



Python数据分析入门

从数据获取到可视化

沈祥壮 著

应用Python 轻松实现数据分析和数据处理

作者简介



沈祥壮，自学Python两年，以数据分析为主线，系统学习了数据的采集、处理、分析和可视化。在研究统计机器学习理论的同时，使用Python语言实现了部分统计学习算法。研究方向包括数据采集、数据挖掘、统计机器学习及图像处理。

内容简介

本书作为数据分析的入门图书，以Python语言为基础，介绍了数据分析的整个流程。本书内容涵盖数据的获取(即网络爬虫程序的设计)、前期数据的清洗和处理、运用机器学习算法进行建模分析，以及使用可视化的方法展示数据及结果。首先，书中不会涉及过于高级的语法，不过还是希望读者有一定的语法基础，这样可以更好地理解本书的内容。其次，本书重点在于应用Python来完成一些数据分析和数据处理的工作，即如何使用Python来完成工作而非专注于Python语言语法等原理的讲解。本书的目的是让初学者不论对数据分析流程本身还是Python语言，都能有一个十分直观的感受，为以后的深入学习打下基础。最后，读者不必须按顺序通读本书，因为各个章节层次比较分明，可以根据兴趣或者需要来自行安排。例如第5章介绍了一些实战的小项目，有趣且难度不大，大家可以在学习前面内容之余来阅读这部分内容。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目(CIP)数据

Python数据分析入门：从数据获取到可视化 / 沈祥壮著. —北京：电子工业出版社，2018.3

ISBN 978-7-121-33653-9

I. ①P... II. ①沈... 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024600号

策划编辑：石倩

责任编辑：牛勇

印刷：

装订：

出版发行：电子工业出版社

北京市海淀区万寿路173信箱 邮编100036

开本：720×1000 1/16 印张：16.5 字数：290千字

版次：2018年3月第1版

印次：2018年3月第1次印刷

印数：2500册 定价：59.00元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010)88254888，88258888。

质量投诉请发邮件至zlts@phei.com.cn，盗版侵权举报请发邮件至dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

前言

Python作为一门优秀的编程语言，近年来受到很多编程爱好者的青睐。一是因为**Python**本身具有简捷优美、易学易用的特点；二是由于互联网的飞速发展，我们正迎来大数据的时代，而 **Python** 无论是在数据的采集与处理方面，还是在数据分析与可视化方面都有独特的优势。我们可以利用 **Python** 便捷地开展与数据相关的项目，以很低的学习成本快速完成项目的研究。本书本着实用性的目的，着眼于整个数据分析的流程，介绍了从数据采集到可视化的大致流程。希望借此为**Python**初学者打开数据分析领域的大门，初窥数据分析的奥秘。

本书的主要内容

第**1**章主要讲解了在**Ubuntu**和**Windows**系统下，**Python**集成开发环境的搭建。考虑到初学者容易为安装第三方库犯难，又介绍了三种简单实用的方法来安装这些常见的库。接着对几个后面要用到的高级语法进行了简单介绍，为之后的应用打下基础。

第**2**章集中讲解了数据采集的流程，即网络爬虫程序的设计与实现。首先本章没有拘泥于使用**Python**的内置库**urllib**库进行实现，而是直接介绍了**requests**和其他更加简捷强大的库来完成程序的设计。在进阶内容中，对常见的编码问题、异常处理、代理**IP**、验证码、机器人协议、模拟登录，以及多线程等相关问题给出了解决的方案。

第**3**章讲解数据的清洗问题。在具体讲解清洗数据之前，先介绍了**TXT**、**XLSX**、**JSON**、**CSV**等各种文件的导入和导出的方法，并介绍了**Python**与**MySQL**数据库交互的方式。接着介绍了**NumPy**和**pandas**库的基本使用方法，这是我们用于数据处理的科学计算的两个强大的工具。最后综合以上的学习介绍了数据的去重、缺失值的填补等经典的数据清洗方法。

第**4**章首先讲解探索性数据分析的应用，并且简单介绍了机器学习基本知识。然后演示如何应用 **sklearn** 库提供的决策树和最邻近算法来处理分类问题，并尝试根据算法原理手动实现最邻近算法。最后介绍如何使用 **pandas**、**matplotlib** 和**seaborn**这三个库来实现数据的可视化。

第5章是综合性学习的章节，讲解了三个小项目的完整实现过程，旨在通过操作生活中真正的数据来强化前面基础内容的学习。

本书的读者对象

本书面向想从事数据工作的**Python**初学者。由于本书并不对**Python**的基础语法做详细的讲解，所以希望读者有一定的语法基础。

测试环境及代码

我们使用的语法是基于**Python 3**的，具体是**Python 3.6**，用到的第三方库也已经全面支持此版本，所以读者不必担心相关的版本问题；测试环境为 **Ubuntu 16.04 LTS 64-Bit**。本书中使用的全部代码及相关数据已经托管至**Github**，

读者可以进入 <https://github.com/shenxiangzhuang/PythonDataAnalysis> 进行下载。

联系作者

虽然本书只是入门级图书，但是限于笔者水平有限，难免会存在一些错误，有些地方的表述可能也不是那么准确。非常欢迎读者指出本书的不当之处或提出建设性的意见。笔者的电子邮件地址是 datahonor@gmail.com。



致谢

在本书的撰写过程中受到过很多人的帮助，这里特别感谢刘松学长，感谢学长对笔者本人长久以来的帮助，从他那里我学到了很多关于**Python** 语言、机器学习以及计算机视觉等相关知识。另外，特别感谢**IT** 工作者谢满锐先生对本书的细心审校，也感谢他为本书的进一步修改提出建设性意见。同时，感谢电子工业出版社石倩、杨嘉媛编辑的帮助。最后，本书参阅了大量的国内外的文献，这里对有关作者表示衷心的感谢。

读者服务

轻松注册成为博文视点社区用户（www.broadview.com.cn），扫码直达本书页面。

- 提交勘误：您对书中内容的修改意见可在 [提交勘误](#) 处提交，若被采纳，将获赠博文视点社区积分（在您购买电子书时，积分可用来抵扣相应金额）。

- 交流互动：在页面下方 [读者评论](#) 处留下您的疑问或观点，与我们和其他读者一同学习交流。

页面入口：<http://www.broadview.com.cn/33653>

目录

[作者简介](#)

[前言](#)

[1 准备](#)

[1.1 开发环境搭建](#)

[1.1.1 在Ubuntu系统下搭建Python集成开发环境](#)

[1.1.2 在Windows系统下搭建Python集成开发环境](#)

[1.1.3 三种安装第三方库的方法](#)

[1.2 Python基础语法介绍](#)

[1.2.1 `if name == 'main'`](#)

[1.2.2 列表解析式](#)

[1.2.3 装饰器](#)

[1.2.4 递归函数](#)

[1.2.5 面向对象](#)

[1.3 The Zen of Python](#)

[参考文献](#)

[2 数据的获取](#)

[2.1 爬虫简介](#)

[2.2 数据抓取实践](#)

[2.2.1 请求网页数据](#)

[2.2.2 网页解析](#)

[2.2.3 数据的存储](#)

[2.3 爬虫进阶](#)

[2.3.1 异常处理](#)

[2.3.2 robots.txt](#)

[2.3.3 动态UA](#)

[2.3.4 代理IP](#)

[2.3.5 编码检测](#)

[2.3.6 正则表达式入门](#)

[2.3.7 模拟登录](#)

[2.3.8 验证码问题](#)

[2.3.9 动态加载内容的获取](#)

[2.4 爬虫总结](#)

[参考文献](#)

[3 数据的存取与清洗](#)

[3.1 数据存取](#)

[3.1.1 基本文件操作](#)

[3.1.2 CSV文件的存取](#)

[3.1.3 JSON文件的存取](#)

[3.1.4 XLSX文件的存取](#)

[3.1.5 MySQL数据库文件的存取](#)

[3.2 NumPy](#)

[3.2.1 NumPy简介](#)

[3.2.2 NumPy基本操作](#)

[3.3 pandas](#)

[3.3.1 pandas简介](#)

[3.3.2 Series与DataFrame的使用](#)

[3.3.3 布尔值数组与函数应用](#)

[3.4 数据的清洗](#)

[3.4.1 编码问题](#)

[3.4.2 缺失值的检测与处理](#)

[3.4.3 去除异常值](#)

[3.4.4 去除重复值与冗余信息](#)

[3.4.5 注意事项](#)

[参考文献](#)

[4 数据的分析及可视化](#)

[4.1 探索性数据分析](#)

[4.1.1 基本流程](#)

[4.1.2 数据降维](#)

[4.2 机器学习入门](#)

[4.2.1 机器学习简介](#)

[4.2.2 决策树——机器学习算法的应用](#)

[4.3 手动实现KNN算法](#)

[4.3.1 特例——最邻近分类器](#)

[4.3.2 KNN算法的完整实现](#)

[4.4 数据可视化](#)

[4.4.1 高质量作图工具——matplotlib](#)

[4.4.2 快速作图工具——pandas与matplotlib](#)

[4.4.3 简捷作图工具——seaborn与matplotlib](#)

[4.4.4 词云图](#)

[参考文献](#)

[5 Python与生活](#)

[5.1 定制一个新闻提醒服务](#)

[5.1.1 新闻数据的抓取](#)

[5.1.2 实现邮件发送功能](#)

[5.1.3 定时执行及本地日志记录](#)

[5.2 Python与数学](#)

[5.2.1 估计 \$\pi\$ 值](#)

[5.2.2 三门问题](#)

[5.2.3 解决LP与QP问题（选读）](#)

[5.3 QQ群聊天记录数据分析](#)

欢迎访问：电子书学习和下载网站 (<https://www.shgis.com>)

文档名称：《Python数据分析入门：从数据获取到可视化》沈祥壮 著.pdf

请登录 <https://shgis.com/post/4046.html> 下载完整文档。

手机端请扫码查看：

