Python网络数据采集 (图灵程序设计丛书)

作者：【美】米切尔（Ryan Mitchell）

版权信息

书名：Python网络数据采集

作者：[美] Ryan Mitchell

译者：陶俊杰　陈小莉

ISBN：978-7-115-41629-2

版权声明

O'Reilly Media, Inc. 介绍

O'Reilly Media 通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自 1978 年开始，O'Reilly 一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly 的发展充满了对创新的倡导、创造和发扬光大。

O'Reilly 为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了 Make 杂志，从而成为 DIY 革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly 的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly 现在还将先锋专家的知识传递给普通的计算机用户。无论是通过书籍出版、在线服务或者面授课程，每一项 O'Reilly 的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O'Reilly Radar 博客有口皆碑。”

——*Wired*

“O'Reilly 凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——*Business 2.0*

“O'Reilly Conference 是聚集关键思想领袖的绝对典范。”

——*CRN*

“一本 O'Reilly 的书就代表一个有用、有前途、需要学习的主题。”

——*Irish Times*

“Tim 是位特立独行的商人，他不光放眼于最长远、最广阔的视野，并且切实地按照 Yogi Berra 的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去，Tim 似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——*Linux Journal*

译者序

每时每刻，搜索引擎和网站都在采集大量信息，非原创即采集。采集信息用的程序一般被称为网络爬虫（Web crawler）、网络铲（Web scraper，可类比考古用的洛阳铲）、网络蜘蛛（Web spider），其行为一般是先“爬”到对应的网页上，再把需要的信息“铲”下来。O'Reilly 这本书的封面图案是一只穿山甲，图灵公司把这本书的中文版定名为“Python 网络数据采集”。当我们看完这本书的时候，觉得网络数据采集程序也像是一只辛勤采蜜的小蜜蜂，它飞到花（目标网页）上，采集花粉（需要的信息），经过处理（数据清洗、存储）变成蜂蜜（可用的数据）。网络数据采集可以为生活加点儿蜜，亦如本书作者所说，“网络数据采集是为普通大众所喜闻乐见的计算机巫术”。

网络数据采集大有所为。在大数据深入人心的时代，网络数据采集作为网络、数据库与机器学习等领域的交汇点，已经成为满足个性化网络数据需求的最佳实践。搜索引擎可以满足人们对数据的共性需求，即“我来了，我看见”，而网络数据采集技术可以进一步精炼数据，把网络中杂乱无章的数据聚合成合理规范的形式，方便分析与挖掘，真正实现“我征服”。工作中，你可能经常为找数据而烦恼，或者眼睁睁看着眼前的几百页数据却只能长恨咫尺天涯，又或者数据杂乱无章的网站中满是带有陷阱的表单和坑爹的验证码，甚至需要的数据都在网页版的 PDF 和网络图片中。而作为一名网站管理员，你也需要了解常用的网络数据采集手段，以及常用的网络表单安全措施，以提高网站访问的安全性，所谓道高一尺，魔高一丈……一念清净，烈焰成池，一念觉醒，方登彼岸，本书试图成为解决这些问题的一念，让你茅塞顿开，船登彼岸。

网络数据采集并不是一门语言的独门秘籍，Python、Java、PHP、C#、Go 等语言都可以讲出精彩的故事。有人说编程语言就是宗教，不同语言的设计哲学不同，行为方式各异，“非我族类，其心必异”，但本着美好生活、快乐修行的初衷，我们对所有语言都时刻保持敬畏之心，尊重信仰自由，努力做好自己的功课。对爱好 Python 的人来说，人生苦短，Python 当歌！简洁轻松的语法，开箱即用的模块，强大快乐的社区，总可以快速构建出简单高效的解决方案。使用 Python 的日子总是充满快乐的，本书关于 Python 网络数据采集的故事也不例外。网络数据采集涉及多个领域，内容包罗万象，因此本书覆盖的主题较多，涉及的知识面相对广阔，书中介绍的 Python 模块有 urllib、BeautifulSoup、lxml、Scrapy、PdfMiner、Requests、Selenium、NLTK、Pillow、unittest、PySocks 等，还有一些知名网站的 API、MySQL 数据库、OpenRefine 数据分析工具、PhanthomJS 无头浏览器以及 Tor 代理服务器等内容。每行到一处，皆是风景独好，而且作者也为每一个主题提供了深入研究的参考资料。不过，本书关于多进程（multiprocessing）、并发（concurrency）、集群（cluster）等高性能采集主题着墨不多，更加关注性能的读者，可以参考其他关于 Python 高性能和多核编程的书籍。总之，本书通俗易懂，简单易行，有编程基础的同学都可以阅读。不会 Python？抽一节课时间学一下吧。

网络数据采集也应该有所不为。国内外关于网络数据保护的法律法规都在不断地制定与完善中，本书作者在书中介绍了美国与网络数据采集相关的法律与典型案例，呼吁网络爬虫严格控制网络数据采集的速度，降低被采集网站服务器的负担。恶意消耗别人网站的服务器资源，甚至拖垮别人网站是一件不道德的事情。众所周知，这已经不仅仅是一句“吸烟有害健康”之类的空洞口号，它可能导致更严重的法律后果，且行且珍惜！

语言是思想的解释器，书籍是语言的载体。本书英文原著是作者用英文解释器为自己思想写的载体，而译本是译者根据英文原著以及与作者的交流，用简体中文解释器为作者思想写的载体。读者拿到的中译本，是作者思想经过两层解释器转换的结果，其目的是希望帮助中文读者消除语言障碍，理解作者的思想，与作者产生共鸣，一起面对作者曾经遇到的问题，共同探索解决问题的方法，从而帮助读者提高解决问题的能力，增强直面 bug 的信心。bug 是产品生命中的挑战，好产品是不断面对 bug 并战胜 bug 的结果。译者水平有限，译文 bug 也在所难免，翻译有不到之处，还请各位读者批评指正！

最后要感谢图灵公司朱巍老师的大力支持，让译作得以顺利出版。也要感谢神烦小宝的温馨陪伴，每天 6 点叫我们起床，让业余时间格外宽裕。

译者联系方式——

邮箱：muxuezi@gmail.com，微信号：muxuezi

邮箱：carrieforchen@gmail.com，微信号：陈小莉

陶俊杰

2015 年 10 月

前言

对那些没有学过编程的人来说，计算机编程看着就像变魔术。如果编程是魔术（magic），那么**网络数据采集**（Web scraping）就是巫术（wizardry）；也就是运用“魔术”来实现精彩实用却又不费吹灰之力的“壮举”。

说句实话，在我的软件工程师职业生涯中，我几乎没有发现像网络数据采集这样的编程实践，可以同时吸引程序员和门外汉的注意。虽然写一个简单的网络爬虫并不难，就是先收集数据，再显示到命令行或者存储到数据库里，但是无论你之前已经做过多少次了，这件事永远会让你感到兴奋，同时又有新的可能。

不过遗憾的是，当和别的程序员提起网络数据采集时，我听到了很多关于这件事的误解与困惑。有些人不确定它是不是合法的（其实合法），有人不明白怎么处理那些到处都是 JavaScript、多媒体和 cookie 的新式网站，还有人对 API 和网络爬虫的区别感到困惑。

这本书的初衷是要解决人们对网络数据采集的诸多问题与误解，并对常见的网络数据采集任务提供全面的指导。

从第 1 章开始，我将不断地提供代码示例来演示书中内容。这些代码示例是开源的，无论注明出处与否都可以免费使用（但若注明会让作者感激不尽）。所有的代码示例都在 GitHub 网站上（https://github.com/REMitchell/python-scraping），可以查看和下载。