The Pragmatic Programmers

TURING 图灵程序设计丛书

Python 数据科学入门

快速掌握数据采集与清洗、数据分析、
机器学习等数据科学领域常见任务和工具，
用Python轻松解决数据科学问题

[俄] Dmitry Zinoviev 著
熊子源 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

Python 数据科学入门

Data Science Essentials in Python

[俄] Dmitry Zinoviev 著
熊子源 译

人民邮电出版社
北 京

图书在版编目（CIP）数据

Python数据科学入门 /（俄罗斯）德米特里·齐诺维耶夫（Dmitry Zinoviev）著；熊子源译. -- 北京：人民邮电出版社，2017.11
（图灵程序设计丛书）
ISBN 978-7-115-47060-7

Ⅰ. ①P… Ⅱ. ①德… ②熊… 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(2017)第253125号

内 容 提 要

本书以 Python 语言讲解数据科学基础知识，涵盖了数据采集、清洗、存储、检索、转换、可视化、高级数据分析（网络分析）、统计和机器学习等内容。具体内容包括：数据科学的 Python 核心特性，文本数据、数据库、表格形式的数值数据、series 和 frame、网络数据的使用，数据的绘制，概率与统计，机器学习。

本书面向研究生和本科生、数据科学教员、刚入门的数据科学专业人员，以及那些想拥有一本参考手册来帮助记住所有 Python 函数及参数的开发人员。

◆ 著 [俄] Dmitry Zinoviev
译 熊子源
责任编辑 岳新欣
执行编辑 吴威娜
责任印制 彭志环

◆ 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
北京鑫正大印刷有限公司印刷

◆ 开本：800×1000 1/16
印张：10
字数：237千字 2017年11月第1版
印数：1-4 000册 2017年11月北京第1次印刷
著作权合同登记号 图字：01-2017-6716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：京东工商广登字 20170147 号

站在巨人的肩上

Standing on Shoulders of Giants

iTuring.cn

站在巨人的肩上

Standing on Shoulders of Giants

iTuring.cn

版权声明

献给我集美丽与智慧于一身的妻子安娜。献给我们的孩子们：优雅的芭蕾舞演员尤金妮亚和浪漫的游戏玩家罗曼。也献给2015年夏天我的第一门数据科学课。

我现在必须给你一个小小的科学指引，来扰乱你的思路。

——英国小说家 Marie Corelli

前　言

2015年夏天，我在位于美国波士顿的萨福克大学使用Python教授数据科学入门课程，授课对象是一组经过选拔的本科生，本书的创作灵感正来源于这门课程。该课程是两个系列课程中的第一门课程，重点是数据的获取、清洗、组织和可视化，涉及统计学、机器学习和网络分析等相关内容。

数据的处理涉及庞大的体系和众多的Python模块（例如数据库、自然语言处理框架、JSON和HTML解析器，以及高性能数值数据结构，等等）。我很快意识到，不仅是本科生，甚至是经验丰富的专业人士，也很容易被这些浩瀚的知识所淹没。事实上，不得不承认，与我熟悉的领域相比，在进行数据科学和网络分析领域的研究时，我需要花更多时间去使用`help()`函数和浏览大量Python网络论坛。另外，我有时在课堂上会因为想不起某个函数名或可选参数而尴尬不已。

作为课程的一部分，我针对多类主题编辑了一套极具参考价值的备忘单。这些备忘单最终演变成了这本书。希望本书能够使你从大量函数名和可选参数中解脱出来，专注于数据科学和数据分析本身。

关于本书

本书涵盖了数据采集、清洗、存储、检索、转换、可视化、高级数据分析（网络分析）、统计和机器学习等内容。本书不是数据科学的综述或参考手册，不过你也能在第1章（“什么是数据科学”）找到如何开展数据科学的简要概述。阅读本书需要的先修知识包括数据科学的相关方法、统计学等。

第2章总结了Python数据结构，字符串、文件和与Web相关的函数，正则表达式，以及列表推导式。总结并非用于讲授这些知识，而是供你温习相关知识点。掌握Python对于一个成功的数据科学家而言无疑是非常重要的，你可以找到许多优秀的图书，进一步学习这门语言。

本书的第一部分介绍了如何使用不同类型的文本数据，包括处理结构化和非结构化的文本，使用NumPy和Pandas模块处理数值数据，以及网络分析。还有三章涉及数据分析的三个方面：使

用关系型和非关系型数据库、数据可视化以及简单的预测分析。

本书是一本半叙述半参考性的书。你可以直接按顺序阅读，也可以先找出你关心的函数或概念，然后查阅相关的说明和示例。若是按顺序阅读，而你又有一定的Python编程经验，就可以直接跳过第2章（“数据科学的Python核心”）。如果你不打算使用外部数据库（比如MySQL），也可以忽略第4章（“使用数据库”）。最后，如果你对统计学已经有一定了解，那么完全可以跳过第9章（“概率与统计”）的前两个单元，直接阅读第47单元（“以Python的方式完成统计”）。

关于读者

你可以在此了解自己是否需要本书。

本书面向研究生和本科生、数据科学教员、刚入门的数据科学专业人员（特别是从R语言转为使用Python的人），以及那些想拥有一本参考手册来帮助记住所有Python函数及参数的开发人员。

如果你是他们中的一员，那就不要犹豫，直接开始阅读吧。

关于软件

尽管在是否要从Python 2.7转到Python 3.3或者更高版本这个问题上，目前尚存在一定争论，但我还是支持使用新版本的Python。许多新的Python软件是针对Python 3.3开发的，而且多数遗留软件都已经成功移植到Python 3.3。考虑到这种趋势，选择将过时的Python 2.7恐怕并非明智之举，即使它现在还很流行[①]。

本书所有Python示例需要用到的模块都列在了下表中。

表1 本书使用的软件组件

包	使用的版本	包	使用的版本
BeautifulSoup	4.3.2	community	0.3
json	2.0.9	html5lib	0.999
matplotlib	1.4.3	networkx	1.10.0
nltk	3.1.0	numpy	1.10.1
pandas	0.17.0	pymongo	3.0.2
pymysql	0.6.2	python	3.4.3
scikit-learn	0.16.1	scipy	0.16.0

① Python 2.7是2.x系列的最后一个版本，支持时间到2020年。——译者注

在这些模块中，除了需要单独安装的community版模块[①]和Python解释器外，其他都已经包含在Anaconda发行版中。Anaconda发行版是由Continuum Analytics公司开发的免费软件[②]。

如果你想试用一下数据库（或者你的工作就要用到数据库），那么你还需要下载并安装MySQL[③]和MongoDB[④]这两个数据库。它们都是免费的，能运行在Linux、Mac OS和Windows平台上。

关于引号

Python允许用户使用以下方式表示字符串：'单引号'、"双引号"、'''三个单引号'''，甚至是"""三个双引号"""（其中后两个可表示多行字符串）。然而，不论程序中使用哪种引号，当打印字符串时，通常都使用单引号。

许多其他语言（C、C++、Java）会在不同场合使用单引号和双引号：单引号用于单个字符，双引号用于字符串。本书沿用这种区分方式——对单个字符使用单引号，对字符串使用双引号。

关于本书的论坛

本书的社区论坛可在Pragmatic Programmers网站上找到[⑤]。你可以在论坛中提问、发表评论和提交勘误。

另一个很好的问答资源（不限于本书）是在Stack Exchange网站上新创建的数据科学板块[⑥]。

轮到你了

每章最后都有一个叫作“轮到你了”的单元。这个单元描述了几个项目，你可以自己独立（或与你信任的人一起）完成这些项目，以加强对本书内容的理解。

单个星号（*）标记的项目是最简单的。完成这些项目只需要前面章节中提到的函数方面的知识。预计不超过三十分钟就可以完成“单星项目”。你可以在附录2（“单星项目的解决方案”）中找到参考的解决方法。

两个星号（**）标记的是较难的项目。完成这些项目可能需要一小时甚至更长的时间，当然这也取决于你的编程技能和习惯。“两星项目”需要使用某些中级数据结构和周密的算法。

---

① pypi.python.org/pypi/python-louvain/0.3
② www.continuum.io
③ www.mysql.com
④ www.mongodb.com
⑤ pragprog.com/book/dzpyds
⑥ datascience.stackexchange.com

最后，标记为三个星号（***）的项目是最难的。一些“三星项目”甚至可能没有一个完美的解决方案，因此如果你解不出来也不必沮丧。不过，通过练习“三星项目”，你一定可以成为更出色的程序员和更优秀的数据科学家。而且如果你是教育工作者，可以考虑将“三星项目”作为期中作业。

现在，让我们开始吧！

Dmitry Zinoviev
dzinoviev@gmail.com
2016年8月

致　　谢

衷心感谢Xinxin Jiang教授（萨福克大学）对本书统计部分所提出的宝贵意见。同时感谢Jason Montojo（*Practical Programming: An Introduction to Computer Science Using Python 3*的作者之一）、Amirali Sanatinia（东北大学）、Peter Hampton（阿尔斯特大学）、Anuja Kelkar（卡内基梅隆大学）和Lokesh Kumar Makani（Skyhigh Networks公司），本书的问世离不开他们的宝贵意见。

目　录

学无止境。

——俄罗斯作家Kozma Prutkov

第1章 什么是数据科学

相信你对数据科学已经有了一些了解，不过我们还是可以一起来回顾一下！数据科学是从数据中提取知识的学科。它依赖于计算机科学（数据结构、算法、可视化、大数据支持和通用编程）、统计学（回归和推理），以及领域知识（用于提问和解释结果）。

传统意义上的数据科学涵盖多种不同主题，有些是你可能已经熟悉的，而有些是你将在本书中遇到的。

- **数据库**，提供信息的存储和集成。关于关系型数据库和文档存储的信息请参见第4章。
- **文本分析**和**自然语言处理**，让我们可以通过将定性文本转化成定量变量，实现“用文字计算”。你是否对情感分析工具感兴趣？那么阅读本书的第16单元再合适不过了。
- **数值数据分析**和**数据挖掘**，可搜索出变量之间的不变性和相互关系。这是第5章和第6章的主题。
- **复杂网络分析**，其实并不复杂。所谓复杂网络，是指任意互连实体的集合。第7章介绍了如何将复杂网络分析简单化。
- **数据可视化**，不仅富有美感，而且非常实用，尤其是当你想说服数据赞助商再次提供赞助的时候。如果说一图胜千言，那么第8章的价值就远超过本书的其他部分。
- **机器学习**（包括聚类、决策树、分类和神经网络），试图让计算机学会“思考”，并根据样本数据进行预测。第10章对如何实现这样的功能进行了说明。
- **时间序列处理**，或者更一般地说，**数字信号处理**，是股市分析师、经济学家以及音频和视频领域的研究人员不可或缺的工具。
- **大数据分析**，通常指对频繁生成和获取的大于1TB的非结构化数据（文本、音频、视频）进行分析。大数据如此之大，以至于难以在本书中进行完整的介绍。

不论针对哪种分析类型，数据科学首先是科学，然后才是魔法。因此，它是一个严格遵循以数据采集为起点、以结果报告为终点的基本处理过程。在本章中，你将了解数据科学的基本过程，

包括：常见数据分析研究的步骤、数据的获取来源，以及常见项目报告的结构。

第1单元

数据分析步骤

常见的数据分析研究步骤通常与一般的科学发现顺序一致。

数据科学发现从要解决的问题和要应用的分析方式开始。最简单的分析类型是**描述性**的，通常使用一种可视化的形式给出数据集总量的描述。不论你接下来打算做什么，至少需要描述一下数据！在**探索性**数据分析的过程中，你需要尝试找出现有变量之间的相互关系。基于统计的**推断分析**可以帮助你利用手头上少量的数据样本，对更大的群体进行描述。**预测分析**是从过去的规律中预测未来。**因果分析**能找出相互影响的变量。最后，**机制性**数据分析准确揭示了一个变量如何影响另一个变量。

然而，分析结果的好坏依赖于数据的质量，因此引出了如下问题：什么样的数据集是理想的呢？在理想情况下，什么样的数据能够解决问题呢？另外，理想的数据集可能根本就不存在，或者是很难甚至不可能获取。对于这种情况，一个较小的或者特征不那么丰富的数据集还依然有用吗？

幸运的是，从Web或数据库获取原始数据并不难，有大量基于Python的工具可用于下载和解析这些数据。你可以在第2单元（“数据获取途径”）中进一步了解这些工具。

应该注意到，完美的数据是不存在的。难免会遇到有丢失值、异常值和其他“非标准”项的“脏”数据。“脏”数据的例子包括：未来的出生日期、负年龄和负体重，以及不合理的电子邮件地址（noreply@）。因此，一旦获得了原始数据，接下来就是使用数据清洗工具和统计知识来正则化数据集。

完成上述处理后，就可以使用干净的数据，开展描述性和探索性分析。这一步的成果通常包括散点图（参考第44单元）、直方图和统计总结（参考第46单元）。它们赋予了你对数据独有的感觉——这是一种在后续研究中不可或缺的对数据集（尤其是针对多维数据集）的直观认识。

现在离实现预测只有一步之遥了。你手中的数据模型工具，在经过恰当的训练后，就可以实现预测功能。值得注意的是，不能忽视对模型的质量及其预测精度的评估！

至此，你可以摘掉统计学家和程序员的帽子，换上一顶领域专家的帽子了。你已经得到了一些成果，但它们称得上是领域内的重要成果吗？换句话说，是否有人关心这些成果，还有，这些成果带来了什么不一样的认知？设想一下，你被聘用为一名评论员，来评价自己的工作。你做的哪些是正确的，哪些是错误的？如果再给你一次机会，哪些工作你能做得更好或者不同？你是否

会使用不同的数据，作出不同类型的分析，提出不同的问题，抑或建立一个不同的模型？一定有人会提出这些问题。提前进行思考，对你是大有裨益的。当你还沉浸在这些字里行间时，寻觅答案的征程已然开始。

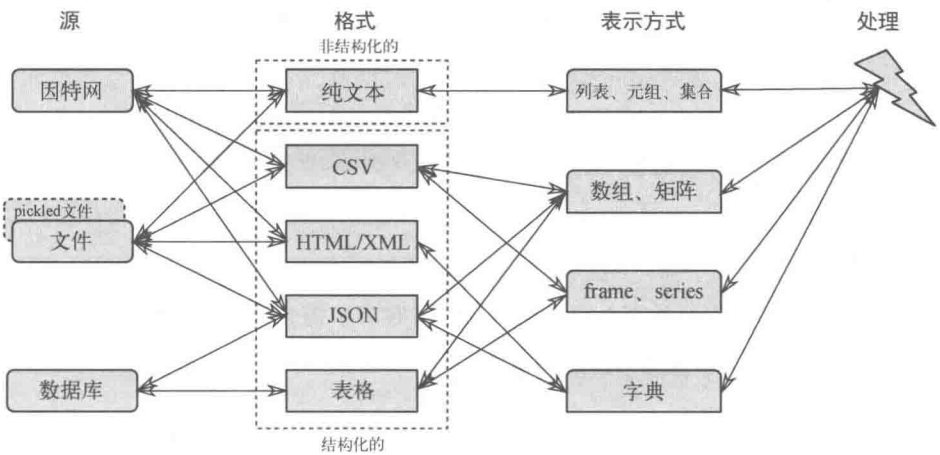
最后，你必须完成一个报告，说明你处理数据的方式及理由、建立了什么模型、可能得出什么结论、可能作出什么预测。本章末尾（第3单元）讲解了报告的结构。

作为一本数据科学领域的Python手册，本书的重点是典型数据分析步骤中早期的、最随意，同时也是最有创意的部分：数据的获取、清洗、组织和分级。本书几乎不涉及数据建模的内容，包括预测数据的建模。（当然，完全抛开数据建模是不合理的，毕竟这是魔法的真正所在！）一般来说，结果解释、质疑和报告非常依赖于特定的领域，这些内容可在专门的教材中找到。

第 2 单元

数据获取途径

数据获取涉及获得包含来自各种输入器件的数据源、从器件中提取数据，以及将其转换为适于进一步处理的表示方式，如下图所示。



数据的三个主要来源是因特网（即万维网）、数据库，以及本地文件（可能是先前手动下载或利用其他软件下载得到的）。某些本地文件可能是通过Python程序生成的，包括序列化的或“pickled”的数据（更详细的解释请参考第12单元）。

来自器件的数据格式多种多样。在后续章节中，你将接触到最流行的数据格式及其对应的数据分析方式和方法。