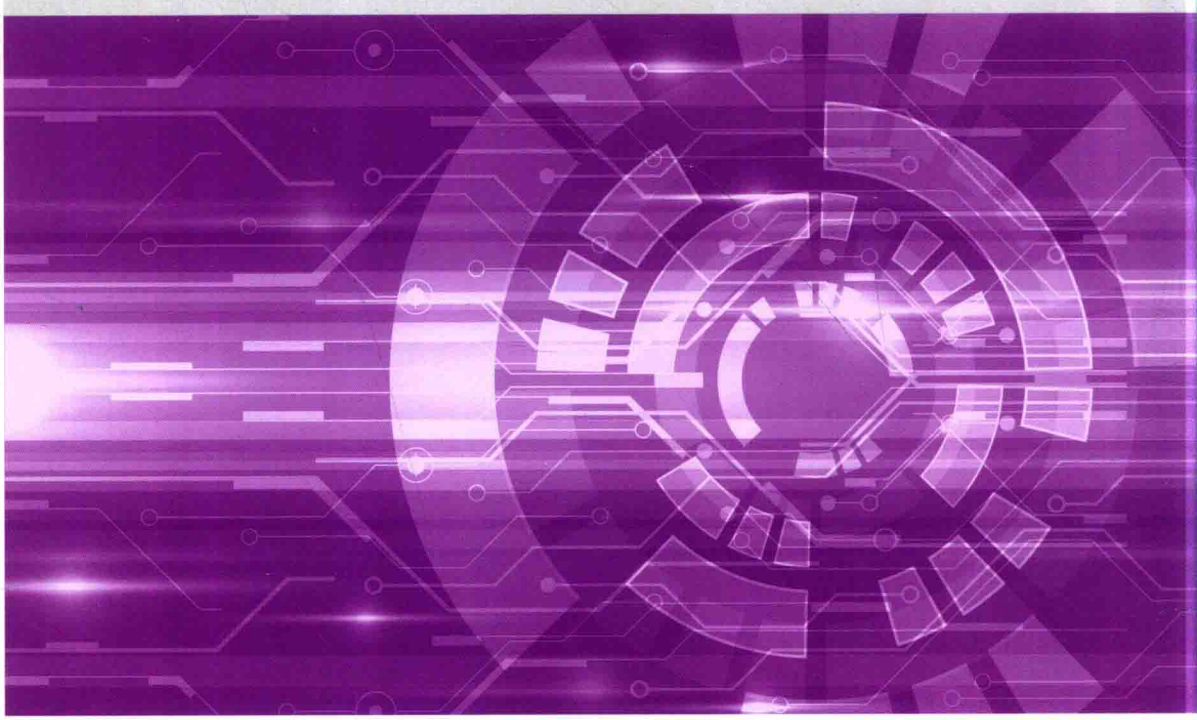
• 大数据应用人才培养系列教材 •

# Python语言

■总主编◎刘　鹏　张　燕　■主编◎李肖俊　■副主编◎钟　涛　刘　河





清华大学出版社

大数据应用人才培养系列教材

# Python 语言

总主编　刘　鹏　张　燕
主　编　李肖俊
副主编　钟　涛　刘　河

清华大学出版社
北　京

## 内 容 简 介

本书以 Win 10 和 Python 3.6.5 搭建 Python 开发基础平台为起点，重点阐述 Python 语言的基础知识和 3 个典型的项目实战案例。全书以理论引导、案例驱动、上机实战为理念打造 Python 语言学习的新模式。具体内容分为两大部分：第一部分以 Python 编程语言基础知识普及为主，分别介绍了 Python 3 概述、基本语法、流程控制、组合数据类型、字符串与正则式、函数、模块、类和对象、异常、文件操作；第二部分以项目实战为核心，以学以致用为导向，以切近生活的案例为依托，分别介绍 Python 爬虫项目实战、Python 数据可视化项目实战、Python 数据分析项目实战。

本书以作者十多年的计算机专业课程教学经验及相应的项目实战心得为依托，力争做到以理论知识为基础、以案例实战为手段、以解决问题为根本的初衷。让读者最大限度地从书中汲取他们所需要的编程知识和实战体验。

本书可作为高等学校尤其是高职院校各专业的 Python 语言启蒙教材，同时也可作为广大 Python 语言爱好者自学的参考书。



图书在版编目（CIP）数据

Python 语言 / 刘鹏，张燕总主编；李肖俊主编. —北京：清华大学出版社，2019
（大数据应用人才培养系列教材）
ISBN 978-7-302-51982-9

Ⅰ. ①P…　Ⅱ. ①刘…　②张…　③李…　Ⅲ. ①软件工具-程序设计-高等职业教育-教材
Ⅳ. ①TP311.561

中国版本图书馆 CIP 数据核字（2019）第 000175 号

**责任编辑：** 贾小红
**封面设计：** 刘　超
**版式设计：** 文森时代
**责任校对：** 马军令
**责任印制：** 宋　林

**出版发行：** 清华大学出版社
**网　　址：** http://www.tup.com.cn，http://www.wqbook.com
**地　　址：** 北京清华大学学研大厦 A 座　　**邮　　编：** 100084
**社 总 机：** 010-62770175　　**邮　　购：** 010-62786544
**投稿与读者服务：** 010-62776969，c-service@tup.tsinghua.edu.cn
**质量反馈：** 010-62772015，zhiliang@tup.tsinghua.edu.cn
**印 装 者：** 北京密云胶印厂
**经　　销：** 全国新华书店
**开　　本：** 185mm×260mm　　**印　　张：** 18　　**字　　数：** 340 千字
**版　　次：** 2019 年 1 月第 1 版　　**印　　次：** 2019 年 1 月第 1 次印刷
**定　　价：** 59.80 元

产品编号：081619-01

# 编写委员会


总主编　刘　鹏　张　燕

主　编　李肖俊

副主编　钟　涛　刘　河

参　编　刘　娅　姜玉玲

# 总　序

短短几年间，大数据就以一日千里的发展速度，快速实现了从概念到落地，直接带动了相关产业的井喷式发展。数据采集、数据存储、数据挖掘、数据分析等大数据技术在越来越多的行业中得到应用，随之而来的就是大数据人才缺口问题的凸显。根据《人民日报》的报道，未来3～5年，中国需要180万数据人才，但目前只有约30万人，人才缺口达到150万之多。

大数据是一门实践性很强的学科，在其金字塔型的人才资源模型中，数据科学家居于塔尖位置，然而该领域对于经验丰富的数据科学家需求相对有限，反而是对大数据底层设计、数据清洗、数据挖掘及大数据安全等相关人才的需求急剧上升，可以说占据了大数据人才需求的80%以上。比如数据清洗、数据挖掘等相关职位，需要源源不断的大量专业人才。

迫切的人才需求直接催热了相应的大数据应用专业。2018年1月18日，教育部公布了“大数据技术与应用”专业备案和审批结果，已有270所高职院校申报开设“大数据技术与应用”专业，其中共有208所职业院校获批“大数据技术与应用”专业。随着大数据的深入发展，未来几年申请与获批该专业的职业院校数量仍将持续走高。同时，对于国家教育部正式设立的“数据科学与大数据技术”本科新专业，除已获批的35所大学之外，2017年申请院校也高达263所。

即使如此，就目前而言，在大数据人才培养和大数据课程建设方面，大部分专科院校仍然处于起步阶段，需要探索的问题还有很多。首先，大数据是个新生事物，懂大数据的老师少之又少，院校缺“人”；其次，院校尚未形成完善的大数据人才培养和课程体系，缺乏“机制”；再次，大数据实验需要为每位学生提供集群计算机，院校缺“机器”；最后，院校没有海量数据，开展大数据教学实验工作缺少“原材料”。

对于注重实操的大数据技术与应用专业专科建设而言，需要重点面向网络爬虫、大数据分析、大数据开发、大数据可视化、大数据运维工程师的工作岗位，帮助学生掌握大数据技术与应用专业必备知识，使其具备大数据采集、存储、清洗、分析、开发及系统维护的专业能力和技能，成为能够服务区域经济的发展型、创新型或复合型技术技能人才。无论是缺“人”、缺“机制”、缺“机器”，还是缺少“原材料”，最终都难以

培养出合格的大数据人才。

其实，早在网格计算和云计算兴起时，我国科技工作者就曾遇到过类似的挑战，我有幸参与了这些问题的解决过程。为了解决网格计算问题，我在清华大学读博期间，于 2001 年创办了中国网格信息中转站网站，每天花几个小时收集和分享有价值的资料给学术界，此后我也多次筹办和主持全国性的网格计算学术会议，进行信息传递与知识分享。2002 年，我与其他专家合作的《网格计算》教材正式面世。

2008 年，当云计算开始萌芽之时，我创办了中国云计算网站（chinacloud.cn）（在各大搜索引擎“云计算”关键词中名列前茅），2010 年出版了《云计算（第 1 版）》，2011 年出版了《云计算（第 2 版）》，2015 年出版了《云计算（第 3 版）》，每一版都花费了大量成本制作并免费分享对应的几十个教学 PPT。目前，这些 PPT 的下载总量达到了几百万次之多。同时，《云计算》一书也成为国内高校的优秀教材，在中国知网公布的高被引图书名单中，《云计算》在自动化和计算机领域排名全国第一。

除了资料分享，在 2010 年，我们在南京组织了全国高校云计算师资培训班，培养了国内第一批云计算老师，并通过与华为、中兴、360 等知名企业合作，输出云计算技术，培养云计算研发人才。这些工作获得了大家的认可与好评，此后我接连担任了工信部云计算研究中心专家、中国云计算专家委员会云存储组组长、中国大数据应用联盟人工智能专家委员会主任等。

近几年，面对日益突出的大数据发展难题，我们也正在尝试使用此前类似的办法去应对这些挑战。为了解决大数据技术资料缺乏和交流不够通透的问题，我们于 2013 年创办了中国大数据网站（thebigdata.cn），投入大量的人力进行日常维护，该网站目前已经在各大搜索引擎的“大数据”关键词排名中名列前茅；为了解决大数据师资匮乏的问题，我们面向全国院校陆续举办多期大数据师资培训班，致力于解决“缺人”的问题。

2016 年年末至今，我们已在南京多次举办全国高校/高职/中职大数据免费培训班，基于《大数据》《大数据实验手册》以及云创大数据提供的大数据实验平台，帮助到场老师们跑通了 Hadoop、Spark 等多个大数据实验，使他们跨过了“从理论到实践，从知道到用过”的门槛。

其中，为了解决大数据实验难问题而开发的大数据实验平台，正在为越来越多的高校教学科研带去方便，帮助解决缺“机器”与缺“原材料”的问题。2016 年，我带领云创大数据的科研人员，应用 Docker 容器技术，成功开发了 BDRack 大数据实验一体机，它打破了虚拟化技

术的性能瓶颈，可以为每一位参加实验的人员虚拟出 Hadoop 集群、Spark 集群、Storm 集群等，自带实验所需数据，并准备了详细的实验手册（包含 42 个大数据实验）、PPT 和实验过程视频，可以开展大数据管理、大数据挖掘等各类实验，并可进行精确营销、信用分析等多种实战演练。

目前，大数据实验平台已经在郑州大学、成都理工大学、金陵科技学院、天津农学院、西京学院、郑州升达经贸管理学院、信阳师范学院、镇江高等职业技术学校等多所院校部署应用，并广受校方好评。该平台也可以云服务的方式在线提供，实验更是增至 85 个，师生通过自学，可用一个月时间成为大数据实验动手的高手。此外，面对席卷而来的人工智能浪潮，我们团队推出的 AIRack 人工智能实验平台、DeepRack 深度学习一体机以及 dServer 人工智能服务器等系列应用，一举解决了人工智能实验环境搭建困难、缺乏实验指导与实验数据等问题，目前已经在清华大学、南京大学、南京农业大学、西安科技大学等高校投入使用。

在大数据教学中，本科院校的实践教学应更加系统性，偏向新技术的应用，且对工程实践能力要求更高。而高职、高专院校则更偏向于技术性和技能训练，理论以够用为主，学生将主要从事数据清洗和运维方面的工作。基于此，我们联合多家高职院校专家准备了《云计算导论》《大数据导论》《数据挖掘基础》《R 语言》《数据清洗》《大数据系统运维》《大数据实践》系列教材，帮助解决“机制”欠缺的问题。

此外，我们也将继续在中国大数据和中国云计算等网站免费提供配套 PPT 和其他资料。同时，持续开放大数据实验平台、免费的物联网大数据托管平台万物云和环境大数据免费分享平台环境云，使资源与数据随手可得，让大数据学习变得更加轻松。

在此，特别感谢我的硕士导师谢希仁教授和博士导师李三立院士。谢希仁教授所著的《计算机网络》已经更新到第 7 版，与时俱进，日臻完美，时时提醒学生要以这样的标准来写书。李三立院士是留苏博士，为我国计算机事业做出了杰出贡献，曾任国家攀登计划项目首席科学家。他的严谨治学带出了一大批杰出的学生。

本丛书是集体智慧的结晶，在此谨向付出辛勤劳动的各位作者致敬！书中难免会有不当之处，请读者不吝赐教。

刘 鹏

于南京大数据研究院

2018 年 5 月

# 前　言

Python 作为胶水语言，其粘合力无与伦比。尤其是站在“大数据+”与“人工智能”的风口之上，可谓是如鱼得水，潜力无限。就如同 Python 语言发明人 Guido van Rossum 曾说：“life is short you need Python.（人生苦短，我用 Python。）”当下的 Python 语言风靡全球，席卷神州大地！Python 凭借其得天独厚的优良基因，使用户如雨后春笋一般涌现出来。

Python 的盛行是时代风口和其内在基因聚合的结果。这是因为 Python 以其开源性、可扩展性为根本抓住了时代的主旋律。尤其是人工智能领域的再次爆发，世界顶尖公司以 Python 为母体推出优秀的机器学习框架（如 Google 的 TensorFlow），更是助推 Python 成为风口上的王者。作者认为用“no Python , no code（无 Python，不代码）”来赞颂 Python 也不为过。然而，Python 的流行过于突然，市场上大部分介绍 Python 的书籍都是外文著作直接翻译过来的，其写作习惯和风格不太适合中国读者的需求，同时国内介绍 Python 的书籍也良莠不齐。

为了使国内读者能够系统地了解新技术、新方法，南京大数据研究院刘鹏教授顺势而为，周密规划，在大数据应用人才培养课程体系中，专门设立了 Python 语言课程，并邀请全国上百家高校中从事一线教学和科研的教师一起，编撰大数据应用人才培养系列丛书，本书即该套丛书之一。

本书以“任务驱动，实战为王”为出发点，详细介绍 Python 语言的基础知识，同时，书中剖析了 3 个典型的切近生活的实战案例，以培养读者解决问题的能力。另外，本书以“理论和实践两手抓，两手都要硬”为根本，在每章的理论学习之后，都有与之匹配的上机实验和课堂练习。将理论和实践融为一体，让读者真正地将理论和实战合二为一，做到学以致用。

本书重点阐述 Python 语言的基础知识和与之相关的 3 个典型的项目实战案例。全书共 13 章，分为两大部分：第一部分以 Python 编程语言基础知识普及为主，分别介绍了 Python 3 概述、基本语法、流程控制、组合数据类型、字符串与正则式、函数、模块、类和对象、异常及文件操作；第二部分以项目实战为核心，以学以致用为导向，以切近生活的案例为依托，分别介绍 Python 爬虫项目实战、Python 数据可视化项目

实战和 Python 数据分析项目实战。其中第一部分：第 1～5 章由钟涛老师编写，第 6～10 章由刘河和刘娅老师编写；第二部分：第 11～13 章项目实战由李肖俊老师编写。

本书的编撰，从提纲的确定到内容的把握与斟酌，到最后的审阅与定稿，得到了南京大数据研究院院长刘鹏教授亲力亲为的大力指导，并提出了诸多建设性的意见。同时，清华大学出版社的王莉编辑和南京云创大数据的武郑浩编辑也评阅了本书书稿，对本书给予了全面的指导和帮助，在此一并致谢。

在此，特别感谢南京大数据研究院院长刘鹏教授，正是由于他洞察时代需求，把握时代脉搏，才有了《Python 语言》这本书的创作需求，才有了我们的创作团队，才有了这本《Python 语言》。

总之，本书是集体智慧的结晶，在此谨向付出辛勤劳动的各位作者致敬！书中难免会有不当之处，请读者不吝赐教。

李肖俊

2019 年 1 月

# 目　录