自动网页爬虫、如何将爬取的内容写进数据库等。其中第12章为基础部分，读者需要着重掌握。

第四篇为项目实战篇（第18~20章），分别讲述了博客类爬虫项目、图片类爬虫项目、模拟登录爬虫项目的编程及实现。其中，也会涉及验证码处理等方面的难点知识，帮助读者通过实际的项目掌握网络爬虫项目的编写。

勘误和支持

由于作者的水平有限，书中难免有一些错误或不准确的地方，恳请各位读者不吝指正。

相关建议各位可以通过微博@韦玮pig或通过QQ公众号a67899或微信公众平台weijc7789（可以直接扫描下方二维码添加）进行反馈，也可以直接向邮箱ceo@iqianyue.com发送邮件，期待能够收到各位读者的意见和建议，欢迎来信。

致谢

感谢机械工业出版社华章公司的副总编杨福川老师与编辑李艺老师，在近一年的时间里，是你们一次次在我遇到困难的时候，给予我鼓励，让我可以坚持写下去。创作一本书是非常艰苦的，除了技术知识等因素之外，还需要非常大的毅力。特别感谢杨福川在写作过程中对我各方面的支持，尤其是对我毅力的培养。

感谢CSDN、51CTO与极客学院，因为你们，让我在这个领域获得了更多的学员与支持。

感谢恩师何云景教授对我创业方面的帮助，因为有您，我才拥有了一个更好的创业开端及工作环境。

特别致谢

最后，需要特别感谢的是我的女友，因为编写这本书，少了很多陪你的时间，感谢你的不离不弃与理解包容。希望未来可以加倍弥补你那些错过吃的美食和那些错过逛的街道。

同时，也要感谢你帮我完成书稿的校对工作，谢谢你的付出与支持。因为有了你默默的付出，我才能坚定地走下去；因为有了你不断的支持，我才可以安心地往前冲。

感谢爷爷从小对我人生观、价值观的培养，您是一个非常有思想的人。

感谢远方的父母、叔叔、姐姐，那些亲情的陪伴是我最珍贵的财富。

谨以此书献给热爱Python的朋友们!

作者介绍:

目录: 目录

前言

第一篇 理论基础篇

第1章 什么是网络爬虫 3

1.1 初识网络爬虫 3

1.2 为什么要学网络爬虫 4

1.3 网络爬虫的组成 5

1.4 网络爬虫的类型 6

1.5 爬虫扩展——聚焦爬虫 7

1.6 小结 8

第2章 网络爬虫技能总览 9

2.1 网络爬虫技能总览图 9

2.2 搜索引擎核心 10

2.3 用户爬虫的那些事儿 11

2.4 小结 12

第二篇 核心技术篇

第3章 网络爬虫实现原理与实现技术 15

3.1 网络爬虫实现原理详解 15

3.2 爬行策略 17

3.3 网页更新策略 18

3.4 网页分析算法 20

3.5 身份识别 21

3.6 网络爬虫实现技术 21

3.7 实例——metaseeker 22

3.8 小结 27

第4章 Urllib库与URLError异常处理 29

4.1 什么是Urllib库 29

4.2 快速使用Urllib爬取网页 30

4.3 浏览器的模拟——Headers属性 34

4.4 超时设置 37

4.5 HTTP协议请求实战 39

4.6 代理服务器的设置 44

4.7 DebugLog实战 45

4.8 异常处理神器——URLError实战 46

4.9 小结 51

第5章 正则表达式与Cookie的使用 52

5.1 什么是正则表达式 52

5.2 正则表达式基础知识 52

5.3 正则表达式常见函数 61

5.4 常见实例解析 64

5.5 什么是Cookie 66

5.6 Cookiejar实战精析 66

5.7 小结 71

第6章 手写Python爬虫 73

6.1 图片爬虫实战 73

6.2 链接爬虫实战 78

6.3 糗事百科爬虫实战 80

6.4 微信爬虫实战 82

6.5 什么是多线程爬虫	89
6.6 多线程爬虫实战	90
6.7 小结	98
第7章 学会使用Fiddler	99
7.1 什么是Fiddler	99
7.2 爬虫与Fiddler的关系	100
7.3 Fiddler的基本原理与基本界面	100
7.4 Fiddler捕获会话功能	102
7.5 使用QuickExec命令行	104
7.6 Fiddler断点功能	106
7.7 Fiddler会话查找功能	111
7.8 Fiddler的其他功能	111
7.9 小结	113
第8章 爬虫的浏览器伪装技术	114
8.1 什么是浏览器伪装技术	114
8.2 浏览器伪装技术准备工作	115
8.3 爬虫的浏览器伪装技术实战	117
8.4 小结	121
第9章 爬虫的定向爬取技术	122
9.1 什么是爬虫的定向爬取技术	122
9.2 定向爬取的相关步骤与策略	123
9.3 定向爬取实战	124
9.4 小结	130
第三篇 框架实现篇	
第10章 了解Python爬虫框架	133
10.1 什么是Python爬虫框架	133
10.2 常见的Python爬虫框架	133
10.3 认识Scrapy框架	134
10.4 认识Crawley框架	135
10.5 认识Portia框架	136
10.6 认识newspaper框架	138
10.7 认识Python-goose框架	139
10.8 小结	140
第11章 爬虫利器——Scrapy安装与配置	141
11.1 在Windows7下安装及配置Scrapy实战详解	141
11.2 在Linux (Centos) 下安装及配置Scrapy实战详解	147
11.3 在MAC下安装及配置Scrapy实战详解	158
11.4 小结	161
第12章 开启Scrapy爬虫项目之旅	162
12.1 认识Scrapy项目的目录结构	162
12.2 用Scrapy进行爬虫项目管理	163
12.3 常用工具命令	166
12.4 实战: Items的编写	181
12.5 实战: Spider的编写	183
12.6 XPath基础	187
12.7 Spider类参数传递	188
12.8 用XMLFeedSpider来分析XML源	191
12.9 学会使用CSVFeedSpider	197
12.10 Scrapy爬虫多开技能	200
12.11 避免被禁止	206
12.12 小结	212
第13章 Scrapy核心架构	214
13.1 初识Scrapy架构	214
13.2 常用的Scrapy组件详解	215

13.3 Scrapy工作流	217
13.4 小结	219
第14章 Scrapy中文输出与存储	220
14.1 Scrapy的中文输出	220
14.2 Scrapy的中文存储	223
14.3 输出中文到JSON文件	225
14.4 小结	230
第15章 编写自动爬取网页的爬虫	231
15.1 实战：items的编写	231
15.2 实战：pipelines的编写	233
15.3 实战：settings的编写	234
15.4 自动爬虫编写实战	234
15.5 调试与运行	239
15.6 小结	242
第16章 CrawlSpider	243
16.1 初识CrawlSpider	243
16.2 链接提取器	244
16.3 实战：CrawlSpider实例	245
16.4 小结	249
第17章 Scrapy高级应用	250
17.1 如何在Python3中操作数据库	250
17.2 爬取内容写进MySQL	254
17.3 小结	259
第四篇 项目实战篇	
第18章 博客类爬虫项目	263
18.1 博客类爬虫项目功能分析	263
18.2 博客类爬虫项目实现思路	264
18.3 博客类爬虫项目编写实战	264
18.4 调试与运行	274
18.5 小结	275
第19章 图片类爬虫项目	276
19.1 图片类爬虫项目功能分析	276
19.2 图片类爬虫项目实现思路	277
19.3 图片类爬虫项目编写实战	277
19.4 调试与运行	281
19.5 小结	282
第20章 模拟登录爬虫项目	283
20.1 模拟登录爬虫项目功能分析	283
20.2 模拟登录爬虫项目实现思路	283
20.3 模拟登录爬虫项目编写实战	284
20.4 调试与运行	292
20.5 小结	294
• • • • •	(收起)

[精通Python网络爬虫：核心技术、框架与项目实战_下载链接1](#)

标签

Python

Python网络爬虫

实战

爬虫

python

网络爬虫

乱

很好的书

评论

先不说这本书的质量，谨慎推断此书作者有黑社会背景或关系，他的爪牙不但私信骚扰骂人甚至电话骚扰打了差评的人，这种书评环境压力与黑社会无异，黑社会都升级了，挺进出版这种文化产业了。请豆瓣官方以及本书出版方华章对作者及其属下的行为予以约束。

奉劝豆瓣同学们：这本书可以看，但是绝对不能打差评，打了差评就等于捅了马蜂窝，黑社会对你及家人的生活骚扰是全方位的，我们毕竟不是独来独往的风清扬，在他们的威逼利诱下，我不得不把评分加一个星星，这好像也是之前大多数一星评价最后的结局：黑社会，惹不起啊。

此书质量:新手可以用来入门，与书名严重不相符，没有requests，没有pyspider，没有模拟，案例简陋，刷五星的小号刷也刷得很不专业，都自己承认和作者是熟人关系。

垃圾

书评全是水军也是厉害

作者水平低到常人難以企及的地步。

小学期的课程做爬虫，从图书馆借的这本事，因为完全是新手且之前没有看评分，所以从头到尾看了一遍。首先是错别字不能忍，多审几遍稿又不会死；其次是，本来以为爬虫就是书里这样，幸亏我看了评论，不然还不知道有很多重要知识本书根本没有涉及。

照葫芦画了几个爬虫，对基础理论解释不深。

对于初学者，还是值得一看的。网络爬虫基础的抓取示例都提供了，包括博客类、图片信息、BBS百科贴文、模拟用户登录等，但是没有浏览器selenium模拟插件及Beautiful Soup网页结构的解析，大部分例子围绕网页查看去定制正则表达式设计；缺少真正的编程算法和原理技巧的解读。给人的感觉，只能做基础的依葫芦画瓢。特别当前网站安全加强的场景下，HTTPS的安全访问要求，很多网页都需要模拟验证获取信息，且内容在JS包定义，不易仿照示例抓取。而爬虫框架主要写Scrapy，内容有点啰嗦，后面章节有点凑数。不管如何，给初学者还是有所启示的。

前面的原理部分还有一点点意思，而后面讲解工具就太简单了。整体来说，入门级教程吧。代码量太多。精通说不上吧。类似的书太多了。哎

开头部分太不友好了，不用requests库有点遗憾。ajax和selenium也没有讲到，scrapy讲的也不是很全面！大体上还可以看看！

还是要看最新的

乱

这都写些啥啊，本科生论文都比他写得好，还好是抽奖送的。亲们别读

糟心的一天

那些打5星的是收了多少钱？

这种书不应该骗骗还没毕业求职困难的大学生 作者搞快餐式培训 目的应该是面向就业 但这样的东西不重过程与实际生产场景 培养消极的代码习惯 太垃圾 又反复仔细看了下 入门都不行 不建议初学者求职者阅读

没什么价值。

基本没用 小学生案例 离企业级需求太远 而且发现目录与内容严重不符 欺骗消费者 建议豆瓣及出版商下架严肃处理；此外作者在受到负面评论后 每天拨打20个以上骚扰电话 如果再发现对广大消费者采取类似报复行为 将保留法律诉讼权 上报公安机关处理

这本书是推荐的，现在国内出的三本爬虫书我都买了，这本确实不错。不知道楼下叫韩某某还有dhcn的两位朋友怎么想的，没有PySpider就是不实战？懂得多就犟？首先说说PySpider这位，此仅仅只是一款框架，书里面已经有Scrapy了，要是把全部涉及的框架都写上，那这本书是文档汇编大全，不是书，PySpider本来就不是必须的内容。另外楼下那位韩某某，你确定你不是故意来黑的？我也是天善那期韦老师的学员，那期口碑怎样你我心里都有数，我就说嘛，为啥你总潜在群里，一到关键点就跑出来装不知道做一些负面的引导，你真名叫韩健，在群里名叫张健，要是不心虚我不知道为啥连姓都能换。最后，如果读都没读过的话，就不必发言了，因为你没有发言权，多尊重他人的劳动，多学习，多赚钱，少装逼。

极力推荐，内容非常实用，而且难得的是，基于Python3编写的，目前这类书非常少，干货非常多，良心之作。

这本书很好，满满的干货。都是具体的例子，可以直接理解知识，很快就能上手。而且是python3写的，很好。学习python爬虫必备的书籍。

书评

大家好，我是《精通Python网络爬虫》的图书作者，非常感谢大家对此书的关注。首先说明一点，我本人很少混豆瓣，偶然间看到豆瓣上居然有这么多人关注本书，观点都比较偏于两极，想想还是统一回复一下大家吧，希望能让大家可以有更客观、更清晰的了解，由于不常混于此，所以回复...

尊敬的各位读者和朋友们，我是本书的策划编辑，没想到本书在京东、当当等上架最及时的网络渠道还没有开始销售，就有这么多读者朋友关注这本书，有些意外，在此感谢大家。

我仔细看了一下大家在评论中关心的问题，我在这里做一个统一回复和说明，希望能解答大家的疑惑。

1.读者...

买了这本书随便翻了一下，发现真的是东拼西凑，写的都是一些浮于表面的东西，实在没什么干货可言，于是上豆瓣看了一下评论，韩某某的评论可以说还算客观：不过豆瓣上还有杨福川老师的解释贴，所以看了之后谁对谁错还真让人困惑了。杨老师一直是让我尊敬的出版人，可以说选书质...

[illegible]

因为对未来大数据的发展感兴趣，我选择了往这条路走。因为公众号——大数据（微信号：hzdashuju）的无私馈赠，我有幸阅读学习了韦玮老师的这本讲述大数据中数据采集这门绝技的书——《精通python网络爬虫》。

进入正题吧。 书目录我就不占篇幅了，有兴趣的读者可以百度。但很...

[精通Python网络爬虫：核心技术、框架与项目实战_下载链接1_](#)