HZ BOOKS
华章IT

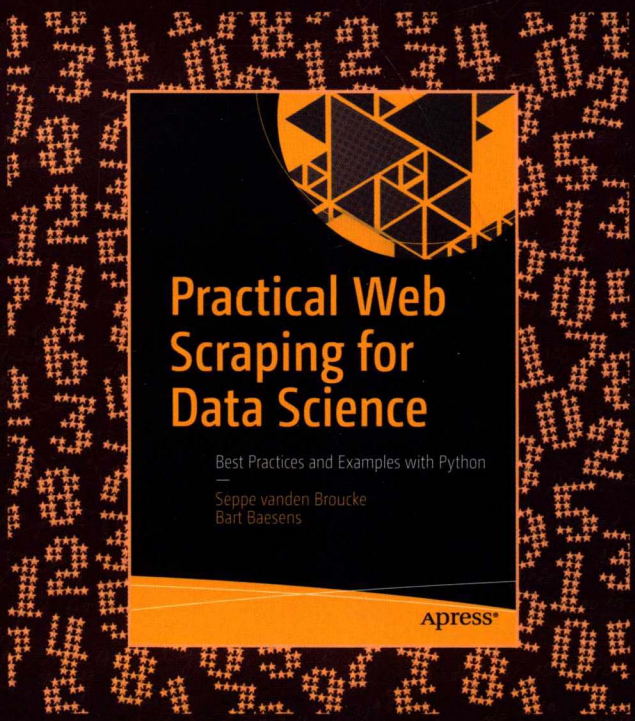
Apress®

数据科学与工程技术丛书

数据科学实战之网络爬取

Python实践和示例

[比] 希普·万登·布鲁克（Seppe vanden Broucke）
巴特·巴森斯（Bart Baesens）
著
罗娜 李福杰 译



PRACTICAL WEB SCRAPING FOR DATA SCIENCE

机械工业出版社
China Machine Press

数据科学与工程技术丛书

PRACTICAL WEB SCRAPING FOR DATA SCIENCE

数据科学实战之网络爬取

Python实践和示例

[比] 希普·万登·布鲁克（Seppe vanden Broucke）
巴特·巴森斯（Bart Baesens）
著
罗娜 李福杰 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

数据科学实战之网络爬取：Python 实践和示例 /（比）希普·万登·布鲁克（Seppe vanden Broucke），（比）巴特·巴森斯（Bart Baesens）著；罗娜，李福杰译．—北京：机械工业出版社，2019.1
（数据科学与工程技术丛书）
书名原文：Practical Web Scraping for Data Science

ISBN 978-7-111-61404-3

I. 数… II. ① 希… ② 巴… ③ 罗… ④ 李… III. 软件工具－程序设计 IV. TP311.561

中国版本图书馆 CIP 数据核字（2018）第 263173 号

本书版权登记号：图字 01-2018-7300

数据科学实战之网络爬取：Python 实践和示例

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：刘　锋　　责任校对：殷　虹
印　　刷：北京市兆成印刷有限责任公司　　版　　次：2019 年 1 月第 1 版第 1 次印刷
开　　本：185mm×260mm　1/16　　印　　张：13.75
书　　号：ISBN 978-7-111-61404-3　　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　88379649　68995259　　读者信箱：hzit@hzbook.com

华章IT
HZBOOKS | Information Technology

译者序

随着大数据时代的到来，互联网上充斥着越来越多的信息。人们期望从互联网上获得有效信息从而为数据分析提供支持，包括分析用户行为、分析产品的不足之处或分析竞争对手的信息等。如何自动获得这些信息或数据成为当务之急。

作为这个时代不可或缺的一部分，网络爬取技术（又称网络爬虫、网页蜘蛛、网络机器人等）通过程序或者脚本自动访问互联网并下载网站内容，再进行过滤、筛选、归纳、整理等，从而实现互联网信息的自动收集整理。在信息收集过程中，网络爬取程序从一个或若干个初始网页的 URL 开始，获得初始网页上的 URL，通过解析器对下载的网页进行解析不断地获取新的 URL 和所需内容，直到满足系统的停止条件，把获取的内容以文件的形式输出，再进一步进行数据分析。而作为开发网络爬取程序的利器，Python 包括 Requests、Beautiful Soup、Selenium 等多个简单易用的库，可以简单、快速、高效地实现大多数场景下的网络数据爬取。本书在介绍 Python 语言的基础上，从零开始实现了基于 Python 的网络爬取，内容由浅及深，同时涵盖了网络协议、超文本标记语言等网络相关基础内容，为理论的落地提供了实际的指导。同时，对于网络爬取技术所带来的问题，包括爬取技术造成的大量 IP 访问网站侵占带宽资源、用户隐私和知识产权等问题，以及企业的“反爬虫”策略，本书亦有涉及。

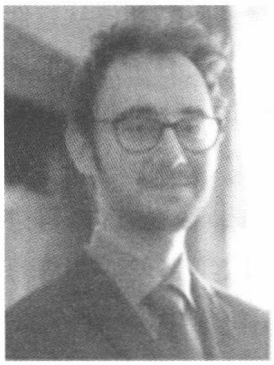
难能可贵的是，作为一本实战类书籍，本书在通俗地介绍大量关于网络爬取技术的同时，还给出了实际网络爬取的若干实例，同时结合机器学习、机器视觉中的若干内容给出了切实可行的程序，以供读者参考使用。

本书内容丰富、案例详实，不仅适合网络爬取初学者阅读，同时对网络爬取的高级内容也有详细介绍。但限于译者水平，对本书中部分内容的理解或中文语言表达难免存在不当之处，敬请读者批评指正，以便能够不断改进。

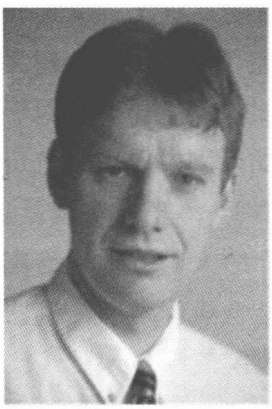
罗娜　李福杰

2018 年 8 月 15 日于上海

作者简介



Seppe vanden Broucke 是比利时鲁汶大学经济与商务学院数据科学方面的助理教授。他的研究兴趣包括商务数据挖掘和分析、机器学习、流程管理和流程挖掘，相关论文发表在知名国际期刊和顶级会议上。Seppe 从事包括高级分析、大数据和信息管理课程方面的教学工作，也经常为工业和商业用户进行培训。工作之余，Seppe 喜欢旅行、阅读（从 Murakami 到 Bukowski 到 Asimov）、听音乐（从 Booka Shade 到 Miles Davis 到 Claude Debussy）、看电影和连续剧（由于没时间现在看得少多了）、玩游戏和关注新闻事件。



Bart Baesens 是比利时鲁汶大学大数据和数据分析方面的教授，也是英国南安普顿大学的讲师。他对大数据及分析、信用风险建模、欺诈检测和营销分析进行了广泛的研究。Bart 撰写了 200 多篇学术论文和若干本书。除了与家人共度时光外，他还是一名布鲁日足球俱乐部的铁杆球迷。Bart 是美食家和业余厨师，他喜欢在他的酒窖里或者在花园里俯瞰红色英式电话亭时喝一杯好酒（他最喜欢的是白维欧尼或红赤霞珠）。Bart 热爱旅行，对第一次世界大战着迷，并阅读了很多关于这个主题的书籍。

技术审校者简介

Mark Furman，MBA、系统工程师、作家、教师和企业家。在过去的16年里，他一直在信息技术领域工作，专注于基于Linux的系统和Python编程，为包括Host Gator、Interland、Suntrust Bank、AT&T和Winn-Dixie在内的多家公司工作。目前，他致力于创客运动，并推出了Tech Forge（techforge.org），以帮助人们创建创客空间并维持现有空间。他拥有俄亥俄大学的工商管理硕士学位。你可以在Twitter @ mfurman上关注他。

前言

恭喜！从选择本书起，你就迈出了进入令人兴奋的网络爬取世界的第一步。感谢你选择本书陪伴你踏上这段旅程。

目标

对于那些不熟悉编程或网络工作机制的人来说，网页爬取通常看起来像个魔法：编写独立探索互联网并收集数据的程序被看作一种神奇的、令人兴奋的甚至可怕的强大力量。实际上，没有太多的编程任务能够像网络爬取那样同时吸引有经验的程序员和新手。第一次看着程序工作，在网络上开始不断收集数据，感觉自己已经避免了工作中的“常规方式”，破解了某种谜题。也许是因为这个原因，网络抓取现在制造了很多头条新闻。

本书将使用 Python 作为编程语言，提供简洁而时髦的网络爬取指南。虽然有很多其他的书籍和在线教程，但对于入门者来说，我们觉得还可以进行这种方式的学习，以提供一个“简短而甜蜜”的指南，而不至于陷入典型的“在 *X* 小时内学会”但重要的细节或最佳实践却因为快速学习而被忽略的陷阱。另外，你会注意到我们将本书称为“数据科学实战之网络爬取”，这是因为我们本身就是数据科学家，在收集数据的过程中发现网络爬取是一个很强大的工具。数据科学项目的第一步是从获得合适的数据集开始，在某些情况下（如果你愿意的话，可以是“理想情况”），数据集由业务合作伙伴、公司的数据仓库或你的学术主管提供，或是从外部数据供应商处购买或获取的结构化格式的数据。但许多真实的项目都需要从网络收集大量信息，就像人们手工从网络上收集数据一样。因此，本书提供以下内容：

- 简明扼要，但同时也详尽地叙述网络爬取的内容；
- 面向数据科学家，展示网络爬取如何嵌入数据科学工作流；

- 采用代码优先方法，无需太多样板文字，让你快速掌握网络爬取；
- 通过使用完善的最佳实践和公开可用的开源 Python 库来实现；
- 比简单的基础内容更进一步，展示如何在现在的网络中进行爬取，包括如何处理 JavaScript、Cookie 和常见的网络反爬取技术；
- 包括有关网络爬取的管理和法律问题的讨论；
- 为进一步阅读和学习提供指导；
- 包括若干大型、完整的实例。

我们希望你能享受阅读本书的乐趣，有如我们写作本书的初衷。如果你有任何疑问、发现了书中的错误或只想联系我们，请随时与我们联系！我们喜欢听取读者的意见，并乐于接收任何想法和问题。

Seppe vanden Broucke, seppe.vandenbroucke@kuleuven.be
Bart Baesens, bart.baesens@kuleuven.be

读者须知

我们在撰写本书时考虑了以数据科学为导向的受众。因此，你可能已经熟悉 Python 或其他一些编程语言或分析工具包，无论是 R、SAS、SPSS，还是其他语言。如果你已经使用过 Python，那么你会在阅读本书时感到更舒服。如果没有，我们将在后面包含 Python 的快速入门，以便你了解基础知识，并提供其他阅读指南。即使你还没有将 Python 用于日常的数据科学任务（很多人会认为你应该这样做），我们也想向你展示 Python 作为一种特别强大的语言，适用于从网络上爬取数据。我们还假设你对网络的工作方式有一些基本了解，也就是说你了解 Web 浏览器的工作方式并知道 URL 是什么。随着书中内容的进展，我们将详细解释相关细节。

总而言之，本书适用于以下目标读者：

- 已经在使用 Python 并希望学习如何使用这种语言来爬取网络数据的数据科学从业者；
- 使用另一种编程语言或工具包，但希望采用 Python 来执行网络爬取部分的数据科学从业者；
- 网络爬取课程的讲师和导师；
- 从事网络爬取项目或旨在提高 Python 技能的学生；

- 需要网络数据实现想法的数据分析师；
- 希望了解网络爬取的全部内容以及如何为其团队带来收益，以及需要考虑相关管理和法律问题的数据科学或商业智能经理。

本书结构

本书可分为如下三个部分：

- 第一部分包括第 1～3 章，将介绍网络爬取及它为什么对数据科学家有用，并讨论网络的关键组件 HTTP、HTML 和 CSS。我们将展示如何使用 Python 中的“requests”和“Beautiful Soup”库编写基本的爬取。
- 第二部分包括第 4～6 章，将深入讨论 HTTP，展示如何使用表单、登录界面和 Cookie。解释如何处理 JavaScript 的繁杂网站，并展示如何实现从简单的网络爬取到高级网络爬虫。
- 第三部分包括第 7～9 章，讨论数据科学背景下网络爬取的管理和法律问题，进一步扩展介绍了其他工具和库。同时，这一部分还列出了有关网络爬取最佳实践的总览和窍门。第 9 章列举了若干网络爬取示例，以显示之前的概念如何组合，并用网络爬取的数据突出显示一些有趣的数据科学用例。

本书易于阅读和实现，因此建议新人从头到尾阅读本书。也就是说，本书的结构是后面的部分会参考前面的内容，以便你想要温习知识或查找特定的概念。

目录

01

第一部分

网络爬取基础

P A R T I

CHAPTER 01

第 1 章
简　介

本章将介绍网络爬取的概念，并强调为什么这种做法对数据科学家有用。在说明各个领域和行业中最近使用的一些有趣的网络爬取案例之后，确保你已经搭建好自己的编程环境并为网页爬取做好准备。

1.1　什么是网络爬取

网络“爬取”/“爬虫”（也称为“网络收集”“网络数据提取”或“网络数据挖掘”），可以定义为“构建一个代理，以自动化的方式从网络上下载、解析和组织数据”。换句话说，用户点击网络浏览器，把其中感兴趣的部分复制粘贴到电子表格中的工作可以通过网络爬取程序实现，并且该程序的执行比人类更快、更准确。

从互联网上进行自动收集数据的时间可能和互联网本身一样久远，但“爬取”这个术语存在的时间也许要比网络还要长。在“网络爬取”成为一个术语流行起来之前，被称为“屏幕抓取”的操作已经被作为从视觉表达中提取数据的一种方式，虽然在计算机时代初期（20 世纪 60 年代到 80 年代），这样的视觉表达仅仅是简单的基于文本格式的“终端”。就像现在一样，当时的人们对在这些终端上“抓取”大量的文本数据并存储这些数据供日后使用充满兴趣。

1.1.1　网络爬取为什么用于数据科学

使用普通网络浏览器浏览网页时可能会遇到多个站点，读者可能会考虑从网站上收集、存储和分析网页上显示的数据。特别是对于“原材料”是数据的数据科学家来说，网络提供了许多有趣的机会：

- 在维基百科网页上可能有一张有趣的表格，你可以通过该表格进行一些统计分析；
- 也许你想要从电影网站获得评论列表来进行文本挖掘、创建推荐引擎或构建预测模型以发现虚假评论；
- 你可能希望得到一个房地产网站上的房产清单，对房产地理信息进行可视化，从而使其更具有吸引力；
- 为丰富数据集，你可能希望能够通过网络上查找的信息获取更多的特征，例如可预测的天气信息、饮料销售量；
- 你可能想知道如何使用网络论坛上的个人资料数据进行社交网络分析；
- 通过监视一个新闻站点，实现对特定感兴趣的主题新闻的趋势分析，可能会更有意思。

网络上面包含了许多有趣的数据源，它们为各种有趣的事情提供了财富宝库。遗憾的是，网络目前的非结构化特性使得并不总是能够以简单的方式收集或导出这些数据。网络浏览器非常善于以一种具有吸引力的方式展示图像、显示动画和网站，但并不能使用一种简单的方式来导出数据，至少在大多数情况下如此。与通过网页浏览器的窗口逐页查看网页内容相比，自动收集丰富的数据集不是更好吗？这也正是网络爬取的优势所在。

如果对网络有一些了解，你可能会想："这不就是应用程序编程接口（API）吗？"事实上，现在很多网站都提供这样一个 API，它允许外部以结构化的方式访问他们的数据存储库，这意味着被计算机程序消费和访问，而不是被人类直接访问（当然，程序是由人类编写的）。例如，Twitter、Facebook、LinkedIn 和 Google 都提供这样的 API，可以搜索和发布推文、获取你的朋友和他们喜欢的列表、查看你与谁联系等。那么，为什么我们仍然需要网络爬取？ API 是访问数据源的好方法，但网站必须提供了这样的 API，公开你想要的功能。一般的经验是先去找这样的 API，如果可以找到的话，在开始构建网络爬取收集数据之前使用这样的 API。例如，你可以轻松使用 Twitter 的 API 来获取最近的推文列表，而不用自己去查找相关数据。尽管如此，仍然有很多原因可以解释为什么网络爬取比使用 API 更可取：

- 要从中提取数据的网站不提供 API；
- 网站的访问是免费的，而提供的 API 不是免费的 ；
- 所提供的 API 速率受限制，也就是说每秒、每天只能通过 API 进行一定次数的访问；
- API 没有公开你想要获得的所有数据，而网站上公开了这些数据。

在所有这些情况下，使用网络爬取可能会更有用。事实上，如果可以在网络浏览器中查看一些数据，那你就能通过一个程序访问和检索它。如果可以通过程序访问它，那你就可以以任何方式存储、处理和使用这些数据。

1.1.2　谁在使用网络爬取

访问和收集网络数据有许多实际应用，其中许多属于数据科学领域。以下列出现实生活中一些有趣的例子：

- Google 的许多产品都得益于 Google 核心业务中对数据的获得。例如，Google 翻译利用存储在网络上的文本来训练和改进翻译。
- 在人力资源和员工分析中很多时候应用网络爬取。例如，位于旧金山的 hiQ 公司通过收集和分析 LinkedIn 网站上的公众档案信息，进行员工的分析并销售这些分析报告。尽管 LinkedIn 公司对此十分不满，但依据规定无法阻止这种行为，可参见 https://www.bloomberg.com/news/features/2017-11-15/the-brutal-fight-to-mine-your-data-and-sell-it-to-your-boss。
- 数字营销者和数字艺术家经常使用网络上的数据进行各种有趣和创造性的项目。比如，Jonathan Harris 和 Sep Kamvar 的“We Feel Fine”，通过爬取以“I feel”短语开头的各种博客、网站，其结果可以直观地显示一天中世界上人们的感受。
- 在另一项研究中，从 Twitter、博客和其他社交媒体爬取的信息被用来构建一个数据集，从而建立一个识别抑郁症和自杀念头的预测模型。这对于援助者来说可能是一个非常宝贵的工具，当然它也需要充分考虑与隐私相关的问题（请参阅 https://www.sas.com/en_ca/insights/articles/analytics/using-big-data-to-predict-suicide-risk-canada.html）。
- Emmanuel Sales 也对 Twitter 进行了数据爬取，他的目标是了解自己的社交圈和发帖的时间轴（请参阅 https://emsal.me/blog/4）。其实他首先考虑使用 Twitter 的 API，但发现 Twitter 限制大量访问数据，如果想要获得 Twitter 上的关注列表，那么只能每 15 分钟访问 15 次，而这样的限制造成了很大的不便。
- 在一篇题为“十亿美元价格的项目：使用在线价格进行衡量和研究”的论文中（参见 http://www.nber.org/papers/w22111），使用网络爬取收集的在线价格信息数据集被用于构建多个国家稳健的日均价格指数。

- 银行和其他金融机构使用网络爬取分析竞争对手。例如，银行经常爬取竞争对手的网站，以了解其分支机构开在哪里或关闭的情况，或追踪竞争对手提供的贷款利率。所有这些都是可以纳入银行内部模型和预测的有益信息。投资公司也经常使用网络爬取来跟踪他们投资组合中资产的新闻报道。
- 社会政治科学家通过挖掘社交网站来追踪人们的情绪和政治倾向。一篇名为“Dissecting Trump's Most Rabid Online Following”的著名文章（请参阅 https://fivethirtyeight.com/features/dissecting-trumps-most-rabid-online-following/）分析了用户在 Reddit 上的讨论，使用语义分析来描述了唐纳德·特朗普的在线关注者和粉丝。
- 一位研究人员能够根据从 Tinder 和 Instagram 爬取的图像信息以及他们的“喜好”来训练深度学习模型，以预测图像是否可能会被视为“有吸引力的”（请参阅 http://karpathy.github.io/2015/10/25/selfie/）。智能手机制造商已经将这些模型纳入他们的照片应用程序中，以帮助你刷新照片。
- 在《The Girl with the Brick Earring》一书中，Lucas Woltmann 从 https://www.bricklink.com 网页中爬取乐高积木信息，以确定最佳的乐高作品图像（请参阅 http://lucaswoltmann.de/art'n'images/2017/04/08/the-girl-with-the-brick-earring.html）（本书的合著者之一是一名狂热的乐高粉丝，所以我们必须包括这个例子）。
- 在“Analyzing 1000+ Greek Wines With Python”一文中，Florese Tselai 从希腊葡萄酒商店中提取了一千种葡萄酒品种的信息（请参阅 https://tselai.com/greek-wines-analysis.html）以分析其来源、评级、类型和浓度（本书的合著者之一是狂热的葡萄酒爱好者，所以我们也要包括这个例子）。
- Lyst，一家位于伦敦的在线时尚市场，通过网络获取关于时尚产品的半结构化信息，然后应用机器学习为消费者提供这些信息，并在中心网站上显示。其他数据科学家也做了类似的收集时尚产品的项目（请参阅 http://talks.lystit.com/dsl-scraping- presentation/）。
- 我们指导了一项研究，利用网络爬取从工作网站提取信息，了解工作中不同的数据科学和相关分析工具的普及程度（剧透：Python 和 R 都在稳步上升）。
- 我们研究小组的另一项研究涉及使用网络爬取来监控新闻媒体和网络论坛，以跟踪公众对比特币的看法。