O'REILLY®

TURING 图灵程序设计丛书

命令行中的数据科学

DATA SCIENCE AT THE COMMAND LINE

[荷] Jeroen Janssens 著

王晓伟 刘峰 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

命令行中的数据科学

Data Science at the Command Line
Facing the Future with Time-Tested Tools

[荷] Jeroen Janssens 著
王晓伟 刘峰 译

人民邮电出版社

图书在版编目（CIP）数据

命令行中的数据科学 / （荷）詹森斯（Janssens, J.）著 ; 王晓伟, 刘峰译. -- 北京 : 人民邮电出版社, 2015.6
（图灵程序设计丛书）
ISBN 978-7-115-39168-1

Ⅰ. ①命… Ⅱ. ①詹… ②王… ③刘… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2015)第087967号

内 容 提 要

本书集实用性和先进性于一身，为数据分析人员使用命令行这个灵活的工具提供了重要参考。作者讲解了众多实用的命令行工具，以及如何使用它们高效地获取、清洗、探索和建模数据。无论你使用 Windows、OS X，还是 Linux，都可以安装包含 80 多个命令行工具的“数据科学工具箱”，迅速建立自己的数据分析环境。无论你是否已经习惯于使用 Python 或 R 语言，都能够通过本书体会到使用命令行的快捷、灵活与伸缩自如。

本书适合各层次的软件开发人员，包括专业和非专业的数据分析人员。

◆ 著 [荷] Jeroen Janssens
译 王晓伟 刘 峰
责任编辑 岳新欣
责任印制 杨林杰

◆ 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
三河市海波印务有限公司印刷

◆ 开本：800×1000 1/16
印张：11.75
字数：242千字 2015年6月第1版
印数：1-3 500册 2015年6月河北第1次印刷
著作权合同登记号 图字：01-2015-2578号

定价：49.00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：京崇工商广字第 0021 号

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*

献给我的妻子 Esther。如果没有她的鼓励、支持和耐心，我不可能完成这本书。

前言

数据科学是个激动人心却又非常年轻的领域。不幸的是，许多个人和公司，总是认为需要利用新技术才能解决数据科学提出的问题。实际上，正如本书所揭示的，许多问题使用命令行就能解决，而且有时候效率要高得多。

大约 5 年前，在攻读博士学位期间，我逐步从使用微软 Windows 转为使用 GUN/Linux。刚开始我有点谨小慎微，因此同时安装了这两个操作系统（也就是双系统启动）。后来，在这两个系统之间切换的需求越来越少，有时我甚至对 Arch Linux 修修补补，能从零开始自己定制操作系统。这时能用的只有命令行，而且想做什么完全随心所欲。很快，我就对使用命令行得心应手。最终，由于业余时间越来越宝贵，我决定使用名为 Ubuntu 的 GNU/Linux 发行版，因为它易于使用并且有庞大的社区。尽管如此，命令行仍然是我完成绝大部分工作的不二选择。

实际上，我后来认识到，命令行不单可以用于安装软件、配置系统以及搜索文件。于是我开始学习诸如 `cut`、`sort` 和 `sed` 这些命令行工具。这些工具都是将数据作为输入，对数据进行处理，然后打印结果。Ubuntu 自带了相当多这样的工具。当明白可以将这些小工具结合起来使用时，我就对它入迷了。

当我拿到博士学位，成为一名数据科学家时，我想充分利用这种方法来做数据科学工作。幸亏有几个新的开源命令行工具，包括 `scrape`、`jq` 和 `json2csv`，我甚至能够使用命令行来完成抓取网站以及处理大量 JSON 数据这样的任务。2013 年 9 月，我写了一篇名为“数据科学的 7 个命令行工具”的博客文章（http://jeroenjanssens.com/2013/09/19/seven-command-line-tools-for-data-science.html）。让我吃惊的是，这篇文章获得很大反响。后来许多人向我推荐其他命令行工具，于是我开始考虑是否可以将这篇文章扩充成书。令人高兴的是，10 个月之后，在许多才华横溢的人的帮助下（参见“致谢”），本书得以付梓。

分享这段个人经历不仅是想介绍本书的由来，更是希望你知道我也是需要学习命令行的。使用命令行与使用图形化用户界面迥然不同，刚开始可能是令人生畏的。但是，既然我能

够学会它，你当然也没问题。不管你目前使用的是什么操作系统，也不管你现在是以什么方式做数据科学的工作，读完本书，你也能够利用命令行的强大能力。即使你已经熟悉命令行，或者甚至已经打算学习 shell 脚本，你仍然可能在书中发现一些有趣技巧或命令行工具，能用于未来的数据科学项目。

从本书可以学到的

书中将对大量数据进行获取、清洗、探索以及建模。我们不会过多介绍如何完成这些数据科学任务，因为对于诸如应该何时及用什么进行统计检验，或者怎样才能将数据可视化做到最好，很容易找到大量参考资料。本书致力于实用性，旨在通过教你用命令行执行数据科学任务，使你更加高效和多产。

尽管书中讨论了 80 多个命令行工具，但这些工具本身并不是最重要的。有些命令行工具存在已久，有些则是新近出现，并且可能最终会被更好的工具所取代。甚至在你阅读本书的时候，有的命令行工具正在创建之中。在过去的 10 个月里，我就已经发现了许多奇妙的命令行工具。遗憾的是，有的工具被发现的时间太晚，无法包含在本书中。总之，命令行工具的新陈代谢是常态。

用工具、管道和数据进行工作的思想才是最重要的。多数命令行工具只做一项任务，并且做得很好。这符合 Unix 的理念，这种理念在书中许多地方都有体现。一旦熟悉了命令行，并且学会了如何将命令行工具结合起来，你就学会了一项非常宝贵的技能。如果还能创建新的工具，那你就出类拔萃了。

怎样阅读本书

一般来说，我们建议你按顺序阅读。书中介绍的概念或者命令行工具，很可能会在下一章中采用。例如在第 9 章中使用了 `parallel` 工具，这个工具在第 8 章进行了深入的讨论。

数据科学是一个宽广的领域，与许多领域都有交叉，例如程序设计、数据可视化以及机器学习。因此，本书触及了许多有趣的话题，遗憾的是无法逐一详尽讨论。在书中我随时都会给出建议的参考资料。如果只是想跟着本书的节奏，并不需要学习这些参考资料，但是如果你感兴趣，可以深入研究一下这些推荐的资料。

本书面向的读者

本书面向的读者只有一个前提条件：你与数据打交道。使用哪种编程语言或者统计计算环境无关紧要。本书会在开头解释所有必要的概念。

你的操作系统是微软 Windows、Mac OS X 还是其他形式的 Unix 系统都无关紧要。书中自

带数据科学工具箱，这是一个容易安装的虚拟环境。它使你可以运行命令行工具，并可以在与本书相同的环境下，跟着一起学习这些示例代码。你不需要浪费时间来研究如何安装所有的命令行工具以及它们依赖的环境。

书中包含一些用 Bash、Python 和 R 编写的代码，如果你有一定的编程经验会有帮助，但这绝不是必要的。

排版约定

本书使用了下列排版约定。

- 楷体
  表示新术语。
- 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。

这个图标表示提示或建议。

这个图标表示一般注记。

这个图标表示警告或提醒。

使用代码示例

补充材料（虚拟机、数据、脚本、定制的命令行工具等）可以从 https://github.com/jeroenjanssens/data-science-at-the-command-line 下载。

本书是要帮你完成工作的。一般来说，如果本书提供了示例代码，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，否则无需联系我们获得许可。比如，用本书的几个代码片段写一个程序就无需获得许可，销售或分发 O'Reilly 图书的示例光盘则需要获得许可；引用本书中的示例代码回答问题无需获得许可，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："*Data Science at the Command Line* by Jeroen H.M. Janssens (O'Reilly). Copyright 2015 Jeroen H.M. Janssens, 978-1-491-94785-2."

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。

Safari® Books Online

Safari Books Online（http://www.safaribooksonline.com）是应运而生的数字图书馆。它同时以图书和视频的形式出版世界顶级技术和商务作家的专业作品。技术专家、软件开发人员、Web 设计师、商务人士和创意专家等，在开展调研、解决问题、学习和认证培训时，都将 Safari Books Online 视作获取资料的首选渠道。

对于组织团体、政府机构和个人，Safari Books Online 提供各种产品组合和灵活的定价策略。用户可通过一个功能完备的数据库检索系统访问 O'Reilly Media、Prentice Hall Professional、Addison-Wesley Professional、Microsoft Press、Sams、Que、Peachpit Press、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 以及其他几十家出版社的上千种图书、培训视频和正式出版之前的书稿。要了解 Safari Books Online 的更多信息，我们网上见。

联系我们

本书有一个专属网页，你可以在那儿找到与代码无关的勘误列表以及其他信息。本书的网站地址是：

http://datascienceatthecommandline.com/

与代码、命令行工具和虚拟机相关的勘误，请通过 GitHub 的问题追踪系统提交：

https://github.com/jeroenjanssens/data-science-at-the-command-line/issues

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

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

请关注 Jeroen 的 Twitter 动态：@jeroenhjanssens

http://twitter.com/jeroenhjanssens

致谢

首先我要感谢 Mike Dewar 和 Mike Loukides，他们相信我在 2013 年 9 月撰写的博客文章“数据科学的 7 个命令行工具”可以扩写成书。感谢 Jared Lander 邀请我在纽约开放统计编程聚会上演讲，我就是在准备那次演讲时有了撰写这篇博客文章的灵感。

我要特别感谢我的技术审阅人 Mike Dewar、Brain Eoff 和 Shanes Reustle，他们不厌其烦地阅读初稿，一丝不苟地测试所有命令，并且提出了宝贵的反馈信息。你们的努力使本书得到了极大的改善。书中若还有错误，全都是我自己的责任。

我有幸与四位优秀的编辑一起工作，他们是 Ann Spencer、Julie Steele、Marie Beaugureau 和 Matt Hacker。感谢你们的指导，我通过你们认识了 O'Reilly 很多才华横溢的朋友，他们是 Huguette Barriere、Sophia DeMartini、Dan Fauxsmith、Yasmina Greco、Rachel James、Jasmine Kwityn、Ben Lorica、Mike Loukides、Andrew Odewahn 和 Christopher Pappas，以及

许多未曾谋面的在幕后英雄。正是这些朋友的共同努力使我与 O'Reilly 的合作十分愉快。

本书讨论了 80 多个命令行工具。毋庸置疑，没有这些工具，本书根本就不可能成行。因此，我要对创建这些工具以及对其有过贡献的作者们表达十二分的谢意。遗憾的是，这些作者人数众多，无法在这里一一列举，只能在附录 A 中提及。特别感谢 Araon Crow、Jehiah Czebotar、Christopher Groskopf、Dima Kogan、Sergey Lisitsyn、Francisco J. Martin 和 Ole Tange，帮助我学习他们了不起的命令行工具。

本书大量使用了数据科学工具箱，它是一个包含书中所有命令行工具的虚拟环境。它是建立在许多巨人的肩膀之上的，有鉴于此，我要感谢在 GNU、Linux、Ubuntu、Amazon Web Services、GitHub、Packer、Ansible、Vagrant 和 VirtualBox 背后的人们，没有他们就没有数据科学工具箱。感谢 Matthew Russell 在我起初开发数据科学工具箱时给我的启发和反馈。他的著作 *Mining the Social Web*（O'Reilly，http://shop.oreilly.com/product/0636920030195.do）也提供了一个虚拟机。

特别感谢我读博士期间的导师 Eric Postma 和 Jaap van den Herik。在这 5 年间他们教我学了很多课程。尽管写书与写博士论文有很大差别，但过去的 10 个月证明了许多课程都非常有帮助。

最后，我要感谢 YPlan 的同事们，感谢我的朋友、家人，特别是我的妻子 Esther，感谢她无私的支持，以及总是在恰当的时间将我从命令行旁边拉开。

目录