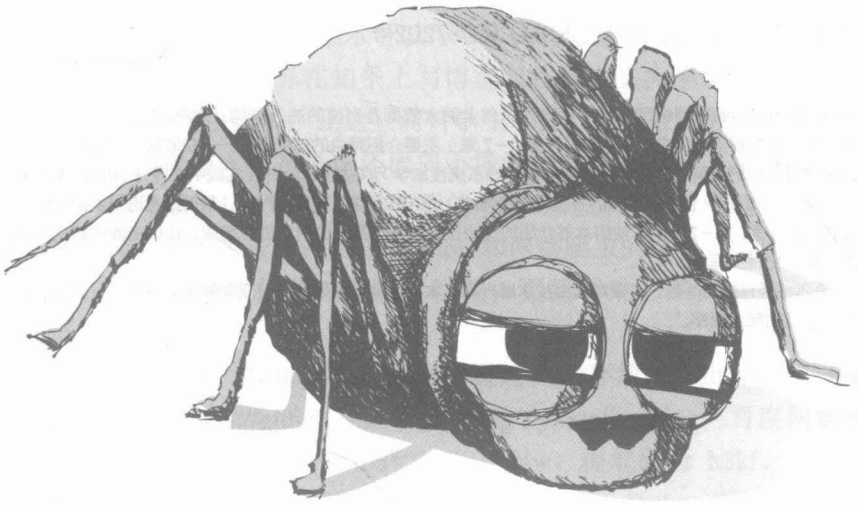
QQ技术交流群（798652826）

Python

网络爬虫从入门到实践 第2版

唐 松 编著

机械工业出版社
China Machine Press



Python 网络爬虫从入门到实践

第2版

唐 松 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 / 唐松编著.—2版.—北京：机械工业出版社，2019.6

ISBN 978-7-111-62687-9

Ⅰ. ①P… Ⅱ. ①唐… 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（2019）第087202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 1~7 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，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 8~13 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 14~17 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。

无论你是否有编程基础，只要对爬虫技术感兴趣，本书就能带领你从入门到实战再到进阶，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　第 2 版

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：周晓娟

印　　刷：中国电影出版社印刷厂　　版　　次：2019 年 6 月第 2 版第 1 次印刷

开　　本：170mm×242mm　1/16　　印　　张：18.25

书　　号：ISBN 978-7-111-62687-9　　定　　价：69.00 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　　读者信箱：hzit@hzbook.com

前　言

近年来，大数据成为业界与学术界的热门话题之一，数据已经成为每个公司极为重要的资产。互联网上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。因此，本书除了面向技术人员外，还面向不懂编程的“小白”，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。过去的一年中，本书第 1 版帮助很多读者开启了 Python 和网络爬虫的世界，因此有幸获得出版社的邀请，在之前版本的基础上进行修改，更新书中的案例以及添加新的内容，形成第 2 版。

本书所有代码均在 Python 3.6 中测试通过，并存放在 Github 和百度网盘上：Github 链接为 https://github.com/Santostang/PythonScraping；百度网盘链接为 https://pan.baidu.com/s/14RA8Srew8tbqVT977JDvNw，提取码为 h2kf。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，本书的教学部分全部基于爬取我的个人博客网（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。

读者对象

（1）对 Python 编程和网络爬虫感兴趣的大专院校师生，需要获取数据进行分析；

（2）打算转行或入行爬虫工程师、数据分析师、数据科学家的人士；

（3）需要使用网络爬虫技术自动获取数据分析的各行业人士。

勘误和支持

由于作者水平和能力有限，编写时间仓促，不妥之处在所难免，希望读者批评指正。本书的读者 QQ 群为 798652826，欢迎读者加群交流。另外，也可以到我的博客 www.santostang.com 反馈意见，欢迎读者和网络爬虫爱好者不吝赐教。

如何阅读本书

本书分为 17 章。

第 1~7 章为基础部分，主要介绍 Python 入门，Python 网络爬虫的获取网页、解析网页和存储数据三个流程，以及 Scrapy 爬虫框架。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 Python 爬虫代码的编写。

第 8~13 章为进阶部分，主要介绍多线程和多进程爬虫、反爬虫、服务器爬虫和分布式爬虫等进阶爬虫技术，这部分为你在爬虫实践中遇到的问题提供了解决方案。

第 14~17 章为项目实践部分，每一章包含一个详细的爬虫案例，每个案例都覆盖之前章节的知识，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。

本书几乎每章都使用案例来学习 Python 网络爬虫，希望告诉读者“通过实战解决实际问题，才能高效地学习新知识”。手输代码，练习案例，才是学习 Python 和网络爬虫的有效方法。

致谢

首先感谢卞诚君老师在我写书过程中给予的指导和帮助。没有他的提议，我不会想到将自己的网络爬虫博客整理成一本书出版，更不会有本书的第 2 版。

从转行数据分析，到申请去康奈尔大学读书，再到回国做数据分析师，我在计算机技术和数据科学的道路上，得到了无数贵人的帮助和提携。首先感谢刘建南教授带我进入了数据挖掘的大门，无私地将数据挖掘、营销知识和经验倾囊相授，您是我的启蒙老师，也是我一生的恩师。

感谢腾讯公司商业分析组和数据服务中心的各位同事，特别感谢我的组长张殿鹏和导师王欢，他们耐心地培养和教导我如何成为一名优秀的数据分析师，让我放手去挑战和尝试不同项目，坚持将数据分析的成果落地。

感谢一路走来，支持我、帮助我的前辈和朋友，包括香港中文大学的教授和朋友——马旭飞教授、李宜威博士、数据科学家周启航、数据分析师赵作栋、数据分析师王礼斌以及好友孙成帅、张蓓等，康奈尔大学的同学——数据科学家汤心韵等、思路富邦有限公司总裁陈智铨、数据科学家吴嘉杰。尤其感谢 IBM 香港 CTO 戴剑寒博士、香港中文大学（深圳）校长讲席教授贾建民博士、TalkingData 腾云大学执行校长杨慧博士和 DaoCloud 首席架构师王天青在百忙中热情地为本书写推荐语。

感谢我的父母、妹妹和女朋友给我一贯的支持和帮助！

唐松

中国深圳

目 录

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，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。

1.1 为什么要学网络爬虫

在数据量爆发式增长的互联网时代，网站与用户的沟通本质上是数据的交换：搜索引擎从数据库中提取搜索结果，将其展现在用户面前；电商将产品的描述、价格展现在网站上，以供买家选择心仪的产品；社交媒体在用户生态圈的自我交互下产生大量文本、图片和视频数据等。这些数据如果得以分析利用，不仅能够帮助第一方企业（拥有这些数据的企业）做出更好的决策，对于第三方企业也是有益的。而网络爬虫技术，则是大数据分析领域的第一个环节。

1.1.1 网络爬虫能带来什么好处

大量企业和个人开始使用网络爬虫采集互联网的公开数据。那么对于企业而言，互联网上的公开数据能够带来什么好处呢？这里将用国内某家知名家电品牌举例说明。

作为一个家电品牌，电商市场的重要性日益凸显。该品牌需要及时了解对手的产品特点、价格以及销量情况，才能及时跟进产品开发进度和营销策略，从而知己知彼，赢得竞争。过去，为了获取对手产品的特点，产品研发部门会手动访问一个个电商产品页面，人工复制并粘贴到 Excel 表格中，制作竞品分析报告。但是这种重复性的手动工作不仅浪费宝贵的时间，一不留神复制少了一个数字还会导致数据错误；此外，竞争对手的销量则是由某一家咨询公司提供报告，每周一次，但是报告缺乏实时性，难以针对快速多变的市场及时调整价格和营销策略。针对上述两个痛点——无法自动化和无法实时获取，本书介绍的网络爬虫技术都能够很好地解决，实现实时自动化获取数据。

上面的例子仅为数据应用的冰山一角。近几年来，随着大数据分析的火热，毕竟有数据才能进行分析，网络爬虫技术已经成为大数据分析领域的第一个环节。

对于这些公开数据的应用价值，我们可以使用 KYC 框架来理解，也就是 Know Your Company（了解你的公司）、Know Your Competitor（了解你的竞争对手）、Know Your Customer（了解你的客户）。通过简单描述性分析，这些公开数据就可以带来很大的商业价值。进一步讲，通过深入的机器学习和数据挖掘，在营销领域可以帮助企业做好 4P（Product：产品创新，Place：智能选址，Price：动态价格，Promotion：个性化营销活动）；在金融领域，大数据征信、智能选股

等应用会让公开数据带来越来越大的价值。

1.1.2　能从网络上爬取什么数据

简单来说，平时在浏览网站时，所有能见到的数据都可以通过爬虫程序保存下来。从社交媒体的每一条发帖到团购网站的价格及点评，再到招聘网站的招聘信息，这些数据都可以存储下来。

1.1.3　应不应该学爬虫

正在准备继续阅读本书的读者可能会问自己：我应不应该学爬虫？

这也是我之前问自己的一个问题，作为一个本科是商学院的学生，面对着技术创新驱动变革的潮流，我还是自学了 Python 的网络爬虫技术，从此踏入了编程的世界。对于编程小白而言，入门网络爬虫并没有想象中那么困难，困难的是你有没有踏出第一步。

我认为，对于任何一个与互联网有关的从业人员，无论是非技术的产品、运营或营销人员，还是前端、后端的程序员，都应该学习网络爬虫技术。

一方面，网络爬虫简单易学、门槛很低。没有任何编程基础的人在认真看完本书的爬虫基础内容后，都能够自己完成简单的网络爬虫任务，从网站上自动获取需要的数据。

另一方面，网络爬虫不仅能使你学会一项新的技术，还能让你在工作的时候节省大量的时间。如果你对网络爬虫的世界有兴趣，就算你不懂编程也不要担心，本书将会深入浅出地为你讲解网络爬虫。

1.2　网络爬虫是否合法

网络爬虫合法吗？

网络爬虫领域目前还属于早期的拓荒阶段，虽然互联网世界已经通过自身的协议建立起一定的道德规范（Robots 协议），但法律部分还在建立和完善中。从目前的情况来看，如果抓取的数据属于个人使用或科研范畴，基本不存在问题；而如果数据属于商业盈利范畴，就要就事而论，有可能属于违法行为，也有可能不违法。

1.2.1 Robots 协议

Robots 协议（爬虫协议）的全称是“网络爬虫排除标准”（Robots Exclusion Protocol），网站通过 Robots 协议告诉搜索引擎哪些页面可以抓取，哪些页面不能抓取。该协议是国际互联网界通行的道德规范，虽然没有写入法律，但是每一个爬虫都应该遵守这项协议。

下面以淘宝网的 robots.txt 为例进行介绍。

这里仅截取部分代码，查看完整代码可以访问 https://www.taobao.com/robots.txt。

```
User-agent:  Baiduspider    #百度爬虫引擎
Allow:  /article     #允许访问/article.htm、/article/12345.com
Allow:  /oshtml
Allow:  /ershou
Disallow:  /product/ #禁止访问/product/12345.com
Disallow:  /          #禁止访问除 Allow 规定页面外的其他所有页面

User-Agent:  Googlebot   #谷歌爬虫引擎
Allow:  /article
Allow:  /oshtml
Allow:  /product     #允许访问/product.htm、/product/12345.com
Allow:  /spu
Allow:  /dianpu
Allow:  /wenzhang
Allow:  /oversea
Disallow:  /
```

在上面的 robots 文件中，淘宝网对用户代理为百度爬虫引擎进行了规定。

以 Allow 项的值开头的 URL 是允许 robot 访问的。例如，Allow：/article 允许百度爬虫引擎访问/article.htm、/article/12345.com 等。

以 Disallow 项为开头的链接是不允许百度爬虫引擎访问的。例如，Disallow：/product/不允许百度爬虫引擎访问/product/12345.com 等。

最后一行，Disallow：/禁止百度爬虫访问除了 Allow 规定页面外的其他所有页面。

因此，当你在百度搜索“淘宝”的时候，搜索结果下方的小字会出现：“由于该网站的 robots.txt 文件存在限制指令（限制搜索引擎抓取），系统无法提供该页面的内容描述”，如图 1-1 所示。百度作为一个搜索引擎，良好地遵守了淘宝网的 robot.txt 协议，所以你是不能从百度上搜索到淘宝内部的产品信息的。

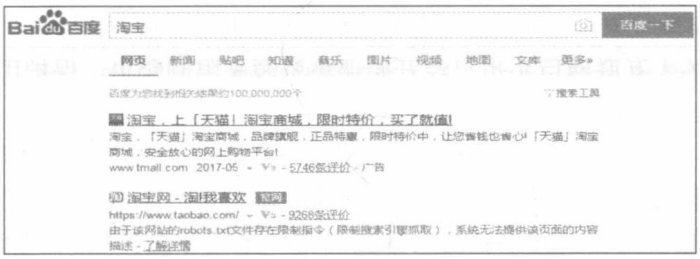


图 1-1　百度搜索提示

淘宝的 Robots 协议对谷歌爬虫的待遇则不一样，和百度爬虫不同的是，它允许谷歌爬虫爬取产品的页面 Allow：/product。因此，当你在谷歌搜索“淘宝 iphone7”的时候，可以搜索到淘宝中的产品，如图 1-2 所示。

图 1-2　谷歌搜索的信息

当你爬取网站数据时，无论是否仅供个人使用，都应该遵守 Robots 协议。

1.2.2　网络爬虫的约束

除了上述 Robots 协议之外，我们使用网络爬虫的时候还要对自己进行约束：过于快速或者频密的网络爬虫都会对服务器产生巨大的压力，网站可能封锁你的 IP，甚至采取进一步的法律行动。因此，你需要约束自己的网络爬虫行为，将请求的速度限定在一个合理的范围之内。

本书中的爬虫仅用于学习、研究用途，请不要用于非法用途。任何由此引发的法律纠纷，请自行负责。

实际上，由于网络爬虫获取的数据带来了巨大的价值，网络爬虫逐渐演变成一场网站方与爬虫方的战争，你的矛长一寸，我的盾便厚一寸。在携程技术微分享上，携程酒店研发部研发经理崔广宇分享过一个“三月爬虫”的故事，也就是每年的三月份会迎来一个爬虫高峰期。因为有大量的大学生五月份交论文，在写论文的