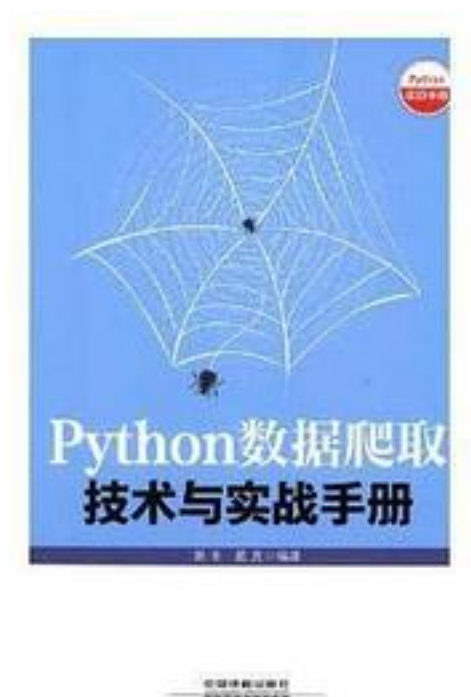


Python数据爬取技术与实战手册



[Python数据爬取技术与实战手册_下载链接1](#)

著者:郭卡

出版者:中国铁道出版社

出版时间:2018-8

装帧:

isbn:9787113245221

海量数据的产生和大数据的高价值利用，让数据爬取变得日益重要。本书为读者介绍了如何使用Python编写网络爬虫批量采集互联网数据，如何处理与保存采集到的信息，以及如何从众多纷乱的数据中提取到真正有用的信息。本书末尾介绍了几种常用的数据可视化工具。让读者能够从头到尾完整地完网络数据的采集与分析项目。本书理论与实例并重，既能够帮助数据从业者快速提升工作效率，又可以帮助大数据爱好者用网络爬虫方便生活。

作者介绍:

郭卡，辽宁师范大学硕士毕业，安徽外国语学院计算机教师，多年从事一线计算机教学及计算机等级培训工作，擅长计算机网络技术和教育学类数据统计分析技术，曾在中文核心期刊发表多篇技术论文。

戴亮，中南大学硕士毕业，数据挖掘从业者；在网络爬虫、数据分析、机器学习等领域有丰富的实战经验，在简书中贡献了许多高质量的技术文章，深受读者好评。

目录: 第1章 最佳拍档：网络爬虫与Python语言

1.1 什么是网络爬虫 1

1.1.1 网络爬虫的定义 2

1.1.2 网络爬虫的工作流程 2

1.1.3 网络爬虫的分类 3

1.1.4 为什么选择用Python编写网络爬虫 4

1.1.5 编写爬虫的注意事项 4

1.2 Python环境配置 5

1.2.1 Python的安装 5

1.2.2 Python第三方库的安装 6

【示例1-1】使用包管理器安装科学计算库numpy 6

【示例1-2】源代码方式安装xlrd库（使用setup.py文件） 7

【示例1-3】源代码方式安装xlrd库（使用whl文件） 8

1.2.3 Python开发工具的选择 8

【示例1-4】将文本编辑器配置成Python开发工具（以Notepad 为例） 12

1.3 Python基本语法 13

1.3.1 Python书写规则 13

1.3.2 Python基本数据类型 18

【示例1-5】以列表a = ['a','a','b','c','d','d','e']为例讲解List的基本操作 21

【示例1-6】以列表a = [1,2,3,4,5,6,7,8]为例讲解数据型列表的属性分析 23

【示例1-7】以字典a为例，讲解字典的基本操作 25

1.3.3 Python独有数据生成方式：推导式 29

1.3.4 函数 30

【示例1-8】局部变量与全局变量重名的运行结果与解决方案 31

1.3.5 条件与循环 34

1.3.6 类与对象 35

【示例1-9】请输出学生信息中某学生的班级、姓名和总分数 35

1.3.7 Python 2代码转为Python 3代码 36

【示例1-10】以文件test.py为例，介绍Python 2代码到Python 3代码的转化 37

第2章 应知应会：网络爬虫基本知识

2.1 网页的构成 38

2.1.1 HTML基本知识 39

2.1.2 网页中各元素的排布 46

【示例2-1】以新浪博客文本为例，学习各类元素的排布规则 46

2.2 正则表达式 48

2.2.1 正则表达式简介 48

2.2.2 Python语言中的正则表达式 49

【示例2-2】正则表达式应用中，当匹配次数达到10万时，预先编译对正则表达式性能的提升 51

2.2.3 综合实例：正则表达式的实际应用——在二手房网站中提取有用信息 52

2.3 汉字编码问题 54

2.3.1 常见编码简介 54

2.3.2 常用编程环境的默认编码 55

2.3.3 网页编码 56

2.3.4 编码转换 56

2.4 网络爬虫的行为准则 57

2.4.1 遵循Robots协议	57
2.4.2 网络爬虫的合法性	59
第3章 静态网页爬取	
3.1 Python常用网络库	61
3.1.1 urllib库	62
【示例3-1】从众多代理IP中选取可用的IP	63
【示例3-2】百度搜索“Python”url演示Parse模块应用	66
3.1.2 综合实例：批量获取高清壁纸	68
3.1.3 requests库	71
【示例3-3】用requests实现豆瓣网站模拟登录	72
3.1.4 综合实例：爬取历史天气数据预测天气变化	74
3.2 网页解析工具	77
3.2.1 更易上手：BeautifulSoup	77
【示例3-4】解析HTML文档（以豆瓣读书《解忧杂货店》为例）	78
3.2.2 更快速度：lxml	81
3.2.3 BeautifulSoup与lxml对比	82
【示例3-5】爬取豆瓣读书中近5年出版的评分7分以上的漫画	82
【示例3-6】BeautifulSoup和lxml解析同样网页速度测试（基于网易新闻首页）	85
3.2.4 综合实例：在前程无忧中搜索并抓取不同编程语言岗位的平均收入	85
第4章 动态网页爬取	
4.1 AJAX技术	89
4.1.1 获取AJAX请求	90
4.1.2 综合实例：抓取简书百万用户个人主页	91
4.2 Selenium操作浏览器	97
4.2.1 驱动常规浏览器	97
4.2.2 驱动无界面浏览器	100
4.2.3 综合实例：模拟登录新浪微博并下载短视频	101
4.3 爬取移动端数据	103
4.3.1 Fiddler工具配置	103
4.3.2 综合实例：Fiddle实际应用——爬取大角虫漫画信息	105
第5章 统一架构与规范：网络爬虫框架	
5.1 最流行的网络爬虫框架：Scrapy	111
5.1.1 安装须知与错误解决方案	111
5.1.2 Scrapy的组成与功能	112
5.2 综合实例：使用Scrapy构建观影指南	118
5.2.1 网络爬虫准备工作	119
5.2.2 编写Spider	121
5.2.3 处理Item	123
5.2.4 运行网络爬虫	124
5.2.5 数据分析	124
5.3 更易上手的网络爬虫框架：Pyspider	126
5.3.1 创建Pyspider项目	127
【示例5-1】利用Pyspider创建抓取煎蛋网项目并测试代码	127
5.3.2 运行Pyspider项目	129
第6章 反爬虫应对策略	
6.1 设置Headers信息	132
6.1.1 User-Agent	133
6.1.2 Cookie	136
6.2 建立IP代理池	138
6.2.1 建立IP代理池的思路	138
6.2.2 建立IP代理池的步骤	138
6.3 验证码识别	140
6.3.1 识别简单的验证码	141
【示例6-1】通过pytesseract库识别8个简单的验证码，并逐步提升准确率	141

6.3.2 识别汉字验证码	146
6.3.3 人工识别复杂验证码	146
6.3.4 利用Cookie绕过验证码	149
第7章 提升网络爬虫效率	
7.1 网络爬虫策略	152
7.1.1 广度优先策略	153
7.1.2 深度优先策略	153
7.1.3 按网页权重决定爬取优先级	154
7.1.4 综合实例：深度优先和广度优先策略效率对比 (抓取慕课网实战课程地址)	154
7.2 提升网络爬虫的速度	158
7.2.1 多线程	159
【示例7-1】使用4个线程同步抓取慕课网实战课程地址（基于深度优先策略）	159
7.2.2 多进程	161
7.2.3 分布式爬取	162
7.2.4 综合实例：利用现有知识搭建分布式爬虫（爬取百度贴吧中的帖子）	162
第8章 更专业的爬取数据存储与处理：数据库	
8.1 受欢迎的关系型数据库：MySQL	170
8.1.1 MySQL简介	170
8.1.2 MySQL环境配置	171
8.1.3 MySQL的查询语法	174
【示例8-1】使用MySQL查询语句从数据表Countries中选取面积大于10000km ² 的欧洲国家	177
8.1.4 使用pymysql连接MySQL数据库	178
8.1.5 导入与导出数据	179
8.2 应对海量非结构化数据：MongoDB数据库	180
8.2.1 MongoDB 简介	180
8.2.2 MongoDB环境配置	182
8.2.3 MongoDB基本语法	186
8.2.4 使用PyMongo连接MongoDB	188
8.2.5 导入/导出JSON文件	189
第9章 Python文件读取	
9.1 Python文本文件读写	190
9.2 数据文件CSV	192
9.3 数据交换格式JSON	193
9.3.1 JSON模块的使用	194
【示例9-1】请用JSON模块将data变量（包含列表、数字和字典的数组）转换成字符串并还原	194
9.3.2 JSON模块的数据转换	195
9.4 Excel读写模块：xlrd	195
9.4.1 读取Excel文件	196
9.4.2 写入Excel单元格	197
9.5 PowerPoint文件读写模块：pptx	197
9.5.1 读取pptx	197
9.5.2 写入pptx	198
9.6 重要的数据处理库：Pandas库	199
9.6.1 使用pandas库处理CSV文件	200
9.6.2 使用pandas库处理JSON文件	200
9.6.3 使用pandas库处理HTML文件	202
【示例9-2】用read_html()将某二手房网站表格中的数据提取出来	203
9.6.4 使用pandas库处理SQL文件	203
9.7 调用Office软件扩展包：win32com	204
9.7.1 读取Excel文件	204
9.7.2 读取Word文件	205

- 9.7.3 读取PowerPoint文件 205
- 9.8 读取PDF文件 206
 - 9.8.1 读取英文PDF文档 206
 - 9.8.2 读取中文PDF文档 208
 - 9.8.3 读取扫描型PDF文档 210
- 9.9 综合实例：自动将网络文章转化为PPT文档 211
- 第10章 通过API获取数据
 - 10.1 免费财经API——TuShare 214
 - 10.1.1 获取股票交易数据 215
 - 【示例10-1】 获取某股票2017年8月份的周K线数据 215
 - 10.1.2 获取宏观经济数据 217
 - 10.1.3 获取电影票房数据 219
 - 10.2 新浪微博API的调用 220
 - 10.2.1 创建应用 220
 - 10.2.2 使用API 222
 - 10.3 调用百度地图API 225
 - 10.3.1 获取城市经纬度 226
 - 【示例10-2】 使用百度地图API获取南京市的经纬度信息 226
 - 10.3.2 定位网络IP 226
 - 【示例10-3】 使用百度API定位IP地址（223.112.112.144） 226
 - 10.3.3 获取全景静态图 227
 - 10.4 调用淘宝API 228
- 第11章 网络爬虫工具
 - 11.1 使用Excel采集网页数据 231
 - 11.1.1 抓取网页中的表格 232
 - 11.1.2 抓取非表格的结构化数据 233
 - 11.2 使用Web Scraper插件 237
 - 11.2.1 安装Web Scraper 237
 - 11.2.2 Web Scraper的使用 238
 - 【示例11-1】 使用Web Scraper爬取当当网小说书目 238
 - 11.3 商业化爬取工具 240
 - 11.3.1 自定义采集 241
 - 【示例11-2】 利用网络爬虫软件八爪鱼自定义采集当当网图书信息 241
 - 11.3.2 网站简易采集 245
 - 【示例11-3】 利用网络爬虫软件八爪鱼的网络简易采集方式抓取房天下网中的合肥新房房价数据 245
- 第12章 数据分析工具：科学计算库
 - 12.1 单一类型数据高效处理：Numpy库 248
 - 12.1.1 ndarray数组 248
 - 【示例12-1】 对一维ndarray数组a进行读取、修改和切片操作 249
 - 【示例12-2】 对多维ndarray数组b进行读取、修改和切片操作 250
 - 【示例12-3】 对多维ndarray数组n进行矩阵运算（拼接、分解、转置、行列式、求逆和点乘） 252
 - 12.1.2 Numpy常用函数 253
 - 【示例12-4】 对多维ndarray数组a进行统计操作 253
 - 【示例12-5】 对一维ndarray数组a进行数据处理操作（去重、直方图统计、相关系数、分段、多项式拟合） 256
 - 12.1.3 Numpy性能优化 257
 - 12.2 复杂数据全面处理：Pandas库 258
 - 12.2.1 Pandas库中的4种基础数据结构 258
 - 12.2.2 Pandas使用技巧 264
 - 【示例12-6】 对比普通for循环遍历与iterrows()遍历方法的速度差异 264
 - 12.3 Python机器学习库：Scikit-learn 268
 - 【示例12-7】 以鸢尾花数据为例，使用Sklearn进行监督学习的基本建模过程（决策树

模型) 269
第13章 掌握绘图软件：将数据可视化
13.1 应用广泛的数据可视化：Excel绘图 271
13.1.1 绘制（对比）柱形图 272
13.1.2 绘制饼图并添加标注 273
13.1.3 其他图形 275
13.1.4 Excel频率分布直方图 276
【示例13-1】利用Excel绘制全国各省市城镇人员平均工资频率分布直方图 276
13.2 适合处理海量数据：Tableau绘图 278
13.2.1 基本操作：导入数据 278
13.2.2 绘制（多重）柱状对比图 279
13.2.3 智能显示：图形转换 281
13.2.4 绘制频率分布直方图 281
【示例13-2】利用Tableau绘制2015年我国城镇就业人员平均工资频率分布直方图 281
13.3 完善的二维绘图库：Matplotlib/Seaborn 283
13.3.1 使用Matplotlib绘制函数图表 283
13.3.2 使用Matplotlib绘制统计图表 285
13.4 优化：Seaborn的使用 289
13.5 综合实例：利用Matplotlib构建合肥美食地图 293
13.5.1 绘制区域地图 293
13.5.2 利用百度地图Web服务API获取美食地址 294
13.5.3 数据分析 298
13.5.4 绘制热力图完善美食地图展示 300
• • • • • ([收起](#))

[Python数据爬取技术与实战手册_下载链接1](#)

标签

python

爬虫

实战

CS

评论

一般，免不了现在大部分这类书的记流水账的俗套，不过覆盖面倒是挺广的，还讲到了

验证码识别和一些反扒的小技巧，收货还是有的，但是很多代码细节没有到位，进度忽快忽慢，让人没有读下去的欲望，前面和后面基本上都是快速翻过去的，python和其他的一些基础知识讲的太多了，明明是一本爬虫书，适合0基础。

爬虫，成为如今初创企业数据的重要来源，那么如何获取数据，从哪里获取，面对众多的反扒，我们应该怎么做呢？本书为读者介绍了如何使用Python编写网络爬虫批量采集互联网数据，如何处理与保存采集到的信息，以及如何从众多纷乱的数据中提取到真正有用的信息。加油，努力学习。

整本书由浅入深，适合有Python基础的初学者循序渐进地学习，书中也配合了相关的实战课程，让读者在学习的过程中能够得到足够的练习。毕竟实践是检验真理的唯一标准。还介绍了各种反爬虫策略，是一本好书！

[Python数据爬取技术与实战手册_下载链接1](#)

书评

书讲的非常好，涵盖了很多知识点，受益匪浅。感谢！
在编程中，总是会遇到很多的问题，令人迷惑，令人不解。这本书就是在为你迷惑的时候准备的，假如你是一个新手，还是去读实践类型的书吧，这本书你很可能看不懂。如果你没有一点编写Python程序的经验的话，书中那些作者在编程...

爬虫，顾名思义，爬取网络上的资料，不论是开源还是不开源，通过技术手段，爬取，在法律允许的范围内。
爬虫，成为如今初创企业数据的重要来源，那么如何获取数据，从哪里获取，面对众多的反扒，我们应该怎么做呢？本书为读者介绍了如何使用Python编写网络爬虫批量采集互联网数...

[Python数据爬取技术与实战手册_下载链接1](#)