

实战Python网络爬虫



[实战Python网络爬虫 下载链接1](#)

著者:黄永祥

出版者:清华大学出版社

出版时间:2019-6-1

装帧:平装

isbn:9787302524892

本书从原理到实践，循序渐进地讲述了使用Python开发网络爬虫的核心技术。全书从逻辑上可分为基础篇、实战篇和爬虫框架篇三部分。基础篇主要介绍了编写网络爬虫所需的基础知识，包括网站分析、数据抓取、数据清洗和数据入库。网站分析讲述如何使用Chrome和Fiddler抓包工具对网站做全面分析；数据抓取介绍了Python爬虫模块Urllib和Requests的基础知识；数据清洗主要介绍字符串操作、正则和BeautifulSoup的使用；数据入库讲述了MySQL和MongoDB的操作，通过ORM框架SQLAlchemy实现数据持久化，进行企业级开发。实战篇深入讲解了分布式爬虫、爬虫软件的开发、12306抢票程序和微博爬取等。框架篇主要讲述流行的爬虫框架Scrapy，并以Scrapy与Selenium、Splash、Redis结合的项目案例，让读者深层次了解Scrapy的使用。此外，本书还介绍了

爬虫的上线部署、如何自己动手开发一款爬虫框架、反爬虫技术的解决方案等内容。

本书使用Python

3.X编写，技术先进，项目丰富，适合欲从事爬虫工程师和数据分析师岗位的初学者、大学生和研究生使用，也很适合有一些网络爬虫编写经验，但希望更加全面、深入理解Python爬虫的开发人员使用。

作者介绍:

黄永祥,CSDN博客专家和签约讲师，多年软件研发经验，主要从事机器人流程系统研发、大数据系统研发、网络爬虫研发以及自动化运维系统研发。擅长使用Python编写高质量代码，对Python有深入研究，热爱分享和新技术的探索。

目录: 第1章 理解网络爬虫 1

1.1 爬虫的定义 1

1.2 爬虫的类型 2

1.3 爬虫的原理 2

1.4 爬虫的搜索策略 4

1.5 爬虫的合法性与开发流程 5

1.6 本章小结 6

第2章 爬虫开发基础 7

2.1 HTTP与HTTPS 7

2.2 请求头 9

2.3 Cookies 10

2.4 HTML 11

2.5 JavaScript 12

2.6 JSON 14

2.7 Ajax 14

2.8 本章小结 15

第3章 Chrome分析网站 16

3.1 Chrome开发工具 16

3.2 Elements标签 17

3.3 Network标签 18

3.4 分析QQ音乐 20

3.5 本章小结 23

第4章 Fiddler抓包 24

4.1 Fiddler介绍 24

4.2 Fiddler安装配置 24

4.3 Fiddler抓取手机应用 26

4.4 Toolbar工具栏 29

4.5 Web Session列表 30

4.6 View选项视图 32

4.7 Quickexec命令行 33

4.8 本章小结 34

第5章 爬虫库Urllib 35

5.1 Urllib简介 35

5.2 发送请求 36

5.3 复杂的请求 37

5.4 代理IP 38

5.5 使用Cookies 39

5.6 证书验证 40

5.7 数据处理 41

5.8 本章小结	42
第6章 爬虫库Requests	43
6.1 Requests简介及安装	43
6.2 请求方式	44
6.3 复杂的请求方式	45
6.4 下载与上传	47
6.5 本章小结	49
第7章 Requests-Cache爬虫缓存	50
7.1 简介及安装	50
7.2 在Requests中使用缓存	50
7.3 缓存的存储机制	53
7.4 本章小结	54
第8章 爬虫库Requests-HTML	55
8.1 简介及安装	55
8.2 请求方式	56
8.3 数据清洗	56
8.4 Ajax动态数据抓取	59
8.5 本章小结	61
第9章 网页操控与数据爬取	62
9.1 了解Selenium	62
9.2 安装Selenium	63
9.3 网页元素定位	66
9.4 网页元素操控	70
9.5 常用功能	73
9.6 实战：百度自动答题	80
9.7 本章小结	85
第10章 手机App数据爬取	86
10.1 Appium简介及原理	86
10.2 搭建开发环境	87
10.3 连接Android系统	92
10.4 App的元素定位	97
10.5 App的元素操控	99
10.6 实战：淘宝商品采集	102
10.7 本章小结	107
第11章 Splash、Mitmproxy与Aiohttp	109
11.1 Splash动态数据抓取	109
11.1.1 简介及安装	109
11.1.2 使用Splash的API接口	112
11.2 Mitmproxy抓包	116
11.2.1 简介及安装	116
11.2.2 用Mitmdump抓取爱奇艺视频	116
11.3 Aiohttp高并发抓取	119
11.3.1 简介及使用	119
11.3.2 Aiohttp异步爬取小说排行榜	123
11.4 本章小结	126
第12章 验证码识别	128
12.1 验证码的类型	128
12.2 OCR技术	129
12.3 第三方平台	131
12.4 本章小结	134
第13章 数据清洗	136
13.1 字符串操作	136
13.1.1 截取	136
13.1.2 替换	137

- 13.1.3 查找 137
- 13.1.4 分割 138
- 13.2 正则表达式 139
 - 13.2.1 正则语法 140
 - 13.2.2 正则处理函数 141
- 13.3 BeautifulSoup数据清洗 144
 - 13.3.1 BeautifulSoup介绍与安装 144
 - 13.3.2 BeautifulSoup的使用示例 146
- 13.4 本章小结 149
- 第14章 文档数据存储 150
 - 14.1 CSV数据的写入和读取 150
 - 14.2 Excel数据的写入和读取 151
 - 14.3 Word数据的写入和读取 154
 - 14.4 本章小结 156
- 第15章 ORM框架 158
 - 15.1 SQLAlchemy介绍与安装 158
 - 15.1.1 操作数据库的方法 158
 - 15.1.2 SQLAlchemy框架介绍 158
 - 15.1.3 SQLAlchemy的安装 159
 - 15.2 连接数据库 160
 - 15.3 创建数据表 162
 - 15.4 添加数据 164
 - 15.5 更新数据 165
 - 15.6 查询数据 166
 - 15.7 本章小结 168
- 第16章 MongoDB数据库操作 169
 - 16.1 MongoDB介绍 169
 - 16.2 MogoDB的安装及使用 170
 - 16.2.1 MongoDB的安装与配置 170
 - 16.2.2 MongoDB可视化工具 172
 - 16.2.3 PyMongo的安装 173
 - 16.3 连接MongoDB数据库 173
 - 16.4 添加文档 174
 - 16.5 更新文档 175
 - 16.6 查询文档 176
 - 16.7 本章小结 178
- 第17章 实战：爬取51Job招聘信息 180
 - 17.1 项目分析 180
 - 17.2 获取城市编号 180
 - 17.3 获取招聘职位总页数 182
 - 17.4 爬取每个职位信息 184
 - 17.5 数据存储 188
 - 17.6 爬虫配置文件 190
 - 17.7 本章小结 191
- 第18章 实战：分布式爬虫——QQ音乐 193
 - 18.1 项目分析 193
 - 18.2 歌曲下载 194
 - 18.3 歌手的歌曲信息 198
 - 18.4 分类歌手列表 201
 - 18.5 全站歌手列表 203
 - 18.6 数据存储 204
 - 18.7 分布式爬虫 205
 - 18.7.1 分布式概念 205
 - 18.7.2 并发库concurrent.futures 206

18.7.3 分布式策略	207
18.8 本章小结	209
第19章 实战：12306抢票爬虫	211
19.1 项目分析	211
19.2 验证码验证	211
19.3 用户登录与验证	214
19.4 查询车次	219
19.5 预订车票	225
19.6 提交订单	227
19.7 生成订单	233
19.8 本章小结	236
第20章 实战：玩转微博	244
20.1 项目分析	244
20.2 用户登录	244
20.3 用户登录（带验证码）	253
20.4 关键词搜索热门微博	259
20.5 发布微博	264
20.6 关注用户	268
20.7 点赞和转发评论	271
20.8 本章小结	277
第21章 实战：微博爬虫软件开发	278
21.1 GUI库及PyQt5的安装与配置	278
21.1.1 GUI库	278
21.1.2 PyQt5安装及环境搭建	279
21.2 项目分析	281
21.3 软件主界面	284
21.4 相关服务界面	288
21.5 微博采集界面	292
21.6 微博发布界面	297
21.7 微博爬虫功能	308
21.8 本章小结	315
第22章 Scrapy爬虫开发	317
22.1 认识与安装Scrapy	317
22.1.1 常见爬虫框架介绍	317
22.1.2 Scrapy的运行机制	318
22.1.3 安装Scrapy	319
22.2 Scrapy爬虫开发示例	320
22.3 Spider的编写	326
22.4 Items的编写	329
22.5 Item Pipeline的编写	330
22.5.1 用MongoDB实现数据入库	330
22.5.2 用SQLAlchemy实现数据入库	332
22.6 Selectors的编写	333
22.7 文件下载	336
22.8 本章小结	339
第23章 Scrapy扩展开发	341
23.1 剖析Scrapy中间件	341
23.1.1 SpiderMiddleware中间件	342
23.1.2 DownloaderMiddleware中间件	344
23.2 自定义中间件	347
23.2.1 设置代理IP服务	347
23.2.2 动态设置请求头	350
23.2.3 设置随机Cookies	353
23.3 实战：Scrapy Selenium爬取豆瓣电影评论	355

- 23.3.1 网站分析 355
- 23.3.2 项目设计与实现 357
- 23.3.3 定义Selenium中间件 359
- 23.3.4 开发Spider程序 360
- 23.4 实战：Scrapy Splash爬取B站动漫信息 362
- 23.4.1 Scrapy_Splash实现原理 363
- 23.4.2 网站分析 363
- 23.4.3 项目设计与实现 365
- 23.4.4 开发Spider程序 367
- 23.5 实战：Scrapy Redis分布式爬取猫眼排行榜 369
- 23.5.1 Scrapy_Redis实现原理 369
- 23.5.2 安装Redis数据库 371
- 23.5.3 网站分析 372
- 23.5.4 项目设计与实现 373
- 23.5.5 开发Spider程序 375
- 23.6 分布式爬虫与增量式爬虫 377
- 23.6.1 基于管道实现增量式 378
- 23.6.2 基于中间件实现增量式 381
- 23.7 本章小结 384
- 第24章 实战：爬取链家楼盘信息 386
- 24.1 项目分析 386
- 24.2 创建项目 389
- 24.3 项目配置 389
- 24.4 定义存储字段 391
- 24.5 定义管道类 392
- 24.6 编写爬虫规则 396
- 24.7 本章小结 400
- 第25章 实战：QQ音乐全站爬取 402
- 25.1 项目分析 402
- 25.2 项目创建与配置 403
- 25.2.1 项目创建 403
- 25.2.2 项目配置 403
- 25.3 定义存储字段和管道类 405
- 25.3.1 定义存储字段 405
- 25.3.2 定义管道类 405
- 25.4 编写爬虫规则 408
- 25.5 本章小结 413
- 第26章 爬虫的上线部署 415
- 26.1 非框架式爬虫部署 415
- 26.1.1 创建可执行程序 415
- 26.1.2 制定任务计划程序 417
- 26.1.3 创建服务程序 421
- 26.2 框架式爬虫部署 424
- 26.2.1 Scrapyd部署爬虫服务 424
- 26.2.2 Gerapy爬虫管理框架 429
- 26.3 本章小结 434
- 第27章 反爬虫的解决方案 435
- 27.1 常见的反爬虫技术 435
- 27.2 基于验证码的反爬虫 436
- 27.2.1 验证码出现的情况 437
- 27.2.2 解决方案 438
- 27.3 基于请求参数的反爬虫 439
- 27.3.1 请求参数的数据来源 439
- 27.3.2 请求参数的查找 440

27.4 基于请求头的反爬虫 441
27.5 基于Cookies的反爬虫 443
27.6 本章小结 447
第28章 自己动手开发爬虫框架 449
28.1 框架设计说明 449
28.2 异步爬取方式 450
28.3 数据清洗机制 455
28.4 数据存储机制 457
28.5 实战：用自制框架爬取豆瓣电影 463
28.6 本章小结 468
• • • • • ([收起](#))

[实战Python网络爬虫_下载链接1](#)

标签

爬虫

Python

计算机

编程

计算机相关

AI

评论

爬虫案例失效速度也特快了，深度一般但内容广博。

看目录就知道这是我想要的虫子

写得可以

[实战Python网络爬虫_下载链接1](#)

书评

没有花很多的篇幅去介绍PYTHON，直接切入主题，介绍关于爬虫的一切，从网页的基础到网页分析工具，从网络抓包到手机数据爬取，还有数据分析清洗的方法，方方面面都包括了进去。但是虽然介绍的内容很多，但是每一处写的都过于简单，不是很深入，对于初学者来说，算是给指出了每...

[实战Python网络爬虫_下载链接1](#)