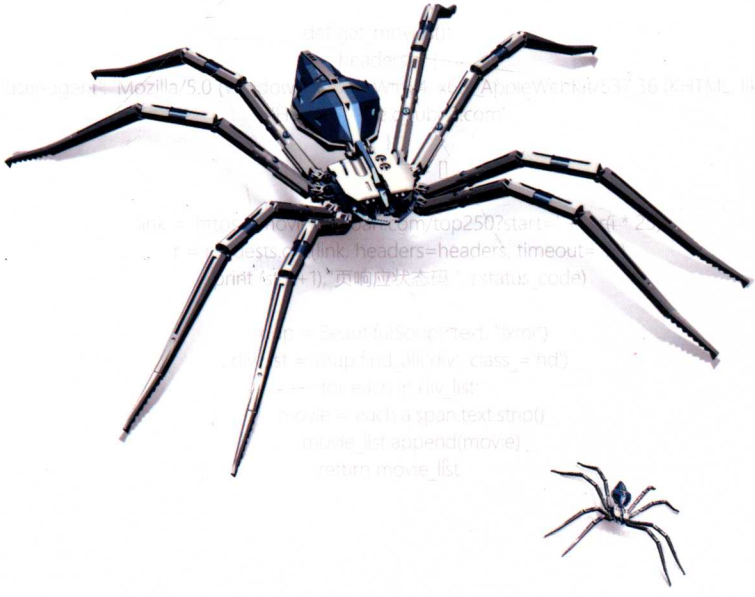
Python 网络爬虫 从入门到实践

打开数据信息收集大门的
金钥匙
范例源代码下载

从基础到应用，解读获取网页、解析网页、存储数据的三大爬虫技术

唐 松 陈智铨 编著

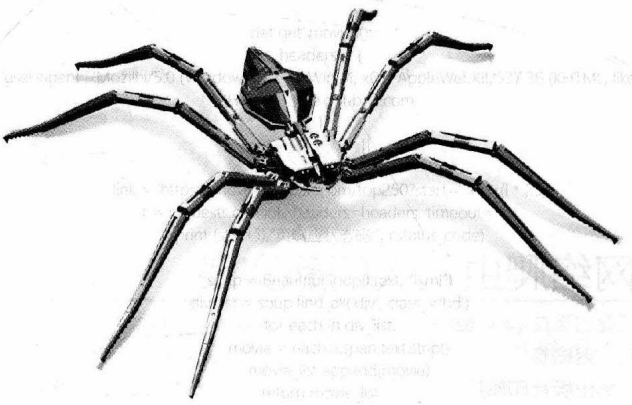


丰富、真实的案例
解构爬虫主流框架
多线程与分布式爬虫进阶实战

机械工业出版社
China Machine Press

Python 网络爬虫 从入门到实践

唐 松 陈智铨 编著



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Python网络爬虫从入门到实践 / 唐松，陈智铨编著. – 北京：机械工业出版社，2017.9

ISBN 978-7-111-57841-3

I. ①P… II. ①唐… ②陈… III. ①软件工具－程序设计 IV. ①TP311.561

中国版本图书馆CIP数据核字（2017）第212791号

使用 Python 编写网络爬虫程序获取互联网上的大数据是当前的热门专题。本书内容包括三部分：基础部分、进阶部分和项目实践部分。基础部分（第 1~6 章）主要介绍爬虫的三个步骤（获取网页、解析网页和存储数据），并通过诸多示例的讲解，让读者从基础内容开始系统性地学习爬虫技术，并在实践中提升 Python 爬虫水平。进阶部分（第 7~12 章）包括多线程的并发和并行爬虫、分布式爬虫、更换 IP 等，帮助读者进一步提升爬虫水平。项目实践部分（第 13~16 章）使用本书介绍的爬虫技术对几个真实的网站进行抓取，让读者能在读完本书后根据自己的需求写出爬虫程序。

无论是否有编程基础，只要是对爬虫技术感兴趣的读者，本书就能带领你从入门到进阶，再到实战，一步步了解爬虫，最终写出自己的爬虫程序。

Python 网络爬虫从入门到实践

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：夏非彼　迟振春　　责任校对：王　叶
印　　刷：中国电影出版社印刷厂　　版　　次：2017 年 9 月第 1 版第 1 次印刷
开　　本：170mm×242mm　1/16　　印　　张：16.25
书　　号：ISBN 978-7-111-57841-3　　定　　价：49.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzit@hzbook.com

推荐序一

我们正处于飞速发展的大数据时代。不同于以往，现如今丰富的数据信息让我们有能力更好地了解消费者、顾客和竞争对手。通过电商网站评论收集可以及时知悉顾客对于产品的看法，通过微博数据收集可以及时洞察潜在消费者的购买意向和需求，通过对手网站信息收集可以及时知晓对手的实时动态，真正做到运筹帷幄之中，决胜千里之外。

本书就是帮助你打开数据信息收集大门的钥匙！本书从最基本的 Python 语言讲起，完整地介绍了爬虫程序的每一个知识模块，同时附有最新案例教大家如何利用学到的知识进行实操，让不了解 Python 语言的人也可以在短时间内掌握爬虫程序的编写，快速成长为爬虫高手。本书条理清晰、层次分明，实用性极强。

作者唐松是一名年轻有为、经验丰富的数据分析专家。他通过这本书和读者分享多年网络爬虫和数据挖掘的经验。这本书是 IT 人士、企业管理人员、市场营销人员和有志于在数据分析方面有所突破的人士值得一读的好书。

香港中文大学市场系教授
刘建南
2017 年 6 月 26 日

推荐序二

“工欲善其事，必先利其器。”

——《论语·卫灵公》

读这本书需要考虑这样一个问题：到底是学习 Python 重要，还是掌握网络爬虫重要，抑或两者一样重要？对于这个问题的回答将直接影响读者最后能从这本书里学到什么。我来给大家一个简单的定义，网络爬虫是“事”，而 Python 是“器”，是用来进行网络爬虫的锋刃。在这个定义下，这一问题就转化成了究竟是“事”重要还是“器”重要。

如果你是管理者，那么这个答案就更接近于“事”。因为管理者在向所在的团队发号施令之前，要先思考你的想法到底能否和所在团队的技术优势相契合。近年来，我们眼见大数据以“迅雷不及掩耳之势”席卷全球，但这场旋风的背后是一个残酷的事实：很多公司在迈入大数据领域后遭遇“滑铁卢”。究其原因，管理层的惰性首当其冲。当管理层只知道在高尔夫球场对大数据高谈阔论时，我们又如何期冀技术部门能够懂得并且做出管理者想要做的“事”呢？因此，对于本书的前 12 章，管理者要能够清晰地回答两个问题：第一个问题是这个章节探讨的是什么问题，第二个问题是为什么要探讨这个问题。举个例子，对于第 1 章，如果你的公司想开发一款新的绿色产品，当你想通过爬取淘宝网上所有绿色产品（如空气净化器）的销量数据来做潜在市场评估时，就要考虑爬虫有哪些潜在的法律纠纷、公司的爬虫合不合法。再举个例子，对于第 6 章，你需要思考数据的存储对公司有什么影响，如何存储数据更有利于公司各个部门（如销售部门）的高效利用，能够更方便地与公司的数据库对接等。

如果你是技术人员、学者或技术“小白”，这本书就是教你如何“利其器”。与其认为是通过 Python 学习网络爬虫，读者不如把这本书看作是通过爬虫来学习 Python。Python 是“器”，就意味着世上也有其他的“器”（如 R、Java 等），而这些工具之间的基本思想往往是类似的。例如，都免不了做逻辑判断（如 if、then）和循环计算（如 for 循环），软件语言之间往往是一通百通。Python 是时下热门的开源软件（这意味着有人源源不断地开发更新且更强大的包

给你用），又不像 C 语言那般底层而需要巨大的学习成本。读者若是能始于爬虫，一步一个脚印，慢慢掌握好 Python 这门语言，将来还可以利用 Python 甚至是其他语言去做一些附加值更高的“事”，如人工智能、统计建模等。很遗憾，这条学习的道路没有什么捷径可走，唯一的方法就是不断尝试、不断失败、不断改进。本书最后 4 章综合考虑登录、爬虫速度、反爬虫、大规模爬虫等现实中遇到的问题，为读者提供一系列非常值得着手操作的实战演练范例。本书语言质朴、逻辑清晰、循序渐进，即使是零基础的读者，在本书的辅导下也可以对 Python 爬虫一步步地从入门到精通。

工欲善其事，必先利其器。希望读者能够借此书获得“事”或“器”，了解技术团队的运作模式，同时希望技术“小白”能通过时常查阅此书不断解决爬虫中遇到的疑惑。

香港中文大学市场系博士生
李宜威
2017 年 6 月 25 日

前言

爬虫程序是 DT（Data Technology，数据技术）收集信息的基础，程序员爬取目标网站的资料后，就可以分析和建立应用了。我们关心的是科技如何给大家带来实效，进而实现目标和理想，不能应用的技术称为魔术，只能用于表演。我们十分关注读者能否把握爬虫概念，所以相关的技术结合不同的实例讲解，希望能指导读者完成整个数据采集的流程。

Python 是一个简单、有效的语言，爬虫所需的获取、存储、整理等流程都可以使用 Python 系统地实现。此外，绝大部分计算机也可以直接使用 Python 语言或简单地安装 Python 系统，相信读者一定能轻松地把 Python 作为爬虫的主要技术。

动其心者，当具有大本大源

DT 的核心是从信息的源头去理解和分析，以做出能打动对方的行动决策方案。由谷歌搜索到现在的大数据时代，爬虫技术的重要性和广泛性一直很突出。程序员理解了信息的获取、存储和整理各方面的基本关系，才有可能系统地收集和应用不同源头和千变万化的网站信息。

数据共享

程序员要建立共利的互联网环境，不能把爬虫作为窃取数据的工具，爬虫必须在合情、合法、合理的情况下获取和应用。尊重数据供应者的知识产权和正常运作才能产生长久共利的环境。保障对方平台的正常运作是每个程序员都应当做到的，因此我们把爬虫的制约放在本书的第 1 章讨论。

自强不息

互联网科技不断更新和进步，网站信息也随之不断改变。爬虫的乐趣在于如何一直高效率、持续不断地从日新月异的网站中获取信息。另外，程序员要不断学习新技术，自我提高，这样在爬虫的过程中才能够理解互联网的运作和结构。

最后，感谢好友唐松给予我一起创作这本书的机会，让我可以分享爬虫技术和当中的乐趣。

思路富邦智能应用有限公司行政总裁
陈智铨

前言二

近年来，大数据成为业界与学术界最火热的话题之一，数据已经成为每个公司极为重要的资产。互联网大量的公开数据为个人和公司提供了以往想象不到的可以获取的数据量。而掌握网络爬虫技术可以帮助你获取这些有用的公开数据集。

执笔本书的起因是我打算在知乎上写博客向香港中文大学市场营销学的研究生讲解 Python 网络爬虫技术，让这些商科学生掌握一些大数据时代重要的技术能力。因此，本书除了面向技术人员外，还面向不懂编程的小白。本书尽量做到浅显易懂，希望能够将网络爬虫学习的门槛降低，让大家都能享受到使用网络爬虫编程的乐趣。

我是从商科自学转到数据科学的，因此编程和数据挖掘能力都是上网自学的。在这个过程中，我深刻地体会到，与不知所云的教学相比，深入浅出的教学对学习效率有很大提升。因此，学习知识最重要的两点是，通过富有逻辑的框架解构学习和通过实战解决实际问题，从而增强学习效果。本书的内容侧重于将网络爬虫技术进行框架性的解构，并使用代码将爬虫技术应用于抓取真实的网站。

本书所有代码均在 Python 3.6 中测试通过，可以从 Github 下载这些代码，地址为 https://github.com/Santostang/PythonScraping；也可以从百度网盘下载，地址为 http://pan.baidu.com/s/1c2w9rck（注意区分数字和字母大小写）。为了方便大家练习 Python 网络爬虫，我专门搭建了一个博客网站用于 Python 网络爬虫的教学，本书教学部分的爬虫全部基于爬取我的个人博客网站（www.santostang.com）。一方面，由于这个网站不会更改设计和框架，因此本书的网络爬虫代码可以一直使用；另一方面，由于这是我自己的博客网站，因此可以避免一些法律上的风险。

本书主要分为三部分：基础部分（第 1~6 章）、进阶部分（第 7~12 章）和项目实践部分（第 13~16 章），以此来针对不同类型的读者。如果你是 Python 爬虫的初学者，那么可以先学习基础部分，这部分每一章的最后都有自我实践题，读者可以通过实践题熟悉编写 Python 爬虫代码。如果你已经对 Python 爬

虫有所了解，但是在实践中遇到了各种问题，那么可以直接学习进阶部分，这部分为你在爬虫实践中遇到的问题提供了解决方案。本书最后的项目实践部分是让你在学习 Python 爬虫后，可以通过在真实网站中练习来消化和吸收 Python 爬虫的知识。

最后，感谢卞诚君老师在我写书过程中给予的指导！感谢我的父母在撰写此书的过程中给予的支持和鼓励！还要感谢李宜威、周启航、吴嘉杰等各位朋友以及刘建南教授等各位前辈在我的数据科学之路上一直给予的支持和无私帮助！

唐　松

2017 年 6 月

目 录

推荐序一
推荐序二
前言一
前言二

第 1 章 网络爬虫入门 1

1.1 为什么要学网络爬虫 2
1.1.1 网络爬虫能带来什么好处 2
1.1.2 能从网络上爬取什么数据 3
1.1.3 应不应该学爬虫 3
1.2 网络爬虫是否合法 3
1.2.1 Robots 协议 4
1.2.2 网络爬虫的约束 5
1.3 网络爬虫的基本议题 6
1.3.1 Python 爬虫的流程 7
1.3.2 三个流程的技术实现 7

第 2 章 编写第一个网络爬虫 8

2.1 搭建 Python 平台 9
2.1.1 Python 的安装 9
2.1.2 使用 pip 安装第三方库 10
2.1.3 使用编译器 Jupyter 编程 11
2.2 Python 使用入门 13
2.2.1 基本命令 13
2.2.2 数据类型 14
2.2.3 条件语句和循环语句 15
2.2.4 函数 16
2.2.5 面向对象编程 17
2.3 编写第一个简单的爬虫 21

2.3.1 第一步：获取页面 22
2.3.2 第二步：提取需要的数据 23
2.3.3 第三步：存储数据 24
2.4 Python 实践：基础巩固 25
2.4.1 Python 基础试题 26
2.4.2 参考答案 27
2.4.3 自我实践题 30

第 3 章 静态网页抓取 31

3.1 安装 Requests 32
3.2 获取响应内容 32
3.3 定制 Requests 33
3.3.1 传递 URL 参数 33
3.3.2 定制请求头 34
3.3.3 发送 POST 请求 35
3.3.4 超时 36
3.4 Requests 爬虫实践：TOP250 电影数据 36
3.4.1 网站分析 37
3.4.2 项目实践 37
3.4.3 自我实践题 39

第 4 章 动态网页抓取 40

4.1 动态抓取的实例 41
4.2 解析真实地址抓取 42
4.3 通过 Selenium 模拟浏览器抓取 47
4.3.1 Selenium 的安装与基本介绍 47
4.3.2 Selenium 的实践案例 48
4.3.3 Selenium 获取文章的所有评论 49
4.3.4 Selenium 的高级操作 52
4.4 Selenium 爬虫实践：深圳短租数据 55
4.4.1 网站分析 55
4.4.2 项目实践 57
4.4.3 自我实践题 60

第 5 章 解析网页 61

5.1 使用正则表达式解析网页 62
5.1.1 re.match 方法 62

5.1.2 re.search 方法 64
5.1.3 re.findall 方法 64
5.2 使用 BeautifulSoup 解析网页 66
5.2.1 BeautifulSoup 的安装 66
5.2.2 使用 BeautifulSoup 获取博客标题 67
5.2.3 BeautifulSoup 的其他功能 68
5.3 使用 lxml 解析网页 72
5.3.1 lxml 的安装 72
5.3.2 使用 lxml 获取博客标题 72
5.3.3 XPath 的选取方法 74
5.4 总结 75
5.5 BeautifulSoup 爬虫实践：房屋价格数据 76
5.5.1 网站分析 76
5.5.2 项目实践 77
5.5.3 自我实践题 79

第 6 章 数据存储 80

6.1 基本存储：存储至 TXT 或 CSV 81
6.1.1 把数据存储至 TXT 81
6.1.2 把数据存储至 CSV 82
6.2 存储至 MySQL 数据库 84
6.2.1 下载安装 MySQL 85
6.2.2 MySQL 的基本操作 88
6.2.3 Python 操作 MySQL 数据库 92
6.3 存储至 MongoDB 数据库 94
6.3.1 下载安装 MongoDB 95
6.3.2 MongoDB 的基本概念 98
6.3.3 Python 操作 MongoDB 数据库 99
6.3.4 RoboMongo 的安装与使用 101
6.4 总结 102
6.5 MongoDB 爬虫实践：虎扑论坛 103
6.5.1 网站分析 103
6.5.2 项目实践 104
6.5.3 自我实践题 110

第 7 章 提升爬虫的速度 111

7.1 并发和并行，同步和异步 112

7.1.1 并发和并行 112
7.1.2 同步和异步 112
7.2 多线程爬虫 113
7.2.1 简单单线程爬虫 114
7.2.2 学习 Python 多线程 114
7.2.3 简单的多线程爬虫 117
7.2.4 使用 Queue 的多线程爬虫 120
7.3 多进程爬虫 122
7.3.1 使用 multiprocessing 的多进程爬虫 122
7.3.2 使用 Pool + Queue 的多进程爬虫 124
7.4 多协程爬虫 127
7.5 总 结 129
7.5.1 回顾多线程、多进程、多协程 129
7.5.2 性能对比 130

第 8 章 反爬虫问题 132

8.1 为什么会被反爬虫 133
8.2 反爬虫的方式有哪些 133
8.2.1 不返回网页 134
8.2.2 返回非目标网页 134
8.2.3 获取数据变难 135
8.3 如何“反反爬虫” 135
8.3.1 修改请求头 135
8.3.2 修改爬虫的间隔时间 136
8.3.3 使用代理 139
8.4 总结 140

第 9 章 解决中文乱码 141

9.1 什么是字符编码 142
9.2 Python 的字符编码 encode 和 decode 144
9.3 解决中文编码问题 146
9.3.1 问题 1：获取网站的中文显示乱码 147
9.3.2 问题 2：非法字符抛出异常 148
9.3.3 问题 3：网页使用 gzip 压缩 149
9.3.4 问题 4：读写文件的中文乱码 150
9.4 总结 152

第 10 章 登录与验证码处理 153
10.1 处理登录表单 154
10.1.1 处理登录表单 154
10.1.2 处理 cookies，让网页记住你的登录 158
10.1.3 完整的登录代码 160
10.2 验证码的处理 162
10.2.1 如何使用验证码验证 163
10.2.2 人工方法处理验证码 164
10.2.3 OCR 处理验证码 167
10.3 总结 169
第 11 章 服务器采集 170
11.1 为什么使用服务器采集 171
11.1.1 大规模爬虫的需要 171
11.1.2 防止 IP 地址被封杀 171
11.2 使用动态 IP 拨号服务器 172
11.2.1 购买拨号服务器 172
11.2.2 登录服务器 172
11.2.3 使用 Python 更换 IP 174
11.2.4 结合爬虫和更换 IP 功能 175
11.3 使用 Tor 代理服务器 176
11.3.1 Tor 的安装 177
11.3.2 Tor 的使用 180
第 12 章 分布式爬虫 184
12.1 安装 Redis 185
12.2 修改 Redis 配置 188
12.2.1 修改 Redis 密码 188
12.2.2 让 Redis 服务器被远程访问 188
12.2.3 使用 Redis Desktop Manager 管理 189
12.3 Redis 分布式爬虫实践 189
12.3.1 安装 Redis 库 190
12.3.2 加入任务队列 190
12.3.3 读取任务队列并下载图片 191
12.3.4 分布式爬虫代码 192
12.4 总结 194

第 13 章　爬虫实践一：维基百科 195

13.1　项目描述 196
13.1.1　项目目标 196
13.1.2　项目描述 196
13.1.3　深度优先和广度优先 198
13.2　网站分析 199
13.3　项目实施：深度优先的递归爬虫 201
13.4　项目进阶：广度优先的多线程爬虫 203
13.5　总结 207

第 14 章　爬虫实践二：知乎 Live 208

14.1　项目描述 209
14.2　网站分析 209
14.3　项目实施 212
14.3.1　获取所有 Live 212
14.3.2　获取 Live 的听众 215
14.4　总结 218

第 15 章　爬虫实践三：百度地图 API 219

15.1　项目描述 220
15.2　获取 API 秘钥 221
15.3　项目实施 222
15.3.1　获取所有拥有公园的城市 224
15.3.2　获取所有城市的公园数据 225
15.3.3　获取所有公园的详细信息 229
15.4　总结 233

第 16 章　爬虫实践四：餐厅点评 234

16.1　项目描述 235
16.2　网站分析 235
16.3　项目实施 237
16.3.1　获取深圳的餐厅列表 237
16.3.2　获取餐厅的详细信息 242
16.4　总结 244

第 1 章

◂ 网络爬虫入门 ▸

网络爬虫就是自动地从互联网上获取程序。想必你听说过这个词汇，但是又不太了解，大家会觉得掌握网络爬虫还是要花一些工夫的，因此这个门槛让你有点望而却步。

我常常觉得计算机和互联网的发明给人类带来了如此大的方便，让人们不用阅读说明书就能知道如何上手，但是偏偏编程的道路又是如此艰辛。因此，本书尽可能地做到浅显易懂，希望能够将网络爬虫学习的门槛降低，大家都能享受到使用网络爬虫编程的快乐。

本书的第 1 章将介绍网络爬虫的基础部分，包括学习网络爬虫的原因、网络爬虫带来的价值、网络爬虫是否合法以及网络爬虫的基本议题和框架。让读者在开始学习爬虫之前理解为什么学习、要学什么内容。