O'REILLY

TURING 图灵程序设计丛书

第2版

Python 网络爬虫权威指南

Web Scraping with Python, 2E

全面介绍网页抓取技术，解决Web数据采集、转换和使用中的诸多常见问题和痛点

[美] 瑞安·米切尔 著

神烦小宝 译

中国工信出版集团 人民邮电出版社 POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

Python网络爬虫权威指南（第2版）

Web Scraping with Python, 2E

[美] 瑞安·米切尔 著
神烦小宝 译

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

O'Reilly Media, Inc.授权人民邮电出版社出版

人民邮电出版社
北　京

图书在版编目（CIP）数据

Python网络爬虫权威指南 /（美）瑞安·米切尔（Ryan Mitchell）著；神烦小宝译. -- 2版. -- 北京：人民邮电出版社，2019.4
（图灵程序设计丛书）
ISBN 978-7-115-50926-0

Ⅰ. ①P… Ⅱ. ①瑞… ②神… 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041375号

内 容 提 要

本书采用简洁强大的 Python 语言，介绍了网页抓取相关技术，并为抓取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页抓取的基本原理：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。

本书适合需要抓取 Web 数据的相关软件开发人员和研究人员阅读。

◆ 著 [美] 瑞安·米切尔
译 神烦小宝
责任编辑 岳新欣
责任印制 周昇亮

◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路11号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
北京鑫正大印刷有限公司印刷

◆ 开本：800×1000 1/16
印张：16.25
字数：384千字 2019年4月第2版
印数：34 301－38 300册 2019年4月北京第1次印刷
著作权合同登记号 图字：01-2018-7366号

定价：79.00元

读者服务热线：(010)51095186转600 印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

站在巨人的肩上

Standing on Shoulders of Giants

iTuring.cn

站在巨人的肩上

Standing on Shoulders of Giants

iTuring.cn

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、在线服务还是面授课程，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 凭借一系列非凡想法（真希望当初我也想到了）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*

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**网页抓取**（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”。

在我的软件工程师职业生涯中，我几乎没有发现像网页抓取这样的编程实践，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页抓取时，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），有些人不明白怎么处理包含大量JavaScript的页面以及如何处理登录问题。很多人困惑于如何开始一个大的网页抓取项目，甚至是到哪里寻找他们需要的数据。本书致力于解决人们关于网页抓取的诸多常见问题，廓清一些误解，并对常见的网页抓取任务提供全面的指导。

网页抓取是一个复杂多变的领域，我会通过介绍高级概念以及详细的示例来尽可能地覆盖你可能会在数据抓取项目中遇到的情形。本书提供了代码示例来演示书中的概念，你可以尝试运行它们来实践。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，作者会感激不尽）。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。

什么是网页抓取

在互联网上进行自动数据抓取这件事和互联网存在的时间差不多一样长。虽然**网页抓取**并不是新术语，但是多年以来，这件事更常见的称谓是**网页抓屏**（screen scraping）、**数据挖掘**（data mining）、**网页收割**（Web harvesting）或其他类似的版本。今天大众好像更倾向

于用“网页抓取”，因此我在本书中使用这个术语，不过我倾向于把遍历多个页面的程序称作**网络爬虫**（Web crawler），或者把网页抓取程序称为**网络机器人**（bot）。

理论上，网页抓取是一种通过多种手段收集网络数据的方式，不光是通过与 API 交互（或者直接与浏览器交互）的方式。最常用的方法是写一个自动化程序向网络服务器请求数据（通常是用 HTML 表单或其他网页文件），然后对数据进行解析，提取需要的信息。

实践中，网页抓取涉及非常广泛的编程技术和手段，比如数据分析、自然语言解析和信息安全等。本书将在第一部分介绍关于网页抓取和网页爬取（crawling）的基础知识，一些高级主题放在第二部分介绍。我建议所有读者仔细学习第一部分，并根据自己的实际需求深入探索第二部分。

为什么要做网页抓取

如果你上网的唯一方式就是用浏览器，那么你其实错过了很多种可能。虽然浏览器可以更方便地执行 JavaScript、显示图片，并且可以以更适合人类阅读的形式展示数据，但是网络爬虫收集和处理大量数据的能力更为卓越。不像狭窄的显示器窗口一次只能让你看一个网页，网络爬虫可以让你一次查看几千甚至几百万个网页。

另外，网络爬虫可以完成传统搜索引擎不能做的事情。用 Google 搜索“飞往波士顿最便宜的航班”，看到的是大量的广告和主流的航班搜索网站。Google 只知道这些网站的网页会显示什么内容，并不知道在航班搜索应用中输入的各种查询的准确结果。但是，设计较好的网络爬虫可以通过抓取大量的网站数据，绘制出飞往波士顿的航班价格随时间变化的图表，告诉你买机票的最佳时间。

你可能会问：“数据不是可以通过 API 获取吗？”（如果你不熟悉 API，请阅读第 12 章。）确实，如果你能找到一个可以解决问题的 API，那会非常给力。它可以非常方便地从一个计算机程序向另一个计算机程序提供格式完好的数据。对于很多类型的数据都可以找到一个 API，比如推文或者维基百科页面。通常，如果有 API 可用，用 API 来获取数据确实比写一个网络爬虫程序更加方便。但是，很多时候你需要的 API 并不存在或者不适用于你的需求，这是因为：

- 你要收集的数据来自不同的网站，没有一个综合多个网站数据的 API；
- 你想要的数据非常小众或不常见，网站不会为你单独创建一个 API；
- 网站没有基础设施或技术能力去创建 API；
- 数据很宝贵 / 被保护起来，不希望广泛传播。

即使 API 已经存在，可能还会有请求内容和次数的限制，API 能够提供的数据类型或者数据格式可能也无法满足你的需求。

这时网页抓取就派上用场了。你在浏览器上看到的内容，大部分都可以通过编写 Python 程序来获取。如果你可以通过程序获取数据，那么就可以把数据存储到数据库里。如果你可以把数据存储到数据库里，自然也就可以将这些数据可视化。

显然，大量的应用场景都会需要这种几乎可以毫无阻碍地获取数据的手段：市场预测、机器语言翻译，甚至医疗诊断领域，通过对新闻网站、文章以及健康论坛中的数据进行抓取和分析，也可以获得很多好处。

甚至在艺术领域，网页抓取也为艺术创作开辟了新方向。由 Jonathan Harris 和 Sep Kamvar 在 2006 年发起的“我们感觉挺好”（We Feel Fine）项目，从大量英文博客中抓取以“I feel”和“I am feeling”开头的短句，最终做成了一个很受大众欢迎的数据可视图，描述了这个世界每天、每分钟的感觉。

无论你现在处于哪个领域，网页抓取都可以让你的工作更高效，帮你提升生产力，甚至开创一个全新的领域。

关于本书

本书不仅介绍了网页抓取，也为抓取、转换和使用新式网络中各种类型的数据提供了全面的指导。虽然本书用的是 Python 编程语言，涉及 Python 的许多基础知识，但这并不是一本 Python 入门书。

如果你完全不了解 Python，那么这本书看起来可能有点儿费劲。请不要将本书用作 Python 的入门书。我尽量按照初、中级 Python 编程水平来编写书中的概念和代码示例，以便让更广泛的读者可以轻松地理解本书。但书中偶尔会包含一些更高级的 Python 编程知识以及一些常见的计算机科学话题。如果你是一位编程高手，那么你可以跳过书中相应的内容。

如果你想更全面地学习 Python，Bill Lubanovic 写的《Python 语言及其应用》[1]是本非常好的教材，只是书有点儿厚。如果不想看书，Jessica McKellar 的教学视频 Introduction to Python 也非常不错。我也非常喜欢我的前教授 Allen Downey 写的《像计算机科学家一样思考 Python》，这本书非常适合编程新手，介绍了计算机科学和软件工程的概念，以及 Python 语言。

技术书通常仅仅专注于一种语言或者一种技术，但是网页抓取是一个相当分散的主题，在实践中会涉及数据库、网络服务器、HTTP 协议、HTML 语言、网络安全、图像处理、数据科学等内容。本书试图从“数据收集”的角度涵盖所有这些内容以及其他话题。当然，本书不会对这些主题做完整的介绍，但是我相信对于入门编写网络爬虫来说足够了。

第一部分深入讲解网页抓取和网页爬取相关内容，并重点介绍全书都要用到的几个 Python

注 1：中文版已经由人民邮电出版社出版，详见 www.ituring.com.cn/book/1560。——编者注

库。可以将这部分内容用作这些库和技术的综合参考（对于一些特殊情形，后面会提供其他参考资料）。这部分内容对于所有编写网络爬虫的人来说都是实用的，不管网络爬虫的目标或者应用场景如何。

第二部分介绍读者在动手编写网络爬虫的过程中可能会觉得有用的一些主题。不过，这些主题可能并不总是适合所有的爬虫。这些主题的范围特别广泛，无法在一章中道尽玄机。因此，文中提供了许多参考资料来方便读者获取更多的信息。

本书结构清晰，你可直接跳到感兴趣的章节中阅读所需的网页抓取技术。如果一个概念或一段代码在之前的章节中出现过，那么我会明确标注出具体的位置。

排版约定

本书使用了下列排版约定。

- **黑体字**
  表示新术语或重点强调的内容。
- 等宽字体（`constant width`）
  表示程序片段，以及正文中出现的变量、函数名、数据库、数据类型、环境变量、语句和关键字等。
- 加粗等宽字体（**`constant width bold`**）
  表示应该由用户输入的命令或其他文本。
- 斜体等宽字体（*`constant width italic`*）
  表示应该由用户输入的值或根据上下文确定的值替换的文本。

该图标表示一般性说明。

该图标表示提示或建议。

该图标表示警告或警示。

使用代码示例

补充材料（代码示例、练习等）可以从 https://github.com/REMitchell/python-scraping 下载。

本书是要帮你完成工作的。一般来说，如果书中提供了示例代码，你可以把它用在你的程序或文档中。除非你使用了很大一部分代码，否则无须联系我们获得许可。比如，用书中的几个代码片段写一个程序无须获得许可，销售或分发 O'Reilly 图书的示例光盘则需要获得许可；引用书中的示例代码回答问题无须获得许可，将书中大量的代码放到你的产品文档中则需要获得许可。

我们很希望但并不强制要求你在引用本书内容时加上引用说明。引用说明一般包括书名、作者、出版社和 ISBN。比如："*Web Scraping with Python*, Second Edition by Ryan Mitchell (O'Reilly). Copyright 2018 Ryan Mitchell, 978-1-491-998557-1."

如果你觉得自己对示例代码的用法超出了上述许可的范围，欢迎你通过 permissions@oreilly.com 与我们联系。

遗憾的是，纸质书很难保持更新。对于网页抓取来说这更是一个挑战，由于本书用到的很多库、网站以及代码可能偶尔会被修改，所以我们的代码示例可能会运行失败或产生意想不到的结果。如果你需要运行代码示例，请从 GitHub 仓库获取代码并运行，而不是从书中直接复制。我和为本书做贡献的读者（可能也包括你）将尽量及时更新 GitHub 仓库的内容。

除了代码示例，书中还提供了用于演示如何安装和运行软件的终端命令。一般来说，这些命令是适用于 Linux 操作系统的，但是通常也适用于拥有正确配置的 Python 环境并安装了 pip 的 Windows 用户。如果无法运行这些终端命令，我提供了针对所有主流操作系统的命令运行说明，并为 Windows 用户提供了一些外部的参考资料。

O'Reilly Safari

Safari（之前称作 Safari Books Online）是一个针对企业、政府、教育者和个人的会员制培训和参考平台。

会员可以访问来自 250 多家出版商的上千种图书、培训视频、学习路径、互动式教程和精选播放列表，这些出版商包括 O'Reilly Media、Harvard Business Review、Prentice Hall Professional、Addison-Wesley Professional、Microsoft Press、Sams、Que、Peachpit Press、Adobe、Focal Press、Cisco Press、John Wiley & Sons、Syngress、Morgan Kaufmann、IBM Redbooks、Packt、Adobe Press、FT Press、Apress、Manning、New Riders、McGraw-Hill、Jones & Bartlett、Course Technology 等。

要了解更多信息，可以访问 http://www.oreilly.com/safari。

联系我们

请把对本书的评价和问题发给出版社。

美国：

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472

中国：

北京市西城区西直门南大街 2 号成铭大厦 C 座 807 室（100035）
奥莱利技术咨询（北京）有限公司

O'Reilly 的每一本书都有专属网页，你可以在那儿找到本书的相关信息，包括勘误表、示例代码以及其他信息。本书的网站地址是：http://shop.oreilly.com/product/0636920078067.do。

对于本书的评论和技术性问题，请发送电子邮件到：bookquestions@oreilly.com。

要了解更多 O'Reilly 图书、培训课程、会议和新闻的信息，请访问以下网站：
http://www.oreilly.com。

我们在 Facebook 的地址如下：http://facebook.com/oreilly。

请关注我们的 Twitter 动态：http://twitter.com/oreillymedia。

我们的 YouTube 视频地址如下：http://www.youtube.com/oreillymedia。

致谢

和那些基于海量用户反馈诞生的优秀产品一样，如果没有许多协作者、支持者和编辑的帮助，本书可能永远都不会出版。首先要感谢 O'Reilly 团队对这个小众主题图书的大力支持，感谢我的朋友和家人阅读初稿并提出宝贵的建议，还要感谢和我一起在 HedgeServ 奋战的同事们帮我分担了很多工作。

尤其要感谢 Allyson MacDonald、Brian Anderson、Miguel Grinberg 和 Eric VanWyk 的建议、指导和偶尔的爱之深责之切。有一些章节和代码示例是根据他们的建议写成的。

还要感谢 Yale Specht 过去 4 年中在本书两个版本上的无尽耐心，他在最初便鼓励我从事这个项目，并在我的写作过程中对文体提出了宝贵的建议。没有他，这本书可能只用一半时间就能写完，但是不会像现在这么实用。

最后，要感谢 Jim Waldo，是他许多年前给一个小孩邮寄了一个 Linux 机箱和 *The Art and Science of C* 那本书，帮她开启了计算机世界的大门。

电子书

扫描如下二维码，即可购买本书电子版。

目录