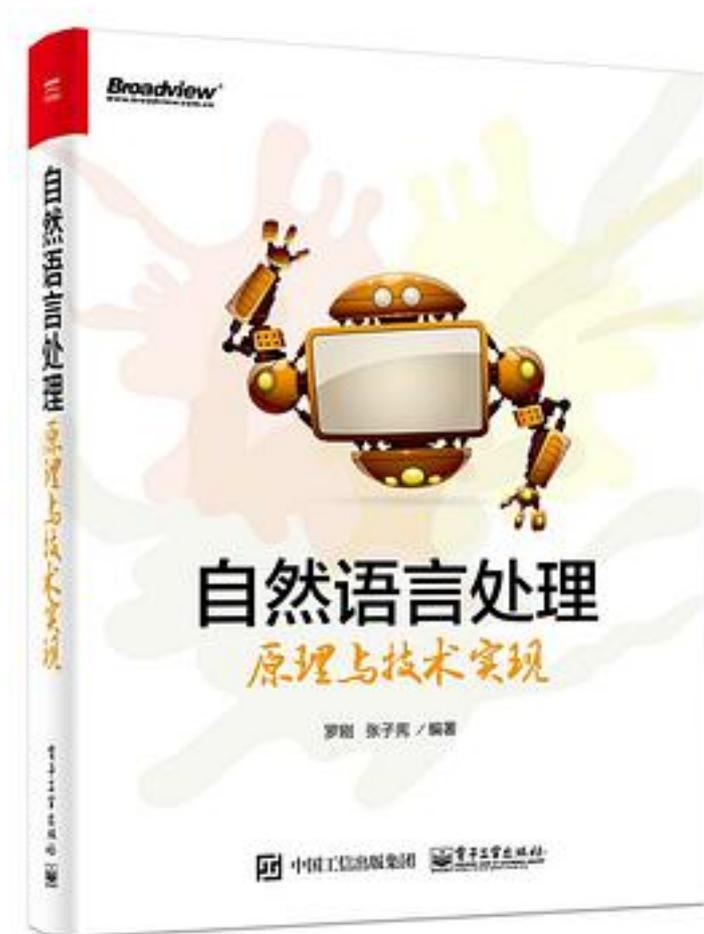


自然语言处理原理与技术实现



[自然语言处理原理与技术实现_下载链接1](#)

著者:罗刚

出版者:电子工业出版社

出版时间:2016-5

装帧:平装

isbn:9787121286209

自然语言处理技术已经深入我们的日常生活。我们经常用到的搜索引擎就用到了自然语言理解等自然语言处理技术。自然语言处理是一门交叉学科，涉及计算机、数学、语言学等领域的知识。

《自然语言处理原理与技术实现》详细介绍中文和英文自然语言处理的原理，并以Java实现，包括中文分词、词性标注、依存句法分析等。其中详细介绍了中文分词和词性标注的过程及相关算法，如隐马尔可夫模型等。在自然语言处理的应用领域主要介绍了信息抽取、自动文摘、文本分类等领域的基本理论和实现过程，此外还有问答系统、语音识别等目前应用非常广泛的领域。在问答系统的介绍中《自然语言处理原理与技术实现》特地介绍了聊天机器人的实现过程，从句子理解、句法分析、同义词提取等方面揭示聊天机器人的实现原理。

《自然语言处理原理与技术实现》详细介绍自然语言处理的各个领域，既有理论，也有实现过程。对于打算从事自然语言处理研究的计算机、数学或语言学领域的专业人士，《自然语言处理原理与技术实现》是难得的入门教材。

作者介绍:

罗刚，猎兔搜索创始人，带领猎兔搜索技术开发团队先后开发出猎兔中文分词系统、猎兔信息提取系统、猎兔智能垂直搜索系统以及网络信息监测系统，实现互联网信息的采集、过滤、搜索和实时监测。曾编写出版《自己动手写搜索引擎》、《自己动手写网络爬虫》、《使用C#开发搜索引擎》，获得广泛好评。在北京和上海等地均有猎兔培训的学员。张子宪，聊城大学教师、中国矿业大学（北京）博士生，研究方向：自动句法分析、机器翻译。

目录: 第1章 应用自然语言处理技术 1

1.1 付出与回报 2

1.1.1 如何开始 2

1.1.2 招聘人员 2

1.1.3 学习 3

1.2 开发环境 3

1.3 技术基础 4

1.3.1 Java 4

1.3.2 规则方法 5

1.3.3 统计方法 5

1.3.4 计算框架 5

1.3.5 文本挖掘 7

1.3.6 语义库 7

1.4 本章小结 9

1.5 专业术语 9

第2章 中文分词原理与实现 11

2.1 接口 12

2.1.1 切分方案 13

2.1.2 词特征 13

2.2 查找词典算法 13

2.2.1 标准Trie树 14

2.2.2 三叉Trie树 18

2.2.3 词典格式 26

2.3 最长匹配中文分词 27

2.3.1 正向最大长度匹配法 28

2.3.2 逆向最大长度匹配法 33

2.3.3 处理未登录串 39

2.3.4 开发分词 43

2.4 概率语言模型的分词方法 45

2.4.1 一元模型 47

2.4.2 整合基于规则的方法 54

2.4.3	表示切分词图	55
2.4.4	形成切分词图	62
2.4.5	数据基础	64
2.4.6	改进一元模型	75
2.4.7	二元词典	79
2.4.8	完全二叉树组	85
2.4.9	三元词典	89
2.4.10	N元模型	90
2.4.11	N元分词	91
2.4.12	生成语言模型	99
2.4.13	评估语言模型	100
2.4.14	概率分词的流程与结构	101
2.4.15	可变长N元分词	102
2.4.16	条件随机场	103
2.5	新词发现	103
2.5.1	成词规则	109
2.6	词性标注	109
2.6.1	数据基础	114
2.6.2	隐马尔可夫模型	115
2.6.3	存储数据	124
2.6.4	统计数据	131
2.6.5	整合切分与词性标注	133
2.6.6	大词表	138
2.6.7	词性序列	138
2.6.8	基于转换的错误学习方法	138
2.6.9	条件随机场	141
2.7	词类模型	142
2.8	未登录词识别	144
2.8.1	未登录人名	144
2.8.2	提取候选人名	145
2.8.3	最长人名切分	153
2.8.4	一元概率人名切分	153
2.8.5	二元概率人名切分	156
2.8.6	未登录地名	159
2.8.7	未登录企业名	160
2.9	平滑算法	160
2.10	机器学习的方法	164
2.10.1	最大熵	165
2.10.2	条件随机场	170
2.11	有限状态机	171
2.12	地名切分	178
2.12.1	识别未登录地名	179
2.12.2	整体流程	185
2.13	企业名切分	187
2.13.1	识别未登录词	188
2.13.2	整体流程	190
2.14	结果评测	190
2.15	本章小结	191
2.16	专业术语	193
第3章	英文分析	194
3.1	分词	194
3.1.1	句子切分	194
3.1.2	识别未登录串	197
3.1.3	切分边界	198

3.2	词性标注	199
3.3	重点词汇	202
3.4	句子时态	203
3.5	本章小结	204
第4章	依存文法分析	205
4.1	句法分析树	205
4.2	依存文法	211
4.2.1	中文依存文法	211
4.2.2	英文依存文法	220
4.2.3	生成依存树	232
4.2.4	遍历	235
4.2.5	机器学习的方法	237
4.3	小结	237
4.4	专业术语	238
第5章	文档排重	239
5.1	相似度计算	239
5.1.1	夹角余弦	239
5.1.2	最长公共子串	242
5.1.3	同义词替换	246
5.1.4	地名相似度	248
5.1.5	企业名相似度	251
5.2	文档排重	251
5.2.1	关键词排重	251
5.2.2	SimHash	254
5.2.3	分布式文档排重	268
5.2.4	使用文本排重	269
5.3	在搜索引擎中使用文本排重	269
5.4	本章小结	270
5.5	专业术语	270
第6章	信息提取	271
6.1	指代消解	271
6.2	中文关键词提取	273
6.2.1	关键词提取的基本方法	273
6.2.2	HITS算法应用于关键词提取	275
6.2.3	从网页中提取关键词	277
6.3	信息提取	278
6.3.1	提取联系方式	280
6.3.2	从互联网提取信息	281
6.3.3	提取地名	282
6.4	拼写纠错	283
6.4.1	模糊匹配问题	285
6.4.2	正确词表	296
6.4.3	英文拼写检查	298
6.4.4	中文拼写检查	300
6.5	输入提示	302
6.6	本章小结	303
6.7	专业术语	303
第7章	自动摘要	304
7.1	自动摘要技术	305
7.1.1	英文文本摘要	307
7.1.2	中文文本摘要	309
7.1.3	基于篇章结构的自动摘要	314
7.1.4	句子压缩	314
7.2	指代消解	314

7.3 Lucene中的动态摘要	314
7.4 本章小结	317
7.5 专业术语	318
第8章 文本分类	319
8.1 地名分类	321
8.2 错误类型分类	321
8.3 特征提取	322
8.4 关键词加权法	326
8.5 朴素贝叶斯	330
8.6 贝叶斯文本分类	336
8.7 支持向量机	336
8.7.1 多级分类	345
8.7.2 规则方法	347
8.7.3 网页分类	350
8.8 最大熵	351
8.9 信息审查	352
8.10 文本聚类	353
8.10.1 K均值聚类方法	353
8.10.2 K均值实现	355
8.10.3 深入理解DBScan算法	359
8.10.4 使用DBScan算法聚类实例	361
8.11 本章小结	363
8.12 专业术语	363
第9章 文本倾向性分析	364
9.1 确定词语的褒贬倾向	367
9.2 实现情感识别	368
9.3 本章小结	372
9.4 专业术语	373
第10章 问答系统	374
10.1 问答系统的结构	375
10.1.1 提取问答对	376
10.1.2 等价问题	376
10.2 问句分析	377
10.2.1 问题类型	377
10.2.2 句型	381
10.2.3 业务类型	381
10.2.4 依存树	381
10.2.5 指代消解	383
10.2.6 二元关系	383
10.2.7 逻辑表示	386
10.2.8 问句模板	386
10.2.9 结构化问句模板	389
10.2.10 检索方式	390
10.2.11 问题重写	395
10.2.12 提取事实	395
10.2.13 验证答案	398
10.2.14 无答案的处理	398
10.3 知识库	398
10.4 聊天机器人	399
10.4.1 交互式问答	401
10.4.2 垂直领域问答系统	402
10.4.3 语料库	405
10.4.4 客户端	405
10.5 自然语言生成	405

10.6 依存句法	406
10.7 提取同义词	410
10.7.1 流程	410
10.8 本章小结	411
10.9 术语表	412
第11章 语音识别	413
11.1 总体结构	414
11.1.1 识别中文	416
11.1.2 自动问答	417
11.2 语音库	418
11.3 语音合成	419
11.3.1 归一化	420
11.4 语音	420
11.4.1 标注	424
11.4.2 相似度	424
11.5 Sphinx	424
11.5.1 中文训练集	426
11.6 Julius	429
11.7 本章小结	429
11.8 术语表	429
参考资源	430
后记	431
• • • • •	(收起)

[自然语言处理原理与技术实现_下载链接1](#)

标签

自然语言处理

语言处理

计算机科学

理论不深入

NLP

算法

代码太简单

评论

也就分词部分可以看看...后面的感觉都在拼凑...

书很烂，不过关于这一块也找不到太多好书。

全书11章 涉及文本挖掘的各个方面 除原理外 还有Java代码 案例以中文举例介绍的比较全面 但较浅

很一般 java相关 理论性不强

这本书是我读过体验最差的，除了开篇对于统计模型还算清楚，后面就是很杂乱的堆砌了，有些原理都没讲清楚，就直接上代码，章节之间的关系也比较混乱，有些章节讲的也太简单了，强烈不推荐

书名说是讲“自然语言处理”，但内容有点不太够，花近一半笔墨讲了中文分词，其他着墨太少且没讲清楚。章节结构关系混乱，很多东西前后不连贯，不成体系。逻辑性好的人，可以拿来当NLP入门读物，特别是中文分词part。如果有其他更好的选择，这本书的优先级一定是排在后面的那个……

给个一颗星吧，好歹也花了不少时间写的。首先书的印刷使用的纸张质量是不错的，但是体验很差，书中内容用三个字总结就是：脏、乱、差。体验比较差，问了作者一个问题就被T出群了。

理论不深入，讲得太泛，很多章节都是一带而过，而很简单的东西又用大量的篇幅去讲

。书中的java代码完全是为了凑页数，

很差，感觉很乱，所以才到豆瓣上来看看大家对本书的评价，以为是自己没基础看不懂。大家有读到什么好的相关书籍吗？急需。先拜谢各位大神了

体验非常差，近半的篇幅讲述了分词这已相对成熟的技术，书中并没有具体代码，几乎所有代码块只有主要方法体，其中调用的其余方法实现方式一概没有。本以为因为篇幅原因只写了重要的，在书的前言中提到读者群提供书中相关代码，下载来看发现只有第一第二章节的部分代码，并且杂乱无章，更像是草稿。作者对于读者所有问题一律不答不回应，后来得知本书的目的仅是为了宣传培训。总之很不推荐

读完本书既没有对自然语言处理领域有那种高屋建瓴的通透感，如同盖房子弄不清框架，又没有对某件具体的任务有充分的解析，如同盖房子但不知道怎么垒砖。读完后，总之一句话，在网上看博文写出来的都比这个强！

[自然语言处理原理与技术实现_下载链接1](#)

书评

[自然语言处理原理与技术实现_下载链接1](#)