

Python 3网络爬虫开发实战



[Python 3网络爬虫开发实战_下载链接1](#)

著者:崔庆才

出版者:人民邮电出版社

出版时间:2018-4

装帧:平装

isbn:9787115480347

本书介绍了如何利用Python 3开发网络爬虫，书中首先介绍了环境配置和基础知识，然后讨论了urllib、requests、正则表达式、Beautiful Soup、XPath、pyquery、数据存储、Ajax数据爬取等内容，接着通过多个案例介绍了

不同场景下如何实现数据爬取，后介绍了pyspider框架、Scrapy框架和分布式爬虫。

本书适合Python程序员阅读。

作者介绍:

崔庆才

北京航空航天大学硕士，静觅博客 (<https://cuiqingcai.com/>) 博主，爬虫博文访问量已过百万，喜欢钻研，热爱生活，乐于分享。欢迎关注个人微信公众号“进击的Coder”。

目录: 第1章 开发环境配置 1

1.1 Python 3的安装 1

1.1.1 Windows下的安装 1

1.1.2 Linux下的安装 6

1.1.3 Mac下的安装 8

1.2 请求库的安装 10

1.2.1 requests的安装 10

1.2.2 Selenium的安装 11

1.2.3 ChromeDriver的安装 12

1.2.4 GeckoDriver的安装 15

1.2.5 PhantomJS的安装 17

1.2.6 aiohttp的安装 18

1.3 解析库的安装 19

1.3.1 lxml的安装 19

1.3.2 Beautiful Soup的安装 21

1.3.3 pyquery的安装 22

1.3.4 tesseract的安装 22

1.4 数据库的安装 26

1.4.1 MySQL的安装 27

1.4.2 MongoDB的安装 29

1.4.3 Redis的安装 36

1.5 存储库的安装 39

1.5.1 PyMySQL的安装 39

1.5.2 PyMongo的安装 39

1.5.3 redis-py的安装 40

1.5.4 RedisDump的安装 40

1.6 Web库的安装 41

1.6.1 Flask的安装 41

1.6.2 Tornado的安装 42

1.7 App爬取相关库的安装 43

1.7.1 Charles的安装 44

1.7.2 mitmproxy的安装 50

1.7.3 Appium的安装 55

1.8 爬虫框架的安装 59

1.8.1 pyspider的安装 59

1.8.2 Scrapy的安装 61

1.8.3 Scrapy-Splash的安装 65

1.8.4 Scrapy-Redis的安装 66

1.9 部署相关库的安装 67

1.9.1 Docker的安装 67

1.9.2 Scrapy的安裝	71
1.9.3 Scrapy-Client的安裝	74
1.9.4 Scrapy API的安裝	75
1.9.5 Scrapyrt的安裝	75
1.9.6 Gerapy的安裝	76
第2章 爬虫基础	77
2.1 HTTP基本原理	77
2.1.1 URI和URL	77
2.1.2 超文本	78
2.1.3 HTTP和HTTPS	78
2.1.4 HTTP请求过程	80
2.1.5 请求	82
2.1.6 响应	84
2.2 网页基础	87
2.2.1 网页的组成	87
2.2.2 网页的结构	88
2.2.3 节点树及节点间的关系	90
2.2.4 选择器	91
2.3 爬虫的基本原理	93
2.3.1 爬虫概述	93
2.3.2 能抓怎样的数据	94
2.3.3 JavaScript渲染页面	94
2.4 会话和Cookies	95
2.4.1 静态网页和动态网页	95
2.4.2 无状态HTTP	96
2.4.3 常见误区	98
2.5 代理的基本原理	99
2.5.1 基本原理	99
2.5.2 代理的作用	99
2.5.3 爬虫代理	100
2.5.4 代理分类	100
2.5.5 常见代理设置	101
第3章 基本库的使用	102
3.1 使用urllib	102
3.1.1 发送请求	102
3.1.2 处理异常	112
3.1.3 解析链接	114
3.1.4 分析Robots协议	119
3.2 使用requests	122
3.2.1 基本用法	122
3.2.2 高级用法	130
3.3 正则表达式	139
3.4 抓取猫眼电影排行	150
第4章 解析库的使用	158
4.1 使用XPath	158
4.2 使用Beautiful Soup	168
4.3 使用pyquery	184
第5章 数据存储	197
5.1 文件存储	197
5.1.1 TXT文本存储	197
5.1.2 JSON文件存储	199
5.1.3 CSV文件存储	203
5.2 关系型数据库存储	207
5.2.1 MySQL的存储	207

- 5.3 非关系型数据库存储 213
 - 5.3.1 MongoDB存储 214
 - 5.3.2 Redis存储 221
- 第6章 Ajax数据爬取 232
 - 6.1 什么是Ajax 232
 - 6.2 Ajax分析方法 234
 - 6.3 Ajax结果提取 238
 - 6.4 分析Ajax爬取今日头条街拍美图 242
- 第7章 动态渲染页面爬取 249
 - 7.1 Selenium的使用 249
 - 7.2 Splash的使用 262
 - 7.3 Splash负载均衡配置 286
 - 7.4 使用Selenium爬取淘宝商品 289
- 第8章 验证码的识别 298
 - 8.1 图形验证码的识别 298
 - 8.2 极验滑动验证码的识别 301
 - 8.3 点触验证码的识别 311
 - 8.4 微博宫格验证码的识别 318
- 第9章 代理的使用 326
 - 9.1 代理的设置 326
 - 9.2 代理池的维护 333
 - 9.3 付费代理的使用 347
 - 9.4 ADSL拨号代理 351
 - 9.5 使用代理爬取微信公众号文章 364
- 第10章 模拟登录 379
 - 10.1 模拟登录并爬取GitHub 379
 - 10.2 Cookies池的搭建 385
- 第11章 App的爬取 398
 - 11.1 Charles的使用 398
 - 11.2 mitmproxy的使用 405
 - 11.3 mitmdump爬取“得到”App电子书信息 417
 - 11.4 Appium的基本使用 423
 - 11.5 Appium爬取微信朋友圈 433
 - 11.6 Appium+mitmdump爬取京东商品 437
- 第12章 pypider框架的使用 443
 - 12.1 pypider框架介绍 443
 - 12.2 pypider的基本使用 445
 - 12.3 pypider用法详解 459
- 第13章 Scrapy框架的使用 468
 - 13.1 Scrapy框架介绍 468
 - 13.2 Scrapy入门 470
 - 13.3 Selector的用法 480
 - 13.4 Spider的用法 486
 - 13.5 Downloader Middleware的用法 487
 - 13.6 Spider Middleware的用法 494
 - 13.7 Item Pipeline的用法 496
 - 13.8 Scrapy对接Selenium 506
 - 13.9 Scrapy对接Splash 511
 - 13.10 Scrapy通用爬虫 516
 - 13.11 Scrapyrt的使用 533
 - 13.12 Scrapy对接Docker 536
 - 13.13 Scrapy爬取新浪微博 541
- 第14章 分布式爬虫 555

14.1 分布式爬虫原理 555
14.2 Scrapy-Redis源码解析 558
14.3 Scrapy分布式实现 564
14.4 Bloom Filter的对接 569
第15章 分布式爬虫的部署 577
15.1 Scrapy分布式部署 577
15.2 Scrapy-Client的使用 582
15.3 Scrapy对接Docker 583
15.4 Scrapy批量部署 586
15.5 Gerapy分布式管理 590
• • • • • [\(收起\)](#)

[Python 3网络爬虫开发实战_下载链接1](#)

标签

Python

爬虫

python

编程

计算机

计算机科学

网络

评论

刚开始看到这么高评价以为是刷的，把书看完才知道自己多虑了，作者水平非常高，通篇干货，无一点水分。很适合有一定开发经验转做爬虫方向的同学

适合有一定python基础，并对爬虫和网页架构有一定了解的人。

目前看过写爬虫写得最好的书了，不仅有方法还会讲原理，解决了我不会处理Ajax请求的疑惑。scrapy框架还没看，暂时不打算深入了。

看了视频教程然后买的书，等了好几个月才上

看书比听课快多了…
作者github放了书稿所有内容和源代码。最后的scrapy框架和分布式爬虫用到再仔细看看吧。

突然发现，我居然读完了一本600页的书。这是第一次读完这么厚的专业类书籍。冥冥中产生了一种自信——不论再厚的书籍，以后都将不会是我的敌人了。

涉及到的工具过多。一开始就讲这么多工具安装会让新人摸不着头脑。

非专业学习，一脸懵逼

本以为这本书够接近实战了，学到最后才明白：实操类的知识还是应该在实际操作中自行摸索解决提高。一步一步跟书走就永远会被牵着鼻子。利用网络基本上可以解决所遇到的90%的问题。话说我学爬虫干嘛ಠ_ಠ？

国人写的，鼓励下给五星

虽然还是编码小白，至少学会了request\beautifulsoup\select三板斧。后面还有针对ajax 验证码识别代理设置和模拟登陆等的详解,小白就有点云里雾里啦。下一步要专门再学下html和css结构，以及selenium。

代码：<https://github.com/Python3WebSpider>

当下介绍scrapy最好的书籍

讲得很全面，也很详细，好书，好参考。

专为爬虫，想看更多python 的有点失望了

挑不出毛病，从只会requests.get到现在学会使用selenium和ajax，希望自己越来越熟练，可以写出一段自己的代码。

适合入门

懒得看Scrapy 前面几种基本够玩了 出成书蛮完蛋的 所谓实战无战可实 如果没有写着玩的兴趣 还是别碰爬虫了 自学的话大概了解一点 再看看最新博客啥的这么学还好一点

质量可以。

实践性强，设计思路对入门者很有启发， practical beginner tutorial

书评

书买来读了几章，感觉说的比较详细，学到一些这方面技能，不过要是能便宜些就更好了。自己也依葫芦画瓢写个爬虫，抓取最近5年豆瓣关于python的书籍评分榜：

《Fluent Python》 2015 9.6 《Python神经网络编程》 2018 9.6 《Django By Example》 2015 9.5 《流畅的Python》 2017...

刚学爬虫的时候，网上都推荐用这本书。看了几章之后无疾而终。转去看官方文档。再次翻开这本书，看的人直冒冷汗。

除了前几章各种库的安装部分可能是作者自己写的。剩下内容都是网上博文按部就班，或者三流英语水平的官方文档翻译。每章避重就轻，把文档用谷歌翻译一遍。其实...

书的目录看起来很全面，可是书的细节处理不是很到位。小的点也讲的不清不楚，感觉错误有点多，有点失望吧，毕竟是我的第一本python。北航硕士跟清华博士还是有差别的。

比如说书中的scrapy一节的scrapy运行机制，本书中讲的是从engine向spideraf发出requests可是之后用的时候yy...

看了高评分才买了这书。但真的写的很差，没用的内容啰嗦很多，开始说要讲更实用的方法时怎么都讲不清楚还自相矛盾，就是一带而过。不知道是不是作者也不知道怎么理解，只是从别处抄了过来。以后再也不能买在读学生写的书了，太浪费时间了。而且现在感觉爬虫不应该看书，应该从...

没想买书也会遇到刷分水军。这书非常失望，啰嗦本是无所谓也希望细点，但这书的啰嗦是浪费时间又讲不清楚的啰嗦。文字非常不流畅，不是照搬照抄拼凑，就是逻辑混乱讲不清楚。读起来就好像一本外国书被一个翻译水平很差的人进行翻译的感觉。可能因为作者还是个学生，又是工科生...

1.此书18年4月底买到，我淘宝，京东，当当都问遍了才在当当上买到，实体书应该是第一批读者，作者将此书前半部分的内容已经公布到网络上，大家可以去崔大大的博客中找到连接地址。

- 2.我看过崔大大的视频，微信公众号及博客中的大部分内容，因此对本书期望值很高；
- 3.此书我现在...

不得不说作者的水平确实很厉害。
最近买了这本书，也买了作者的视频《Python3网络爬虫实战》，书比视频讲解的细致很多，所以还是尽量购买书吧。
视频里有很多知识点一带而过，初学者可能无法理解。但是书的话就能一点一点扣明白。总体来说十分推荐 唯一的缺点就是有些昂贵，要...

[Python 3网络爬虫开发实战_下载链接1](#)