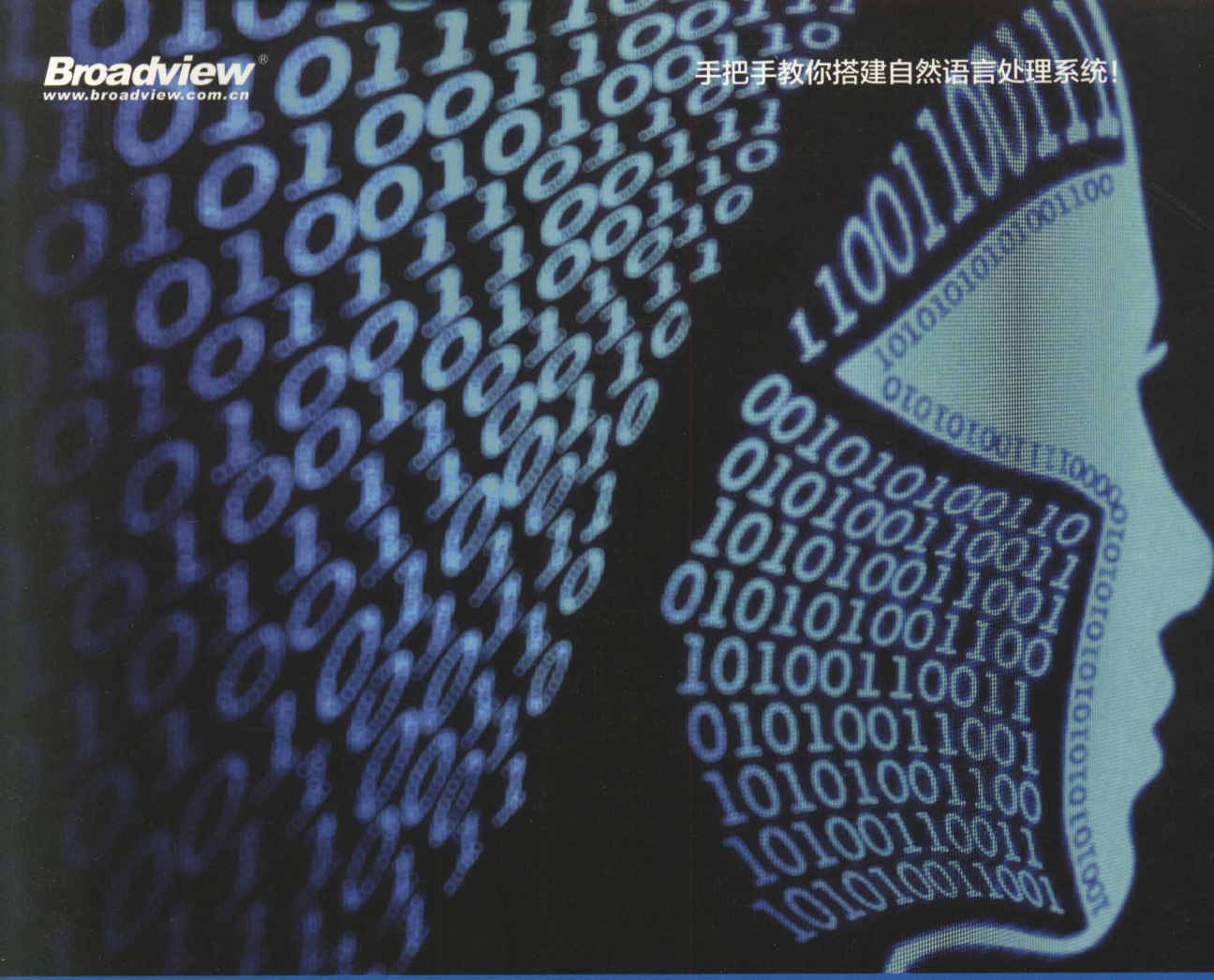
Broadview
www.broadview.com.cn

手把手教你搭建自然语言处理系统！

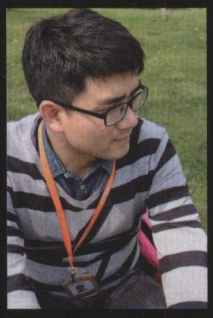


自然语言处理技术
入门与实战

兰红云◎编著

实用技术要点：语义模型详解、自然语言处理系统基础算法和系统案例实战
完整解决方案：问题解决的原理、实现的算法原理、具体算法的实现

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



作者简介

兰红云，湖北襄阳人。曾任职于猎豹移动，现为阿里影业数据挖掘专家，拥有多年算法和数据挖掘的工作经验，申请多项算法专利。研究的方向包括自然语言处理和机器学习。

自然语言处理技术
入门与实战

兰红云◎编著

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

内 容 简 介

本书主要从语义模型详解、自然语言处理系统基础算法和系统案例实战三个方面，介绍了自然语言处理中相关的一些技术。对于每一个算法又分别从应用原理、数学原理、代码实现，以及对当前方法的思考四个方面进行讲解。

本书面向的读者为有志于从事自然语言处理相关工作的在校学生、企事业单位工作人员等人群。本书的结构特点是由浅入深地进行相关内容的介绍，以满足不同层次读者的学习需求。

图书在版编目（CIP）数据

自然语言处理技术入门与实战 / 兰红云编著. —北京：电子工业出版社，2017.10
ISBN 978-7-121-32763-6

Ⅰ. ①自… Ⅱ. ①兰… Ⅲ. ①自然语言处理—研究 Ⅳ. ①TP391

中国版本图书馆 CIP 数据核字(2017)第 233723 号

策划编辑：张慧敏
责任编辑：牛　勇
特约编辑：顾慧芳
印　　刷：三河市华成印务有限公司
装　　订：三河市华成印务有限公司
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：12.5　字数：280 千字
版　　次：2017 年 10 月第 1 版
印　　次：2017 年 10 月第 1 次印刷
定　　价：59.00 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

随着移动互联网的飞速发展，特别是物联网（Internet of Thing，IoT）的飞速发展，人与设备的语言交互场景也越来越多，并且越来越成为核心。这种语言的交互既包括纯文字的，也包括语音的。自然语言处理（Natural Language Processing，NLP）就是以电子计算机、编程语言为工具对人类特有的书面和口头形式的自然语言信息进行各种类型处理和加工的技术。当然，随着技术的不断发展，其处理领域也出现了跨形态的组合。比如通过与图形图像处理技术的结合，可以实现看图说话、在线答题等应用。所以，自然语言处理是一门涉及语言学、计算机科学，当然还有数学的交叉性科学。

自然语言处理的目的是为了让计算机能够理解人的语言，然后做出相应的处理或者应答。根据应用场景的不同，自然语言处理可以分为如下三点：（1）信息抽取，包括自动摘要、自动检索、舆情分析等；（2）语言理解，包括机器翻译、人机对话、语义理解等；（3）跨形态组合，包括看图说话、语音自动合成、辅助教学等。这些应用都是利用自然语言处理技术，对所需要处理的信息进行挖掘和分析，找出人们想要的东西，进而作出响应。而落实到具体的应用，又会衍生出很多不同的应用系统，由此衍生出来的应用系统包括但不限于：信息自动抽取系统、信息自动检索系统、文本信息挖掘系统、机器翻译系统、人机对话系统、图片描述自动生成系统、语音自动识别系统、语音自动合成系统、计算机辅助教学系统，等等。

因为自然语言自身的复杂性，比如：很多歧义、结构复杂多样、表达千变万化……导致其处理方法纷繁复杂，要考虑非常多的情景。所以上述这些系统之间又存在交叉，或者上下关联，或者前后依赖等复杂的关系。而这些复杂的应用对于一个初学者来说，是非常庞杂和难以掌握的，在学习的过程中难免会因为其中某一个细节不能掌握，而影响整个进程的进度；或者虽然理解了算法的数学原理，但是怎么在实际场景中应用，以及当前算法能解决哪些实际问题，还是不了解。在笔者学习的过程中，发现目前出版的一些书籍，或

者是偏理论性的，会介绍很多自然语言处理技术发展的历史，比如符号逻辑的发展轨迹、语义网络的发展轨迹、语言学派和统计学派的“恩怨情仇”，会让初学者在学习的过程中抓不住重点，有时候又感觉它们好像就是一回事；又或者介绍的内容过于偏技术，开篇就把其中涉及的一些数学知识全都介绍一遍，因为这其中有很多数学知识是比较高阶的，比如隐马尔科夫链、条件随机场、数理逻辑推理等，在介绍数学知识的过程中，又难免会涉及相关的证明。本来其数学形式就比较复杂，再加上连环的证明就更难懂了，对于数学基础稍微薄弱一点的读者，就感觉没有学习的欲望和必要了。但是在实际应用中，其实这些烦琐的证明根本不需要，有时候只需要记住一个结论，然后根据自己数据的情况，优化模型中的参数就可以了。所以笔者就想结合自己学习过程中和实际工作中的一些经验和教训，从应用的角度来对自然语言处理中的一些技术进行介绍。在介绍的时候，为求尽量避免烦琐和突兀的数学证明，从应用的角度尽可能简洁明了地对一个算法或者处理系统进行简要的介绍，先让大家对这个方法有一个直观感性的认识，然后再深入了解其中的难点，进而深入学习和攻克难点。

本书采用以应用为主、算法和实现为辅的形式对自然语言处理中的一些技术进行介绍。对于算法数学原理的介绍，都是穿插在每一个应用的介绍中，对每一部分的数学知识进行分别介绍和讲解，没有开篇便对所涉及的数学知识进行一个全面的介绍，这样大家就不会因为某一个部分的数学知识不完备，或者掌握起来有困难而放弃整个知识体系的学习，这样大家就可以独立学习和掌握。同时因为知识遗忘的必然性，笔者将数学知识融入应用中进行介绍，就更容易让读者记住。否则前后脱节之后，就忘记了之前讲解的数学原理，即使在应用中又要重新学习，也并不一定能够知道具体的应用原理。

因此，笔者完全从应用的角度来进行各个内容的组织，没有涉及太多的处理技术起源、变革、发展等历史信息。这一方面是因为各个技术都有自己的长处和缺点，这个是理论学派争论的焦点，但不是应用层面应该关心或者需要表明立场的地方；另一方面作为主要介绍应用实战的书，这里更多的是想让读者了解对于同一个问题目前的一些处理方法和这些方法之间的优劣，以及相互的关联，以便找到解决问题更好的方法，这样也更有利于整个事情的发展。所以从做事情的本身来说，我们需要关心的是事情怎么能够做起来，没有做起来是因为什么，所以我们更多关心的是“术”的事情，而对于“道”的层面更多的是了解，是取众家之长，来“集大成”，而不能剑走偏锋。

目前，随着源工具的不断增加，大家对底层应用的开发需求在逐渐降低，所以本书先从上层应用介绍入手，让读者能够直接用起来，这样更有利于读者边实践边学习，也可以避免大家因为学习底层技术太难而阻碍后期应用的学习。从企业的角度出发，缺的也不是底层通用的处理技术和能力，更多的是缺少对实际业务的处理能力，业务跑起来之后，整个系统便会随着业务的发展而不断发展。所以本书采用以应用贯串始终的方式来进行相关技术的介绍和说明。

具体来说，本书主要从以下三个方面介绍了自然语言处理中相关的一些技术。

1）语义模型详解：主要是从应用的角度介绍自然语言处理中的一些语义处理模型，比如关键词提取、计算词距离、文本自动生成等。

2）自然语言处理系统基础算法：这一部分主要是从基础系统搭建的角度对相关算法进行介绍。包括分词、词性标注、句法分析等。这两部分介绍的内容又分别从使用原理、实现原理、具体的代码实现，以及对当前方法的思考这四个角度进行介绍。

3）系统案例实战：介绍了搭建一个舆情分析和挖掘系统所要涉及的环节、各个环节的算法实现，以及部分实现代码。

本书在写作过程中力求普及并与实践相结合，尽可能地照顾到不同层次不同专业的读者。另外，本书是以应用场景来组织各个内容的，每一个章节都包含一个完整的应用解决方案：问题解决的原理、实现的算法原理、具体算法的实现，所以读者可以根据自己的需要独立地学习各个章节的内容。在各个章节的学习过程中，笔者强烈建议读者在学习具体方法之前，一定要认真地理解所要解决问题的具体场景。要理解当前场景的输入是什么、输出是什么，为什么会是这样的结构，只有弄明白了这些，才会对算法有更深入的理解，也才能更好地使用所学习的算法，做到举一反三。因为算法本身是一种数据处理逻辑，所以只要具有相同处理逻辑的问题都可以用同样的算法，比如最大熵模型发挥了巨大的作用是人们找到了其适用的场景，而不是对模型进行各种变形以让其去适合具体的应用。

本书在写作的过程中参考了很多国内外学者的论文和著作。如果没有他们的出色工作，没有他们极为宝贵的研究成果，本书是写不出来的。在本书出版之际，谨向他们表示衷心的感谢。

在本书写作过程中，笔者常为自己的学识不足而苦恼。自然语言处理作为一门交叉性边缘性学科，涉及语言学、计算机科学、数学等各个方面的知识，笔者学识浅陋，论述之中倘有不当，恳请读者批评指正。有任何意见和建议请发到 392071814@qq.com，不胜感激。

最后，谨向帮助、支持和鼓励我完成本书的我的家人、同事、领导、朋友以及出版社的领导、编辑致以深深的敬意和真挚的感谢！

作者

2017 年 9 月于杭州

目 录

第 1 篇　语义模型详解

第1篇

语义模型详解

在本书的开始，我们想先从应用的角度来向读者介绍自然语言处理，同时对于这些应用的实现技术进行由浅入深的介绍。希望这样能让读者对于自然语言处理技术能解决什么问题，以及对于这些问题现在已经有的解决方法有哪些，有一个初步的了解。这一篇将尽量少用公式，或者使用一些简单的公式，让大家在了解技术实现的同时，不至于被繁杂的公式所吓到。毕竟霍金大神曾说过，每一个公式都将会吓退一半的读者。

第1章

关键词抽取模型

首先我们向读者介绍的是关键词提取模型，关键词提取能让我们快速地了解一篇文章，或者从大量的语料中快速找到其想要说明的主题。特别是在信息爆炸的时代，能够有效提取文本的关键词，则对于快速、及时、高效地获取信息是非常有帮助的。

1.1 TF-IDF 算法实现关键词抽取

TF-IDF 算法是关键词提取算法中基础并且有效的一种算法，因为它的实现简单，并且效果显著，所以应用非常广泛。

1.1.1 场景

假设现在有一批短文本，比如很多条一句话新闻。现在需要提取这些一句话新闻的关键词。有哪些方法可以使你采用呢？这里介绍一种非常基础的，也非常好用的算法，叫做 TF-IDF 算法。

TF-IDF（term frequency - inverse document frequency）是一种用于资讯检索与资讯探勘的常用加权技术。TF-IDF 是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数呈正比地增加，但同时也会随着它在语料库中出现的频率呈反比地下降。

1.1.2 原理

TF-IDF 的主要思想是：如果某个词或短语在一篇文章中出现的频率（Term Frequency，TF）高，并且在其他文章中很少出现，即反文档频率（Inverse Document Frequency，IDF）低，则认为此词或者短语具有很好的类别区分能力，适合用来分类。那么对于这篇文章来说，这个词也就可以算作该文章的一个关键性的词语。基于上述思想，就提出了 TF-IDF 算法，具体计算公式如下：

$$tfidf_{i,j} = tf_{i,j} \times idf_i$$

其中，

$tfidf_{i,j}$：是指词 i 相对于文档 j 的重要性值。

$tf_{i,j}$：指的是某一个给定的词语在指定文档中出现的次数占比。即给定的词语在该文档中出现的频率。这个数字是对**词数**（term count）的归一化，以防止它偏向长的文档。计算公式如下：

$$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

其中

$n_{i,j}$是该词在文件d_j中出现的次数。

$\sum_k n_{k,j}$是在文件d_j中所有字词的出现次数之和。

idf_i：指的是词 i 的逆向文档频率，是用总文档数目除以包含指定词语的文档的数目，再将得到的商取对数得到。这是一种度量词语重要性的指标。计算公式如下：

$$idf_i = log\frac{|D|}{|\{j: t_i \epsilon d_j\}|}$$

其中

|D|为语料库中的文档总数。

$|\{j: t_i \in d_j\}|$为包含词语t_i的文档数目。

至此，我们对 TF-IDF 算法有了一个初步的了解，下面从一个具体的例子来看看这个算法在实际例子中的应用。

1.1.3 实例

在开篇的场景部分，我们提了一个场景，对短文本进行关键词提取。这里就以这样的一个例子进行算法说明。

先看看测试数据（以下数据摘自百度百科）。

文档 1：程序员（Programmer）是从事程序开发、维护的专业人员。一般将程序员分为程序设计人员和程序编码人员，但两者的界限并不非常清楚，特别是在中国。软件从业人员分为初级程序员、高级程序员、系统分析员和项目经理四大类。

文档 2：现在网络流行上把男程序员称为“程序猿”，女程序员称为“程序媛”。目前从事 IT 技术行业的大多数为男性，女性多数从事其他（如：会计，行政，人力资源等）种类的工作，在 IT 技术里女程序员是很受欢迎的，因此现在人们爱称女程序员为“程序媛”。

因为 TF-IDF 对词的顺序不关心，所以分词部分就不作说明了。假设我们对上述两个文