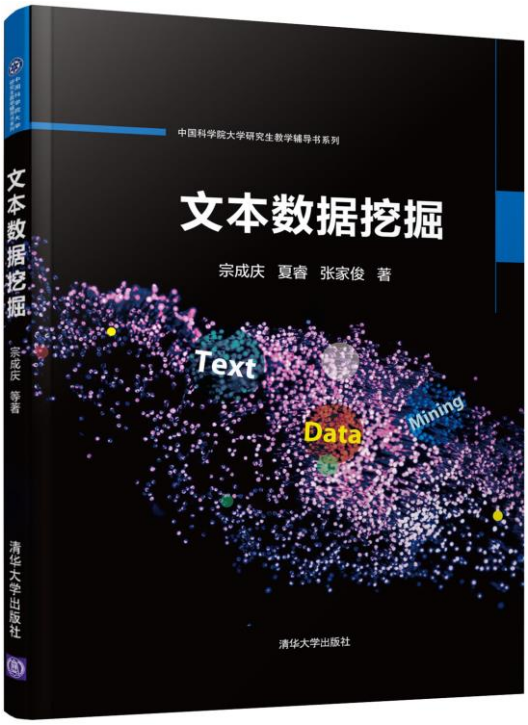
《文本数据挖掘》由清华大学出版社出版

文本数据挖掘是通过机器学习、自然语言处理和推理分析等方法，根据文本内容完成信息抽取、关系发现、热点预测、文本分类和自动摘要等具体任务的信息处理技术。随着互联网和移动通信技术的快速发展和普及应用，这项技术已在众多领域得到了广泛应用。三位作者历时两年多，全面梳理了该领域技术发展的“前生今世”，向读者展现了一个全新的视角。数据挖掘领域国际著名学者、伊利诺伊大学厄巴纳香槟分校Abel Bliss特聘教授韩家炜先生为该书作序。



正如韩家炜教授在序中所言：“我们生活在大数据时代，现实世界中 80%以上的信息是以自然语言文本形式（如书籍、新闻报道、研究论文、社交媒体和网页等）记载的非结构化数据。尽管数据挖掘和机器学习已经成为数据分析的主要手段，但是大部分数据挖掘方法只能处理结构化的或半结构化的数据。与结构化的数据挖掘任务相比较，非结构化的文本挖掘具有更大的挑战性，而且这项技术能够在将海量数据转化为结构化知识的过程中发挥巨大的作用。目前已经有不少关于数据挖掘、机器学习和统计自然语言处理的专著和教材，但是，尚没有一部系统介绍文本挖掘重要主题和最新方法的学术专著，这本《文本数据挖掘》很好地填补了这一空缺。”

内容简介：

该书全面介绍了与文本数据挖掘相关的基本概念、理论模型和实现算法，包括数据预处理、文本表示、文本分类、文本聚类、主题模型、情感分析与观点挖掘、话题检测与跟踪、信息抽取以及文本自动摘要等。

开篇从文本预处理（包括英文的和中文的文本预处理）方法介绍开始，随后给出文本表示方法，包括向量空间模型和词汇、短语、句子及文档的分布式表示，都从统计建模和深度学习建模两个角度进行了阐述。之后针对文本分类问题介绍了特征选择方法、统计学习方法和深度神经网络方法。接下来是文本聚类，包括简单的类别相似性度量和各种聚类

算法以及性能评价方法。在对上述文本挖掘基础理论和方法进行介绍之后，该书用 5 章介绍了文本挖掘技术的具体应用，包括主题模型、情感分析和观点挖掘、主题发现与跟踪、信息抽取及自动文摘。这些都是目前文本挖掘领域活跃的前沿研究课题，该书不但给予了全面而透彻的介绍，而且在传统方法和最新进展（包括深度学习方法）之间进行了很好的平衡。

宗成庆教授已经撰写和出版的《统计自然语言处理》在本领域享有盛名，拥有广泛的读者。这本新作与《统计自然语言处理》的覆盖范围完全不同，它所呈现的是关于文本挖掘的新主题，是对已有著作的扩展和补充。无论是对于自然语言处理领域的初学者，还是相关技术的研发人员，两部著作配合阅读必将从中大获裨益。

清华大学出版社官方旗舰店天猫（https://m.tb.cn/h.eU3kWvv）、京东其他网店或新华书店均匀销售。

作者简介：

宗成庆：中国科学院自动化所研究员、博士生导师，中国科学院大学岗位教授。主要从事自然语言处理、机器翻译、人机对话系统和文本数据挖掘等相关研究，主持国家级项目 10 余项，发表论文 200 余篇，出版专著《统计自然语言处理》一部和译著两部。2013 年当选国际计算语言学委员会（ICCL）委员，目前担任亚洲自然语言处理学会（AFNLP）主席和中国中文信息学会副理事长等职务，是学术期刊 ACM TALLIP 副主编、《自动化学报》副主编和 IEEE Intelligent Systems 等期刊的编委，曾任国际顶级学术会议 ACL-IJCNLP 2015 程序委员会主席，IJCAI 2017、IJCAI-ECAI 2018 和 AAAI 2019 领域主席等。获国家科技进步奖二等奖、钱伟长中文信息处理科学技术奖一等奖和中国电子学会科技进步奖一等奖，获北京市优秀教师、中科院优秀导师等荣誉称号。享受政府特殊津贴。

夏睿：南京理工大学计算机学院教授、博士生导师。主要从事自然语言处理、文本数据挖掘、情感分析与观点挖掘等领域的研究。在国际知名学术期刊和会议上发表论文 40 余篇，主持国家和省部级科研项目近 10 项。担任多个国际一流学术会议的领域主席、高级程序委员会委员和程序委员会委员。2014 年入选南京理工大学“紫金之星”人才计划，2016 年获得首届江苏省优青项目资助，2017 年入选南京理工大学青年拔尖人才计划并破格晋升为教授。

张家俊：中科院自动化所模式识别国家重点实验室副研究员，研究方向为自然语言处理、机器翻译和跨语言跨模态信息处理等。担任中国中文信息学会机器翻译专委会副主任等学术职务，在国际知名学术期刊和会议上发表论文 60 余篇，曾四次获得最佳论文奖。担任多个国际一流学术会议的领域主席和高级程序委员会委员。曾获中国中文信息学会钱伟长中文信息处理科学技术奖一等奖和汉王青年创新奖一等奖。2015 年入选首届中国科协“青年人才托举工程”计划。

全书目录：