

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

本书由北京图灵文化发展有限公司发行数字版。版权所有，侵权必究。

您购买的图灵电子书仅供您个人使用，未经授权，不得以任何方式复制和传播本书内容。

我们愿意相信读者具有这样的良知和觉悟，与我们共同保护知识产权。

如果购买者有侵权行为，我们可能对该用户实施包括但不限于关闭该帐号等维权措施，并可能追究法律责任。

图灵社区会员 人民邮电出版社 (zhanghaichuan@ptpress.com.cn) 专享 尊重版权

[版权声明](#)

[O'Reilly Media, Inc. 介绍](#)

[业界评论](#)

[译者序](#)

[前言](#)

[什么是网络数据采集](#)

[为什么要做网络数据采集](#)

[关于本书](#)

[排版约定](#)

[使用代码示例](#)

[Safari® Books Online](#)

[联系我们](#)

[致谢](#)

[第一部分 创建爬虫](#)

[第 1 章 初见网络爬虫](#)

[1.1 网络连接](#)

[1.2 BeautifulSoup简介](#)

[1.2.1 安装BeautifulSoup](#)

[1.2.2 运行BeautifulSoup](#)

[1.2.3 可靠的网络连接](#)

[第 2 章 复杂 HTML 解析](#)

[2.1 不是一直都要用锤子](#)

[2.2 再端一碗BeautifulSoup](#)

[2.2.1 BeautifulSoup的find\(\)和findAll\(\)](#)

[2.2.2 其他BeautifulSoup对象](#)

[2.2.3 导航树](#)

[2.3 正则表达式](#)

[2.4 正则表达式和BeautifulSoup](#)

[2.5 获取属性](#)

[2.6 Lambda表达式](#)

[2.7 超越BeautifulSoup](#)

[第 3 章 开始采集](#)

[3.1 遍历单个域名](#)

[3.2 采集整个网站](#)

[收集整个网站数据](#)

[3.3 通过互联网采集](#)

[3.4 用Scrapy采集](#)

[第 4 章 使用 API](#)

[4.1 API概述](#)

[4.2 API通用规则](#)

[4.2.1 方法](#)

[4.2.2 验证](#)

[4.3 服务器响应](#)

[API调用](#)

[4.4 Echo Nest](#)

[几个示例](#)

[4.5 Twitter API](#)

[4.5.1 开始](#)

[4.5.2 几个示例](#)

[4.6 Google API](#)

[4.6.1 开始](#)

[4.6.2 几个示例](#)

[4.7 解析JSON数据](#)

[4.8 回到主题](#)

[4.9 再说一点API](#)

[第 5 章 存储数据](#)

[5.1 媒体文件](#)

[5.2 把数据存储到CSV](#)

[5.3 MySQL](#)

[5.3.1 安装MySQL](#)

[5.3.2 基本命令](#)

[5.3.3 与Python整合](#)

[5.3.4 数据库技术与最佳实践](#)

[5.3.5 MySQL里的“六度空间游戏”](#)

[5.4 Email](#)

[第 6 章 读取文档](#)

[6.1 文档编码](#)

[6.2 纯文本](#)

[文本编码和全球互联网](#)

[6.3 CSV](#)

[读取CSV文件](#)

[6.4 PDF](#)

[6.5 微软Word和.docx](#)

[第二部分 高级数据采集](#)

[第 7 章 数据清洗](#)

[7.1 编写代码清洗数据](#)

[数据标准化](#)

[7.2 数据存储后再清洗](#)

[OpenRefine](#)

[第 8 章 自然语言处理](#)

[8.1 概括数据](#)

[8.2 马尔可夫模型](#)

[维基百科六度分割：终结篇](#)

[8.3 自然语言工具包](#)

[8.3.1 安装与设置](#)

[8.3.2 用NLTK做统计分析](#)

[8.3.3 用NLTK做词性分析](#)

[8.4 其他资源](#)

[第 9 章 穿越网页表单与登录窗口进行采集](#)

[9.1 Python Requests库](#)

[9.2 提交一个基本表单](#)

[9.3 单选按钮、复选框和其他输入](#)

[9.4 提交文件和图像](#)

[9.5 处理登录和cookie](#)

[HTTP基本接入认证](#)

[9.6 其他表单问题](#)

[第 10 章 采集 JavaScript](#)

[10.1 JavaScript简介](#)

[常用JavaScript库](#)

[10.2 Ajax和动态HTML](#)

[在Python中用Selenium执行JavaScript](#)

[10.3 处理重定向](#)

[第 11 章 图像识别与文字处理](#)

[11.1 OCR库概述](#)

[11.1.1 Pillow](#)

[11.1.2 Tesseract](#)

[11.1.3 NumPy](#)

[11.2 处理格式规范的文字](#)

[从网站图片中抓取文字](#)

[11.3 读取验证码与训练Tesseract](#)

[训练Tesseract](#)

[11.4 获取验证码提交答案](#)

[第 12 章 避开采集陷阱](#)

[12.1 道德规范](#)

[12.2 让网络机器人看起来像人类用户](#)

[12.2.1 修改请求头](#)

[12.2.2 处理cookie](#)

[12.2.3 时间就是一切](#)

[12.3 常见表单安全措施](#)

[12.3.1 隐含输入字段值](#)

[12.3.2 避免蜜罐](#)

[12.4 问题检查表](#)

[第 13 章 用爬虫测试网站](#)

[13.1 测试简介](#)

[什么是单元测试](#)

[13.2 Python单元测试](#)

[测试维基百科](#)

[13.3 Selenium单元测试](#)

[与网站进行交互](#)

[13.4 Python单元测试与Selenium单元测试的选择](#)

[第 14 章 远程采集](#)

[14.1 为什么要用远程服务器](#)

[14.1.1 避免IP地址被封杀](#)

[14.1.2 移植性与扩展性](#)

[14.2 Tor代理服务器](#)

[PySocks](#)

[14.3 远程主机](#)

[14.3.1 从网站主机运行](#)

[14.3.2 从云主机运行](#)

[14.4 其他资源](#)

[14.5 勇往直前](#)

[附录 A Python 简介](#)

[安装与“Hello,World!”](#)

[附录 B 互联网简介](#)

[附录 C 网络数据采集的法律与道德约束](#)

[C.1 商标、版权、专利](#)

[版权法](#)

[C.2 侵犯动产](#)

[C.3 计算机欺诈与濫用法](#)

[C.4 robots.txt和服务协议](#)

[C.5 三个网络爬虫](#)

[C.5.1 eBay起诉Bidder's Edge与侵犯动产](#)

[C.5.2 美国政府起诉Auerheimer与《计算机欺诈与濫用法》](#)

[C.5.3 Field起诉Google: 版权和robots.txt](#)

[作者简介](#)

[封面介绍](#)

版权声明

© 2015 by Ryan Mitchell.

Simplified Chinese Edition, jointly published by O'Reilly Media, Inc. and Posts & Telecom Press, 2016. Authorized translation of the English edition, 2015 O'Reilly Media, Inc., the owner of all rights to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

英文原版由 O'Reilly Media, Inc. 出版，2015。

简体中文版由人民邮电出版社出版，2016。英文原版的翻译得到 O'Reilly Media, Inc. 的授权。此简体中文版的出版和销售得到出版权和销售权的所有者——O'Reilly Media, Inc. 的许可。

版权所有，未得书面许可，本书的任何部分和全部不得以任何形式复制。

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……一念清静，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、bml、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、PhantomJS 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，用简体中文解释器作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 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。

欢迎访问：电子书学习和下载网站 (<https://www.shgis.cn>)

文档名称：《Python网络数据采集》【美】米切尔 (Ryan Mitchell).epub

请登录 <https://shgis.cn/post/1847.html> 下载完整文档。

手机端请扫码查看：

