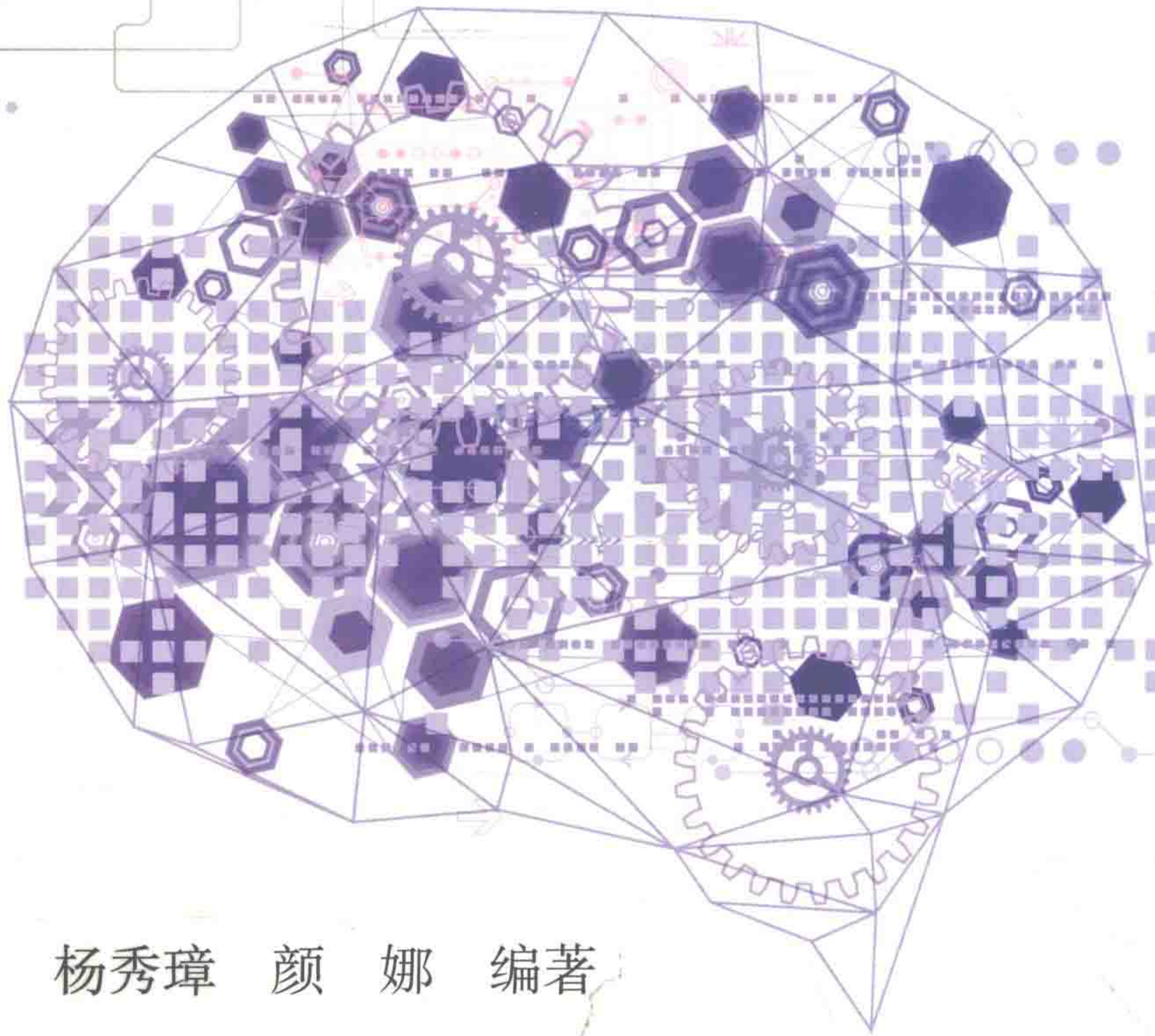
Python

网络数据爬取及分析

从入门到精通

（分析篇）



杨秀璋　颜　娜　编著

北京航空航天大学出版社
BEIHANG UNIVERSITY PRESS

Python 网络数据爬取及分析从入门到精通(分析篇)

杨秀璋　颜　娜　编著

北京航空航天大学出版社

图书在版编目(CIP)数据

Python 网络数据爬取及分析从入门到精通. 分析篇 / 杨秀璋，颜娜编著. -- 北京 ：北京航空航天大学出版社，2018.5

ISBN 978-7-5124-2713-6

Ⅰ. ①P… Ⅱ. ①杨… ②颜… 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 101752 号

Python 网络数据爬取及分析从入门到精通(分析篇)

杨秀璋　颜　娜　编著

责任编辑　孙兴芳

*

北京航空航天大学出版社出版发行

北京市海淀区学院路 37 号(邮编 100191)　http://www.buaapress.com.cn

发行部电话：(010)82317024　传真：(010)82328026

读者信箱：emsbook@buaacm.com.cn　邮购电话：(010)82316936

涿州市新华印刷有限公司印装　各地书店经销

*

开本：710×1 000　1/16　印张：16.75　字数：357 千字

2018 年 6 月第 1 版　2018 年 6 月第 1 次印刷

ISBN 978-7-5124-2713-6　定价：59.80 元

序 一

作为与秀璋同窗同寝的10年老友，有幸见证秀璋与颜娜相识相知相爱。此书可以说是他们爱的结晶。秀璋是深受朋友信任的好兄弟，亦是深受学生爱戴的好老师，似乎有着用不完的热情，这种热情，带给我们这个社会一丝丝的温暖，在人与人之间传递着。当初在博客上不断写文章，并耐心解答网友们的各种问题，还帮助许多网友学习编程，指导他们的作业甚至毕业论文，所以，“当教师”这颗种子早已埋下。毕业后的秀璋，拿着同学们羡慕的北京IT行业某网络公司的录取通知书，却毅然决然踏上返乡的路，这一走，走进了大山里的贵州，成了一名受人尊敬的人民教师。生活平淡而辛苦，而乐观的秀璋却收获了爱情，此也命也。

拒绝了无数聚会的邀请，见证了无数贵阳凌晨的灯火，秀璋和颜娜孜孜不倦写下这本书，作为朋友，着实替他们高兴。作为见证这本书从下笔到问世的读者，作为一个Python爱好者及有一定数据分析功底的学生，读这本书真是如晤老友——有大量的网络数据分析实例，从Python常用数据分析库到可视化分析，再到回归分析、聚类分析、分类分析、关联规则、文本预处理，并普及了词云热点与主题分布分析、复杂网络和基于数据库的分析。

本书配以专业但不晦涩的语言，将原本枯燥的学术知识娓娓道来，此时的秀璋不是老师，而是一个熟悉的老友，用大家听得懂的话，解释着您需要了解的一切。同时，当您学习完Python网络数据分析之后，还推荐您继续学习本套书中的另一本书——《Python网络数据爬取及分析从入门到精通(爬取篇)》，进而更好地掌握与Python相关的知识。

总之，再多赞美的语言，都比不上滴滴汗水凝结的成功带来的满足与喜悦。愿您合上书时，亦能感受到秀璋和颜娜的真诚。

大疆公司　宋籍文

2017年11月1日于深圳

序　二

当我被秀璋邀请为本书写序时，我首先感到的是惊讶和荣幸。秀璋是我最好的朋友之一，在本科和硕士学习期间，我们一起在北京理工大学度过了六年的美好时光。秀璋是一个真诚而严谨的人，在学习、工作，甚至游戏中，他都力争完美，很开心看到他完成了这本著作。

在大学期间，每个人都知道他有当老师的梦想，之后他也确实回到了家乡贵州，做着他喜欢的事情。我希望他能在教育领域保持着那份激情和初心，即使这是一个漫长而艰难的过程，但我相信他会用他的热情和爱意克服一切困难，教书育人。

这本书就像他的一个“孩子”，他花了很多时间和精力撰写而成。它是一本关于Python 数据分析的技术书，包括很多有用的实例，比如利用线性回归预测价格、利用K-Means 聚类分析篮球运动员、利用决策树分析鸢尾花、利用词云技术和主题模型分析文本数据等。现在我们都知道一些与计算机科学相关的热门术语，如机器学习、大数据、人工智能等。而许多像 SAP 这样的公司也在关注这些新兴的技术，从海量信息中挖掘出有价值的信息，以便将来为客户提供更好的软件解决方案和服务，关注为公司决策提供支撑。

但我们从哪里开始学习这些新知识呢？我想您可以从读这本书开始。在本书中，秀璋介绍了各种常见的数据分析方法及实例，通过这些方法我们能够构建自己感兴趣的应用或研究领域，例如舆情分析、价格预测、商品推荐、社区发现等。本书既可以当作 Python 数据分析的入门教程，也可以当作指导手册或科普书。对于初学者来说，学习本书中的内容并不难，它就是一步步的教程，包括基本的 Python 数据分析常用库、可视化绘图、回归分析、聚类分析、分类分析、主题分布、复杂网络、数据预处理等。书中有许多生动而有趣的案例，以及详细的图形指南和代码注释，绝不会让您感到无聊。

本书是学习 Python 数据分析的不二选择。同时推荐您继续学习本套书中的另一本书——《Python 网络数据爬取及分析从入门到精通（爬取篇）》，让您的数据分析研究更加自如，挖掘自己感兴趣的数据集之后再进行深入分析。

如果您真的是 Python、网络爬虫、数据分析或大数据的忠实粉丝，请不要犹豫，学习 Python 就从这本书开始吧！如果您对人工智能、机器学习、深度学习感兴趣，也请把这本书当作您的入门教程吧！

SAP 工程师　数字商务服务　徐溥

2017 年 11 月 23 日于美国

序　三

杨老师是我认识的人里最忠于自己内心的人。在青春年少时他便抱定自己的理想，多年来一直不忘初心、心无旁骛地朝着目标踽踽前行，既仰望星空，又脚踏实地，直到达成所愿。

相较于大多数与梦想渐行渐远的人们，他是幸运的，这幸运离不开他多年的努力与坚持。年少时，他可能从未想过自己会成为一名“程序猿”，误打误撞进入编程领域，从此在代码的世界里乐此不疲，越走越远。对于他而言，重要的是学有所成，继承父亲遗志，做一名传道授业解惑的教师。为此，他勤奋学习，纵然辛劳却乐在其中；他乐于助人，以帮助、辅导他人学习技术为傲，从不求回报；他常有危机感，担心自己学得还不够，不足以为人传道授业解惑；他也常常感叹，为自己能在普及编程知识上做一点贡献而感到自豪。这些，成为他五年来坚持在 CSDN 更新博客的坚强动力，也是他在北京航空航天大学出版社多番邀请下，终于下定决心要倾自己所学写一套书的初衷。

因为工作调整的缘故，2017 年杨老师异常忙碌，加班是家常便饭，写这套书几乎占据了他全部的休息时间。很多个安静的夜里，家人酣睡，他却敲击着键盘，灵感如火花四溅，脑海里的知识渐渐凝聚成书。

8 年编程积累，近 300 篇博文厚积薄发，Python 系列专栏荣获“2017 年 CSDN 博客十大专栏”，得到网友们的充分肯定。历时一年倾囊而出、潜心创作，本套书凝聚了他诸多心血，同时也是他学习 Python 语言的阶段性总结。本套书简单易懂，包含了网络数据爬取和数据分析两方面知识。杨老师充分考虑初学者可能会遇到的困难和问题，深入浅出，理论结合案例，力求让每位读者在合上书后，都能真正学有所得，熟练掌握 Python 语言、网络爬虫和数据分析。同时，因为力求丰富完善，内容较多，故本套书分为两本出版，一本即为本书，重点涵盖了可视化分析、回归分析、聚类分析、分类分析、关联规则挖掘、数据处理、主题分布、复杂网络等技术，并且每一章节都通过实例代码和图表步骤进行详细讲解。另一本为《Python 网络数据爬取及分析从入门到精通(爬取篇)》，主要介绍 Python 网络数据爬取。建议大家将两本书结合起来学习。

杨老师是一个善良、纯粹而又执着的人，日常交往中人们很容易在他身上建立起信任感，他对得失的毫不计较，对教育事业的虔诚，对他人的真挚友善，对知识的尊重与渴求，无不深深打动着身边的人。程序员有很多种，他可能并不是技术最厉害的，但他选择了一条更为艰难的路，学习积累，潜心创作，教书育人，用一篇篇文章、一个个精彩的案例去帮助更多人。

作为长期陪伴他左右的人,我敬他、恋他,同时从心底深深感激他为我倾注的一切。历经一年,与他一起查阅资料、一起校稿、一起默默付出,整套书终于要问世了。作为整套书的第一个读者,我深深地知道他对整套书所倾注的炽热情感与心血,每一段文字、每一行代码都闪现着我们生活和工作中合作的点点滴滴,希望您在阅读过程中,也能体会到我们满满的诚意。

此生幸事莫过于得一知己共白首! 也希望所有的读者能包容本书的不足之处,如果此书能激发您对数据挖掘与分析的兴趣,给您的学习和工作带来些灵感和帮助,我们将不胜欢喜。编程路漫漫,期待与各位读者的交流与学习,共同进步。

颜　娜

2018 年 3 月 14 日于贵阳

前　言

随着数据分析和人工智能风暴的来临，Python 也变得越来越火热。它就像一把利剑，使我们能随心所欲地去做各种分析与研究。在研究机器学习、深度学习与人工智能之前，我们有必要静下心来学习一下 Python 的基础知识、基于 Python 的网络数据爬取及分析，这些知识点都将为我们后续的开发和研究打下扎实的基础。同时，由于市面上缺少以实例为驱动，全面详细介绍 Python 网络爬虫及数据分析的书，本套书很好地填补了这一空白，它通过 Python 语言来教读者编写网络爬虫并教大家针对不同的数据集做算法分析。本套书既可以作为 Python 数据爬取及分析的入门教材，也可以作为实战指南，其中包括多个经典案例。

它究竟是一套什么样的书呢？对您学习网络数据爬取及分析是否有帮助呢？

本套书是以实例为主、使用 Python 语言讲解网络数据爬虫及分析的书和实战指南。本套书结合图表、代码、示例，采用通俗易懂的语言，介绍了 Python 基础知识、数据爬取、数据分析、数据预处理、数据可视化、数据库存储、算法评估等多方面知识，每一部分知识都从安装过程、导入扩展库到算法原理、基础语法，再结合实例详细讲解。本套书适合计算机科学、软件工程、信息技术、统计数学、数据科学、数据挖掘、大数据等专业的学生学习，也适合对网络数据爬取、数据分析、文本挖掘、统计分析等领域感兴趣的读者阅读，同时也可作为数据挖掘、数据分析、数据爬取、机器学习、大数据等技术相关课程的教材或实验指南。

本套书分为两篇——爬取篇和分析篇。其中，爬取篇详细讲解了正则表达式、BeautifulSoup、Selenium、Scrapy、数据库存储相关的爬虫知识，并通过实例让读者真正学会如何分析网站、抓取自己所需的数据；分析篇详细讲解了 Python 数据分析常用库、可视化分析、回归分析、聚类分析、分类分析、关联规则挖掘、文本预处理、词云分析及主题模型、复杂网络和基于数据库的分析。爬取篇突出爬取，分析篇侧重分析，为了更好地掌握相关知识，建议读者将两本书结合起来学习。

为什么本套书会选择 Python 作为数据爬取和数据分析的编程语言呢？

随着大数据、数据分析、深度学习、人工智能的迅速发展，网络数据爬取和网络数据分析也变得越来越热门。由于 Python 具有语法清晰、代码友好、易读易学等特点，同时拥有强大的第三方库支持，包括网络爬取、信息传输、数据分析、绘图可视化、机器学习等库函数，所以本套书选择 Python 作为数据爬取和数据分析的编程语言。

首先,Python 既是一种解释性编程语言,又是一种面向对象的语言,其操作性和可移植性较高,因而被广泛应用于数据挖掘、文本爬取、人工智能等领域。就作者看来,Python 最大的优势在于效率。有时程序员或科研工作者的工作效率比机器的效率更为重要,对于很多复杂的功能,使用较清晰的语言能给程序员减轻更多的负担,从而大大提高代码质量,提高工作效率。虽然 Python 底层运行速度要比 C 语言慢,但 Python 清晰的结构能节省程序员的时间,简单易学的特点也降低了编程爱好者的门槛,所以说“人生苦短,我学 Python”。

其次,Python 可以应用在网络爬虫、数据分析、人工智能、机器学习、Web 开发、金融预测、自动化测试等多个领域,并且都有非常优秀的表现,从来没有一种编程语言可以像 Python 这样同时扎根在这么多领域。另外,Python 还支持跨平台操作,支持开源,拥有丰富的第三方库。尤其随着人工智能的持续火热,Python 在 IEEE 发布的 2017 年最热门语言中排名第一,同时许多程序爱好者、科技工作者也都开始认识 Python,使用 Python。

接下来作者将 Python 和其他常用编程语言进行简单对比,以突出其优势。相比于C#,Python 是一种跨平台的、支持开源的解释型语言,可以运行在 Windows、Linux 等平台上;而 C# 则相反,其平台受限,不支持开源,并且需要编译。相比于 Java,Python 更简洁,学习难度也相对低很多,而 Java 则过于庞大复杂。相比于 C 和 C++,Python 的语法简单易懂,代码清晰,是一种脚本语言,使用起来更为灵活;而 C 和 C++通常要和底层硬件打交道,语法也比较晦涩难懂。

目前,Python 3.x 版本已经发布并正在普及,本套书却选择了 Python 2.7 版本,并贯穿整套书的所有代码,这又是为什么呢?

在 Python 发布的版本中,Python 2.7 是比较经典的一个版本,其兼容性较高,各方面的资料和文章也比较完善。该版本适用于多种信息爬取库,如 Selenium、BeautifulSoup 等,也适用于各种数据分析库,如 Sklearn、Matplotlib 等,所以本套书选择 Python 2.7 版本;同时结合官方的 Python 解释器和 Anaconda 集成软件进行详细介绍,也希望读者喜欢。Python 3.x 版本已经发布,具有一些更便捷的地方,但大部分功能和语法都与 Python 2.7 是一致的,作者推荐大家结合 Python 3.x 进行学习,并可以尝试将本套书中的代码修改为 Python 3.x 版本,以加深印象。

同时,作者针对不同类型的读者给出一些关于如何阅读和使用本套书的建议。

如果您是一名没有任何编程基础或数据分析经验的读者,建议您在阅读本套书时,先了解对应章节的相关基础知识,并手动敲写每章节对应的代码进行学习;虽然本套书是循序渐进深入讲解的,但是为了您更好地学习数据爬取和数据分析知识,独立编写代码是非常必要的。

如果您是一名具有良好的计算机基础、Python 开发经验或数据挖掘、数据分析背景的读者，则建议您独立完成本套书中相应章节的实例，同时爬取自己感兴趣的数据集并深入分析，从而提升您的编程和数据分析能力。

如果您是一名数据挖掘或自然语言处理相关行业的研究者，建议您从本套书中找到自己感兴趣的章节进行学习，同时也可以将本套书作为数据爬取或数据分析的小字典，希望给您带来一些应用价值。

如果您是一名老师，则推荐您使用本套书作为网络数据爬取或网络数据分析相关课程的教材，您可以按照本套书中的内容进行授课，也可以将本套书中相关章节布置为学生的课后习题。个人建议老师在讲解完基础知识之后，把相应章节的任务和数据集描述布置给学生，让他们实现对应的爬取或分析实验。但切记，一定要让学生自己独立实现书中的代码，以扩展他们的分析思维，从而培育更多数据爬取和数据分析领域的人才。

如果您只是一名对数据爬取或数据分析感兴趣的读者，则建议您简单了解本书的结构、每章节的内容，掌握数据爬取和数据分析的基本流程，作为您学习 Web 数据挖掘和大数据分析的参考书。

无论如何，作者都希望本套书能给您普及一些网络数据爬取相关的知识，更希望您能爬取自己所需的语料，结合本套书中的案例分析自己研究的内容，给您的研究课题或论文提供一些微不足道的思路。如果本套书让您学会了 Python 爬取网络数据的方法，作者就更加欣慰了。

最后，完成本套书肯定少不了很多人的帮助和支持，在此送上最诚挚的谢意。

本套书确实花费了作者很多心思，包括多年来从事 Web 数据挖掘、自然语言处理、网络爬虫等领域的研究，汇集了作者 5 年来博客知识的总结。本套书在编写期间得到了许多 Python 数据爬取和数据分析爱好者，作者的老师、同学、同事、学生，以及互联网一些“大牛”的帮助，包括张老师（北京理工大学）、籍文（大疆创新科技公司）、徐溥（SAP 公司）、俊林（阿里巴巴公司）、容神（北京理工大学）、峰子（华为公司）、田一（南京理工大学）、王金（重庆邮电大学）、罗炜（北京邮电大学）、胡子（中央民族大学）、任行（中国传媒大学）、青哥（老师）、兰姐（电子科技大学）、小何幸（贵州财经大学）、小民（老师）、任瑶（老师）等，在此表示最诚挚的谢意。同时感谢北京理工大学和贵州财经大学对作者多年的教育与培养，感谢 CSDN 网站、博客园网站、阿里云栖社区等多年来对作者博客和专栏的支持。

由于本套书是结合作者关于 Python 实际爬取网络数据和分析数据的研究，以及多年撰写博客经历而编写的，所以书中难免会有不足或讲得不够透彻的地方，敬请广大读者谅解。如果您发现书中的错误，请联系作者，联系方式：1455136241@qq.

com,https://blog.csdn.net/eastmount(博客地址)。

最后,以作者离开北京选择回贵州财经大学信息学院任教的一首诗结尾吧!

贵州纵美路迢迢,未付劳心此一遭。
收得破书三四本,也堪将去教尔曹。
但行好事,莫问前程。
待随满天桃李,再追学友趣事。

作　者
2018 年 2 月 24 日

目　录

第 1 章

网络数据分析概述

Web 数据分析是一门多学科融合的学科，它涉及统计学、数据挖掘、机器学习、数据科学、知识图谱等领域。数据分析是指用适当的统计方法对所收集数据进行分析，通过可视化手段或某种模型对其进行理解分析，从而最大化挖掘数据的价值，形成有效的结论。本章主要普及网络数据分析(Web Data Analysis)的基本概念，讲述数据分析流程和相关技术，同时讲解 Python 环境下数据分析的环境配置与常用数据集等。

1.1 数据分析

网络数据分析是指采用合适的统计分析方法，建立正确的分析模型，对 Web 网络数据进行分析，提取有价值的信息和结论，挖掘出数据的价值，从而造福社会和人类。数据分析可以帮助人们做出预测和预判，以便采取适当行动解决问题。

数据分析的目的是从海量数据或无规则数据集中把有价值的信息挖掘出来，把隐藏的信息提炼出来，并总结出所研究数据的内在规律，从而帮助用户进行决策、预测和判断。

数据分析通常包括前期准备、数据爬取、数据预处理、数据分析、可视化绘图及分析评估 6 个步骤，如图 1.1 所示。

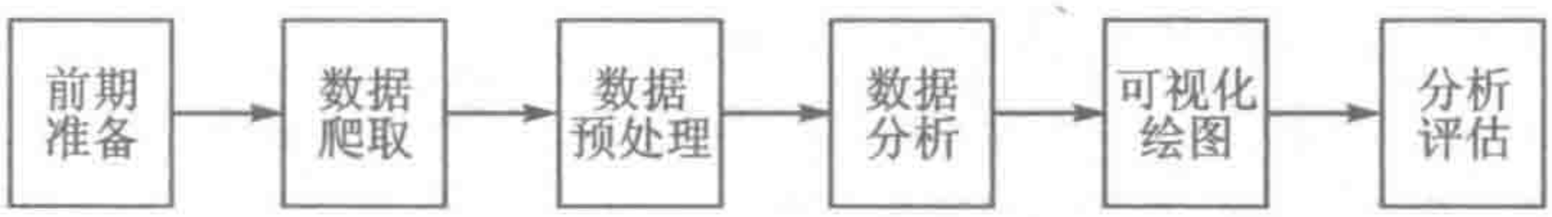


图 1.1 数据分析流程

① 前期准备。在获取数据之前，先要决定本次数据分析的目标，这些目标需要进行大量的数据收集和前期准备，判断整个实验是否能向着正确的方向进行。