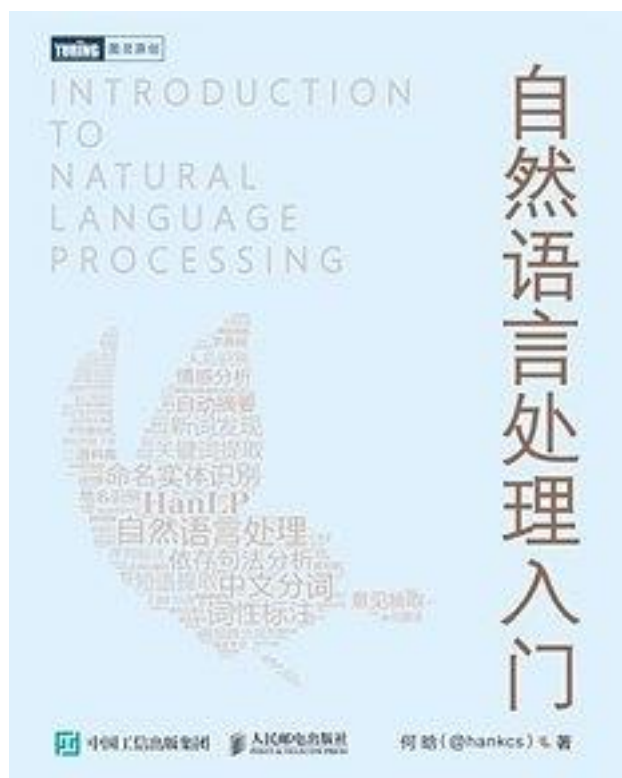


自然语言处理入门



[自然语言处理入门_下载链接1](#)

著者:何晗

出版者:人民邮电出版社

出版时间:2019-10

装帧:平装

isbn:9787115519764

这是一本务实的入门书，助你零起点上手自然语言处理。

HanLP

作者何晗汇集多年经验，从基本概念出发，逐步介绍中文分词、词性标注、命名实体识别、信息抽取、文本聚类、文本分类、句法分析这几个热门问题的算法原理与工程实现。书中通过对多种算法的讲解，比较了它们的优缺点和适用场景，同时详细演示生产级成熟代

码，助你真正将自然语言处理应用在生产环境中。

随着本书的学习，你将从普通程序员晋级为机器学习工程师，最后进化到自然语言处理工程师。

作者介绍:

何晗 (@hankcs)

自然语言处理类库 HanLP 作者 (GitHub 加星超过 14 600)， “码农场” 博主 (日活跃读者数超过 3000)， 埃默里大学计算机博士生，研究方向是句法分析、语义分析与问答系统。

HanLP 和 “码农场” 是 NLP 领域实用的学习资源，何晗大约每周处理一次 HanLP GitHub 上的 Issues。

目录: 第1章新手上路1

1.1 自然语言与编程语言.2

1.1.1 词汇量.2

1.1.2 结构化.2

1.1.3 歧义性.3

1.1.4 容错性.3

1.1.5 易变性.4

1.1.6 简略性.4

1.2 自然语言处理的层次.4

1.2.1 语音、图像和文本..5

1.2.2 中文分词、词性标注和命名实体识别.5

1.2.3 信息抽取.6

1.2.4 文本分类与文本聚类..6

1.2.5 句法分析.6

1.2.6 语义分析与篇章分析..7

1.2.7 其他高级任务.7

1.3 自然语言处理的流派.8

1.3.1 基于规则的专家系统..8

1.3.2 基于统计的学习方法..9

1.3.3 历史.9

1.3.4 规则与统计.11

1.3.5 传统方法与深度学习.11

1.4 机器学习..12

1.4.1 什么是机器学习.13

1.4.2 模型..13

1.4.3 特征..13

1.4.4 数据集..15

1.4.5 监督学习..16

1.4.6 无监督学习..17

1.4.7 其他类型的机器学习算法..18

1.5 语料库.19

1.5.1 中文分词语料库.19

1.5.2 词性标注语料库.19

1.5.3 命名实体识别语料库.20

1.5.4 句法分析语料库.20

1.5.5 文本分类语料库.20

1.5.6语料库建设.	21
1.6开源工具..	21
1.6.1主流NLP工具比较..	21
1.6.2Python接口	23
1.6.3Java接口.	28
1.7总结.	31
第2章词典分词	32
2.1什么是词..	32
2.1.1词的定义..	32
2.1.2词的性质--齐夫定律..	33
2.2词典.	34
2.2.1HanLP词典.	34
2.2.2词典的加载.	34
2.3切分算法..	36
2.3.1完全切分..	36
2.3.2正向最长匹配.	37
2.3.3逆向最长匹配.	39
2.3.4双向最长匹配.	40
2.3.5速度评测..	43
2.4字典树	46
2.4.1什么是字典树.	46
2.4.2字典树的节点实现	47
2.4.3字典树的增删改查实现..	48
2.4.4首字散列其余二分的字典树.	50
2.4.5前缀树的妙用.	53
2.5双数组字典树	55
2.5.1双数组的定义.	55
2.5.2状态转移..	56
2.5.3查询..	56
2.5.4构造*	57
2.5.5全切分与最长匹配	60
2.6AC自动机..	60
2.6.1从字典树到AC自动机	61
2.6.2goto表	61
2.6.3output表..	62
2.6.4fail表	63
2.6.5实现..	65
2.7基于双数组字典树的AC自动机.	67
2.7.1原理..	67
2.7.2实现..	67
2.8HanLP的词典分词实现	71
2.8.1DoubleArrayTrieSegment	72
2.8.2AhoCorasickDoubleArrayTrie-Segment.	73
2.9准确率评测.	74
2.9.1准确率..	74
2.9.2混淆矩阵与TP/FN/FP/TN..	75
2.9.3精确率..	76
2.9.4召回率..	76
2.9.5F1值..	77
2.9.6中文分词中的P、R、F1计算..	77
2.9.7实现..	78
2.9.8第二届国际中文分词评测..	79
2.9.9OOVRecallRate与IVRecallRate.	81
2.10字典树的其他应用.	83

2.10.1停用词过滤..83	83
2.10.2简繁转换87	87
2.10.3拼音转换90	90
2.11总结.91	91
第3章二元语法与中文分词.92	92
3.1语言模型..92	92
3.1.1什么是语言模型92	92
3.1.2马尔可夫链与二元语法..94	94
3.1.3n元语法..95	95
3.1.4数据稀疏与平滑策略96	96
3.2中文分词语料库.96	96
3.2.11998年《人民日报》语料库PKU.97	97
3.2.2微软亚洲研究院语料库MSR98	98
3.2.3繁体中文分词语料库98	98
3.2.4语料库统计.99	99
3.3训练.100	100
3.3.1加载语料库..101	101
3.3.2统计一元语法..101	101
3.3.3统计二元语法..103	103
3.4预测..104	104
3.4.1加载模型104	104
3.4.2构建词网107	107
3.4.3节点间的距离计算111	111
3.4.4词图上的维特比算法.112	112
3.4.5与用户词典的集成115	115
3.5评测..118	118
3.5.1标准化评测..118	118
3.5.2误差分析118	118
3.5.3调整模型119	119
3.6日语分词122	122
3.6.1日语分词语料..122	122
3.6.2训练日语分词器.123	123
3.7总结..124	124
第4章隐马尔可夫模型与序列标注.125	125
4.1序列标注问题.125	125
4.1.1序列标注与中文分词.126	126
4.1.2序列标注与词性标注.127	127
4.1.3序列标注与命名实体识别128	128
4.2隐马尔可夫模型..129	129
4.2.1从马尔可夫假设到隐马尔可夫模型129	129
4.2.2初始状态概率向量.130	130
4.2.3状态转移概率矩阵.131	131
4.2.4发射概率矩阵..132	132
4.2.5隐马尔可夫模型的三个基本用法..133	133
4.3隐马尔可夫模型的样本生成133	133
4.3.1案例--医疗诊断.133	133
4.3.2样本生成算法..136	136
4.4隐马尔可夫模型的训练..138	138
4.4.1转移概率矩阵的估计.138	138
4.4.2初始状态概率向量的估计139	139
4.4.3发射概率矩阵的估计.140	140
4.4.4验证样本生成与模型训练141	141
4.5隐马尔可夫模型的预测..142	142
4.5.1概率计算的前向算法.142	142

4.5.2搜索状态序列的维特比算法..	143
4.6隐马尔可夫模型应用于中文分词.	147
4.6.1标注集	148
4.6.2字符映射	149
4.6.3语料转换	150
4.6.4训练	151
4.6.5预测	152
4.6.6评测	153
4.6.7误差分析	154
4.7二阶隐马尔可夫模型*	154
4.7.1二阶转移概率张量的估计	155
4.7.2二阶隐马尔可夫模型中的维特比算法	156
4.7.3二阶隐马尔可夫模型应用于中文分词	158
4.8总结..	159
第5章感知机分类与序列标注.	160
5.1分类问题	160
5.1.1定义	160
5.1.2应用	161
5.2线性分类模型与感知机算法	161
5.2.1特征向量与样本空间.	162
5.2.2决策边界与分离超平面	164
5.2.3感知机算法..	167
5.2.4损失函数与随机梯度下降*	169
5.2.5投票感知机和平均感知机	171
5.3基于感知机的人名性别分类	174
5.3.1人名性别语料库.	174
5.3.2特征提取	174
5.3.3训练	175
5.3.4预测	176
5.3.5评测	177
5.3.6模型调优	178
5.4结构化预测问题..	180
5.4.1定义	180
5.4.2结构化预测与学习的流程	180
5.5线性模型的结构化感知机算法..	180
5.5.1结构化感知机算法.	180
5.5.2结构化感知机与序列标注	182
5.5.3结构化感知机的维特比解码算法..	183
5.6基于结构化感知机的中文分词..	186
5.6.1特征提取	187
5.6.2多线程训练..	189
5.6.3特征裁剪与模型压缩*.	190
5.6.4创建感知机分词器.	192
5.6.5准确率与性能..	194
5.6.6模型调整与在线学习*.	195
5.6.7中文分词特征工程*.	197
5.7总结..	199
第6章条件随机场与序列标注.	200
6.1机器学习的模型谱系	200
6.1.1生成式模型与判别式模型	201
6.1.2有向与无向概率图模型	202
6.2条件随机场..	205
6.2.1线性链条件随机场.	205
6.2.2条件随机场的训练*	207

6.2.3对比结构化感知机	210
6.3条件随机场工具包	212
6.3.1CRF++的安装	212
6.3.2CRF++语料格式	213
6.3.3CRF++特征模板	214
6.3.4CRF++命令行训练	215
6.3.5CRF++模型格式*	216
6.3.6CRF++命令行预测	217
6.3.7CRF++代码分析*	218
6.4HanLP中的CRF++API	220
6.4.1训练分词器	220
6.4.2标准化评测	220
6.5总结	221
第7章词性标注	222
7.1词性标注概述	222
7.1.1什么是词性	222
7.1.2词性的用处	223
7.1.3词性标注	223
7.1.4词性标注模型	223
7.2词性标注语料库与标注集	224
7.2.1《人民日报》语料库与PKU标注集	225
7.2.2国家语委语料库与863标注集	231
7.2.3《诛仙》语料库与CTB标注集	234
7.3序列标注模型应用于词性标注	236
7.3.1基于隐马尔可夫模型的词性标注	237
7.3.2基于感知机的词性标注	238
7.3.3基于条件随机场的词性标注	240
7.3.4词性标注评测	241
7.4自定义词性	242
7.4.1朴素实现	242
7.4.2标注语料	243
7.5总结	244
第8章命名实体识别	245
8.1概述	245
8.2基于规则的命名实体识别	246
8.3命名实体识别语料库	250
8.4基于层叠隐马尔可夫模型的角色标注框架	252
8.5基于序列标注的命名实体识别	260
8.6自定义领域命名实体识别	266
8.7总结	268
第9章信息抽取	270
9.1新词提取	270
9.2关键词提取	276
9.3短语提取	283
9.4关键句提取	284
9.5总结	287
第10章文本聚类	288
10.1概述	288
10.2文档的特征提取	291
10.3k均值算法	293
10.4重复二分聚类算法	300
10.5标准化评测	303
10.6总结	305
第11章文本分类	306

11.1 文本分类的概念 306
11.2 文本分类语料库 307
11.3 文本分类的特征提取 308
11.4 朴素贝叶斯分类器 312
11.5 支持向量机分类器 317
11.6 标准化评测 320
11.7 情感分析 321
11.8 总结 323
第12章 依存句法分析 324
12.1 短语结构树 324
12.1.3 宾州树库和中文树库 326
12.2 依存句法树 327
12.3 依存句法分析 333
12.4 基于转移的依存句法分析 334
12.5 依存句法分析 API 340
12.6 案例：基于依存句法树的意见抽取 342
12.7 总结 344
第13章 深度学习与自然语言处理 345
13.1 传统方法的局限 345
13.2 深度学习与优势 348
13.3 word2vec 353
13.4 基于神经网络的高性能依存句法分析器 360
13.5 自然语言处理进阶 363
自然语言处理学习资料推荐 365
• • • • • ([收起](#))

[自然语言处理入门_下载链接1](#)

标签

NLP

自然语言处理

人工智能

机器学习

入门

AI

计算机

技术

评论

这本书这么多大佬推荐，但是很没意思啊

简单明白易懂，最喜欢这种极简风格，书的排版也不错，是双色的，赏心悦目，今年最喜欢的一本技术书。

喵完了mantch的读书笔记。可以说是入门中的入门，且偏研究方向。至于实战方向，并不是简单代码能够解决的。可以想见HanLP的作者，肯定擅长理论+工程，但对业务场景的理解却极为有限。如何落地NLP，如何通过实践将NLP的各项技术统筹起来，均未曾提及。极客时间《NLP实战高手课》中提及了「结构化」数据挖掘。「理解」是为了获取「信息」。从「非结构化」数据中提取出「结构化」数据，并提取出信息，进而转化成决策，乃是一大应用场景。NLP只是一种技术和工具，如何用好兵器，打造好武器库，是需要思考的问题。

作为一本能马上上手的NLP入门书确实不错。

应该叫《HanLP自然语言入门》。我并不想如果脱离HanLP 我就不懂 NLP 了，以及我并不想看这本书还顺带看那些占了30%的 Java 代码，虽然也是解释 HanLP 的实现思路，但是理论没懂怎么入门。我只是冲着书名，想入门，这书没达到，心里预期落差有点大。不如还是啃manning去了。

与其说入门，不如说实战。与其说实战不如说是推销。
反正和入门两个字关系真的不大。
中文的书感觉都是这个德性，堆知识，累公式，无讲解，self-contained更是无从说起

不用尬吹，实际上没那么好，作为一本入门书，理论偏少（占全书不到30%），大部分是HanLP的示例代码（Java一份、Python一份），给什么都不懂的人科普NLP的确不错，但还远达不到入门书的理论水平，入门仅靠这一本书是不够的。
(另外发现有刷分嫌疑，一星不谢)

總的來講是一本教材，把nlp會用到的相關技術，模型之類的都簡單講了一遍。如果完全沒有接觸過！nlp領域的人，看這本書入門應該是挺適合的

利益相关：本书的执行编辑英子。认识我的人可能还是有的，我就不匿了。不匿的另外一个原因是，我想为自己执行的书，也是为自己认可的内容和作者站台。这本书的定位是想成为——“从来不了解 NLP，但是想了解这个领域，或者想入门这个领域的读者”的靠谱首选参考书——也就是这本书封面书背上的 Slogan，你一定能读懂的 NLP 图书。不论是作者的内容，还是图书的排版装帧设计，还是图书周边服务的提供，我们都尽可能地做到了最好。如果有任何问题，欢迎随时找我。图灵社区
<https://www.ituring.com.cn/book/2706> 这个页面可以联系我。

自然语言处理的书本来就不多，读起来这么轻松的还是头一本，学了不少入门的基础知识。听说作者还在读博，但是已经维护了 HanLP 这种项目，挺佩服的。

买来作者的这本书看了大约两个星期了，不得不说这本书是真的很适合入门。相比宗成庆老师的统计自然语言处理这种综述一类的书，里面只要一涉及到数学知识我就看不下去了。。。何啥大大的这本书对我这种数学不太行的人友好多了，而且计算机的人嘛，书里面没有代码看着也不舒服，一边看书一边敲敲java,python代码可以说是很不错了

非常好的入门书籍。我是一个刚转计算机的门外汉，自己学过一些基本的数据结构和算法。想学NLP时，面对如海的资料，苦于不知道从哪里开始。正好看到了图灵刚出版的这本书，买来一读觉得很受益。感觉整个NLP的知识体系得以搭建，而且从书中学到很多工程中实用的技巧。读代码时顺便把java也熟悉了。自己上手做NLP任务时，随书的代码和示例中可以学到很多，其中遇到的问题还可以在hankcs的bbs中与作者和其他读者讨论，也会得到耐心的解答。这样的社区让我非常受益。而且作者一直在维护HanLP，现在还推出了2.0版。遇到这么好的开源NLP工具真的很幸运

比较适合我这种零起点的人看，听说是偏向于科普的，可能不适合所有人
详见：<https://bbs.hankcs.com/t/topic/1123>

绝对入门中文自然语言处理的最短路线，语言简练，原理公式来龙去脉讲得一清二楚。前面说中文书都不行的是跪久了站不起来的奴儿吧！

通俗易懂，有理有据，不局限于实战，目前为止看过的最佳NLP入门书。

作为NLP入门书非常棒！配套代码我用的是GitHub最新才行：<https://github.com/hankcs/pyhanlp>

作为一个NLP小白（背景：金融转统计转码），无意中入手了本书（封面很好看颜值高，排名高，购买多），非常适合一个人（弱鸡，泛指毫无CS背景，想要转码挣大钱的小朋友）自学。作者大大的逻辑非常严谨，结合code已经可以初入门槛。了解了一下大佬的背景，更加推荐数学基础薄弱，想要了解NLP的小伙伴。结合大佬的GitHub和Blog食用更佳，<http://hankcs.com> Github:<https://github.com/hankcs> Blog:<https://blog.hankcs.com> 中文：<https://www.hankcs.com> 跟着看完一遍也敲了一遍Code，推荐大家去大神的网站深入学习。很幸运可以用这本书作为NLP启蒙书

易懂

我是Java码农，想转大数据。没什么数学基础，但这本NLP书也能看懂，属于我读过的最好的入门教程

盗号评五星？还能再缺点德吗？ 我送您个1星收好别客气

[自然语言处理入门_下载链接1](#)

书评

[自然语言处理入门_下载链接1](#)