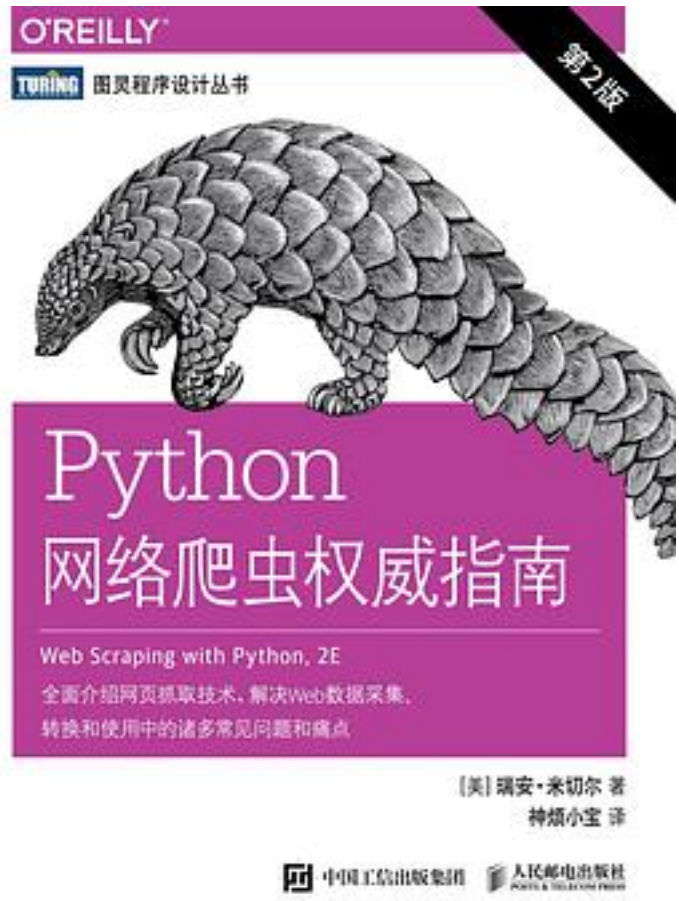
Python网络爬虫权威指南（第2版）



Python网络爬虫权威指南（第2版）_下载链接1_

著者:[美] 瑞安 · 米切尔

出版者:人民邮电出版社

出版时间:2019-4

装帧:平装

isbn:9787115509260

本书采用简洁强大的Python语言，全面介绍网页抓取技术，解答诸多常见问题，是掌握从数据爬取到数据清洗全流程的系统实践指南。书中内容分为两部分。第一部分深入讲解网页抓取的基础知识，重点介绍BeautifulSoup、Scrapy等Python库的应用。第二

部分介绍网络爬虫编写相关的主题，以及各种数据抓取工具和应用程序，帮你深入互联网的每个角落，分析原始数据，获取数据背后的故事，轻松解决遇到的各类网页抓取问题。第2版全面更新，新增网络爬虫模型、Scrapy和并行网页抓取相关章节。

- 解析复杂的HTML页面

- 使用Scrapy框架开发爬虫

- 学习存储数据的方法

- 从文档中读取和提取数据

- 清洗格式糟糕的数据

- 自然语言处理

- 通过表单和登录窗口抓取数据

- 抓取JavaScript及利用API抓取数据

- 图像识别与文字处理

- 避免抓取陷阱和反爬虫策略

- 使用爬虫测试网站

作者介绍:

瑞安·米切尔（Ryan Mitchell）

数据科学家、软件工程师，有丰富的网络爬虫和数据分析实战经验，目前就职于美国格理集团，经常为网页数据采集项目提供咨询服务，并在美国东北大学和美国欧林工程学院任教。

目录: 前言 xi
第一部分 创建爬虫
第1章 初见网络爬虫 3
1.1 网络连接 3
1.2 BeautifulSoup 简介 5
1.2.1 安装BeautifulSoup 6
1.2.2 运行BeautifulSoup 8
1.2.3 可靠的网络连接以及异常的处理 9
第2章 复杂HTML 解析 13
2.1 不是一直都要用锤子 13
2.2 再端一碗BeautifulSoup 14
2.2.1 BeautifulSoup 的find() 和find_all() 16
2.2.2 其他BeautifulSoup 对象 18
2.2.3 导航树 18
2.3 正则表达式 22
2.4 正则表达式和BeautifulSoup 25
2.5 获取属性 26
2.6 Lambda 表达式 26

第3章 编写网络爬虫 28
3.1 遍历单个域名 28
3.2 抓取整个网站 32
3.3 在互联网上抓取 36
第4章 网络爬虫模型 41
4.1 规划和定义对象 41
4.2 处理不同的网站布局 45
4.3 结构化爬虫 49
4.3.1 通过搜索抓取网站 49
4.3.2 通过链接抓取网站 52
4.3.3 抓取多种类型的页面 54
4.4 关于网络爬虫模型的思考 55
第5章 Scrapy 57
5.1 安装Scrapy 57
5.2 创建一个简易爬虫 59
5.3 带规则的抓取 60
5.4 创建item 64
5.5 输出item 66
5.6 item 管线组件 66
5.7 Scrapy 日志管理 69
5.8 更多资源 70
第6章 存储数据 71
6.1 媒体文件 71
6.2 把数据存储到CSV 74
6.3 MySQL 75
6.3.1 安装MySQL 76
6.3.2 基本命令 78
6.3.3 与Python 整合 81
6.3.4 数据库技术与最佳实践 84
6.3.5 MySQL 里的“六度空间游戏” 86
6.4 Email 88
第二部分 高级网页抓取
第7章 读取文档 93
7.1 文档编码 93
7.2 纯文本 94
7.3 CSV 98
7.4 PDF 100
7.5 微软Word 和.docx 102
第8章 数据清洗 106
8.1 编写代码清洗数据 106
8.2 数据存储后再清洗 111
第9章 自然语言处理 115
9.1 概括数据 116
9.2 马尔可夫模型 119
9.3 自然语言工具包 124
9.3.1 安装与设置 125
9.3.2 用NLTK 做统计分析 126
9.3.3 用NLTK 做词性分析 128
9.4 其他资源 131
第10章 穿越网页表单与登录窗口进行抓取 132
10.1 Python Requests 库 132
10.2 提交一个基本表单 133
10.3 单选按钮、复选框和其他输入 134
10.4 提交文件和图像 136

10.5 处理登录和cookie 136
10.6 其他表单问题 139
第11章 抓取JavaScript 140
11.1 JavaScript 简介 140
11.2 Ajax 和动态HTML 143
11.2.1 在Python 中用Selenium 执行JavaScript 144
11.2.2 Selenium 的其他webdriver 149
11.3 处理重定向 150
11.4 关于JavaScript 的最后提醒 151
第12章 利用API 抓取数据 152
12.1 API 概述 152
12.1.1 HTTP 方法和API 154
12.1.2 更多关于API 响应的介绍 155
12.2 解析JSON 数据 156
12.3 无文档的API 157
12.3.1 查找无文档的API 159
12.3.2 记录未被记录的API 160
12.3.3 自动查找和记录API 160
12.4 API 与其他数据源结合 163
12.5 再说一点API 165
第13章 图像识别与文字处理 167
13.1 OCR 库概述 168
13.1.1 Pillow 168
13.1.2 Tesseract 168
13.1.3 NumPy 170
13.2 处理格式规范的文字 171
13.2.1 自动调整图像 173
13.2.2 从网站图片中抓取文字 176
13.3 读取验证码与训练Tesseract 178
13.4 获取验证码并提交答案 183
第14章 避开抓取陷阱 186
14.1 道德规范 186
14.2 让网络机器人看着像人类用户 187
14.2.1 修改请求头 187
14.2.2 用JavaScript 处理cookie 189
14.2.3 时间就是一切 191
14.3 常见表单安全措施 191
14.3.1 隐含输入字段值 192
14.3.2 避免蜜罐 192
14.4 问题检查表 194
第15章 用爬虫测试网站 196
15.1 测试简介 196
15.2 Python 单元测试 197
15.3 Selenium 单元测试 201
15.4 单元测试与Selenium 单元测试的选择 205
第16章 并行网页抓取 206
16.1 进程与线程 206
16.2 多线程抓取 207
16.2.1 竞争条件与队列 209
16.2.2 threading 模块 212
16.3 多进程抓取 214
16.3.1 多进程抓取 216
16.3.2 进程间通信 217
16.4 多进程抓取的另一种方法 219

第17章 远程抓取 221
17.1 为什么要用远程服务器 221
17.1.1 避免IP 地址被封杀 221
17.1.2 移植性与扩展性 222
17.2 Tor 代理服务器 223
17.3 远程主机 224
17.3.1 从网站主机运行 225
17.3.2 从云主机运行 225
17.4 其他资源 227
第18章 网页抓取的法律与道德约束 228
18.1 商标、版权、专利 228
18.2 侵害动产 230
18.3 计算机欺诈与滥用法 232
18.4 robots.txt 和服务协议 233
18.5 3 个网络爬虫 236
18.5.1 eBay 起诉Bidder’s Edge 侵害其动产 236
18.5.2 美国政府起诉Auernheimer 与《计算机欺诈与滥用法》 237
18.5.3 Field 起诉Google：版权和robots.txt 239
18.6 勇往直前 239
关于作者 241
关于封面 241
・・・・・・(收起)

Python网络爬虫权威指南（第2版）_下载链接1_

标签

爬虫

Python

编程

python

计算机

再版

6产品・开发

計算機

评论

那个叫小宝的翻译，你说你抄袭第一版一样的也就罢了，108页程序里边变量名从第一版的item改成了word你在109页第十行还是写成item是不是太恶心了。编辑和校对也没仔细工作，差评。

主要库是urllib、request、selenium、bs4、pymysql，简单介绍了下scrapy框架，阅读难度不是很高，代码实例非常实用。

急需爬虫一只，这只是web爬虫

内容不深却很多，包括一般网络知识、常用的模块和框架介绍、数据处理和存储、自然语言处理、图像识别与文字处理、测试、甚至于道德法律规范。对于爬虫的各方面都有介绍，很值得一看。

还行吧

---
python版本更替原因，这本书有些案例代码无法实现

维基百科爬不了啊？！怎么办？

对于初学者跳跃有些大，而且涉及文本分词那里其实根本不会用到。
对于有基础的人又有些简单了。

Python网络爬虫权威指南（第2版）_下载链接1_

书评

第三章有好几个地方出现“分号”，但又实在不明白哪里有分号，只好查了原文。
原文是 colons，也就是冒号。 写在这里，给其他同学提个醒。： 这是冒号；
这是分号 公平地说，原书中也有一些低级错误，比如第七章开始不久，有个函数里把input 写成了content，中文版照抄了...

作者显然是此行达人，踩坑踩多了都是直接上经验。
书里的代码很优美、正规并且很简洁，运用了大量的递归算法和正则表达式。但是有些地方译者翻译的有误，比如第31页，倒数第六行冒号翻译成了分号，显然运行了源码并且对比了wiki网站才会知道这是误翻译。 另外，作者源码也有错...

5.3.2 基本命令 第二段第一句话： 除了用户自定义变量名（MySQL 5.x
版本是不区分大小写的，MySQL 5.0 之前的版本是不区分大小写的），MySQL
语句是不区分大小写的。（wtf？？？？？？？ 5.4 Email
查询圣诞节的代码缩进错误（sendMail函数和while都错了，会造成死循环！ 8.2...

诚然，这本书里面提到的一些python库不一定是最好的，但是整个爬虫的思路，还是非常值得大家借鉴。
其实python的语法，以及爬虫的代码段，都不难，就是写爬虫的过程中，需要注意的事项和有可能踩到的坑，是我比较看中的。
书中提到了一点，就是修改浏览器的header，默认貌似...

最近刚学了python3，看了一些讲语法的书籍和练手的题目，感觉这本书是一个比较好的系统的利用python完成从数据爬取到数据清洗整个流程的实践过程。觉得自己很有必要实践一下。刚刚看了下试读章节，15年出的英文版，难得的用python3进行工程实践而不只是讲语法的书。

1.可以尝试使用Google API 2.对于容易被封杀的站点使用tor来匿名
3.使用Tesseract识别验证码，可以训练特殊字体提高识别率
4.爬取整个网站的外链链接是件容易的事情 5.使用selenium作为测试网站的框架
6.注意cookie和request header的使用，努力让网站不把你当做爬虫对待

第177页的代码从逻辑上就不对啊，import的pytesseract就没用，而是通过subprocess调用，这应该是第一版的思路，不过我也搞不清这是作者还是译者的锅，把代码改成如下更合理 import time from urllib.request import urlretrieve from PIL import Image import pytesseract from...

我看过了 我看...

Python网络爬虫权威指南（第2版）_下载链接1_