

自然语言处理



[自然语言处理 下载链接1](#)

著者:唐聃

出版者:电子工业出版社

出版时间:2018-7

装帧:平装

isbn:9787121343902

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。

《自然语言处理理论与实战》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。《自然语言处理理论与实战》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。《自然语言处理理

论与实战》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。

《自然语言处理理论与实战》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《自然语言处理理论与实战》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。

作者介绍：

1

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）。

2

白宁超，工学硕士，四川省计算机研究院开发工程师，曾参与国家自然基金项目和四川省科技支撑计划等多个省级项目。

3 冯暄

高级工程师，四川省计算机研究院信息化工程研究所所长。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、三等奖1项。

目录: 第1章基础入门 1

1.1 什么是自然语言处理 1

1.1.1 自然语言处理概述 1

1.1.2 自然语言处理的发展历史 3

1.1.3 自然语言处理的工作原理 6

1.1.4 自然语言处理的应用前景 7

1.2 开发工具与环境 7

1.2.1 Sublime Text 和Anaconda 介绍 7

1.2.2 开发环境的安装与配置 8

1.3 实战：第一个小程序的诞生 13

1.3.1 实例介绍 13

1.3.2 源码实现 13

第2章快速上手Python 15

2.1 初识Python 编程语言 15

2.1.1 Python 概述 15

2.1.2 Python 能做什么 17

2.1.3 Python 的语法和特点 19

2.2 Python 进阶 24

2.2.1 Hello World 24

2.2.2 语句和控制流 24

2.2.3 函数 27

2.2.4 List 列表 29

2.2.5 元组 32

2.2.6 set 集合 33

2.2.7 字典 33

2.2.8 面向对象编程：类 34

2.2.9 标准库 36

2.3 Python 深入——第三方库 36

2.3.1 Web 框架	36
2.3.2 科学计算	37
2.3.3 GUI	37
2.3.4 其他库	37
第3章 线性代数	39
3.1 线性代数介绍	39
3.2 向量	40
3.2.1 向量定义	40
3.2.2 向量表示	42
3.2.3 向量定理	42
3.2.4 向量运算	43
3.3 矩阵	47
3.3.1 矩阵定义	47
3.3.2 矩阵表示	48
3.3.3 矩阵运算	48
3.3.4 线性方程组	51
3.3.5 行列式	51
3.3.6 特征值和特征向量	55
3.4 距离计算	56
3.4.1 余弦距离	56
3.4.2 欧氏距离	57
3.4.3 曼哈顿距离	58
3.4.4 明可夫斯基距离	59
3.4.5 切比雪夫距离	61
3.4.6 杰卡德距离	62
3.4.7 汉明距离	63
3.4.8 标准化欧式距离	64
3.4.9 皮尔逊相关系数	65
第4章 概率论	67
4.1 概率论介绍	67
4.2 事件	68
4.2.1 随机试验	68
4.2.2 随机事件和样本空间	69
4.2.3 事件的计算	70
4.3 概率	71
4.4 概率公理	73
4.5 条件概率和全概率	76
4.5.1 条件概率	76
4.5.2 全概率	77
4.6 贝叶斯定理	78
4.7 信息论	79
4.7.1 信息论的基本概念	79
4.7.2 信息度量	80
第5章 统计学	85
5.1 图形可视化	85
5.1.1 饼图	85
5.1.2 条形图	88
5.1.3 热力图	91
5.1.4 折线图	93
5.1.5 箱线图	96
5.1.6 散点图	99
5.1.7 雷达图	102
5.1.8 仪表盘	104
5.1.9 可视化图表用法	106

5.2 数据度量标准	108
5.2.1 平均值	108
5.2.2 中位数	108
5.2.3 众数	110
5.2.4 期望	111
5.2.5 方差	112
5.2.6 标准差	113
5.2.7 标准分	114
5.3 概率分布	115
5.3.1 几何分布	115
5.3.2 二项分布	116
5.3.3 正态分布	118
5.3.4 泊松分布	121
5.4 统计假设检验	123
5.5 相关和回归	125
5.5.1 相关	125
5.5.2 回归	127
5.5.3 相关和回归的联系	130
第6章 语言学	132
6.1 语音	132
6.1.1 什么是语音	132
6.1.2 语音的三大属性	133
6.1.3 语音单位	134
6.1.4 记音符号	135
6.1.5 共时语流音变	136
6.2 词汇	137
6.2.1 什么是词汇	137
6.2.2 词汇单位	137
6.2.3 词的构造	138
6.2.4 词义及其分类	140
6.2.5 义项与义素	141
6.2.6 语义场	142
6.2.7 词汇的构成	143
6.3 语法	143
6.3.1 什么是语法	143
6.3.2 词类	144
6.3.3 短语	148
6.3.4 单句	150
6.3.5 复句	152
第7章 自然语言处理	155
7.1 自然语言处理的任务和限制	155
7.2 自然语言处理的主要技术范畴	156
7.2.1 语音合成	156
7.2.2 语音识别	156
7.2.3 中文自动分词	157
7.2.4 词性标注	158
7.2.5 句法分析	158
7.2.6 文本分类	159
7.2.7 文本挖掘	160
7.2.8 信息抽取	161
7.2.9 问答系统	161
7.2.10 机器翻译	162
7.2.11 文本情感分析	163
7.2.12 自动摘要	164

7.2.13 文字蕴涵	165
7.3 自然语言处理的难点	165
7.3.1 语言环境复杂	165
7.3.2 文本结构形式多样	166
7.3.3 边界识别限制	166
7.3.4 词义消歧	167
7.3.5 指代消解	168
7.4 自然语言处理展望	169
第8章语料库	173
8.1 语料库浅谈	173
8.2 语料库深入	174
8.3 自然语言处理工具包：NLTK	176
8.3.1 NLTK 简介	176
8.3.2 安装NLTK	177
8.3.3 使用NLTK	180
8.3.4 在Python NLTK 下使用Stanford NLP	186
8.4 获取语料库	194
8.4.1 国内外著名语料库	195
8.4.2 网络数据获取	197
8.4.3 NLTK 获取语料库	200
8.5 综合案例：走进大秦帝国	208
8.5.1 数据采集和预处理	208
8.5.2 构建本地语料库	208
8.5.3 大秦帝国语料操作	209
第9章中文自动分词	216
9.1 中文分词简介	216
9.2 中文分词的特点和难点	218
9.3 常见中文分词方法	219
9.4 典型中文分词工具	220
9.4.1 HanLP 中文分词	220
9.4.2 其他中文分词工具	223
9.5 结巴中文分词	224
9.5.1 基于Python 的结巴中文分词	224
9.5.2 结巴分词工具详解	227
9.5.3 结巴分词核心内容	230
9.5.4 结巴分词基本用法	233
第10章数据预处理	241
10.1 数据清洗	241
10.2 分词处理	242
10.3 特征构造	242
10.4 特征降维与选择	243
10.4.1 特征降维	243
10.4.2 特征选择	243
10.5 简单实例	244
10.6 本章小结	249
第11章马尔可夫模型	250
11.1 马尔可夫链	250
11.1.1 马尔可夫简介	250
11.1.2 马尔可夫链的基本概念	251
11.2 隐马尔可夫模型	253
11.2.1 形式化描述	253
11.2.2 数学形式描述	255
11.3 向前算法解决HMM 似然度	256
11.3.1 向前算法定义	256

11.3.2 向前算法原理	256
11.3.3 现实应用：预测成都天气的冷热	258
11.4 文本序列标注案例：Viterbi 算法	259
第12章 条件随机场	263
12.1 条件随机场介绍	263
12.2 简单易懂的条件随机场	265
12.2.1 CRF 的形式化表示	265
12.2.2 CRF 的公式化表示	266
12.2.3 深度理解条件随机场	268
第13章 模型评估	269
13.1 从统计角度介绍模型概念	269
13.1.1 算法模型	269
13.1.2 模型评估和模型选择	270
13.1.3 过拟合与欠拟合的模型选择	272
13.2 模型评估与选择	275
13.2.1 模型评估的概念	275
13.2.2 模型评估的评测指标	275
13.2.3 以词性标注为例分析模型评估	276
13.2.4 模型评估的几种方法	278
13.3 ROC 曲线比较学习器模型	279
第14章 命名实体识别	281
14.1 命名实体识别概述	281
14.2 命名实体识别的特点与难点	284
14.3 命名实体识别方法	284
14.4 中文命名实体识别的核心技术	286
14.5 展望	295
第15章 自然语言处理实战	296
15.1 GitHub 数据提取与可视化分析	296
15.1.1 了解GitHub 的API	296
15.1.2 使用NetworkX 作图	299
15.1.3 使用NetworkX 构建兴趣图	301
15.1.4 NetWorkX 部分统计指标	304
15.1.5 构建GitHub 的兴趣图	305
15.1.6 可视化	318
15.2 微博话题爬取与存储分析	320
15.2.1 数据采集	320
15.2.2 数据提取	329
15.2.3 数据存储	332
15.2.4 项目运行与分析	333
附录A Python 与其他语言调用	337
附录B Git 项目上传简易教程	339
参考文献	341
· · · · · (收起)	

[自然语言处理](#) [下载链接1](#)

标签

NLP

自然语言处理

实用

比较适合初学者

容易上手

浅显易懂

架构清晰

代码齐全

评论

实体书，无笔记 书中错误不少，囫囵吞枣，枣还有不少是坏的

数学概念清楚，同时比较清楚地介绍了马尔科夫模型及条件随机场/数据预处理等概念(HMM/CRF推导需要再了解下) 理论清晰，没啥实战指导，没什么代码，

前一半是废话

基本全是废话和原理推导 这书分数分布很奇怪，怀疑五星都是作者的学生打的

我错了 我不该看这本书

很空泛的一本书，看了前8章，都还是很空泛。一大半在讲数学相关知识，一章讲了点综述，还有一章介绍工具怎么使用。。。里面错误也不少

面面俱到，所涉不深；代码极少，python为主；尾篇还夹杂了爬虫，浪费纸张。给7分吧

很不幸购买了这本书，不仅废话多，错误多。没什么营养，作者是怀着什么样的心态写了这本书呢

八个字形容：删繁就简

，举重若轻。刚刚入手一本，不得不说，这本书适合我这种入门机器学习的小白学习。特别是急于看到效果想要实战一下的。是一本较为实用的入门书籍

很好的一本书，从入门到精通，深入浅出，涉及各方面的知识，对自然语言这块有很好的帮助，案例+讲解，初学者也可以很快上手

这本书总体架构比较清晰，

针对初学者来说真挺合适的。首先，书中系统的描述了自然语言处理所涉及到的理论知识、技术、工具等，可以让你对NLP有个系统性的知识了解。再次，书中所提到的很多技术知识点，都有较详细的例子以及核心代码，可以让你充分的学习并操练nlp知识技术（不像有些书籍，写得很泛，看似很高端，但看过之后，很多基础的知识都可能没帮助你搞懂，意思何在？）。最后，感觉这本书比较偏工程，因为文中提到了很多实际项目工程中的NLP处理方式或者处理技巧，这无疑是对没有系统学习过nlp理论知识的开发者来说，对其工作应该会有很大的帮助。

结合Python的语言特性，循序渐进介绍了很多自然语言处理方面的知识，以及如何使用nltk进行NLP。再次体会到Python在文本处理方面有着很多先天优势。如果你要是自然语言学或语言信息处理相关专业的学生，又对python与nltk感兴趣的话，强烈推荐这本书，可以当做入门读物来看。

这本书非常适合初学者，其最大的特点就是不仅仅介绍理论知识，而是从自然语言处理

的角度出发，详细介绍整个体系，并从零开始介绍需要的数学知识，工程基础，语言知识等等。涵盖了包括线性代数、概率论、统计学等多方面的知识。并且结合实际，给出了很多实际应用案例及代码。总体来说非常适合自学机器学习、数据挖掘、文本分析等的初学者。

对于一个想学习大数据相关的知识，却又茫然不知所措的初学者来说很适用，这本书深入浅出的讲解了专业性的知识，有理论有示例有代码，提升了我的理论认知又让我手动操作增强了实践性，就觉着这书很赞，极力推荐啊。

很不错的一本书，内容全面，思路清晰，读起来易懂，适合成长期的学者阅读。

非常适合对NLP一知半解却急于想入门的朋友，这书无疑是指路明灯。对我这个文科生而言，章节设置上至少让我知道了要去补哪些数理知识。看到语言学章节时的亲切感也让我十分欣喜。语言学内容基本按着黄廖本《现代汉语》来的，虽无太多新意，但例子都算举得恰当，让我这个中文口的挑不出太多毛病。
书中部分地方有讹误空缺之处，希望之后能改进。

适合初学者，理论知识介绍详细，从自然语言处理的角度出发，读者可以详细了解整个体系，所以总体来说非常适合自学机器学习的初学者。

条理清晰，易懂，值得看

总体来说吧，这本书还算可以。之前买过几本自然语言处理中文书，看不懂太深奥了。很多没有代码，对于我这样的新手不太容易。这本书案例跟代码都一样。不过书本上下载的代码地址有问题，我还是联系到官网找到的。不足的就是连接有的不对，这一点不爽。

之前一直做java技术开发，接触一点python。想学自然语言处理，不知道如何入门，这本书非常有用，真的很赞。

书评

什么是大数据，什么是自然语言处理，如果对这两个概念不是很了解的初学者，这本书是一个不错的选择。作者从一个基础的学者的角度出发，给读者从基础到实践的流程化讲解，不仅适合代码水平一流的码农，更适合动手能力一般的学生。理论与实践相结合，使得讲解更加清楚明了，也使...

我也是计算机专业的哈。就书的目录而言，读起来一目了然，结构清晰，这在机器学习和开始就得了不少。本人还是机器学习的小白，很多数学基础还在慢慢深入了解。这本书涉及了自然语言处理，帮助我学习机器学习的同时，还可以给予我机器学习和自然语言两者间关联规则的启发。所...

该书讲解的非常细致，对于初学者来说是本很好的入门级教材，书中的案例讲解，由浅入深，加强理论知识的同时，增强实践性操作，作为对代码的动手能力较差的个人，这本书弥补了我的不足，慢慢引导我学习，多动手敲代码，多练习，作为入门级教材，极力推荐！在此衷心希望作者继续...

这本书特别适合自然语言学习的初学者，它的架构清晰，循序渐进，学习的同时帮我重新复习了大学拉下的高等数学，本书从最基础的环境搭建做起，理论知识和案例编码的互相支持，使得枯燥的理论和概念变得通俗易懂，书中案例代码能下载的到，只要静下心来慢慢学，学习自然语言...

这本书内容全面，本书几乎涵盖了自然语言处理所有相关重要的知识点，逻辑很清晰，目前在读，适合对NLP学习处于成长期的学者阅读，同样也适合当个学习路径看。作为数学专业出身，对代码的动手能力较差，这本书也恰好弥补了我的不足，想学习好NLP，就要多动手，多练习，最后希望...

非常适合对NLP一知半解却急于想入门的朋友，这书无疑是指路明灯。对我这个文科生而言，章节设置上至少让我知道了要去补哪些数理知识。看到语言学章节时的亲切感也让我十分欣喜。语言学内容基本按着黄廖本《现代汉语》来的，虽无太多新意，但例子都算举得恰当，让我这个中文口...

作为一个颜控，不得不说最先吸引我的是简洁的书面设计，配色很舒服，看了下出版社，电子工业出版社，很多教材都是出自这个出版社，内容肯定不会差到哪儿去。因为对自然语言颇感兴趣，于是去书店寻找相关书籍，恰巧看到了这本，当时粗略翻了几页，内容深入浅出，可读性很强，于...

该书从基础讲起，对于初学者而言是本很好的教材，由浅入深，讲的非常的通俗易懂，而且还一步一步的讲解了每个算法模型的相关代码，该书使用的编程语言是目前人工智能方向最火的Python编程语言，具有很好的实践性与发展性，详细的讲解了Python编程语言的基础知识，对编程知识较...

正在看，暂时看到第五章，就目前所看的内容来说，关于基础知识的介绍还是很详尽的，也很容易让人理解，浅显易懂。随手翻阅了后面的介绍自然语言处理的内容，描述与代码相匹配，应该会很实用的，最后还有例子的介绍，学完应该对自然语言处理不说百分之百的学会，至少能学会百分...

[自然语言处理 下载链接1](#)