PACKT PUBLISHING

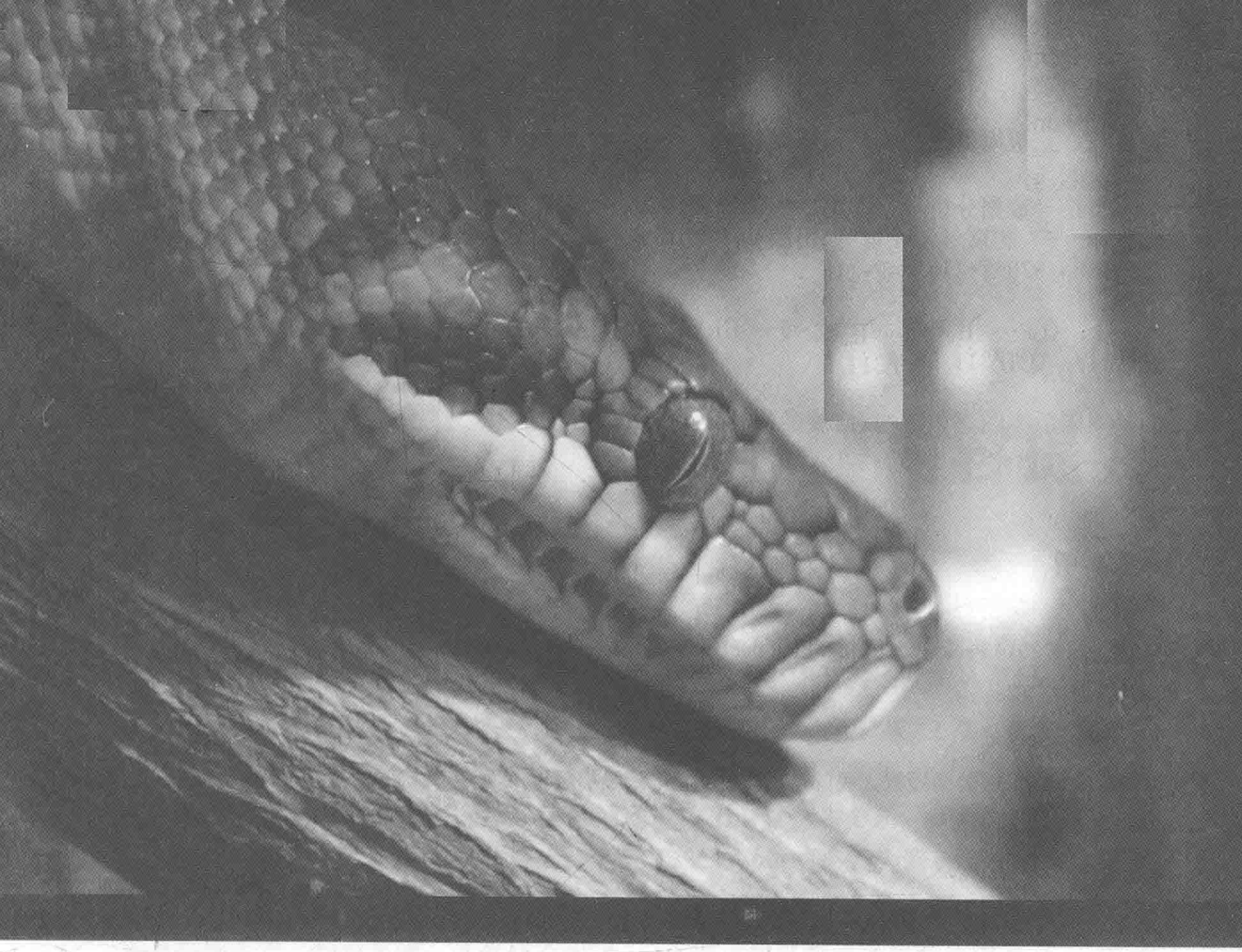
学习流行的开源Python模块，掌握强大的数据分析技术

Python 数据分析

Python Data Analysis

[印尼] Ivan Idris 著
韩波 译

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



Python 数据分析

[印尼] Ivan Idris 著
韩波 译

人民邮电出版社
北京

图书在版编目（CIP）数据

Python数据分析 /（印尼）伊德里斯（Idris, I.）著；
韩波译. -- 北京 : 人民邮电出版社, 2016.2
ISBN 978-7-115-41122-8

Ⅰ. ①P… Ⅱ. ①伊… ②韩… Ⅲ. ①软件工具－程序设计 Ⅳ. ①TP311.56

中国版本图书馆CIP数据核字(2016)第000506号

版权声明

◆ 著 ［印尼］Ivan Idris
译 韩 波
责任编辑 陈冀康
执行编辑 胡俊英
责任印制 张佳莹 焦志炜

◆ 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
北京天宇星印刷厂印刷

◆ 开本：800×1000 1/16
印张：20.75
字数：434 千字 2016 年 2 月第 1 版
印数：1－2 500 册 2016 年 2 月北京第 1 次印刷
著作权合同登记号 图字：01-2015-6184 号

定价：59.00 元

读者服务热线：(010)81055410 印装质量热线：(010)81055316
反盗版热线：(010)81055315

内容提要

作为一种高级程序设计语言，Python 凭借其简洁、易读及可扩展性日渐成为程序设计领域备受推崇的语言。同时，Python 语言的数据分析功能也逐渐为大众所认可。

本书是一本介绍如何用 Python 进行数据分析的学习指南。全书共 12 章，从 Python 程序库入门、NumPy 数组、matplotlib 和 pandas 开始，陆续介绍了数据加工、数据处理和数据可视化等内容。同时，本书还介绍了信号处理、数据库、文本分析、机器学习、互操作性和性能优化等高级主题。在本书的最后，还采用 3 个附录的形式为读者补充了一些重要概念、常用函数以及在线资源等重要内容。

本书示例丰富、简单易懂，非常适合对 Python 语言感兴趣或者想要使用 Python 语言进行数据分析的读者参考阅读。

前言

“数据分析是 Python 的杀手锏。”

——佚名

数据分析在自然科学、生物医学和社会科学领域有着悠久的历史。目前，如雷贯耳的大数据虽然尚没有严格的定义，但是它对数据分析工作的影响是毋庸置疑的。下面列举几个与大数据相关的趋势。

- 世界人口持续增长。
- 越来越多的数据被搜集和存储。
- 电脑芯片集成的晶体管数量不可能无限增长。
- 政府、科学界、工业界和个人对数据洞察力的需求与日俱增。

随着数据科学的炒作，数据分析也呈现流行之势。与数据科学类似，数据分析也致力于从数据中提取有效信息。为此，我们需要用到统计学、机器学习、信号处理、自然语言处理和计算机科学领域中的各种技术。

在 `http://www.xmind.net/m/WvfC/`页面上，可以找到一幅描绘与数据分析相关 Python 软件的脑图。首先要知道的是，Python 生态系统已经非常完备，具有诸如 NumPy、SciPy 和 matplotlib 等著名的程序包。当然，这没有什么好奇怪的，因为 Python 自 1989 年就诞生了。Python 易学、易用，并且与其他程序设计语言相比语法简练，可读性非常强，即使从未接触过 Python 的人，也可以在几天内掌握该语言的基本用法，对熟悉其他编程语言的人来说尤其如此。你无需太多的基础知识，就能顺畅地阅读本书。此外，关于 Python 的书籍、课程和在线教程也非常多。

本书内容

作为学习教程，本书将从 NumPy、SciPy、matplotlib 和 pandas 着手，这些开源程序包对于数据加工、数据处理和可视化而言非常有用。如果能够将这些工具结合起来使用，其功效足以与 MATLAB、Mathematica 和 R 相媲美。

本书还将为读者介绍更高级的主题，包括信号处理、数据库、文本分析、机器学习、互操作性和性能优化。

第 1 章“Python 程序库入门”手把手地指导读者正确安装配置 Python 数值计算软件。同时，本章还会展示如何创建一个小程序。

第 2 章“NumPy 数组”介绍 NumPy 和数组的基础知识。通过阅读本章，读者能够基本掌握 NumPy 数组及其相关函数。

第 3 章“统计学与线性代数”对线性代数和统计函数做了简要回顾。

第 4 章“pandas 入门”阐述 pandas 的基本功能，其中涉及 pandas 的数据结构与相应的操作。

第 5 章“数据的检索、加工与存储”介绍如何获取不同格式的数据，以及原始数据的清洗和存储方法。

第 6 章“数据可视化”介绍如何利用 matplotlib 绘制数据图。

第 7 章“信号处理与时间序列”利用太阳黑子周期数据来实例讲解时间序列和信号处理，同时还会介绍一些相关的统计模型。本章使用的主要工具是 NumPy/SciPy。

第 8 章“应用数据库”介绍各种数据库和有关 API 的知识，其中包括关系数据库和 NoSQL 数据库。

第 9 章“分析文本数据和社交媒体”考察基于文本数据的情感分析和主题抽取。同时，本章还将会为读者展示一个网络分析方面的实例。

第 10 章“预测性分析与机器学习”通过一个例子来说明人工智能在天气预报上面的应用，这主要借助于 scikit-learn。不过，有些机器学习算法在 scikit-learn 中尚未实现，所以有时还要求助其他 API。

第 11 章“Python 生态系统的外部环境和云计算”将提供各种实例，来说明如何集成非 Python 编写的现有代码。此外，本章还将为读者演示如何在云中部署应用。

第 12 章“性能优化、性能分析与并发性”为读者介绍通过性能分析（Profling）和 Cython 等关键技术来改善性能的各种技巧。

此外，我们还将讨论与分布式多核系统有关的一些框架。

附录 A“重要概念”将对本书中涉及的重要概念进行简要介绍。

附录 B“常用函数”概述本书中用到的各种函数。

附录 C“在线资源”给出相关文档、论坛、文章及其他重要信息的网络链接。

本书需要的资源

本书中的示例代码可以在大部分现代操作系统上运行；所有章节中的代码，都需要用到 Python 2 和 pip 软件。为了安装 Python，可以先到 `https://wiki.python.org/moin/BeginnersGuide/Download` 页面下载；对于 pip，可以到 `http://pip.readthedocs.org/en/latest/installing.html` 页面下载。软件的具体安装方法会在相应章节介绍，大部分情况下，我们都需要以管理员权限来执行下列命令：

```
$ pip install <some software>
```

下面是运行本书示例代码所需的软件及其相应的版本号码：

- NumPy 1.8.1
- SciPy 0.14.0
- matplotlib 1.3.1
- IPython 2.0.0
- pandas Version 0.13.1
- tables 3.1.1
- numexpr 2.4
- openpyxl 2.0.3
- XlsxWriter 0.5.5
- xlrd 0.9.3
- feedparser 5.1.3

- Beautiful Soup 4.3.2
- StatsModels 0.6.0
- SQLAlchemy 0.9.6
- Pony 0.5.1
- dataset 0.5.4
- MongoDB 2.6.3
- PyMongo 2.7.1
- Redis server 2.8.12
- Redis 2.10.1
- Cassandra 2.0.9
- Java 7
- NLTK 2.0.4
- scikit-learn 0.15.0
- NetworkX 1.9
- DEAP 1.0.1
- theanets 0.2.0
- Graphviz 2.36.0
- pydot2 1.0.33
- Octave 3.8.0
- R 3.1.1
- rpy2 2.4.2
- JPype 0.5.5.2
- Java 7
- SWIG 3.02
- PCRE 8.35

- Boost 1.56.0
- gfortran 4.9.0
- GAE for Python 2.7
- gprof2dot 2014.08.05
- line_profler beta
- Cython 0.20.0
- cytoolz 0.7.0
- Joblib 0.8.2
- Bottleneck 0.8.0
- Jug 0.9.3
- MPI 1.8.1
- mpi4py 1.3.1

当然，你的软件版本不必与这里的完全相同。通常情况下，应该选用最新版本。

上面列出的某些软件只是用于某个示例，因此安装前，请先检查一下这个软件是否仅限用于某个示例代码。

对于通过 pip 安装的 Python 程序包，卸载方法如下所示：

```
$ pip uninstall <some software>
```

目标读者

本书的目标读者是对 Python 和数学有基本了解，并且想进一步学习如何利用 Python 软件进行数据分析的朋友。我们力争让本书简单易懂，但无法保证所有主题都面面俱到。如果需要，可以经由 Khan Academy、Coursera 或者维基百科来复习自己的数学知识。

下列 Packt 出版社的书籍是推荐给读者的进阶读物：

- *Building Machine Learning Systems with Python*, Willi Richert and Luis Pedro Coelho (2013)
- *Learning Cython Programming*, Philip Herron (2013)
- *Learning NumPy Array*, Ivan Idris (2014)

- *Learning scikit-learn: Machine Learning in Python*, Raúl Garreta and Guillermo Moncecchi (2013)
- *Learning SciPy for Numerical and Scientifc Computing*, Francisco J. Blanco-Silva (2013)
- *Matplotlib for Python Developers*, Sandro Tosi (2009)
- *NumPy Beginner's Guide - Second Edition*, Ivan Idris (2013)
- *NumPy Cookbook*, Ivan Idris (2012)
- *Parallel Programming with Python*, Jan Palach (2014)
- *Python Data Visualization Cookbook*, Igor Milovanović (2013)
- *Python for Finance*, Yuxing Yan (2014)
- *Python Text Processing with NLTK 2.0 Cookbook*, Jacob Perkins (2010)

排版约定

本书中，不同类型的信息会采用不同的排版样式，以示区别。下面针对各种排版样式及其含义进行举例说明。

文本、数据库表名、文件夹名、文件名、文件扩展名和路径名、伪 URL、用户输入和推特句柄（Twitter handles）中出现的代码文字，会显示："请注意，`numpysum()` 无需使用 `for` 循环"。

代码段会显示：

```
def pythonsum(n):
    a = range(n)
    b = range(n)
    c = []

    for i in range(len(a)):
        a[i] = i ** 2
        b[i] = i ** 3
        c.append(a[i] + b[i])

    return c
```

所有的命令行输入或者输出内容会显示：

```
$ yum install python-numpy
```

新术语及**重要词汇**使用粗体字表示。对于在屏幕中看到的文字，如菜单或者对话框中的文字，排版形式为“单击 Next 按钮进入下一屏”。

警告或者重要的注释在此显示。

提示和小技巧在此显示。

读者反馈

我们欢迎读者对本书进行反馈，希望了解你对本书的看法：你喜欢哪些方面或不喜欢哪些方面。在帮助本社推出真正符合读者需要的图书方面，反馈信息至关重要。

如果想为我们提供一般反馈，请向 feedback@packtpub.com 邮箱发送电子邮件，并在邮件的标题中指出相应的书名即可。

如果某些主题是你擅长的领域，并且有意著书或撰稿，请进入 www.packtpub.com/authors，进一步阅读作者指南。

客户支持

你已经是 Packt 出版社的尊贵用户，为了让你的订购物超所值，我们将为你提供一些增值服务。

下载示例代码

访问 http://www.packtpub.com 网站并登录账户后，读者便可以下载所有已购 Packt 出版社图书的示例代码。如果是在其他地方购买的本书，可以访问 http://www.packtpub.com/support 并注册，通过电子邮件获取相应的代码。

勘误

虽然我们已经非常谨慎，尽力确保内容的正确性，但还是难免出错。如果你在书中发

现了不管是文字，还是代码方面的错误，并且通知我们，我们将感激不尽。这样做，能让其他读者免受这些错误的困扰，而且还能帮助我们改善本书的后续版本。如果发现了任何错误，请访问 http://www.packtpub.com/submit-errata，选择书名，单击“**勘误提交表**”链接，然后输入勘误的详细资料。一旦这些错误被确认，你的提交就会被接受，勘误信息就会上传到我们的网站，或者添加到该书勘误区中的已发现错误清单中。从 http://www.packtpub.com/support 选择书名，可以看到该书目前的所有勘误。

关于盗版行为

对各种媒体而言，互联网上受版权保护的各种材料都长期面临非法复制的问题。Packt 出版社非常重视版权保护和版权许可，如果你在网上看到本社图书任何形式的非法复制，请立刻向我们提供网络地址信息，以便我们及时采取补救措施。

请通过 copyright@packtpub.com 联系我们，并提供疑似盗版材料的链接信息。

感谢你帮助我们保护作者的权益，从而使我们能够提供更有价值的内容。

疑问解答

如果你对本书有任何疑问，可以通过 questions@packtpub.com 联系我们，我们将尽力为你解答。

作者简介

Ivan Idris，实验物理学硕士，学位论文侧重于应用计算机科学。毕业后，他曾经效力于多家公司，从事 Java 开发、数据仓库开发以及 QA 分析等方面的工作；目前，他的兴趣主要集中在商业智能、大数据和云计算等专业领域。

Ivan Idris 以编写简洁可测试的程序代码以及撰写有趣的技术文章为乐，同时也是 Packt 出版社 *NumPy Beginner's Guide-Second Edition*、*NumPy Cookbook* 和 *Learning NumPy Array* 等书籍的作者。读者可以访问 ivanidris.net 获取更多关于他的信息。

借此机会，我要向 Packt 出版社为本书的出版付出努力的众位审稿人和团队成员致以深深的谢意，是他们的付出令本书得以与读者见面；同时，还要感谢我的老师、教授和同事，感谢他们将科学和程序设计方面的知识传授给我。最后，还要向我的父母、妻子和孩子以及朋友们给予的支持表示万分感谢。

技术评审简介

Amanda Casari，数据科学家和工程师，来自西雅图地区。Amanda 拥有佛蒙特大学（University of Vermont）的电气工程学硕士学位和复杂系统研究证书，以及美国海军学院（United States Naval Academy）的系统工程学学士学位，具有 10 年以上的从业经验，职业生涯从海军军官、分析师、环境保护随行领队，直到集成工程师。她的研究兴趣主要集中在揭露天然系统的各种特性，并以此更新和优化人造复杂网络。同时，她还热衷于努力让数学和科学变得更加平易近人。

我非常感谢家人对我旅行计划的支持，以及本书审读过程中给予我的各种鼓励。N.Manukyan 对所有数据的热忱和 C.Stone 别出心裁的早餐总是让人难以忘怀；同时，向康乃馨登山俱乐部和 P.Nathan 对我各种爱好的亲切鼓励表示感谢。

Thomas A.Dyar（Tom）是美国北卡罗来纳州三角科技园区 BD Technologies（www.bd.com）公司基因科学团队的高级数据科学家，一直致力于为传染病和肿瘤诊断应用提供各种语境下基因数据的处理算法，其中语境包括从靶向面板（targeted panels）到整个基因组。他的专业领域包括如下几类：一是科学编程，涉及的语言有 Java、Python 和 R；二是机器学习，包括神经网络和核方法；三是数据分析和可视化。他的主要爱好是使用云资源开发海量数据驱动的解决方案，并将其概念化。

Tom 的职业生涯早期是从事软件方面的工作，为航天和石油化学工业开发用于过程控制的神经网络和专家系统工具。此外，他还在 MIT 从事过用于中风康复研究的分布式虚拟环境方面的工作，并在 BD 进行细胞生物学实验的高吞吐量图像处理自动化方面的研究。

Tom 毕业于波士顿大学纯应用数学专业，并且还是 ACM 和 IEEE 协会成员。

Dr. Hari Shanker Gupta 是算法交易系统开发领域中的一名资深量化研究员。在此之前，他在印度班加罗尔的印度科技大学获得博士后学位，同时，他还获得过该校的应用数学和科学计算博士学位。他的数学硕士学位是从贝勒纳斯印度教大学取得的，在研究生期间，因优异成绩而获得过该校 4 枚金质奖章。

Hari 在数学和科学计算领域知名期刊上发表过 5 篇研究论文，他的工作领域包括数学、统计学和计算等。他的工作经验涉及数值方法、偏微分方程、数学金融、随机积分、数据分析、有限差分和有限元方法。他对数学软件 MATLAB、统计学程序语言 R、Python 和 C 语言也非常精通。

同时，他还是 Packt 出版社 *Introduction to R for Quantitative Finance* 一书的技术评审。

Puneet Narula 在银行和金融领域有 8 年以上的从业经验，不过，其在技术领域有着过人的天赋和无限的热忱，使他重新回归了数据和分析世界的怀抱。他做了一项艰难的决定：放弃稳定的银行工作，最终选择追逐自己的梦想。

他于 2013 年获得都柏林理工学院的数据分析硕士学位，从此进入分析和数据科学的世界。目前，Puneet 在 Web Reservations International 从事 PPC 数据分析工作。

在 Web Reservations International（WRI），Puneet 每天都要面对海量的点击流数据，为了处理这些数据，需要综合运用 RapidMiner、R 和 Python。

非常感谢 Silviu Preoteasa 自始至终的支持和鼓励。

Alan J.Salmoni 以解读数据为乐，并且是 Salstat 网站（http://www.salstat.com）的创始人。他自 2001 年开始使用 Python 从事数据分析，并且为大学生和研究生讲授统计学。除了陪伴家人外，他的大部分时间都用在了自然语言处理的文本统计模型上面。

Alan 还创办了一家专门提供数据分析和用户体验分析服务的公司：Thought Into Design。

在此，我要向妻子 Jell 和女儿 Louise 的耐心致以深深的谢意。

目录