

# 《Python网络数据采集》

## 图书基本信息

书名：《Python网络数据采集》

13位ISBN编号：978711541629X

出版时间：2016-3-1

作者：米切尔 (Ryan Mitchell)

页数：200

译者：陶俊杰,陈小莉

版权说明：本站所提供下载的PDF图书仅提供预览和简介以及在线试读，请支持正版图书。

更多资源请访问：[www.tushu000.com](http://www.tushu000.com)

# 《Python网络数据采集》

## 内容概要

本书采用简洁强大的Python语言，介绍了网络数据采集，并为采集新式网络中的各种数据类型提供了全面的指导。第一部分重点介绍网络数据采集的基本原理：如何用Python从网络服务器请求信息，如何对服务器的响应进行基本处理，以及如何以自动化手段与网站进行交互。第二部分介绍如何用网络爬虫测试网站，自动化处理，以及如何通过更多的方式接入网络。

# 《Python网络数据采集》

## 作者简介

Ryan Mitchell

数据科学家、软件工程师，目前在波士顿LinkeDrive公司负责开发公司的API和数据分析工具。此前，曾在Abine公司构建网络爬虫和网络机器人。她经常做网络数据采集项目的咨询工作，主要面向金融和零售业。另著有Instant Web Scraping with Java。

## 书籍目录

译者序	ix
前言	xi
第一部分 创建爬虫	
第1章 初见网络爬虫	2
1.1 网络连接	2
1.2 BeautifulSoup简介	4
1.2.1 安装BeautifulSoup	5
1.2.2 运行BeautifulSoup	7
1.2.3 可靠的网络连接	8
第2章 复杂HTML解析	11
2.1 不是一直都要用锤子	11
2.2 再端一碗BeautifulSoup	12
2.2.1 BeautifulSoup的find()和findAll()	13
2.2.2 其他BeautifulSoup对象	15
2.2.3 导航树	16
2.3 正则表达式	19
2.4 正则表达式和BeautifulSoup	23
2.5 获取属性	24
2.6 Lambda表达式	24
2.7 超越BeautifulSoup	25
第3章 开始采集	26
3.1 遍历单个域名	26
3.2 采集整个网站	30
3.3 通过互联网采集	34
3.4 用Scrapy采集	38
第4章 使用API	42
4.1 API概述	43
4.2 API通用规则	43
4.2.1 方法	44
4.2.2 验证	44
4.3 服务器响应	45
4.4 Echo Nest	46
4.5 Twitter API	48
4.5.1 开始	48
4.5.2 几个示例	50
4.6 Google API	52
4.6.1 开始	52
4.6.2 几个示例	53
4.7 解析JSON数据	55
4.8 回到主题	56
4.9 再说一点API	60
第5章 存储数据	61
5.1 媒体文件	61
5.2 把数据存储到CSV	64
5.3 MySQL	65
5.3.1 安装MySQL	66
5.3.2 基本命令	68

5.3.3	与Python整合	71
5.3.4	数据库技术与最佳实践	74
5.3.5	MySQL里的“六度空间游戏”	75
5.4	Email	77
第6章	读取文档	80
6.1	文档编码	80
6.2	纯文本	81
6.3	CSV	85
6.4	PDF	87
6.5	微软Word和.docx	88
第二部分	高级数据采集	
第7章	数据清洗	94
7.1	编写代码清洗数据	94
7.2	数据存储后再清洗	98
第8章	自然语言处理	103
8.1	概括数据	104
8.2	马尔可夫模型	106
8.3	自然语言工具包	112
8.3.1	安装与设置	112
8.3.2	用NLTK做统计分析	113
8.3.3	用NLTK做词性分析	115
8.4	其他资源	119
第9章	穿越网页表单与登录窗口进行采集	120
9.1	Python Requests库	120
9.2	提交一个基本表单	121
9.3	单选按钮、复选框和其他输入	123
9.4	提交文件和图像	124
9.5	处理登录和cookie	125
9.6	其他表单问题	127
第10章	采集JavaScript	128
10.1	JavaScript简介	128
10.2	Ajax和动态HTML	131
10.3	处理重定向	137
第11章	图像识别与文字处理	139
11.1	OCR库概述	140
11.1.1	Pillow	140
11.1.2	Tesseract	140
11.1.3	NumPy	141
11.2	处理格式规范的文字	142
11.3	读取验证码与训练Tesseract	146
11.4	获取验证码提交答案	151
第12章	避开采集陷阱	154
12.1	道德规范	154
12.2	让网络机器人看起来像人类用户	155
12.2.1	修改请求头	155
12.2.2	处理cookie	157
12.2.3	时间就是一切	159
12.3	常见表单安全措施	159
12.3.1	隐含输入字段值	159

12.3.2	避免蜜罐	160
12.4	问题检查表	162
第13章	用爬虫测试网站	164
13.1	测试简介	164
13.2	Python单元测试	165
13.3	Selenium单元测试	168
13.4	Python单元测试与Selenium单元测试的选择	172
第14章	远程采集	174
14.1	为什么要用远程服务器	174
14.1.1	避免IP地址被封杀	174
14.1.2	移植性与扩展性	175
14.2	Tor代理服务器	176
14.3	远程主机	177
14.3.1	从网站主机运行	178
14.3.2	从云主机运行	178
14.4	其他资源	179
14.5	勇往直前	180
附录A	Python简介	181
附录B	互联网简介	184
附录C	网络数据采集的法律与道德约束	188
作者简介		200
封面介绍		200

## 精彩短评

1、还算不错，因为接触爬虫有段时间了，所以前半部分就当梳理和复习了。

从第二部分里还是可以学到点不错的第三方库的，不过想要更深入的学习得将这些第三方库吃透才行。

2、上周简单翻了一遍，接下来看 Requests。20161118

3、2017.3.18

对于python刚入门的来说还是挺友好的

介绍了很多新奇的东西

4、编程还是很多年前的事了，嵌入式，汇编，C是当年我的长项，没有复杂晦涩的语法和长名词需要学习，之前从来没有学过Python，但这本书我看了两遍，一方面为Python的极简魅力所折服，另一方面为本书之内容深深地吸引了，他不但带领你从头到尾的学习了一遍如何使用Python访问网络，又如何使用Python和数据库、语义处理所连接，内容全面简单易懂，更重要的是他像其他工具书那样只讲一堆第三方库怎么用，而是把很多重要的功能点带你用极简的方式重新实现了一遍，哪怕你是从头开始研究如何实现中文分词，这样的领路人都是足够了呢。

强力推荐极客入门阅读。

5、很有意思

6、挺好的，系统的学习了一下爬虫。

7、这本书主要讲了两方面的内容，一是网络爬虫，爬取网站数据的基本原理，包括使用python的beautiful软件库，爬取网站数据的基本原理，复杂web页面的解析方法，爬取网络数据的数据存储管理方法，不同编码类型，不同，文档格式的数据读取方法，使用api采集web页面获取标准xml，json格式数据等。

二是高级数据采集方法，主要包括非规范格式数据的数据清洗，穿越登录表，认证页面，cookie，JavaScript脚本等网站数据的采集方法，以及网站图片文字识别等技术。

8、可以跳过代码去看，能对python的数据抓取相关工具有一个初步的认识

9、由于写了一段时间的爬虫，很多东西我已经知道，所以只花了一天时间就翻完了，书中讲了一些反爬虫机制以及如何应对，也有和爬虫关系不太大的话题比如数据库存储等

10、介绍了BeautifulSoup(HTML解析)，scrapy(爬虫)，nltk(几个简单的nlp例子)和request(login, cookie)相关用法和示例

11、总体来说，讲了一个普通爬虫遇到的很多事情。最喜欢的还是图像文字处理那部分。

12、很薄，入门中的入门，适合非计算机专业的数据分析师学习。

13、缺点是选择的一些库并不是现在最合适的 然后非常适合入门 提供了一个完整的思路框架

14、python爬虫入门。

15、过于入门...主要用bs4。且略微旧，scrapy已经支持python3了。

16、拿来入门可以，主要用bs4和正则表达式之类的来抓数据，生产环境里我一般用requests和lxml，如果网站反爬策略厉害，就用大杀器selenium，webdriver

17、简易的入门，架构显得奇怪

18、太水了，没有干货，某章还有一句JS是一门过时的技术

19、ORELLY的书我觉得到最后作为收藏正好，无论是浅显入门的还是那些略专业的工具书

20、很简单

21、适合入门，感觉是各种爬虫库文档的精简版

22、非常好，适合初学者入门。

23、自己抓数据有段时间了，这本书讲了爬虫的用途、思路和常用的工具（很多我都用过），如果我早一点发现这本书就好了（自己躺坑还是太累了

24、基础入门

25、全面但浅简，适合了解与入门。

26、不知道为什么评分那么低。框架的话 scrapy 可能更流行，这本书从出版到现在也应该有些年头了，但爬虫涉及的方方面面都至少指明了学习方向。是本好书。

- 27、可以的
- 28、看了三遍。为什么作者很喜欢用内置函数名来命名变量
- 29、适合入门的书籍，如果你对HTTP，HTML，Python这些都不了解的话
- 30、bs介绍的很详细
- 31、对爬虫的一些关键细节讲解的不错。整体上讲，本书内容全，有条理，但还不够细。
- 32、广而不精
- 33、对于小白来讲已经很满足，基本介绍到了爬虫的方方面面，还有代码示例和配套的网站。
- 34、入门书，但需要其它的一些基础，比如Python，HTML/CSS
- 35、数据采集的启迪书，不得不说这本书写得很浅显易懂，也出来得很及时，比较全但是深度还是略显不够
- 36、爬虫入门作，BeautifulSoup护佑着你，不包含进阶知识
- 37、读完就觉得流畅，干货也多，看得出是一个一线高手的实力总结。也确实是作者说的那种适合有一定基础的人看的书，后面的很多章节其实都可以独立成一本书。
- 38、打开了新世界的大门
- 39、买个电子版方便很多，可以查找代码。
- 40、挺好的 有实例有供爬的网站 系统的介绍
- 41、每一点都写的不详细
- 42、初学爬虫挺有用的
- 43、终于有本针对3.X的爬虫书了~
- 44、本书对于爬虫技术介绍的很全面，而且大多数内容是点到为止。颇有一种“师父领进门修行在个人”的意味。从代码和写作风格可以看出确实是作者多年来的经验之谈。全书的文字风格随意但是又不缺严谨，有幽默的风格。很适合，初学者或是像我这样没有系统地学习过爬虫技术的人。
- 45、篇幅不大，三四天看完。对涉及的内容点到即止，属于带你入门，深究靠自己再去了解。对网络数据采集没有研究的人来说还是不错的，至少让你知道了有什么工具解决什么问题。
- 46、一般，讲的很浅，给个大概认识吧
- 47、水
- 48、第一本爬虫书，扫清恐惧~
- 49、爬虫入门经典 知识点很全面
- 50、难得用python3讲解如何爬虫，对使用python3的人来说是一大福音。网上大部分的教程使用python2进行爬虫，和python3用的库有较大不同。但是python就是这样，年轻又有活力，意味着不仅仅是python本身，包括第三方库更新的也特别快。尽管这是一本2016年出版的书，但是书中的代码并不能完全复用，因为有些用到的库已经更新了：或者是接口，或者是输出，和书中的不尽相同。



## 精彩书评

1、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，中文版照抄了下来。第97页那段代码，如果你不明白它是怎么做到的，请翻到105页查看。

## 版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:[www.tushu000.com](http://www.tushu000.com)