华章IT

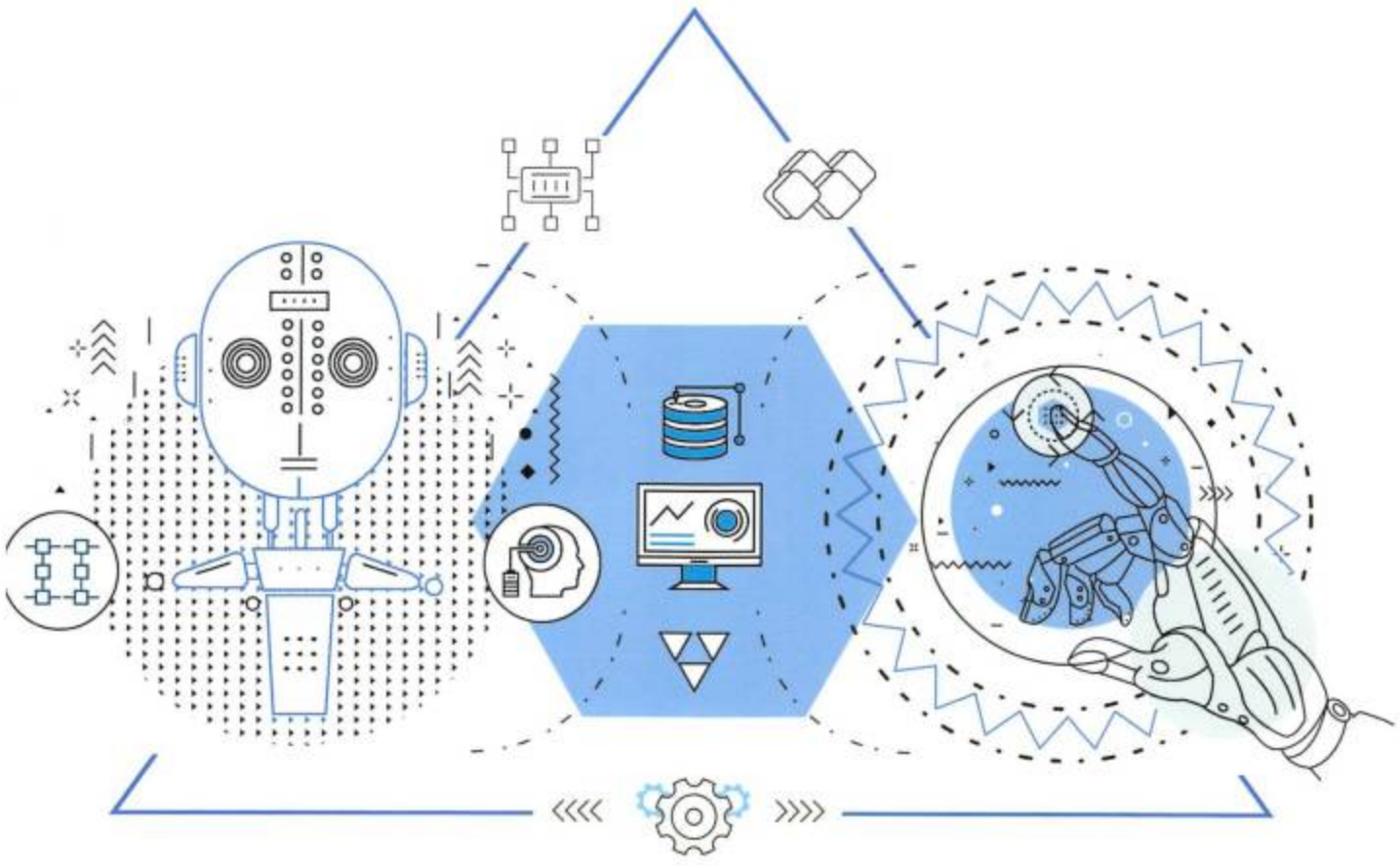
智能系统与技术丛书

天善智能
TIANSHAN SOFT

Natural Language Processing
Core Technology and Algorithm with Python

Python自然语言处理实战
核心技术与算法

涂铭 刘祥 刘树春 著



机械工业出版社
China Machine Press

内容简介

自然语言处理是一门融语言学、计算机科学、数学于一体的学科，比较复杂，学习门槛高，但本书巧妙地避开了晦涩难懂的数学公式和证明，即便没有数学基础，也能零基础入门。

本书专注于中文的自然语言处理，以Python及其相关框架为工具，以实战为导向，详细讲解了自然语言处理的各种核心技术、方法论和经典算法。三位作者在人工智能、大数据和算法领域有丰富的积累和经验，是阿里巴巴、前明略数据和七牛云的资深专家。同时，本书也得到了阿里巴巴达摩院高级算法专家、七牛云AI实验室Leader等专家的高度评价和鼎力推荐。

全书一共11章，在逻辑上分为2个部分：

第一部分（第1、2、11章）

主要介绍了自然语言处理中需要了解的基础知识、前置技术、Python科学包、正则表达式以及Solr检索等。

第二部分（第3~10章）

第3~5章讲解了词法分析相关的技术，包括中文分词技术、词性标注与命名实体识别、关键词提取算法等。

第6章讲解了句法分析技术，该部分目前理论研究较多，工程实践中使用门槛相对较高，且效果多是依赖结合业务知识进行规则扩展，因此本书未做深入探讨。

第7章讲解了常用的向量化方法，这些方法常用于各种NLP任务的输入。

第8章讲解了情感分析相关的概念、场景以及一般做情感分析的流程，情感分析在很多行业都有应用。

第9章介绍了机器学习的重要概念，重点突出了NLP常用的分类算法、聚类算法，同时还介绍了几个案例。

第10章介绍了NLP中常用的一些深度学习算法，这些方法比较复杂，但是非常实用，需要读者耐心学习。

华章IT
HZBOOKS | Information Technology

智能系统与技术丛书

Natural Language Processing
Core Technology and Algorithm with Python

Python自然语言处理实战
核心技术与算法

涂铭 刘祥 刘树春 著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Python 自然语言处理实战：核心技术与算法 / 涂铭，刘祥，刘树春著 . —北京：机械工业出版社，2018.4
（智能系统与技术丛书）

ISBN 978-7-111-59767-4

I. P… II. ①涂… ②刘… ③刘… III. 软件工具 - 自然语言处理 - 教材 IV. ① TP311.56 ② TP391

中国版本图书馆 CIP 数据核字（2018）第 088115 号

Python 自然语言处理实战：核心技术与算法

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：张锡鹏　　责任校对：殷　虹
印　　刷：北京诚信伟业印刷有限公司　　版　　次：2018 年 5 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：18.75
书　　号：ISBN 978-7-111-59767-4　　定　　价：69.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzit@hzbook.com

FOREWORD I

序　一

不知不觉间，我们已经进入了“人工智能”时代，如今随处可见基于自然语言处理技术的聊天机器人，回想以前都是靠人工服务，现在都依靠机器人回答大部分的常见问题了。

过去几年，深度学习架构和算法在图像识别和语音处理等领域取得了重大的进步。而在 NLP（自然语言处理）领域，起初并没有太大的进展。不过现在，NLP 领域取得的一系列进展已证明深度学习技术将会对自然语言处理做出重大贡献。一些常见的任务如实体命名识别，词类标记及情感分析等，NLP 都能提供最新的结果，并超越了传统方法。另外，在机器翻译领域的应用上，深度学习技术所取得的进步应该是最显著的。

记得在上学期间感觉 NLP 这个领域很新鲜、很空白，决定尝试做一下，读完博士，感觉 NLP 比我最初接触时理解的 NLP 更新鲜，更值得挖掘。NLP 很多问题都没有正式定义，或者说很难用统一的标准去训练机器、很难搞 benchmark dataset，这可能也是 AI 的一大挑战。

我认为现在比较成熟的 AI 方向都是相对确定的。比如语音识别，拿来一段语音，就知道说的是什么话；比如 vision，猫的照片就是猫，这个人脸的照片就是这个人。NLP 有一些问题就没这么明确。比如文本摘要，到底哪一个摘要是最好的呢？机器翻译，到底哪一个译文是最好的呢？复杂一些的情感分析，这篇报道的作者到底有没有在暗讽这个人？如果一个问题有明确的答案，比如 Waston——专门参加开心辞典回答问题，算法训练起来轻松一些。但如果一个问题本身的答案并无明确的高下之分，那算法也无可奈何。

定义新问题，以较小的代价搜集新的数据集，开发新的 evaluation method，这些与研究新算法一样有趣、有挑战性。举个简单的例子。我们想让机器自动识别出来讽刺的语气，那么去哪里找讽刺的话呢？我们有现成的分析情感的工具，再利用这些有 #sarcasm 标签的推文，可以训练一个识别“什么情况是倒霉情况”的分类器。以后就可以用这个倒霉识别器去识别没有标签的讽刺句子了，bootstrap 一下把数据集搞大，这就是一个最初级的讽刺方面的数据集。

NLP 圈里很多人喜欢搞新的数据集，这个现象有利有弊，但这说明了有很多空白问题需要定义，有很多空白资源需要创建。本书从各个方面着手，帮助读者理解 NLP 的过程，提供了各种实战场景，结合现实项目背景，帮助读者理解 NLP 中的数据结构和算法以及目前主流的 NLP 技术与方法论，结合信息检索技术与大数据应用等流行技术，最终完成对 NLP 的学习和掌握。

阿里巴巴达摩院高级算法专家　黄英

2018 年 1 月 17 于杭州

FOREWORD 2

序　二

近年来，几乎整个人工智能界的研究者们都注意到一个技术名词——自然语言处理（NLP）。NLP 作为人工智能领域的一个重要分支，现在已经发展成为人工智能研究中的热点方向。最近几十年来，随着软硬件协同发展，数据爆炸性增长，信息过载的问题越来越严重，全部依赖人来分析和驱动的传统方式，面对海量信息的局面显得越来越捉襟见肘。这样的情况下，能够自动化处理大规模文本相关的数据的 NLP，即将成为未来人工智能发展技术的新趋势和方向。

自然语言处理作为机器学习与语言学、统计学等的综合学科，不仅知识内容多，发展迅速，而且非常依赖于工程能力。目前，统计学以及数据驱动的方法在 NLP 中占据着统治地位。同时，最近几年深度学习不断被引入 NLP 领域，越来越多的知识需要读者去学习。这时候急需一本能够从全局梳理 NLP 的书籍，帮助 NLP 学习者快速入门。传统的 NLP 书籍对于具体问题的方法讲解有足够的思路，但是要么是基于英文语料的讲解，要么通篇都是理论，面对复杂的中文语料环境缺乏实践性。

本书的作者通过对前人传统 NLP 技术以及新兴的深度学习方法深入梳理，形成自己理解的 NLP 解决之道。本书在内容上平衡了理论和技术，在每章的理论之后都配备了实践课，方便读者能够动手加深理解，避免成为只会夸夸其谈的 NLP 理论“专家”。本书可以帮助研究者，特别是初学者，加强对 NLP 的理论与技术的学习，授人以鱼的同时授人以渔，帮助读者灵活解决实际工作当中遇到的各种 NLP 问题。

七牛云 AI 实验室 Leader，10 余年人工智能和深度学习研究　林亦宁

PREFACE

前　言

为什么要写这本书

这是一本关于中文自然语言处理（简称 NLP）的书，NLP 是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。NLP 是一门融语言学、计算机科学、数学于一体的科学。本书偏重实战，不仅系统介绍了 NLP 涉及的知识点，同时也教会读者如何实际应用与开发。围绕这个主题，本书从章节规划到具体的讲述方式，具有以下两个特点：

第一个特点是本书的主要目标读者定位为高校相关专业的大学生（统计学、计算机技术）、NLP 爱好者，以及不具备专业数学知识的人群。NLP 是一系列学科的集合体，其中包含了语言学、机器学习、统计学、大数据以及人工智能等方面，尤其依赖数学知识才能深入理解其原理。因此本书对专业知识的讲述过程必须绕过复杂的数学证明，从问题的前因后果、创造者思考的过程、概率或几何解释代替数学解释等一系列迂回的路径去深入模型的本源，这可能多少会牺牲一些严谨性，但是却能换来对大多数人更为友好的阅读体验。

第二个特点是本书是一本介绍中文自然语言处理的书，中文分词相对于英文分词来说更为复杂，读者将通过例子来学习，体会到能够通过实践验证自己想法的价值，我们提供了丰富的来自 NLP 领域的案例。在本书的内容编制上，从知识点背景介绍到原理剖析，辅以实战案例，所有的代码会在书中详细列出或者上传 Github 方便读者下载与调