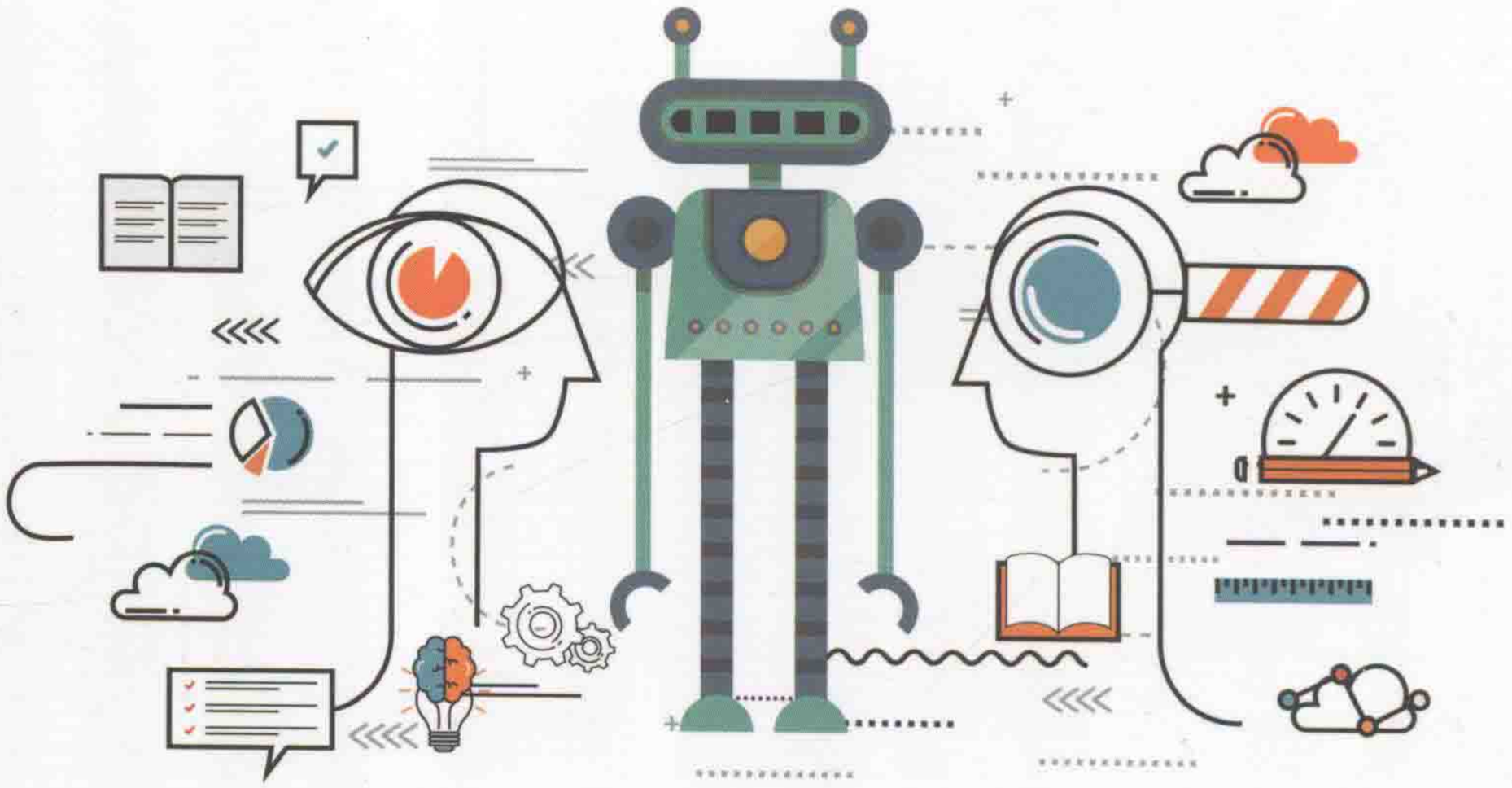
Broadview®
www.broadview.com.cn



自然语言处理
理论与实战

唐聃 白宁超 冯暄 / 等著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

自然语言处理

理论与实战

唐聃 白宁超 冯暄 卿鸿宾 文俊 著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

自然语言处理是什么？谁需要学习自然语言处理？自然语言处理在哪些地方应用？相关问题一直困扰着不少初学者。针对这一情况，作者结合教学经验和工程应用编写此书。本书讲述自然语言处理相关学科知识和理论基础，并介绍使用这些知识的应用和工具，以及如何在实际环境中使用它们。由于自然语言处理的特殊性，其是一门多学科交叉的学科，初学者难以把握知识的广度和宽度，对侧重点不能全面掌握。本书针对以上情况，经过科学调研分析，选择以理论结合实例的方式将内容呈现出来。其中涉及开发工具、Python语言、线性代数、概率论、统计学、语言学等工程上常用的知识介绍，然后介绍自然语言处理的核心理论和案例解析，最后通过几个综合性的例子完成自然语言处理的学习和深入。本书旨在帮助读者快速、高效地学习自然语言处理和人工智能技术。

本书适用于具备一定编程基础的计算机专业、软件工程专业、通信专业、电子技术专业和自动化专业的大学二年级以上的学生、科研工作者和相关技术人员。一些做工程应用的自然语言处理工程师，也可以通过阅读本书补充理论知识，理论知识的魅力在于遇到工程难题时，可以知道其背后的原因，快速、准确地解决问题。

图书在版编目（CIP）数据

自然语言处理理论与实战 / 唐聃等著. —北京：电子工业出版社，2018.6
ISBN 978-7-121-34390-2

Ⅰ. ①自… Ⅱ. ①唐… Ⅲ. ①自然语言处理—研究 Ⅳ. ①TP391

中国版本图书馆 CIP 数据核字（2018）第 122905 号

责任编辑：陈晓猛
印　　刷：三河市君旺印务有限公司
装　　订：三河市君旺印务有限公司
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：787×980　1/16　　印张：22.5　　字数：432 千字
版　　次：2018 年 6 月第 1 版
印　　次：2018 年 6 月第 1 次印刷
定　　价：79.00 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

本书讲述自然语言处理重要的相关学科知识和理论基础，并介绍使用这些知识的应用和工具，以及如何在实际环境中使用它们。市面上出版的自然语言处理书籍不多，且大多数讨论的是其背后的深奥原理，很少涉及基础知识和编程实现。自然语言处理是一门多学科交叉的学科，初入门的读者难以把握知识的广度和宽度，尤其对侧重点不能全面掌握。本书针对以上情况，经过科学调研分析，选择以理论结合实例的方式呈现知识点。首先介绍开发工具、Python 语言、线性代数、概率论、统计学、语言学等工程上常用的知识，然后介绍自然语言处理的核心理论和案例解析，最后通过几个综合性的例子完成自然语言处理的学习和深入。本书旨在帮助读者快速高效地学习自然语言处理和人工智能技术。

读者对象

自然语言处理是什么？谁需要学习自然语言处理？自然语言处理在哪些地方应用？本书就是对这几个问题的回答。自然语言处理领域主要探讨：如何处理及运用自然语言；自然语言认知（让计算机“懂”人类的语言）；自然语言生成系统（将计算机数据转化为自然语言）和自然语言理解系统（将自然语言转化为计算机程序更易于处理的形式）。自然语言处理在我们身边应用得非常广泛，其中包括：语音的自动合成与识别、机器翻译、自然语言理解、人机对话、信息检索、文本分类、自动文摘，等等。此外，自然语言处理也是人工智能、机器学习、深度学习的基础，重要程度不言而喻。如果读者有一定的编程基础，那么将有助于本书的阅读。如果读者不具备线性代数、概率论、统计学、语言学的知识，则可从本书中快速学习常见的工程应用知识；如果读者具备线性代数、概率论、统计学、语言学的知识，则更利于本书的阅读，可以对知识进行查全补充。此外，本身使用 Python 语言进行编程，假设读者具备 Python 知识，则可以跳过第 2 章，也更有利于本书的阅读。本书对于具备一定编程基础的计算机专业、软件工程专业、通信专业、电子技术专业和自动化专业的大学二年级以上的学生都是适宜的。一些做工程应用的自然语言处理工程师，也可以通过阅读本书补充理论知识。理论知识的最大魅力在于遇到工程难题时，可以知道其背后的原因，快速准确地解决问题。本书整体难度适宜，适合作为自学用书或课程教材。

本书结构

本书共四大部分 15 章，第一部分为基础部分，从第 1 章至第 6 章，主要介绍在自然语言交叉学科中，工程应用常用的学科知识，包括自然语言处理概述、Python 基础知识和环境搭建、线性代数、概率论、统计学、语言学。第二部分为理论部分，从第 7 章至第 14 章，主要介绍自然语言处理常用的理论知识，包括自然语言处理任务限制、技术范畴、语料库、中文自动分词、数据预处理、马尔可夫模型、条件随机场、模型评估和命名实体识别。第三部分为实战部分，第 15 章通过 GitHub 数据提取与可视化分析、微博话题爬取与存储分析，综合介绍网络爬虫、中文分词、数据处理、模型选择、数据分析、自然语言处理工具和数据可视化等技术点，这些技术也适用于以机器学习为代表的人工智能领域。本书各章节的具体内容介绍如下。

- ◎ 第 1 章基础入门：随着人工智能的快速发展，自然语言处理和机器学习技术的应用愈加广泛。然而身为初学者，要想快速入门这些前沿技术总是存在着各种各样的困难。为使读者对该领域的整体概况有一个系统明晰的认识，本章主要从发展历程、研究现状、应用前景等角度概要介绍自然语言处理及相关的机器学习技术。
- ◎ 第 2 章快速上手 Python：Python 作为一门简洁优美且功能强大的语言，越来越受到编程人员的青睐，在工业界和学术界也非常受欢迎。本书的全部代码都是通过 Python 实现的，之所以选择 Python 语言，是因为其可以跨平台跨应用开发，因此本章旨在帮助读者快速领略 Python 的概貌。如果读者已经具备 Python 基础，则可略过此章。
- ◎ 第 3 章线性代数：机器学习是计算机科学、统计学、数学和信息论等多个领域交叉的学科。线性代数又是数学的一个重要分支，对机器学习有着直接的影响。诸如算法建模、参数设置、验证策略、识别欠拟合和过拟合，等等。读者往往知道线性代数很有用，常常全书通读，造成时间不足和效率较低，归因于对线性代数在机器学习中的重点和用途不明。本章主要以简明的方式介绍常用的线性代数知识，并使读者知道线性代数常用于哪些方面。
- ◎ 第 4 章概率论：机器学习与深度学习是多学科交叉的科学技术，其中数学尤为重要，是很多形式化模型向数学建模的必经过程。继线性代数核心知识的介绍之后，本章着重介绍概率论的相关知识。
- ◎ 第 5 章统计学：在数据科学中，统计学的地位尤为显著。这是一门在数据分析的基础上，研究如何测定、收集、整理、归纳和分析数据规律，以便给出正确消息的学科。通过揭示数据背后的规律和隐藏信息，给相关角色提供参照价值，以做出相应的决策。其在数据挖掘、自然语言处理、机器学习中都被广泛应用。本章首先介绍常见的图形可视化的概念和使用，继而介绍数据度量标准、概率分布、统计假设检验、相关和回归，

以短小精悍的篇章使读者掌握基本的统计知识。

◎ 第 6 章语言学：本章主要从语音、词汇、语法三个角度对现代汉语进行一个简单概要的勾勒，在以往传统的语言学教材中一般还有“文字”“修辞”两节内容，因篇幅有限、与全书关联不强，在此删繁就简，未给读者一一呈现。需要注意的是，语言学本身是一门十分庞杂的学科，知识体系与研究方法或因语言不同而有区别，或因派别主义不同而有区别。但无论是何种语言，或是何门何派，在进行自然语言处理时我们要面临的永远是一个个真实的语料和具体的语言现象。理论是用来指导实践、拓宽我们研究思路的，究竟最后采用何种理论，这只是一个“白猫黑猫”的问题。

◎ 第 7 章自然语言处理：本章开篇直击要点，即自然语言处理的任务和限制。进而介绍其所涉及的主要技术范畴，并对这些技术方向进行介绍。在针对当前自然语言处理的难点进行详细剖析后，最终对 2017 年以后自然语言处理的发展进行展望。

◎ 第 8 章语料库：大数据发展的基石就是数据量的快速增加，无论是自然语言处理、数据挖掘、文本处理，还是机器学习领域，都是在此基础上通过规则或统计方法进行模型构建的。但是不是数据量足够大就叫大数据了呢？是不是数据量足够多就构成语料库了呢？带着这些疑问，本章将带你走进语料库的世界，对语料知识进行一次全面而深入的了解。

◎ 第 9 章中文自动分词：中文分词技术属于自然语言处理的技术范畴，中文分词是其他中文信息处理的基础，搜索引擎只是中文分词的一个应用。诸如机器翻译（MT）、语音合成、自动分类、自动摘要、自动校对，等等。

◎ 第 10 章数据预处理：数据预处理的整个步骤流程在自然语言处理的工程中要比其在机器学习的工程中精简一些，最大的区别就在于数据清洗和特征构造这两个至关重要的过程。在自然语言处理中特征构造是否良好，很大程度上取决于所构造的特征数据集的数据特性与文本内容语义吻合程度的高低。比如，文本情感分类和文本内容分类都属于分类范畴，但对于同一种算法（参数都调整到最优），在两个不同分类的业务下，得到的结果可能会相差很大。通过仔细分析，我们不难发现造成这种差异的根本原因就是构造出来的特征数据集的数据模式没有很好地契合文本的真实语义，这也是自然语言处理的最大难点。

◎ 第 11 章马尔可夫模型：笔者最早接触马尔可夫模型的定义源于吴军先生的《数学之美》一书，起初觉得深奥难懂且没什么用处。直到学习自然语言处理时，才真正使用到马尔可夫模型，并体会到此模型的奇妙之处。马尔可夫模型在处理序列分类时具有强大的功能，解决诸如词类标注、语音识别、句子切分、字素音位转换、局部句法剖

析、语块分析、命名实体识别、信息抽取等问题。此外它还广泛应用于自然科学、工程技术、生物科技、公用事业、信道编码等多个领域。

◎ 第 12 章条件随机场：条件随机场常用于序列标注、数据分割等自然语言处理任务中，此外在中文分词、中文人名识别和歧义消解等任务中也有应用。本书基于笔者在做语句识别序列标注过程中对条件随机场产生的了解。主要内容源于自然语言处理、机器学习、统计学习方法和部分网上资料对 CRF 的相关介绍，最后由笔者进行大量研究整理后汇总成知识体系。本章首先介绍条件随机场的相关概念，然后结合实例以期让读者深入理解条件随机场的应用。

◎ 第 13 章模型评估：本章源于基于 HMM 模型序列标注的一个实验，在实验完成之后，迫切想知道采用的序列标注模型好坏，有哪些指标可以度量。于是就产生了对这一专题进度的学习总结，这样也便于其他人参考。本章依旧简明扼要地梳理出模型评估核心指标，以期达到实用的目的。

◎ 第 14 章命名实体识别：命名实体识别在自然语言处理中占据着非常重要的地位，也是不可逾越的学术问题。命名实体识别的学术理论和研究方法众多，本章侧重整体介绍。首先阐述命名实体识别的背景知识和研究概况，介绍中文命名实体识别的特点与难点，辅以案例加深理解；然后对命名实体识别当前的研究方法和核心技术进行详细介绍；最后展望其在未来人工智能方面的发展前景。

◎ 第 15 章自然语言处理实战：自然语言处理技术是理论与实践相结合的一门学科，通过前面基础理论知识的介绍，读者对其理论有所认识，但其究竟有何用、怎么用却不深刻。本章通过实例演练，一方面对前面几章的知识进行复习回顾，另一方面利于加深理解研发的相关工作。本章的第一个案例以 GitHub 为例，实现数据提取和可视化；第二个案例以微博话题为例，实现数据采集、提取、存储与分析。

勘误

由于笔者能力有限，时间仓促，书中难免有错漏，欢迎读者批评指正。

联系方式：nlpjiaocheng@sina.com。

作者介绍

唐聃 教授，中科院工学博士。现工作于成都信息工程大学软件工程学院。研究方向包括自然语言处理、信息安全、数据分析。曾参与多项国家 863 项目和中科院知识创新工程项

目、省科技厅和教育厅项目；2016 年入选中国科学院西部之光人才计划（中国科学院西部青年学者）。

白宁超 四川省计算机研究院软件开发工程师，曾参与多项四川省科技厅项目。其自然语言处理系列博文曾被 CSDN、博客园、阿里云栖等多个技术社区转载。

冯暄 高级工程师，硕士学位。现任四川省计算机研究院信息化工程研究所所长。研究方向包括物联网、多源信息融合、软件工程。主持或参与国家级、省级科研项目 16 项。获得四川省科技进步奖二等奖 2 项、四川省科技进步奖三等奖 1 项。

卿鸿宾 四川大学中文系在校生。研究方向包括应用语言学、计算语言学、韵律句法学等。常年从事文学创作与文字工作，2017 年作品《黄昏速写》发表于《子曰书院》微信公众号，取得了不错的反响。

文俊 硕士学位，现工作于成都广播电视台橙视传媒大数据中心，大数据算法工程师。研究方向主要包括数据挖掘、机器学习、自然语言处理、深度学习及云计算。

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/34390

目录