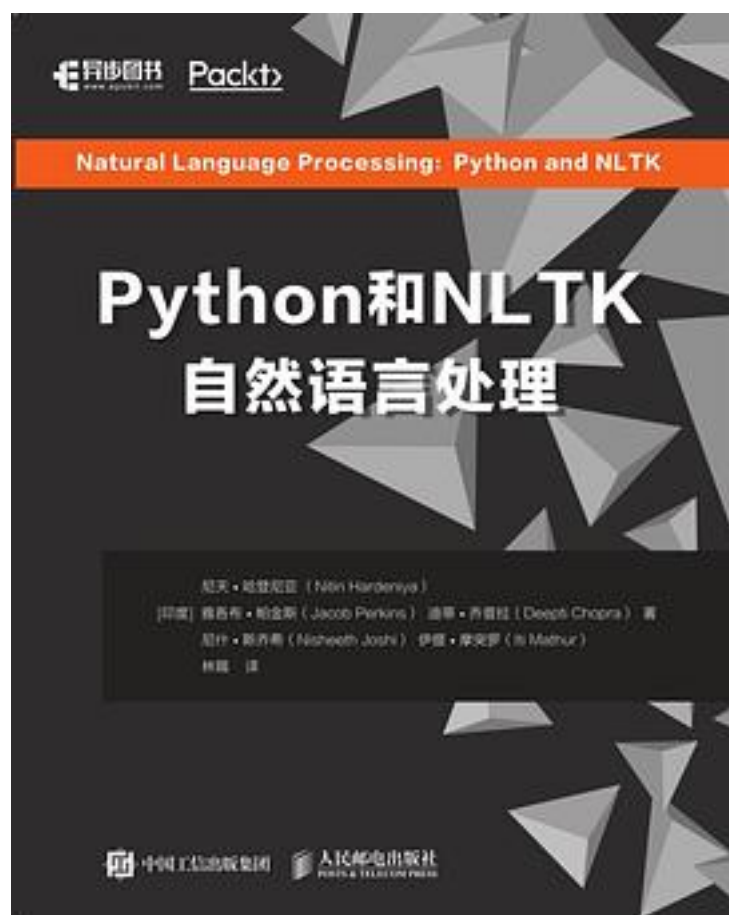


Python和NLTK自然语言处理



[Python和NLTK自然语言处理_下载链接1](#)

著者:[印度] 尼天·哈登尼亚 (Nitin Hardeniya)

出版者:人民邮电出版社

出版时间:2019-4

装帧:平装

isbn:9787115503343

NLTK是自然语言处理领域中非常受欢迎和广泛使用的Python库。NLTK的优点在于其简单性，其中大多数复杂的自然语言处理任务使用几行代码即可完成。本书旨在讲述如何用Python和NLTK解决各种自然语言处理任务并开发机器学习方面的应用。本书介绍了NLTK的基本模块，讲述了采用NLTK实现自然语言处理的大量技巧，讨论了一些文本处

理方法和语言处理技术，展示了使用Python实现NLP项目的大量实践经验。本书主要内容包括文本挖掘/NLP任务中所需的所有预处理步骤，如何使用Python 3的NLTK 3进行文本处理，如何通过Python开展NLP项目。本书适合NLP和机器学习领域的爱好者、Python程序员以及机器学习领域的研究人员阅读。

作者介绍:

Nitin

Hardeniya是一位数据科学家，精通Python、NLTK、机器学习，与多家知名公司都有长期的合作。他的业务范围比较广，擅长解决不同领域的各种业务问题。他发表过5篇专利。

目录: 模块1 NLTK基础知识

第1章 自然语言处理简介 3

1.1 为什么要学习NLP 4

1.2 从Python的基本知识开始 7

1.2.1 列表 7

1.2.2 自助 8

1.2.3 正则表达式 9

1.2.4 词典 11

1.2.5 编写函数 11

1.3 NLTK 13

1.4 试一试 18

1.5 本章小结 18

第2章 文本的整理和清洗 19

2.1 文本整理 19

2.2 文本清洗 21

2.3 句子拆分器 22

2.4 标记解析 22

2.5 词干提取 24

2.6 词形还原 25

2.7 停用词删除 26

2.8 生僻字删除 27

2.9 拼写校正 27

2.10 试一试 28

2.11 本章小结 28

第3章 词性标注 30

3.1 什么是词性标注 30

3.1.1 斯坦福标注器 33

3.1.2 深入了解标注器 34

3.1.3 序列标注器 35

3.1.4 布里尔标注器 37

3.1.5 基于标注器的机器学习 37

3.2 命名实体识别 38

3.3 试一试 40

3.4 本章小结 41

第4章 对文本的结构进行语法分析 42

4.1 浅层语法分析与深层语法分析 42

4.2 语法分析的两种方法 43

4.3 为什么需要语法分析 43

4.4 不同类型的语法分析器	45
4.4.1 递归下降的语法分析器	45
4.4.2 移位归约语法分析器	45
4.4.3 图表语法分析器	45
4.4.4 正则表达式语法分析器	46
4.5 依存分析	47
4.6 组块化	49
4.7 信息抽取	51
4.7.1 命名实体识别	52
4.7.2 关系抽取	52
4.8 本章小结	53
第5章 NLP应用	54
5.1 构建第一个NLP应用	54
5.2 其他的NLP应用	58
5.2.1 机器翻译	58
5.2.2 统计机器翻译	59
5.2.3 信息检索	59
5.2.4 语音识别	61
5.2.5 文本分类	62
5.2.6 信息提取	63
5.2.7 问答系统	64
5.2.8 对话系统	64
5.2.9 词义消歧	64
5.2.10 主题建模	64
5.2.11 语言检测	65
5.2.12 光学字符识别	65
5.3 本章小结	65
第6章 文本分类	66
6.1 机器学习	67
6.2 文本分类	68
6.3 采样	70
6.3.1 朴素贝叶斯	73
6.3.2 决策树	75
6.3.3 随机梯度下降	76
6.3.4 逻辑回归	77
6.3.5 支持向量机	78
6.4 随机森林算法	79
6.5 文本聚类	79
6.6 文本的主题建模	81
6.7 参考资料	83
6.8 本章小结	83
第7章 网络爬取	85
7.1 网络爬虫	85
7.2 编写第一个爬虫程序	86
7.3 Scrapy中的数据流	89
7.3.1 Scrapy命令行界面	89
7.3.2 项	94
7.4 站点地图蜘蛛	96
7.5 项管道	97
7.6 外部参考	98
7.7 本章小结	99
第8章 与其他Python库一同使用NLTK	100

- 8.1 NumPy 100
 - 8.1.1 ndarray 101
 - 8.1.2 基本操作 102
 - 8.1.3 从数组中提取数据 103
 - 8.1.4 复杂的矩阵运算 103
- 8.2 SciPy 107
 - 8.2.1 线性代数 108
 - 8.2.2 特征值和特征向量 108
 - 8.2.3 稀疏矩阵 109
 - 8.2.4 优化 110
- 8.3 Pandas 111
 - 8.3.1 读取数据 112
 - 8.3.2 时序数据 114
 - 8.3.3 列转换 115
 - 8.3.4 噪声数据 116
- 8.4 Matplotlib 117
 - 8.4.1 subplot 118
 - 8.4.2 添加轴 119
 - 8.4.3 散点图 120
 - 8.4.4 柱状图 120
 - 8.4.5 3D图 121
- 8.5 外部参考 121
- 8.6 本章小结 121
- 第9章 使用Python进行社交媒体挖掘 122
 - 9.1 数据收集 122
 - 9.2 数据提取 126
 - 9.3 地理可视化 128
 - 9.3.1 影响者检测 129
 - 9.3.2 Facebook 130
 - 9.3.3 影响者的朋友 134
 - 9.4 本章小结 135
- 第 10章 大规模的文本挖掘 136
 - 10.1 在Hadoop上使用Python的不同方法 136
 - 10.1.1 Python的流 137
 - 10.1.2 Hive/Pig UDF 137
 - 10.1.3 流包装器 137
 - 10.2 在Hadoop上运行NLTK 138
 - 10.2.1 UDF 138
 - 10.2.2 Python流 140
 - 10.3 在Hadoop上运行Scikit-learn 141
 - 10.4 PySpark 144
 - 10.5 本章小结 146
- 模块2 使用Python 3的NLTK 3进行文本处理
- 第 1章 标记文本和WordNet的基础 149
 - 1.1 引言 149
 - 1.2 将文本标记成句子 150
 - 1.2.1 准备工作 150
 - 1.2.2 工作方式 151
 - 1.2.3 工作原理 151
 - 1.2.4 更多信息 151
 - 1.2.5 请参阅 152

- 1.3 将句子标记成单词 152
 - 1.3.1 工作方式 152
 - 1.3.2 工作原理 153
 - 1.3.3 更多信息 153
 - 1.3.4 请参阅 154
- 1.4 使用正则表达式标记语句 154
 - 1.4.1 准备工作 155
 - 1.4.2 工作方式 155
 - 1.4.3 工作原理 155
 - 1.4.4 更多信息 155
 - 1.4.5 请参阅 156
- 1.5 训练语句标记生成器 156
 - 1.5.1 准备工作 156
 - 1.5.2 工作方式 156
 - 1.5.3 工作原理 157
 - 1.5.4 更多信息 158
 - 1.5.5 请参阅 158
- 1.6 在已标记的语句中过滤
停用词 158
 - 1.6.1 准备工作 158
 - 1.6.2 工作方式 159
 - 1.6.3 工作原理 159
 - 1.6.4 更多信息 159
 - 1.6.5 请参阅 160
- 1.7 查找WordNet中单词的
Synset 160
 - 1.7.1 准备工作 160
 - 1.7.2 工作方式 160
 - 1.7.3 工作原理 161
 - 1.7.4 更多信息 161
 - 1.7.5 请参阅 163
- 1.8 在WordNet中查找词元和
同义词 163
 - 1.8.1 工作方式 163
 - 1.8.2 工作原理 163
 - 1.8.3 更多信息 163
 - 1.8.4 请参阅 165
- 1.9 计算WordNet和Synset的
相似度 165
 - 1.9.1 工作方式 165
 - 1.9.2 工作原理 165
 - 1.9.3 更多信息 166
 - 1.9.4 请参阅 167
- 1.10 发现单词搭配 167
 - 1.10.1 准备工作 167
 - 1.10.2 工作方式 167
 - 1.10.3 工作原理 168
 - 1.10.4 更多信息 168
 - 1.10.5 请参阅 169
- 第 2 章 替换和校正单词 170
 - 2.1 引言 170
 - 2.2 词干提取 170
 - 2.2.1 工作方式 171
 - 2.2.2 工作原理 171

2.2.3 更多信息	171
2.2.4 请参阅	173
2.3 使用WordNet进行词形还原	173
2.3.1 准备工作	173
2.3.2 工作方式	173
2.3.3 工作原理	174
2.3.4 更多信息	174
2.3.5 请参阅	175
2.4 基于匹配的正则表达式替换 单词	175
2.4.1 准备工作	175
2.4.2 工作方式	175
2.4.3 工作原理	176
2.4.4 更多信息	177
2.4.5 请参阅	177
2.5 移除重复字符	177
2.5.1 准备工作	177
2.5.2 工作方式	178
2.5.3 工作原理	178
2.5.4 更多信息	179
2.5.5 请参阅	179
2.6 使用Enchant进行拼写校正	180
2.6.1 准备工作	180
2.6.2 工作方式	180
2.6.3 工作原理	181
2.6.4 更多信息	181
2.6.5 请参阅	183
2.7 替换同义词	183
2.7.1 准备工作	183
2.7.2 工作方式	183
2.7.3 工作原理	184
2.7.4 更多信息	184
2.7.5 请参阅	185
2.8 使用反义词替换否定形式	186
2.8.1 工作方式	186
2.8.2 工作原理	187
2.8.3 更多信息	187
2.8.4 请参阅	188
第3章 创建自定义语料库	189
3.1 引言	189
3.2 建立自定义语料库	190
3.2.1 准备工作	190
3.2.2 工作方式	190
3.2.3 工作原理	191
3.2.4 更多信息	192
3.2.5 请参阅	192
3.3 创建词汇表语料库	192
3.3.1 准备工作	192
3.3.2 工作方式	193
3.3.3 工作原理	193
3.3.4 更多信息	194
3.3.5 请参阅	194
3.4 创建已标记词性单词的 语料库	195

3.4.1 准备工作	195
3.4.2 工作方式	195
3.4.3 工作原理	196
3.4.4 更多信息	196
3.4.5 请参阅	199
3.5 创建已组块短语的语料库	199
3.5.1 准备工作	199
3.5.2 工作方式	199
3.5.3 工作原理	201
3.5.4 更多信息	201
3.5.5 请参阅	203
3.6 创建已分类文本的语料库	203
3.6.1 准备工作	204
3.6.2 工作方式	204
3.6.3 工作原理	204
3.6.4 更多信息	205
3.6.5 请参阅	206
3.7 创建已分类组块语料库 读取器	206
3.7.1 准备工作	206
3.7.2 工作方式	207
3.7.3 工作原理	208
3.7.4 更多信息	209
3.7.5 请参阅	213
3.8 懒惰语料库加载	213
3.8.1 工作方式	213
3.8.2 工作原理	214
3.8.3 更多信息	214
3.9 创建自定义语料库视图	215
3.9.1 工作方式	215
3.9.2 工作原理	216
3.9.3 更多信息	217
3.9.4 请参阅	218
3.10 创建基于MongoDB的 语料库读取器	218
3.10.1 准备工作	219
3.10.2 工作方式	219
3.10.3 工作原理	220
3.10.4 更多信息	221
3.10.5 请参阅	221
3.11 在加锁文件的情况下编辑 语料库	221
3.11.1 准备工作	221
3.11.2 工作方式	221
3.11.3 工作原理	222
第4章 词性标注	224
4.1 引言	224
4.2 默认标注	225
4.2.1 准备工作	225
4.2.2 工作方式	225
4.2.3 工作原理	226
4.2.4 更多信息	227
4.2.5 请参阅	228
4.3 训练一元组词性标注器	228

- 4.3.1 工作方式 228
- 4.3.2 工作原理 229
- 4.3.3 更多信息 230
- 4.3.4 请参阅 231
- 4.4 回退标注的组合标注器 231
 - 4.4.1 工作方式 231
 - 4.4.2 工作原理 232
 - 4.4.3 更多信息 232
 - 4.4.4 请参阅 233
- 4.5 训练和组合N元标注器 233
 - 4.5.1 准备工作 233
 - 4.5.2 工作方式 233
 - 4.5.3 工作原理 234
 - 4.5.4 更多信息 235
 - 4.5.5 请参阅 236
- 4.6 创建似然单词标签的模型 236
 - 4.6.1 工作方式 236
 - 4.6.2 工作原理 237
 - 4.6.3 更多信息 237
 - 4.6.4 请参阅 238
- 4.7 使用正则表达式标注 238
 - 4.7.1 准备工作 238
 - 4.7.2 工作方式 238
 - 4.7.3 工作原理 239
 - 4.7.4 更多信息 239
 - 4.7.5 请参阅 239
- 4.8 词缀标签 239
 - 4.8.1 工作方式 239
 - 4.8.2 工作原理 240
 - 4.8.3 更多信息 240
 - 4.8.4 请参阅 241
- 4.9 训练布里尔标注器 241
 - 4.9.1 工作方式 241
 - 4.9.2 工作原理 242
 - 4.9.3 更多信息 243
 - 4.9.4 请参阅 244
- 4.10 训练TnT标注器 244
 - 4.10.1 工作方式 244
 - 4.10.2 工作原理 244
 - 4.10.3 更多信息 245
 - 4.10.4 请参阅 246
- 4.11 使用WordNet进行标注 246
 - 4.11.1 准备工作 246
 - 4.11.2 工作方式 247
 - 4.11.3 工作原理 248
 - 4.11.4 请参阅 248
- 4.12 标注专有名词 248
 - 4.12.1 工作方式 248
 - 4.12.2 工作原理 249
 - 4.12.3 请参阅 249
- 4.13 基于分类器的标注 249
 - 4.13.1 工作方式 250

- 4.13.2 工作原理 250
- 4.13.3 更多信息 251
- 4.13.4 请参阅 252
- 4.14 使用NLTK训练器训练标注器 253
 - 4.14.1 工作方式 253
 - 4.14.2 工作原理 254
 - 4.14.3 更多信息 258
 - 4.14.4 请参阅 260
- 第5章 提取组块 261
 - 5.1 引言 261
 - 5.2 使用正则表达式组块和隔断 262
 - 5.2.1 准备工作 262
 - 5.2.2 工作方式 262
 - 5.2.3 工作原理 263
 - 5.2.4 更多信息 265
 - 5.2.5 请参阅 267
 - 5.3 使用正则表达式合并和拆分组块 267
 - 5.3.1 工作方式 267
 - 5.3.2 工作原理 269
 - 5.3.3 更多信息 270
 - 5.3.4 请参阅 271
 - 5.4 使用正则表达式扩展和删除组块 271
 - 5.4.1 工作方式 271
 - 5.4.2 工作原理 272
 - 5.4.3 更多信息 273
 - 5.4.4 请参阅 273
 - 5.5 使用正则表达式进行部分解析 273
 - 5.5.1 工作方式 273
 - 5.5.2 工作原理 274
 - 5.5.3 更多信息 275
 - 5.5.4 请参阅 276
 - 5.6 训练基于标注器的组块器 276
 - 5.6.1 工作方式 276
 - 5.6.2 工作原理 277
 - 5.6.3 更多信息 278
 - 5.6.4 请参阅 279
 - 5.7 基于分类的分块 279
 - 5.7.1 工作方式 279
 - 5.7.2 工作原理 282
 - 5.7.3 更多信息 282
 - 5.7.4 请参阅 283
 - 5.8 提取命名实体 283
 - 5.8.1 工作方式 283
 - 5.8.2 工作原理 284
 - 5.8.3 更多信息 284
 - 5.8.4 请参阅 285
 - 5.9 提取专有名词组块 285
 - 5.9.1 工作方式 286
 - 5.9.2 工作原理 286

5.9.3 更多信息	286
5.10 提取部位组块	287
5.10.1 工作方式	288
5.10.2 工作原理	290
5.10.3 更多信息	290
5.10.4 请参阅	290
5.11 训练命名实体组块器	290
5.11.1 工作方式	290
5.11.2 工作原理	292
5.11.3 更多信息	292
5.11.4 请参阅	293
5.12 使用NLTK训练器训练 组块器	293
5.12.1 工作方式	293
5.12.2 工作原理	294
5.12.3 更多信息	295
5.12.4 请参阅	299
第6章 转换组块与树	300
6.1 引言	300
6.2 过滤句子中无意义的 单词	301
6.2.1 准备工作	301
6.2.2 工作方式	301
6.2.3 工作原理	302
6.2.4 更多信息	302
6.2.5 请参阅	303
6.3 纠正动词形式	303
6.3.1 准备工作	303
6.3.2 工作方式	303
6.3.3 工作原理	305
6.3.4 请参阅	306
6.4 交换动词短语	306
6.4.1 工作方式	306
6.4.2 工作原理	307
6.4.3 更多信息	307
6.4.4 请参阅	307
6.5 交换名词基数	308
6.5.1 工作方式	308
6.5.2 工作原理	309
6.5.3 请参阅	309
6.6 交换不定式短语	309
6.6.1 工作方式	309
6.6.2 工作原理	310
6.6.3 更多信息	310
6.6.4 请参阅	310
6.7 单数化复数名词	310
6.7.1 工作方式	310
6.7.2 工作原理	311
6.7.3 请参阅	311
6.8 链接组块变换	311
6.8.1 工作方式	311
6.8.2 工作原理	312
6.8.3 更多信息	312
6.8.4 请参阅	313

6.9 将组块树转换为文本	313
6.9.1 工作方式	313
6.9.2 工作原理	314
6.9.3 更多信息	314
6.9.4 请参阅	314
6.10 平展深度树	314
6.10.1 准备工作	315
6.10.2 工作方式	315
6.10.3 工作原理	316
6.10.4 更多信息	317
6.10.5 请参阅	318
6.11 创建浅树	318
6.11.1 工作方式	318
6.11.2 工作原理	320
6.11.3 请参阅	320
6.12 转换树标签	320
6.12.1 准备工作	320
6.12.2 工作方式	321
6.12.3 工作原理	322
6.12.4 请参阅	322
第7章 文本分类	323
7.1 引言	323
7.2 词袋特征提取	324
7.2.1 工作方式	324
7.2.2 工作原理	325
7.2.3 更多信息	325
7.2.4 请参阅	327
7.3 训练朴素贝叶斯分类器	327
7.3.1 准备工作	327
7.3.2 工作方式	328
7.3.3 工作原理	329
7.3.4 更多信息	330
7.3.5 请参阅	333
7.4 训练决策树分类器	334
7.4.1 工作方式	334
7.4.2 工作原理	335
7.4.3 更多信息	335
7.4.4 请参阅	337
7.5 训练最大熵分类器	337
7.5.1 准备工作	337
7.5.2 工作方式	337
7.5.3 工作原理	338
7.5.4 更多信息	339
7.5.5 请参阅	340
7.6 训练scikit-learn分类器	340
7.6.1 准备工作	341
7.6.2 工作方式	341
7.6.3 工作原理	342
7.6.4 更多信息	343
7.6.5 请参阅	345
7.7 衡量分类器的精准率和召回率	346

7.7.1 工作方式	346
7.7.2 工作原理	347
7.7.3 更多信息	348
7.7.4 请参阅	349
7.8 计算高信息量单词	349
7.8.1 工作方式	350
7.8.2 工作原理	351
7.8.3 更多信息	352
7.8.4 请参阅	354
7.9 使用投票组合分类器	354
7.9.1 准备工作	355
7.9.2 工作方式	355
7.9.3 工作原理	356
7.9.4 请参阅	356
7.10 使用多个二元分类器 分类	357
7.10.1 准备工作	357
7.10.2 工作方式	357
7.10.3 工作原理	361
7.10.4 更多信息	362
7.10.5 请参阅	363
7.11 使用NLTK训练器训练 分类器	363
7.11.1 工作方式	363
7.11.2 工作原理	364
7.11.3 更多信息	365
7.11.4 请参阅	371
第8章 分布式进程和大型数据集的 处理	372
8.1 引言	372
8.2 使用execnet进行分布式 标注	372
8.2.1 准备工作	373
8.2.2 工作方式	373
8.2.3 工作原理	374
8.2.4 更多内容	375
8.2.5 请参阅	377
8.3 使用execnet进行分布式 组块	377
8.3.1 准备工作	377
8.3.2 工作方式	377
8.3.3 工作原理	378
8.3.4 更多内容	379
8.3.5 请参阅	379
8.4 使用execnet并行处理 列表	379
8.4.1 工作方式	379
8.4.2 工作原理	380
8.4.3 更多内容	381
8.4.4 请参阅	381
8.5 在Redis中存储频率分布	382
8.5.1 准备工作	382
8.5.2 工作方式	382
8.5.3 工作原理	384

- 8.5.4 更多内容 385
- 8.5.5 请参阅 386
- 8.6 在Redis中存储条件频率分布 386
 - 8.6.1 准备工作 386
 - 8.6.2 工作方式 386
 - 8.6.3 工作原理 387
 - 8.6.4 更多内容 388
 - 8.6.5 请参阅 388
- 8.7 在Redis中存储有序字典 388
 - 8.7.1 准备工作 388
 - 8.7.2 工作方式 388
 - 8.7.3 工作原理 390
 - 8.7.4 更多内容 391
 - 8.7.5 请参阅 392
- 8.8 使用Redis和execnet进行分布式单词评分 392
 - 8.8.1 准备工作 392
 - 8.8.2 工作方式 392
 - 8.8.3 工作原理 393
 - 8.8.4 更多内容 396
 - 8.8.5 请参阅 396
- 第9章 解析特定的数据类型 397
 - 9.1 引言 397
 - 9.2 使用dateutil解析日期和时间 398
 - 9.2.1 准备工作 398
 - 9.2.2 工作方式 398
 - 9.2.3 工作原理 399
 - 9.2.4 更多信息 399
 - 9.2.5 请参阅 399
 - 9.3 时区的查找和转换 400
 - 9.3.1 准备工作 400
 - 9.3.2 工作方式 400
 - 9.3.3 工作原理 402
 - 9.3.4 更多信息 402
 - 9.3.5 请参阅 403
 - 9.4 使用lxml从HTML中提取URL 403
 - 9.4.1 准备工作 403
 - 9.4.2 工作方式 403
 - 9.4.3 工作原理 404
 - 9.4.4 更多信息 404
 - 9.4.5 请参阅 405
 - 9.5 清理和剥离HTML 405
 - 9.5.1 准备工作 405
 - 9.5.2 工作方式 405
 - 9.5.3 工作原理 405
 - 9.5.4 更多信息 406
 - 9.5.5 请参阅 406
 - 9.6 使用BeautifulSoup转换HTML实体 406
 - 9.6.1 准备工作 406

- 9.6.2 工作方式 406
- 9.6.3 工作原理 407
- 9.6.4 更多信息 407
- 9.6.5 请参阅 407
- 9.7 检测和转换字符编码 407
 - 9.7.1 准备工作 408
 - 9.7.2 工作方式 408
 - 9.7.3 工作原理 409
 - 9.7.4 更多信息 409
 - 9.7.5 请参阅 410
- 附录A 宾州treebank词性标签 411
- 模块3 使用Python掌握自然语言处理
 - 第1章 使用字符串 417
 - 1.1 标记化 417
 - 1.1.1 将文本标记为句子 418
 - 1.1.2 其他语言文字的标记化 418
 - 1.1.3 将句子标记为单词 419
 - 1.1.4 使用TreebankWordTokenizer进行标记化 420
 - 1.1.5 使用正则表达式进行标记化 421
 - 1.2 规范化 424
 - 1.2.1 消除标点符号 424
 - 1.2.2 转化为小写和大写 425
 - 1.2.3 处理停用词 425
 - 1.2.4 计算英语中的停用词 426
 - 1.3 替代和纠正标记 427
 - 1.3.1 使用正则表达式替换单词 427
 - 1.3.2 使用一个文本替换另一个文本的示例 428
 - 1.3.3 在标记化之前进行替代 428
 - 1.3.4 处理重复的字符 428
 - 1.3.5 删除重复字符的示例 429
 - 1.3.6 使用单词的同义词替换单词 430
 - 1.4 在文本上应用齐夫定律 431
 - 1.5 相似性量度 431
 - 1.5.1 使用编辑距离算法应用相似性量度 432
 - 1.5.2 使用杰卡德系数应用相似性量度 434
 - 1.5.3 使用史密斯-沃特曼算法应用相似性量度 434
 - 1.5.4 其他字符串相似性指标 435
 - 1.6 本章小结 436
 - 第2章 统计语言模型 437
 - 2.1 单词频率 437
 - 2.1.1 对给定文本进行最大似然估计 441
 - 2.1.2 隐马尔可夫模型估计 448
 - 2.2 在MLE模型上应用平滑 450
 - 2.2.1 加一平滑法 450

- 2.2.2 古德-图灵算法 451
- 2.2.3 聂氏估计 456
- 2.2.4 威滕 贝尔估计 457
- 2.3 为MLE指定回退机制 457
- 2.4 应用数据插值获得混合和匹配 458
- 2.5 应用困惑度评估语言模型 458
- 2.6 在建模语言中应用梅特罗波利斯-黑斯廷斯算法 459
- 2.7 在语言处理中应用吉布斯采样 459
- 2.8 本章小结 461
- 第3章 词语形态学—试一试 462
- 3.1 词语形态学 462
- 3.2 词根还原器 463
- 3.3 词形还原 466
- 3.4 开发用于非英语语言的词根还原器 467
- 3.5 词语形态分析器 469
- 3.6 词语形态生成器 471
- 3.7 搜索引擎 471
- 3.8 本章小结 475
- 第4章 词性标注—识别单词 476
- 4.1 词性标注 476
- 4.2 创建POS标注的语料库 482
- 4.3 选择某个机器学习算法 484
- 4.4 涉及n元组方法的统计建模 486
- 4.5 使用POS标注的语料库开发组块器 491
- 4.6 本章小结 494
- 第5章 解析—分析训练数据 495
- 5.1 解析 495
- 5.2 构建树库 496
- 5.3 从树库中提取上下文无关文法的规则 501
- 5.4 从CFG中创建概率上下文无关的文法 507
- 5.5 CYK图解析算法 509
- 5.6 厄雷图解析算法 510
- 5.7 本章小结 516
- 第6章 语义分析—意义重大 517
- 6.1 语义分析 517
- 6.1.1 NER简介 521
- 6.1.2 使用隐马尔可夫模型的NER系统 525
- 6.1.3 使用机器学习工具包训练NER 530
- 6.1.4 使用POS标注的NER 531
- 6.2 从Wordnet中生成同义词集ID 534
- 6.3 使用Wordnet消除歧义 537
- 6.4 本章小结 541
- 第7章 情感分析—我很高兴 542

- 7.1 情感分析 542
- 7.2 使用机器学习的情感分析 548
- 7.3 本章小结 572
- 第8章 信息检索—访问信息 573
- 8.1 信息检索 573
 - 8.1.1 停用词删除 574
 - 8.1.2 利用向量空间模型进行信息检索 576
- 8.2 向量空间评分以及与查询操作器交互 583
- 8.3 利用隐含语义索引开发IR系统 586
- 8.4 文本摘要 587
- 8.5 问答系统 588
- 8.6 本章小结 589
- 第9章 话语分析—知识就是信仰 590
- 9.1 话语分析 590
 - 9.1.1 使用定中心理论进行话语分析 595
 - 9.1.2 回指解析 596
- 9.2 本章小结 601
- 第 10章 NLP系统的评估—性能分析 602
- 10.1 对NLP系统进行评估的需求 602
 - 10.1.1 NLP工具（POS标注器、词干还原器和形态分析器）的评估 603
 - 10.1.2 使用黄金数据评估解析器 613
- 10.2 IR系统的评估 614
- 10.3 错误识别的指标 614
- 10.4 基于词汇匹配的指标 615
- 10.5 基于语法匹配的指标 619
- 10.6 使用浅层语义匹配的指标 620
- 10.7 本章小结 621
- 参考书目 622
- • • • • (收起)

[Python和NLTK自然语言处理 下载链接1](#)

标签

自然语言处理

编程

人工智能

python

评论

[Python和NLTK自然语言处理 下载链接1](#)

书评

[Python和NLTK自然语言处理 下载链接1](#)