




Python数据分析 (影印版)

Python Data Analysis

Ivan Idris 著

[PACKT]
PUBLISHING

学习如何运用流行的开源Python模块实现强大的数据分析技术

 东南大学出版社
SOUTHEAST UNIVERSITY PRESS

Python 数据分析(影印版)

Ivan Idris 著

南京 东南大学出版社

图书在版编目(CIP)数据

Python 数据分析:英文/(印尼)伊德里斯(Idris, I.)
著. —影印本. —南京:东南大学出版社, 2016.1

书名原文:Python Data Analysis

ISBN 978-7-5641-6064-7

I. ①P… II. ①伊… III. ①软件工具—程序设计—英文 IV. ①TP311.56

中国版本图书馆 CIP 数据核字(2015)第 243362 号

© 2014 by PACKT Publishing Ltd

Reprint of the English Edition, jointly published by PACKT Publishing Ltd and Southeast University Press, 2016.
Authorized reprint of the original English edition, 2015 PACKT Publishing Ltd, the owner of all rights to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

英文原版由 PACKT Publishing Ltd 出版 2014。

英文影印版由东南大学出版社出版 2016。此影印版的出版和销售得到出版权和销售权的所有者——PACKT Publishing Ltd 的许可。

版权所有,未得书面许可,本书的任何部分和全部不得以任何形式重制。

Python 数据分析(影印版)

出版发行:东南大学出版社

地 址:南京四牌楼 2 号 邮编:210096

出 版 人:江建中

网 址:<http://www.seupress.com>

电子邮件:press@seupress.com

印 刷:常州市武进第三印刷有限公司

开 本:787 毫米×980 毫米 16 开本

印 张:21.75

字 数:426 千字

版 次:2016 年 1 月第 1 版

印 次:2016 年 1 月第 1 次印刷

书 号:ISBN 978-7-5641-6064-7

定 价:68.00 元

Credits

Author

Ivan Idris

Project Coordinator

Shipra Chawhan

Reviewers

Amanda Casari

Thomas A. Dyar

Dr. Hari Shanker Gupta

Puneet Narula

Alan J. Salmoni

Proofreaders

Simran Bhogal

Maria Gould

Ameesha Green

Commissioning Editor

Akram Hussain

Indexers

Hemangini Bari

Mariammal Chettiyar

Rekha Nair

Tejal Soni

Acquisition Editor

Owen Roberts

Graphics

Sheetal Aute

Content Development Editor

Prachi Bisht

Production Coordinators

Adonia Jones

Manu Joseph

Komal Ramchandani

Technical Editor

Pankaj Kadam

Cover Work

Manu Joseph

Copy Editors

Roshni Banerjee

Sarang Chari

Adithi Shetty

About the Author

Ivan Idris has an MSc degree in Experimental Physics. His graduation thesis had a strong emphasis on Applied Computer Science. After graduating, he worked for several companies as Java developer, data warehouse developer, and QA analyst. His main professional interests are Business Intelligence, Big Data, and Cloud Computing.

Ivan Idris enjoys writing clean, testable code and interesting technical articles. He is the author of *NumPy Beginner's Guide - Second Edition*, *NumPy Cookbook*, and *Learning NumPy Array*, all by Packt Publishing. You can find more information and a blog with a few NumPy examples at ivanidris.net.

I would like to take this opportunity to thank the reviewers and the team at Packt Publishing for making this book possible. Also, my thanks go to my teachers, professors, and colleagues, who taught me about science and programming. Last but not least, I would like to acknowledge my parents, family, and friends for their support.

About the Reviewers

Amanda Casari is currently a data scientist and engineer in the Seattle area. Amanda received her MSEE degree and Certificate of Study in Complex Systems from the University of Vermont and a BS degree in Systems Engineering from the United States Naval Academy. She has more than 10 years of professional experience, ranging from naval officer, analyst, conservation trip leader to integration engineer. Her research interests focus on discovering attributes of natural systems to update and optimize man-made complex networks. Amanda is passionate about making Mathematics and Science approachable to everyone.

I would like to thank my family for supporting our journey and inspiring me during this effort, N. Manukyan for all of her data enthusiasm, C. Stone for creative breakfasts, the Carnation Climbing Club, and P. Nathan for kindly encouraging my myriad interests.

Thomas A. Dyar (Tom) is a senior data scientist in the Genomic Sciences group at BD Technologies (www.bd.com), Research Triangle Park, North Carolina, where he develops algorithms to process genomic data in a variety of contexts—from targeted panels to whole genomes—for infectious disease and oncology diagnostics applications. His areas of expertise are scientific programming in Java, Python, and R; machine learning, including neural networks and kernel methods; and data analysis and visualization. His primary interests are in conceptualizing and developing large-scale data-driven solutions using Cloud resources.

Tom started his career in software, developing neural networks and expert systems tools for process control in the aerospace and petrochemical industries. He has also worked on distributed virtual environments for stroke rehabilitation at MIT and automated image processing for high-throughput cell biology experiments at BD.

Tom earned his BA degree in Pure & Applied Mathematics from Boston University and is a member of the ACM and IEEE associations.

Dr. Hari Shanker Gupta is a senior quantitative research analyst working in the area of algorithmic trading system development. Prior to this, he was a post-doctoral fellow at the Indian Institute of Science (IISc), Bangalore, India. He obtained his PhD in Applied Mathematics and Scientific Computation from IISc. He completed his MSc in Mathematics from Banaras Hindu University (BHU), Varanasi, India. During his MSc, he was awarded four gold medals for outstanding performance at BHU.

Hari has published five research papers in reputed journals in the field of Mathematics and Scientific Computation. He has experience working in the areas of Mathematics, Statistics, and Computation. His experience includes working in numerical methods, partial differential equations, mathematical finance, stochastic calculus, data analysis, finite difference, and finite element methods. He is very comfortable with the mathematics software, MATLAB; the statistics programming language, R; Python; and the programming language, C.

He has reviewed the book *Introduction to R for Quantitative Finance*, Packt Publishing.

Puneet Narula has over 8 years of experience in the Banking and Finance industry, but his aptitude and passion for the technology sector has brought him back into the world of data and analytics. Leaving behind a stable career in banking was a very tough decision, but following his dreams was even more important to him. He completed his MSc degree in Data Analytics from Dublin Institute of Technology in 2013 to enter the world of analytics and data science. Currently, Puneet is working with Web Reservations International as a PPC data analyst.

At Web Reservations International (WRI), Puneet works with massive clickstream data from both direct and affiliate sources. The technologies used for the analysis is a combination of RapidMiner, R, and Python.

I want to thank Silviu Preoteasa for all his support and motivation at all times.

Alan J. Salmoni enjoys making sense of data and is the author of Salstat (<http://www.salstat.com>). He has been using Python for data analysis since 2001 and has taught statistics to undergraduates and postgraduates. When not with his family, he spends time generating large statistical models of text for natural language processing.

Alan owns a company, Thought Into Design, which specializes in data analysis and user experience.

I would like to thank my wife, Jell, and my daughter, Louise, for their patience.

Table of Contents

Preface	1
Chapter 1: Getting Started with Python Libraries	9
Software used in this book	10
Installing software and setup	10
On Windows	10
On Linux	12
On Mac OS X	13
Building NumPy SciPy, matplotlib, and IPython from source	14
Installing with setuptools	15
NumPy arrays	16
A simple application	16
Using IPython as a shell	19
Reading manual pages	22
IPython notebooks	22
Where to find help and references	23
Summary	23
Chapter 2: NumPy Arrays	25
The NumPy array object	25
The advantages of NumPy arrays	26
Creating a multidimensional array	27
Selecting NumPy array elements	27
NumPy numerical types	28
Data type objects	30
Character codes	30
The dtype constructors	31
The dtype attributes	31

One-dimensional slicing and indexing	32
Manipulating array shapes	32
Stacking arrays	35
Splitting NumPy arrays	39
NumPy array attributes	41
Converting arrays	48
Creating array views and copies	48
Fancy indexing	50
Indexing with a list of locations	52
Indexing NumPy arrays with Booleans	53
Broadcasting NumPy arrays	55
Summary	58
Chapter 3: Statistics and Linear Algebra	59
NumPy and SciPy modules	59
Basic descriptive statistics with NumPy	63
Linear algebra with NumPy	66
Inverting matrices with NumPy	66
Solving linear systems with NumPy	68
Finding eigenvalues and eigenvectors with NumPy	69
NumPy random numbers	71
Gambling with the binomial distribution	72
Sampling the normal distribution	74
Performing a normality test with SciPy	75
Creating a NumPy-masked array	78
Disregarding negative and extreme values	80
Summary	83
Chapter 4: pandas Primer	85
Installing and exploring pandas	86
pandas DataFrames	87
pandas Series	90
Querying data in pandas	94
Statistics with pandas DataFrames	97
Data aggregation with pandas DataFrames	99
Concatenating and appending DataFrames	103
Joining DataFrames	105
Handling missing values	108
Dealing with dates	110
Pivot tables	113
Remote data access	114
Summary	117

Chapter 5: Retrieving, Processing, and Storing Data	119
Writing CSV files with NumPy and pandas	120
Comparing the NumPy .npy binary format and pickling pandas DataFrames	122
Storing data with PyTables	124
Reading and writing pandas DataFrames to HDF5 stores	126
Reading and writing to Excel with pandas	129
Using REST web services and JSON	131
Reading and writing JSON with pandas	132
Parsing RSS and Atom feeds	134
Parsing HTML with Beautiful Soup	135
Summary	142
Chapter 6: Data Visualization	143
matplotlib subpackages	144
Basic matplotlib plots	144
Logarithmic plots	146
Scatter plots	148
Legends and annotations	150
Three-dimensional plots	153
Plotting in pandas	155
Lag plots	158
Autocorrelation plots	159
Plot.ly	160
Summary	163
Chapter 7: Signal Processing and Time Series	165
statsmodels subpackages	166
Moving averages	167
Window functions	168
Defining cointegration	170
Autocorrelation	173
Autoregressive models	176
ARMA models	179
Generating periodic signals	181
Fourier analysis	184
Spectral analysis	186
Filtering	187
Summary	189
Chapter 8: Working with Databases	191
Lightweight access with sqlite3	192
Accessing databases from pandas	194

SQLAlchemy	196
Installing and setting up SQLAlchemy	196
Populating a database with SQLAlchemy	198
Querying the database with SQLAlchemy	200
Pony ORM	201
Dataset – databases for lazy people	202
PyMongo and MongoDB	204
Storing data in Redis	206
Apache Cassandra	207
Summary	210
Chapter 9: Analyzing Textual Data and Social Media	211
Installing NLTK	212
Filtering out stopwords, names, and numbers	214
The bag-of-words model	216
Analyzing word frequencies	217
Naive Bayes classification	219
Sentiment analysis	222
Creating word clouds	225
Social network analysis	230
Summary	232
Chapter 10: Predictive Analytics and Machine Learning	233
A tour of scikit-learn	235
Preprocessing	236
Classification with logistic regression	238
Classification with support vector machines	240
Regression with ElasticNetCV	242
Support vector regression	245
Clustering with affinity propagation	248
Mean Shift	250
Genetic algorithms	252
Neural networks	257
Decision trees	259
Summary	261
Chapter 11: Environments Outside the Python Ecosystem and Cloud Computing	263
Exchanging information with MATLAB/Octave	264
Installing rpy2	265
Interfacing with R	265
Sending NumPy arrays to Java	268
Integrating SWIG and NumPy	269

Integrating Boost and Python	272
Using Fortran code through f2py	274
Setting up Google App Engine	275
Running programs on PythonAnywhere	276
Working with Wakari	277
Summary	278
Chapter 12: Performance Tuning, Profiling, and Concurrency	279
Profiling the code	280
Installing Cython	284
Calling C code	288
Creating a process pool with multiprocessing	290
Speeding up embarrassingly parallel for loops with Joblib	293
Comparing Bottleneck to NumPy functions	294
Performing MapReduce with Jug	296
Installing MPI for Python	298
IPython Parallel	299
Summary	303
Appendix A: Key Concepts	305
Appendix B: Useful Functions	311
matplotlib	311
NumPy	312
pandas	313
Scikit-learn	314
SciPy	315
scipy.fftpack	315
scipy.signal	315
scipy.stats	315
Appendix C: Online Resources	317
Index	319

Preface

"*Data analysis is Python's killer app.*"

– *Unknown*

Data analysis has a rich history in the natural, biomedical, and social sciences. You may have heard of *Big Data*. Although, it's hard to give a precise definition of Big Data, we should be aware of its impact on data analysis efforts. Currently, we have the following trends associated with Big Data:

- The world's population continues to grow
- More and more data is collected and stored
- The number of transistors that can be put on a computer chip cannot grow indefinitely
- Governments, scientists, industry, and individuals have a growing need to learn from data

Data analysis has gained popularity lately due to the hype around *Data Science*. Data analysis and Data Science attempt to extract information from data. For that purpose, we use techniques from statistics, machine learning, signal processing, natural language processing, and computer science.

A mind map visualizing Python software that can be used for data analysis can be found at <http://www.xmind.net/m/Wvfc/>. The first thing that we should notice is that the Python ecosystem is very mature. It includes famous packages such as NumPy, SciPy, and matplotlib. This should not come as a surprise since Python has been around since 1989. Python is easy to learn and use, less verbose than other programming languages, and very readable. Even if you don't know Python, you can pick up the basics within days, especially if you have experience in another programming language. To enjoy this book, you don't need more than the basics. There are plenty of books, courses, and online tutorials that teach Python.

What this book covers

This book starts as a tutorial on NumPy, SciPy, matplotlib, and pandas. These are open source Python packages useful for numerical work, data wrangling, and visualization. Combined, they can compete with MATLAB, Mathematica, and R. The second half of the book teaches more advanced topics such as signal processing, databases, text analysis, machine learning, interoperability, and performance tuning.

Chapter 1, Getting Started with Python Libraries, guides us to achieve a successful installation of the numerical Python software and set it up step by step. Also, we will create a small application.

Chapter 2, NumPy Arrays, introduces us to NumPy fundamentals and arrays. By the end of this chapter, we will have basic understanding of NumPy arrays and the associated functions.

Chapter 3, Statistics and Linear Algebra, gives a quick overview of linear algebra and statistical functions.

Chapter 4, pandas Primer, provides a tutorial on basic pandas functionality where we learn about pandas data structures and operations.

Chapter 5, Retrieving, Processing, and Storing Data, explains how to acquire data in various formats and how to clean raw data and store it.

Chapter 6, Data Visualization, teaches how to plot data with matplotlib.

Chapter 7, Signal Processing and Time Series, contains time series and signal processing examples using sunspot cycles data. The examples mostly use NumPy/SciPy, along with statsmodels in at least one example.

Chapter 8, Working with Databases, provides information about various databases (relational and NoSQL) and related APIs.

Chapter 9, Analyzing Textual Data and Social Media, analyzes texts for sentiment analysis and topics extraction. A small example is also given of network analysis.

Chapter 10, Predictive Analytics and Machine Learning, explains artificial intelligence with weather prediction as a running example and mostly uses scikit-learn. However, some machine learning algorithms are not covered by scikit-learn, so for those, we use other APIs.

Chapter 11, Environments Outside the Python Ecosystem and Cloud Computing, gives various examples on how to integrate existing code not written in Python. Also, setup in the Cloud will be demonstrated.

Chapter 12, Performance Tuning, Profiling, and Concurrency, gives hints on improving performance with profiling and Cythoning as key techniques. For multicore, distributed systems, we discuss the relevant frameworks too.

Appendix A, Key Concepts, serves as a glossary containing short descriptions of key concepts found throughout the book.

Appendix B, Useful Functions, gives an overview of functions used in the book.

Appendix C, Online Resources, lists links to documentation, forums, articles, and other important information.

What you need for this book

The code examples in this book should work on most modern operating systems. For all chapters, Python 2 and pip is required. To install Python, go to <https://wiki.python.org/moin/BeginnersGuide/Download>. To install pip, go to <http://pip.readthedocs.org/en/latest/installing.html>. Instructions to install software are given throughout the chapters. Most of the time, we need to run the following command with admin privileges:

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

The following is a list of software used for the examples and versions used for testing purposes:

- 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
- BeautifulSoup 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
- theano 0.2.0
- Graphviz 2.36.0
- pydot2 1.0.33
- Octave 3.8.0
- R 3.1.1
- rpy2 2.4.2
- JPytype 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_profiler 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