

Python数据分析

(第2版)

Python Data Analysis
Second Edition

[美] 阿曼多·凡丹戈 (Armando Fandango) 著
韩波 译

Python数据分析

(第2版)

Python Data Analysis

Second Edition

[美] 阿曼多·凡丹戈 (Armando Fandango) 著

韩波 译

人民邮电出版社

北京

图书在版编目 (C I P) 数据

Python数据分析 / (美) 阿曼多·凡丹戈
(Armando Fandango) 著 ; 韩波译. -- 2版. -- 北京 :
人民邮电出版社, 2018.6
ISBN 978-7-115-48117-7

I. ①P… II. ①阿… ②韩… 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(2018)第052547号

版权声明

Copyright ©2017 Packt Publishing. First published in the English language under the title *Python Data Analysis, Second Edition*.

All rights reserved.

本书由英国 **Packt Publishing** 公司授权人民邮电出版社出版。未经出版者书面许可,对本书的任何部分不得以任何方式或任何手段复制和传播。

版权所有,侵权必究。

◆ 著 [美] 阿曼多·凡丹戈 (Armando Fandango)
译 韩 波
责任编辑 胡俊英
责任印制 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>
固安县铭成印刷有限公司印刷

◆ 开本: 800×1000 1/16
印张: 18.25
字数: 362千字 2018年6月第2版
印数: 12 001—14 400册 2018年6月河北第1次印刷

著作权合同登记号 图字: 01-2017-8632 号

定价: 69.00 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

Python 作为一种高级程序设计语言，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。

本书就是一本介绍如何用 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 数组和 Pandas 入门开始，陆续介绍了数据的检索、数据加工与存储、数据可视化等内容。同时，本书还介绍了信号处理与时间序列、应用数据库、分析文本数据与社交媒体、预测性分析与机器学习、Python 生态系统的外部环境和云计算、性能优化及分析、并发性等内容。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。

本书延续了上一版示例丰富、简单易懂的优点，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。

作者简介

Armando Fandango 是 Epic 工程咨询集团首席数据科学家，负责与国防和政府机构有关的保密项目。Armando 是一位技术精湛的技术人员，拥有全球创业公司和大型公司的工作经历和高级管理经验。他的工作涉及金融科技、证券交易所、银行、生物信息学、基因组学、广告技术、基础设施、交通运输、能源、人力资源和娱乐等多个领域。

Armando 在预测分析、数据科学、机器学习、大数据、产品工程、高性能计算和云基础设施等项目中工作了十多年。他的研究兴趣横跨机器学习、深度学习和科学计算等领域。

我要特别感谢我的妻子在编写本书过程中给予我的支持。同时，我还要感谢 UCF 的 Paul Wiegand 博士，他总是鼓励我寻求机会。最后，我非常感谢 Packt 团队的 Tushar、Sumeet、Amrita、Deepti 以及其他工作人员，感谢他们为本书的面世所付出的巨大努力。

技术审稿人简介

Joran Beasley 毕业于爱达荷大学，获得了计算机科学学士学位。7 年以来，他一直从事基于 Python 的桌面应用程序编程，同时将其应用于监控农业的大规模传感器网络。目前，他居住于爱达荷州莫斯科市，作为软件工程师供职于 METER 集团。

我要感谢我的妻子 Nicole，当我把大量的时间倾注在写作上的时候，她在默默地抚养我们的两个孩子。

Ratan Kumar 在过去的几年里一直在利用各种语言和技术从事软件编程。2013 年以来，Ratan Kumar 在 Web 服务领域使用 Python 语言，他发现无论是在个人项目还是专业项目方面，Python 都是最优雅、最高效、最易于接受的编程语言之一。Ratan 目前住在班加罗尔，是一个旨在简化股票市场投资的小型团队的核心成员。

前言

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。随着数据科学的发展，数据分析也呈现流行之势，几乎已经渗透到工业的方方面面。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。

在第 1 章中，我们将给出一幅描绘与数据分析相关的 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 在 1989 年就诞生了。Python 易学、易用，而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

第 1 章“Python 程序库入门”手把手地指导读者正确安装配置 Python 和基础的 Python 数值分析软件库。同时，本章还会展示如何通过 NumPy 创建一个小程序以及如何利用 Matplotlib 来绘制简单的图形。

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“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 4 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 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 和 Pandas 的绘图函数来实现数据的可视化。

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 和 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使用 Python。

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（Profiling）和 Cython 等关键技术来改善性能的各种技巧，同时还为读者介绍多核和分布式系统方面的相关框架。

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”概述本书中用到的程序库中的各种函数，以便于读者查阅。

读者须知

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 3.5.0 版本以上的 Python 和 pip3 软件。Python 3.5.x 的下载地址为 <https://www.python.org/downloads/>，这个页面上不仅提供了用于 Windows 和 Mac OS X 的安装程序，也提供了用于 Linux、UNIX 和 Mac OS X 系统的 Python 源代码。多数情况下，我们都得通过运行以下命令来以管理员权限安装各种本书要用到的 Python 库：

```
$ pip3 install <some library>
```

以下是用于本书中的示例的 Python 库的清单。

- numpy
- scipy
- Pandas
- Matplotlib
- ipython
- jupyter
- notebook
- readline
- scikit-learn
- rpy2
- Quandl
- statsmodels
- feedparser
- beautifulsoup4
- lxml
- numexpr
- tables
- openpyxl
- xlswriter
- xlrd
- pony
- dataset
- pymongo
- redis
- python3-memcache
- cassandra-driver
- sqlalchemy
- nltk
- networkx
- theano
- nose_parameterized
- pydot2
- deap
- JPype1
- gprof2dot
- line_profiler
- cython
- cytoolz
- joblib
- bottleneck
- jug
- mpi4py

除了各种 Python 库之外，我们还需要以下软件。

- Redis server
- Cassandra
- Java 8
- Graphviz
- Octave
- R
- SWIG
- PCRE
- Boost
- gfortran
- MPI

通常情况下，对于以上的程序化和软件，我们应该选用最新版本。



以上列出的某些软件只是用于某个示例，因此在安装之前，请先检查该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：

```
$ pip3 uninstall <some library>
```

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，读者可以经由 Khan Academy、Coursera 之类的在线资源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 Scientific 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：“如果您的当前登录账户没有足

够权限的话，则需要在这条命令前面加上 `sudo`。”

代码段显示格式如下。

```
def pythonsum(n):  
    a = list(range(n))  
    b = list(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为如下格式。

```
$ pip3 install numpy scipy pandas matplotlib jupyter notebook
```



警告或者重要的提示在此显示。



小技巧在此显示。

资源与支持

本书由异步社区出品，社区（<https://www.epubit.com/>）为您提供相关资源和后续服务。

配套资源

本书提供如下资源：

- 本书源代码；
- 书中彩图文件。

要获得以上配套资源，请在异步社区本书页面中点击 **配套资源**，跳转到下载界面，按提示进行操作即可。注意：为保证购书读者的权益，该操作会给出相关提示，要求输入提取码进行验证。

如果您是教师，希望获得教学配套资源，请在社区本书页面中直接联系本书的责任编辑。

提交勘误

作者和编辑尽最大努力来确保书中内容的准确性，但难免会存在疏漏。欢迎您将发现的问题反馈给我们，帮助我们提升图书的质量。

当您发现错误时，请登录异步社区，按书名搜索，进入本书页面，点击“提交勘误”，输入勘误信息，单击“提交”按钮即可。本书的作者和编辑会对您提交的勘误进行审核，确认并接受后，您将获赠异步社区的 100 积分。积分可用于在异步社区兑换优惠券、样书或奖品。



扫码关注本书

扫描下方二维码，您将会在异步社区微信服务号中看到本书信息及相关的服务提示。



与我们联系

我们的联系邮箱是 contact@epubit.com.cn。

如果您对本书有任何疑问或建议，请您发邮件给我们，并请在邮件标题中注明本书书名，以便我们更高效地做出反馈。

如果您有兴趣出版图书、录制教学视频，或者参与图书翻译、技术审校等工作，可以发邮件给我们；有意出版图书的作者也可以到异步社区在线提交投稿（直接访问 www.epubit.com/selfpublish/submission 即可）。

如果您是学校、培训机构或企业，想批量购买本书或异步社区出版的其他图书，也可以发邮件给我们。

如果您在网上发现有针对异步社区出品图书的各种形式的盗版行为，包括对图书全部或部分内容的非授权传播，请您将怀疑有侵权行为的链接发邮件给我们。您的这一举动是对作者权益的保护，也是我们持续为您提供有价值的内容的动力之源。

关于异步社区和异步图书

“异步社区”是人民邮电出版社旗下 IT 专业图书社区，致力于出版精品 IT 技术图书和相关学习产品，为作译者提供优质出版服务。异步社区创办于 2015 年 8 月，提供大量精品 IT 技术图书和电子书，以及高品质技术文章和视频课程。更多详情请访问异步社区官网 <https://www.epubit.com>。

“异步图书”是由异步社区编辑团队策划出版的精品 IT 专业图书的品牌，依托于人民邮电出版社近 30 年的计算机图书出版积累和专业编辑团队，相关图书在封面上印有异步图书的 LOGO。异步图书的出版领域包括软件开发、大数据、AI、测试、前端、网络技术 etc。



异步社区



微信服务号

目录

第 1 章 Python 程序库入门.....1

1.1 安装 Python 3.....3

1.1.1 安装数据分析程序库.....3

1.1.2 Linux 平台或 Mac OS X 平台.....3

1.1.3 Windows 平台.....4

1.2 将 IPython 用作 shell.....4

1.3 学习手册页.....6

1.4 Jupyter Notebook.....7

1.5 NumPy 数组.....8

1.6 一个简单的应用.....8

1.7 从何处寻求帮助和参考资料.....11

1.8 查看 Python 库中包含的模块.....12

1.9 通过 Matplotlib 实现数据的 可视化.....12

1.10 小结.....14

第 2 章 NumPy 数组.....15

2.1 NumPy 数组对象.....16

2.2 创建多维数组.....17

2.3 选择 NumPy 数组元素.....17

2.4 NumPy 的数值类型.....18

2.4.1 数据类型对象.....20

2.4.2 字符码.....20

2.4.3 dtype 构造函数.....21

2.4.4 dtype 属性.....22

2.5 一维数组的切片与索引.....23

2.6 处理数组形状.....23

2.6.1 堆叠数组.....25

2.6.2 拆分 NumPy 数组.....28

2.6.3 NumPy 数组的属性.....30

2.6.4 数组的转换.....34

2.7 创建数组的视图和拷贝.....35

2.8 花式索引.....36

2.9 基于位置列表的索引方法.....38

2.10 用布尔型变量索引 NumPy 数组.....39

2.11 NumPy 数组的广播.....41

2.12 小结.....44

2.13 参考资料.....44

第 3 章 Pandas 入门.....45

3.1 Pandas 的安装与概览.....46

3.2 Pandas 数据结构之 DataFrame.....47	第 5 章 数据的检索、加工与存储92
3.3 Pandas 数据结构之 Series49	5.1 利用 NumPy 和 pandas 对 CSV 文件进行写操作.....92
3.4 利用 Pandas 查询数据.....52	5.2 二进制.npy 与 pickle 格式.....94
3.5 利用 Pandas 的 DataFrame 进行 统计计算.....56	5.3 使用 PyTables 存储数据.....97
3.6 利用 Pandas 的 DataFrame 实现 数据聚合.....58	5.4 Pandas DataFrame 与 HDF5 仓库 之间的读写操作.....99
3.7 DataFrame 的串联与附加 操作.....62	5.5 使用 Pandas 读写 Excel 文件.....102
3.8 连接 DataFrames.....63	5.6 使用 REST Web 服务和 JSON ...103
3.9 处理缺失数据问题65	5.7 使用 Pandas 读写 JSON105
3.10 处理日期数据67	5.8 解析 RSS 和 Atom 订阅106
3.11 数据透视表70	5.9 使用 Beautiful Soup 解析 HTML108
3.12 小结71	5.10 小结.....114
3.13 参考资料71	5.11 参考资料.....114
第 4 章 统计学与线性代数72	第 6 章 数据可视化115
4.1 用 NumPy 进行简单的描述性 统计计算.....72	6.1 Matplotlib 的子库.....116
4.2 用 NumPy 进行线性代数运算.....75	6.2 Matplotlib 绘图入门116
4.2.1 用 NumPy 求矩阵的逆.....75	6.3 对数图.....118
4.2.2 用 NumPy 解线性 方程组77	6.4 散点图.....119
4.3 用 NumPy 计算特征值和特征 向量.....78	6.5 图例和注解121
4.4 NumPy 随机数80	6.6 三维图.....123
4.4.1 用二项式分布进行博弈.....81	6.7 Pandas 绘图125
4.4.2 正态分布采样83	6.8 时滞图.....127
4.4.3 用 SciPy 进行正态检验.....84	6.9 自相关图129
4.5 创建掩码式 NumPy 数组86	6.10 Plot.ly.....130
4.6 忽略负值和极值88	6.11 小结.....132
4.7 小结91	第 7 章 信号处理与时间序列133
	7.1 statsmodels 模块.....134
	7.2 移动平均值134

7.3 窗口函数	136	9.4 词袋模型	180
7.4 协整的定义	138	9.5 词频分析	181
7.5 自相关	140	9.6 朴素贝叶斯分类	183
7.6 自回归模型	142	9.7 情感分析	186
7.7 ARMA 模型	145	9.8 创建词云	189
7.8 生成周期信号	147	9.9 社交网络分析	193
7.9 傅里叶分析	149	9.10 小结	195
7.10 谱分析	152		
7.11 滤波	153	第 10 章 预测性分析与机器学习	197
7.12 小结	155	10.1 预处理	198
第 8 章 应用数据库	156	10.2 基于逻辑回归的分类	201
8.1 基于 sqlite3 的轻量级访问	157	10.3 基于支持向量机的分类	202
8.2 通过 Pandas 访问数据库	159	10.4 基于 ElasticNetCV 的回归 分析	205
8.3 SQLAlchemy	161	10.5 支持向量回归	207
8.3.1 SQLAlchemy 的安装和 配置	161	10.6 基于相似性传播算法的聚类 分析	210
8.3.2 通过 SQLAlchemy 填充 数据库	162	10.7 均值漂移算法	211
8.3.3 通过 SQLAlchemy 查询 数据库	164	10.8 遗传算法	213
8.4 Pony ORM	166	10.9 神经网络	217
8.5 Dataset: 懒人数据库	167	10.10 决策树	219
8.6 PyMongo 与 MongoDB	168	10.11 小结	222
8.7 利用 Redis 存储数据	170		
8.8 利用 memcache 存储数据	171	第 11 章 Python 生态系统的外部环境和 云计算	223
8.9 Apache Cassandra	172	11.1 与 MATLAB/Octave 交换 信息	224
8.10 小结	174	11.2 安装 rpy2	225
第 9 章 分析文本数据和社交媒体	176	11.3 连接 R	225
9.1 安装 NLTK	177	11.4 为 Java 传递 NumPy 数组	228
9.2 NLTK 简介	177	11.5 集成 SWIG 和 NumPy	229
9.3 滤除停用字、姓名和数字	178	11.6 集成 Boost 和 Python	233
		11.7 通过 f2py 使用 Fortran 代码	235

11.8	PythonAnywhere 云	236	并发性	254	
11.9	小结	238	12.6	比较 Bottleneck 函数与 NumPy 函数	255
第 12 章	性能优化、性能分析与 并发性	239	12.7	通过 Jug 实现 MapReduce	257
12.1	代码的性能分析	240	12.8	安装 MPI for Python	259
12.2	安装 Cython	245	12.9	IPython Parallel	260
12.3	调用 C 代码	248	12.10	小结	263
12.4	利用 multiprocessing 创建 进程池	252	附录 A	重要概念	264
12.5	通过 Joblib 提高 for 循环的		附录 B	常用函数	269