

Python爬虫开发 从入门到实战（微课版）



[Python爬虫开发 从入门到实战（微课版）_下载链接1](#)

著者:谢乾坤

出版者:人民邮电出版社

出版时间:2018-9-1

装帧:平装

isbn:9787115490995

本书较为全面地介绍了定向爬虫的开发过程、各种反爬虫机制的破解方法和爬虫开发的

相关技巧。全书共13章，包括绪论、Python基础、正则表达式与文件操作、简单的网页爬虫开发、高性能HTML内容解析、Python与数据库、异步加载与请求头、模拟登录与验证码、抓包与中间人爬虫、Android原生App爬虫、Scrapy、Scrapy高级应用、爬虫开发中的法律和道德问题等。除第1、12、13章外的其他章末尾都有动手实践，以帮助读者巩固本章和前面章节所学的内容。针对书中的疑难内容，还配有视频讲解，以便更好地演示相关操作。

作者介绍:

谢乾坤，网易高级数据挖掘工程师。前极客学院爬虫课程讲师，以网名『青南』与『kingname』开设爬虫开发课程，受众超过10万。多年爬虫开发经验，对爬虫开发有独到的理解。曾在多家知名公司从事爬虫开发工作，并为国内外众多基金公司与投资公司提供数据支持。

目录: 第1章 绪论 1

1.1 爬虫 2

1.2 爬虫可以做什么 2

1.2.1 收集数据 2

1.2.2 尽职调查 3

1.2.3 刷流量和秒杀 3

1.3 爬虫开发技术 4

第2章 Python基础 5

2.1 Python的安装和运行 6

2.1.1 在Windows中安装Python 6

2.1.2 在Mac OS中安装Python 7

2.1.3 在Linux中安装Python 8

2.2 Python开发环境 8

2.2.1 PyCharm介绍与安装 8

2.2.2 运行代码 9

2.3 Python的数据结构和控制结构 12

2.3.1 整数、浮点数和变量 12

2.3.2 字符串、列表、元组 13

2.3.3 数据的读取 14

2.3.4 字典与集合 16

2.3.5 条件语句 17

2.3.6 for循环与while循环 19

2.4 函数与类 21

2.4.1 函数 21

2.4.2 类与面向对象编程 28

2.5 阶段案例——猜数游戏 32

2.5.1 需求分析 32

2.5.2 核心代码构建 33

2.5.3 调试与运行 33

2.6 本章小结 34

2.7 动手实践 34

第3章 正则表达式与文件操作 35

3.1 正则表达式 36

3.1.1 正则表达式的基本符号 36

3.1.2 在Python中使用正则表达式 38

3.1.3 正则表达式提取技巧 41

3.2 Python文件操作 44

3.2.1 使用Python读/写文本文件 44

3.2.2 使用Python读/写CSV文件	46
3.3 阶段案例——半自动爬虫开发	49
3.3.1 需求分析	49
3.3.2 核心代码构建	50
3.3.3 调试与运行	51
3.4 本章小结	54
3.5 动手实践	54
第4章 简单的网页爬虫开发	55
4.1 使用Python获取网页源代码	56
4.1.1 Python的第三方库	56
4.1.2 requests介绍与安装	56
4.1.3 使用requests获取网页源代码	57
4.1.4 结合requests与正则表达式	59
4.2 多线程爬虫	60
4.2.1 多进程库（multiprocessing）	60
4.2.2 开发多线程爬虫	61
4.3 爬虫的常见搜索算法	62
4.3.1 深度优先搜索	62
4.3.2 广度优先搜索	63
4.3.3 爬虫搜索算法的选择	64
4.4 阶段案例——小说网站爬虫开发	64
4.4.1 需求分析	64
4.4.2 核心代码构建	65
4.4.3 调试与运行	68
4.5 本章小结	68
4.6 动手实践	68
第5章 高性能HTML内容解析	69
5.1 HTML基础	70
5.2 XPath	71
5.2.1 XPath的介绍	71
5.2.2 lxml的安装	71
5.2.3 XPath语法讲解	73
5.2.4 使用Google Chrome浏览器辅助构造XPath	77
5.3 Beautiful Soup4	81
5.3.1 BS4的安装	81
5.3.2 BS4语法讲解	82
5.4 阶段案例——大麦网演出爬虫	85
5.4.1 需求分析	85
5.4.2 核心代码构建	85
5.4.3 调试与运行	86
5.5 本章小结	87
5.6 动手实践	87
第6章 Python与数据库	88
6.1 MongoDB	89
6.1.1 MongoDB的安装	89
6.1.2 PyMongo的安装与使用	94
6.1.3 使用RoboMongo执行MongoDB命令	101
6.2 Redis	102
6.2.1 环境搭建	102
6.2.2 Redis交互环境的使用	103
6.2.3 Redis-py	104
6.3 MongoDB的优化建议	105
6.3.1 少读少写少更新	105
6.3.2 能用Redis不用MongoDB	106

6.4 阶段案例	107
6.4.1 需求分析	107
6.4.2 核心代码构建	107
6.4.3 调试与运行	108
6.5 本章小结	108
6.6 动手实践	108
第7章 异步加载与请求头	109
7.1 异步加载	110
7.1.1 AJAX技术介绍	110
7.1.2 JSON介绍与应用	110
7.1.3 异步GET与POST请求	111
7.1.4 特殊的异步加载	113
7.1.5 多次请求的异步加载	114
7.1.6 基于异步加载的简单登录	117
7.2 请求头 (Headers)	118
7.2.1 请求头的作用	118
7.2.2 伪造请求头	119
7.3 模拟浏览器	122
7.3.1 Selenium介绍	123
7.3.2 Selenium安装	124
7.3.3 Selenium的使用	124
7.4 阶段案例	128
7.4.1 需求分析	128
7.4.2 核心代码构建	128
7.4.3 调试与运行	130
7.5 本章小结	131
7.6 动手实践	131
第8章 模拟登录与验证码	132
8.1 模拟登录	133
8.1.1 使用Selenium模拟登录	133
8.1.2 使用Cookies登录	135
8.1.3 模拟表单登录	137
8.2 验证码	139
8.2.1 肉眼打码	139
8.2.2 自动打码	141
8.3 阶段案例——自动登录果壳网	144
8.3.1 需求分析	144
8.3.2 核心代码构建	145
8.3.3 运行与调试	146
8.4 本章小结	147
8.5 动手实践	147
第9章 抓包与中间人爬虫	148
9.1 数据抓包	149
9.1.1 Charles的介绍和使用	149
9.1.2 App爬虫和小程序爬虫	156
9.2 中间人爬虫	163
9.2.1 mitmproxy的介绍和安装	163
9.2.2 mitmproxy的使用	164
9.2.3 使用Python定制mitmproxy	165
9.3 阶段案例——Keep热门	170
9.3.1 需求分析	170
9.3.2 核心代码构建	170
9.3.3 调试运行	172
9.4 本章小结	172

- 9.5 动手实践 172
- 第 10 章 Android 原生 App 爬虫 173
 - 10.1 实现原理 174
 - 10.1.1 环境搭建 175
 - 10.1.2 使用 Python 操纵手机 178
 - 10.1.3 选择器 180
 - 10.1.4 操作 181
 - 10.2 综合应用 188
 - 10.2.1 单设备应用 188
 - 10.2.2 多设备应用（群控） 191
 - 10.3 阶段案例——BOSS 直聘爬虫 196
 - 10.3.1 需求分析 196
 - 10.3.2 核心代码构建 196
 - 10.3.3 调试与运行 197
 - 10.4 本章小结 197
 - 10.5 动手实践 198
- 第 11 章 Scrapy 199
 - 11.1 Scrapy 的安装 200
 - 11.1.1 在 Windows 下安装 Scrapy 200
 - 11.1.2 在 Linux 下安装 Scrapy 202
 - 11.1.3 在 Mac OS 下安装 Scrapy 202
 - 11.2 Scrapy 的使用 203
 - 11.2.1 创建项目 203
 - 11.2.2 在 Scrapy 中使用 XPath 207
 - 11.3 Scrapy 与 MongoDB 213
 - 11.3.1 items 和 pipelines 的设置 213
 - 11.3.2 在 Scrapy 中使用 MongoDB 215
 - 11.4 Scrapy 与 Redis 218
 - 11.4.1 Scrapy_redis 的安装和使用 218
 - 11.4.2 使用 Redis 缓存网页并自动去重 218
 - 11.5 阶段案例——博客爬虫 220
 - 11.5.1 需求分析 220
 - 11.5.2 核心代码构建 221
 - 11.5.3 调试与运行 226
 - 11.6 本章小结 228
 - 11.7 动手实践 228
- 第 12 章 Scrapy 应用 229
 - 12.1 中间件（Middleware） 230
 - 12.1.1 下载器中间件 230
 - 12.1.2 爬虫中间件 242
 - 12.2 爬虫的部署 246
 - 12.2.1 Scrapy 介绍与使用 246
 - 12.2.2 权限管理 253
 - 12.3 分布式架构 258
 - 12.3.1 分布式架构介绍 258
 - 12.3.2 如何选择 Master 259
 - 12.4 阶段案例 259
 - 12.5 本章小结 259
- 第 13 章 爬虫开发中的法律和道德问题 260
 - 13.1 法律问题 261
 - 13.1.1 数据采集的法律问题 261
 - 13.1.2 数据的使用 261
 - 13.1.3 注册及登录可能导致的法律问题 261
 - 13.1.4 数据存储 261

13.1.5 内幕交易 261
13.2 道德协议 262
13.2.1 robots.txt协议 262
13.2.2 爬取频率 262
13.2.3 不要开源爬虫的源代码 262
13.3 本章小结 262
• • • • • ([收起](#))

[Python爬虫开发 从入门到实战（微课版）_下载链接1](#)

标签

爬虫

Python

计算机

python

计算科学

Scrapy

程序设计

黑客

评论

简单易懂

系统讲解了使用Python开发爬虫的常用知识点，最后还有对Scrapy框架和爬虫相关法

律的介绍。

scrapy和分布式爬虫部分讲的比较好！

【2020读书】 39:
这是一本实操指南。语言：python；工具：Scrapy；数据库：MangoDB；缓存：Redis；用到的包：re（正则表达式）；request；lxml；编程工具：pythoncham；只记得这些条框了，写个小爬虫练习的时候再记录。

唉...还是有点用的 但是实战书的问题在于当书出版的时候 他喵的网站就改版了啊

尽请期待

除了书和微课，还有微信群可以沟通

很棒，深入浅出，大量实例，方便上手。附带的视频课程还是很直观的，不过我还是喜欢直接看文字……

作者很牛逼，一直在装逼。对内容对小白不太友好，下载个源码都很麻烦，还有书的字太小了。

挺好的入门python书，我一个三本文科生能读懂实践，不错！

[Python爬虫开发 从入门到实战（微课版）_下载链接1](#)

书评

序言 这篇文章没有代码，请放心阅读。

多年以后，面对人工智能研究员那混乱不堪的代码，我会想起第一次和S君相见的那个遥远的下午。那时的B公司，还是一个仅有6个人的小团队，Mac和显示器在桌上依次排开，大家坐在一起，不需要称呼姓名，转过脸去，对方就知道你在和他说话。一...

[Python爬虫开发 从入门到实战（微课版）_下载链接1](#)