PACKT PUBLISHING

掌握和使用正确的数据可视化方法

Python 数据可视化编程实战

Python Data Visualization Cookbook

[爱尔兰] Igor Milovanović 著
颛清山 译

人民邮电出版社
POSTS & TELECOM PRESS

Python
数据可视化编程实战

[爱尔兰] Igor Milovanović 著
颛清山 译

人民邮电出版社
北京

图书在版编目（CIP）数据

Python数据可视化编程实战 /（爱尔兰）米洛万诺维奇（Milovanovic,I.）著 ；颛清山译. -- 北京 ：人民邮电出版社，2015.5
ISBN 978-7-115-38439-3

Ⅰ. ①P… Ⅱ. ①米… ②颛… 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(2015)第057566号

版权声明

◆ 著　[爱尔兰] Igor Milovanović
译　颛清山
责任编辑　陈冀康
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
三河市海波印务有限公司印刷

◆ 开本：800×1000　1/16
印张：16.75
字数：318 千字　2015 年 5 月第 1 版
印数：1－3 000 册　2015 年 5 月河北第 1 次印刷
著作权合同登记号　图字：01-2013-9037 号

定价：49.00 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

本书是一本使用 Python 实现数据可视化编程的实战指南，介绍了如何使用 Python 最流行的库，通过 60 余种方法创建美观的数据可视化效果。

全书共 8 章，分别介绍了准备工作环境、了解数据、绘制并定制化图表、学习更多图表和定制化、创建 3D 可视化图表、用图像和地图绘制图表、使用正确的图表理解数据以及更多的 matplotlib 知识。

本书适合那些对 Python 编程有一定基础的开发人员阅读，可以帮助读者从头开始了解数据、数据格式、数据可视化，并学会使用 Python 可视化数据。

译者序

图形可视化是展示数据的一个非常好的手段，好的图表自己会说话。毋庸多言，在Python的世界里，matplotlib是最著名的绘图库，它支持几乎所有2D绘图和部分3D绘图，被广泛地应用在科学计算和数据可视化领域。但是介绍matplotlib的中文书籍很少，大部分书籍只是在部分章节中提到了matplotlib的基本用法，因此在内容和深度上都力有不逮。本书则是一本专门介绍matplotlib的译著。

matplotlib是一个开源项目，由John Hunter发起。关于matplotlib的由来，有一个小故事。John Hunter和他研究癫痫症的同事借助一个专有软件做脑皮层电图分析，但是他所在的实验室只有一份该电图分析软件的许可。他和许多一起工作的同事不得不轮流使用该软件的硬件加密狗。于是，John Hunter便有了开发一个工具来替代当前所使用的软件的想法。当时MATLAB被广泛应用在生物医学界中，John Hunter等最初是想开发一个基于MATLAB的版本，但是由于MATLAB的一些限制和不足，加上他本身对Python非常熟悉，于是就有了matplotlib的诞生。

所以，无论从名字上，还是从matplotlib提供的函数名称、参数及使用方法都与MATLAB非常相似。对于一个MATLAB开发人员，使用起来会相当得心应手。即使对不熟悉MATLAB的开发人员（譬如我），对其函数的使用也能够一目了然，而且matplotlib有着非常丰富的文档和实例，加上本书的介绍，学习起来将会非常轻松。

matplotlib命令提供了交互绘图的方式，在Python的交互shell中，我们可以执行matplotlib命令来实时地绘制图形并对其进行修改。生成的图像可以保存成许多格式，这取决于其所使用的后端，但绝大多数后端都支持如png、pdf、ps、eps和svg等格式。

在之前的项目中，我使用 Python 的 Locust 工具进行性能测试，该工具非常出色，然而在对获取到的性能数据的分析上，没有提供太多的功能。于是我决定使用 matplotlib 进行性能数据的分析和可视化。从绘制最简单的柱状图、线形图，到引入散点图、直方图，我渐渐对 matplotlib 有了进一步的了解，也对它提供了如此强大的功能却又不失易用性而着迷。

虽然本书是一本 cookbook，然而它并不仅仅局限在讲解如何绘制各种图形上，更重要的是，本书让我们了解了如何用正确的图形把数据可视化出来，也就是“do the right thing in the right way”。在翻译本书的过程中，我意识到，如果我当时手头有这么一本书，将会少走不少弯路。本书包括了非常多的图形介绍以及丰富的示例，我相信，读完本书以后，读者将能应对各种常见的数据可视化问题。

在这里，我要特别感谢一下我的妻子董秋影，在精神和专业知识上她都给予了我莫大的帮助，没有她就没有这本译稿的完成。她从事医疗图像算法工作，对各种图形和算法以及 MATLAB 都有很深的了解，本书的每一章都经过了她认真的审阅校对。此外，感谢彭明伟先生完成了第 1 章的初稿翻译工作。最后，感谢人民邮电出版社陈冀康老师专业细心地审核，和陈老师合作很轻松、很开心。

由于译者水平有限，错误和失误在所难免，如有任何意见和建议，请不吝指正，我将感激不尽。我的邮箱：zhuanqingshan@163.com。

颛清山

2014 年 12 月于北京

作者简介

Igor Milovanović 是一个在Linux系统和软件工程领域有深厚背景的经验丰富的开发人员。具备创建可扩展数据驱动分布式富软件系统的技术。

他是一个高性能系统设计的布道者，对软件架构和开发方法论有着浓厚的兴趣。他一直坚持倡导促进高质量软件的方法论，如测试驱动开发、一键部署和持续集成。

他也拥有坚实的产品开发知识。拥有领域经验知识，并参加过官方培训，他能够在业务和开发人员之间很好地传递业务知识和业务流程。

非常感谢我的未婚妻，她允许我把大量的时间花费在工作上而没有陪伴她，并在我无休止地谈论本书时甘愿做一个热心的听众。我也想感谢我的哥哥，他一直是我坚强的后盾。感谢我的父母，给予我各种发展自己的空间，让我成为今天的自己。

如果没有开发Python、matplotlib和所有本书中使用的库的开源社区的巨大能量，我不可能写出这本书。我深深地感谢所有这些项目背后的人们。感谢你们！

评阅者简介

Tarek Amr 从东安格里亚大学获得了数据挖掘和信息检索专业的研究生学位。他在软件开发领域有近 10 年的经验。自从 2007 年开始，他一直在 Global Voices Online (GVO)义务工作，目前他是埃及 Open Knowledge Foundation (OKFN)的大使。他热衷于开放数据、Government 2.0、数据可视化、数据新闻、机器学习和自然语言处理。

Tarek 的 Twitter 账号是@gr33ndata，主页是 http://tarekamr.appspot.com/。

Jayesh K. Gupta 是 Matlab Toolbox for Biclustering Analysis (MTBA)的首席开发人员。他目前是一名 IIT Kanpur 的在读研究生和研究员。他的兴趣是模式识别。他对基础科学也有浓厚的兴趣，认为它们是自然界中的模式分析工具。来到 IIT 之后，他看到这种分析是如何借助机器学习算法广泛应用在各种不同的应用程序中的。他相信通过机器智能来强化人类的想法是增进人类知识的最好方式之一。他是一个长期的技术爱好者和自由软件的布道者。他的网名是 rejuvyesh。他也是一名狂热的读者，从 Goodreads 可以获得他读过的书籍的信息。从 Bitbucket 和 GitHub 可以找到他的项目。所有的链接都在 http://home.iitk.ac.in/~jayeshkg/上，也可以通过 a2z.jayesh@gmail.com 联系他。

Kostiantyn Kucher 出生在乌克兰敖德萨。2012 年他在敖德萨国立理工大学获得了计算机科学专业的硕士学位。他使用 Python、Matplotlib 和 PIL 从事机器学习和图像识别的工作。

目前，Kostiantyn 是一名计算机科学专业信息可视化方向的博士研究生。他在 Andreas Kerren 博导的指导下，在瑞典林奈大学计算机科学系的 ISOVIS 小组进行研究。

Kenneth Emeka Odoh 从事高级的数据可视化技术研究工作。他的研究兴趣是通过可

视的线索指导用户得出研究结果的探索性研究。

Kenneth 精通 Python 编程。2012 年他曾在芬兰的 Pycon 大会做演讲，主题是 Django 中的数据可视化。

他目前是加拿大里贾纳大学的一名研究员，通晓多种编程语言，有 C、C++、Python 和 Java 的应用开发经验。

编写代码之余，Kenneth 还参加了坎皮恩学院圣歌合唱团。

前言

最好的数据是我们能看到并理解的数据。作为一个开发人员，我们想创造并构建出最全面且容易理解的可视化图形。然而这并非总是很简单，我们需要找出数据，读取它、清理它、揣摩它，然后使用恰当的工具将其可视化。本书通过简单（和不那么简单）直接的方法解释了如何读取、清理和可视化数据的流程。

本书对怎样读取本地数据、远程数据、CSV、JSON 以及关系型数据库中的数据，都进行了讲解。

通过 matplotlib，我们能用一行简单的 Python 代码绘制出一些简单的图表，但是进行更高级的绘图还需要除 Python 之外的其他知识。我们需要理解信息理论和人类的审美学来生成最吸引人的可视化效果。

本书讲解在 Python 中使用 matplotlib 绘图的一些练习、使用情况，以及对于不同图表特性应该使用的方法的一些最佳实践。

本书的写作及代码开发均基于 Ubuntu 12.03，使用了 Python 2.7、IPython 0.13.2、virtualenv 1.9.1、matplotlib 1.2.1、NumPy 1.7.1 和 SciPy 0.11.0。

本书涵盖内容

第 1 章，准备工作环境，包括一些安装方法，以及如何在你的平台上安装所需的 Python 包和库的一些建议。

第 2 章，了解数据，介绍通用的数据格式，以及如何读写，如 CSV、JSON、XSL 或

者关系型数据库。

第 3 章，绘制图表及定制化，着手绘制简单的图表并介绍图表的定制化。

第 4 章，学习更多图表和定制化，继续上一章内容，介绍更多的高级表格和网格定制化。

第 5 章，3D 可视化，介绍三维数据的可视化，如 3D 柱状图、3D 直方图，以及 matplotlib 动画。

第 6 章，用图像和地图绘制图表，涵盖图像处理、在地图上投射数据，以及创建 CAPTCHA 测试图像。

第 7 章，使用正确的图表理解数据，涵盖一些更高级绘图技术的讲解和方法，如频谱图和相关性。

第 8 章，更多的 matplotlib 知识，介绍一些图表如甘特图、箱线图，并且介绍如何在 matplotlib 中使用 LaTeX 渲染文本。

准备工作

学习本书时，需要你在自己的操作系统上安装 Python2.7.3 或最新版本。本书使用 Ubuntu12.03 系统上的默认 Python 版本（2.7.3）。

本书中用到的另一个软件包是 IPython，它是一个交互式的 Python 环境，功能非常强大、灵活。可以通过基于 Linux 平台的包管理工具或者用于 Windows 和 Mac OS 系统的预安装文件安装。

一般来说，如果你对于 Python 安装和相关软件安装不熟悉，强烈推荐你使用预打包的 Python 科学发行包如 Anaconda、Enthought Python 发行包或者 Python（X,Y）进行安装。

其他所需的软件主要包括 Python 包，可全部通过 Python 安装管理器 pip 进行安装。pip 本身通过 Python 的 easy_install 安装工具安装。

谁适合阅读本书

本书是为那些通常已经了解 Python 编程的开发人员编写的。如果你听说过数据可视化但又不知道从何入手，本书会从头开始指导你了解数据、数据格式、数据可视化，以及如

何使用 Python 可视化数据。

你需要知道一些一般的编程概念，如果你有编程经验，会非常有用。然而，本书中的代码几乎是逐行讲解的。阅读本书不需要任何数学知识，书中介绍的每一个概念都有详细的讲解，并且提供了一些参考资料以供进一步的兴趣阅读。

约定

在本书中，不同的信息由一些不同风格的文字来区分。这里有一些文字风格的例子，以及它们的含义解释。

书中的代码文字显示如下："我们把小演示程序封装在 `DemoPIL` 类中，这样可以共享示例函数 `run_fixed_filters_demo` 的代码，并能很容易地对其进行扩展。"

代码块设置如下：

```
def _load_image(self, imfile):
    self.im = mplimage.imread(imfile)
```

当我们想要让你关注代码块中的某一特定部分时，相关的行或元素将设置为粗体：

```
# tidy up tick labels size
all_axes = plt.gcf().axes
for ax in all_axes:
    for ticklabel in ax.get_xticklabels() + ax.get_yticklabels():
        ticklabel.set_fontsize(10)
```

所有的命令行输入或者输出的写法如下。

```
$ sudo python setup.py install
```

新术语和**关键词**将显示为粗体。例如，在屏幕上、菜单或对话框中的文字将会显示为："然后我们为火柴杆图设置一个标签和基线位置，默认值为 **0**。"

警告或者重要的说明出现在这样的一个文本框中。

提示和技巧像这样显示。

读者反馈

欢迎读者向我们反馈意见。请让我们知道你对本书的看法——哪些是你喜欢或者不喜欢的。读者反馈对我们非常重要，可以帮我们完善一些你非常关心的内容。

如果给我们发送一般的反馈，可以简单地发电子邮件到 feedback@packtpub.com，并请在消息标题中提及书名。

如果你对某个话题有经验，并且有兴趣写作或者想为一本书做贡献，请参考我们的作者指南 www.packtpub.com/authors。

支持

既然你已经是 Packt 书籍的读者，我们有许多辅助材料可以帮助你从本书中得到最大的收获。

下载示例代码

可以在 http://www.packtpub.com 网站上你的账户中下载你所购买的所有图书的示例代码文件。如果你在其他地方购买本书，可以访问 http://www.packtpub.com/support 页面进行登记，文件会通过邮件直接发送给你。

勘误

尽管我们已经竭尽全力确保本书内容的准确，但是错误在所难免。如果你在书中发现了错误——可能是文本或者代码中的错误——如果你能把它报告给我们，我们将万分感谢。如此，这样可以减轻其他读者的痛苦，并且可以帮我们改进该书的后续版本。如果你找到任何错误，请访问 http://www.packtpub.com/submit-errata 并报告给我们，选择你的图书，点击 **errata submission form** 链接，并加入你勘误的详细内容。一旦你的勘误通过验证，你的提交将被接受，勘误将会上传到我们的网站，或者添加到位于该标题的勘误部分的已有勘误列表中。可以在 http://www.packtpub.com/support 上选择你的标题来查看所有现有的勘误信息。

著作权侵害

互联网上的版权侵害是一个跨越所有媒介的持续的问题。在 Packt，我们很认真地看待版权和许可保护。如果你不经意在互联网上得到了关于我们作品的任何形式的不合法的副

本，请及时给我们提供其地址或者网站名称，以便我们及时补救。

请通过 copyright@packtpub.com 联系我们，同时请提供涉嫌侵权材料的链接。

非常感激你帮助保护我们的作者，让我们尽力提供更有价值的内容。

问题

如果你对本书有任何疑问，可以通过 questions@packtpub.com 联系我们，我们会竭尽全力提供帮助。

目录