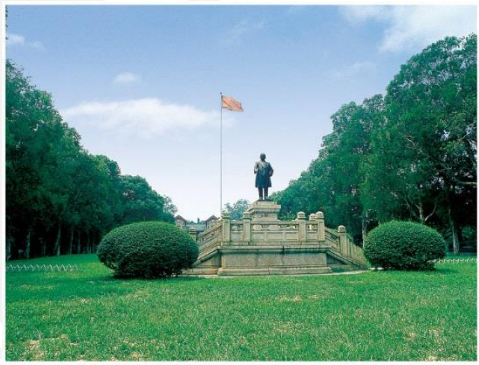
中文分词



- 中文分词是很多中文自然语言处理应用重要的基础，如：情感分析、机器翻译、文本摘要等。
- 分词的效果直接影响模型的准确率。
- 中文分词技术已经比较成熟，不同分词器各有优劣，差异较小。

目录

清洗数据

- 内容提取
- 去除超链接
- 乱码

分词的颗粒度问题

- 可分的复合词
- 不可分的专业名词、人名等。

如：“北京大学”、“中山大学”、“爱因斯坦”、“自然语言处理”

分词的颗粒度问题

- 应用不同，汉语分词的颗粒度大小应该不同。
- 在机器翻译中，一般而言，大粒度翻译效果好。如：“联想公司”作为一个整体，翻译为“Lenovo”，如果分开，则很可能翻译失败。
- 在网页搜索中，小粒度比大粒度效果好。如：“清华大学”，当用户查询“清华”时，是找不到清华大学的，这是有问题的。
- 在不同的应用中，经常是一种词的切分比另一种更有效。根据具体应用选择适宜颗粒度的分词器。

分词的颗粒度问题

```
>>> import jieba
>>> text = “现如今，机器学习和深度学习带动人工智能飞速的发展。”

>>> # 精确模式：精确模式试图将句子最精确地切开.
>>> "/".join(jieba.cut(text, cut_all=False))
'现如今/，/机器/学习/和/深度/学习/带动/人工智能/飞速/的/发展/。'

>>> # 全模式：把句子中所有的可能是词语的都扫描出来。
>>> "/".join(jieba.cut(text, cut_all=True))
'现如今/如今///机器/学习/和/深度/学习/带动/动人/人工/人工智能/智能/飞速/
的/发展//'

>>> # 搜索引擎模式：在精确模式的基础上，对长词再次切分，
>>> # 提高召回率，适合用于搜索引擎分词。
>>> "/".join(jieba.cut_for_search(text))
'如今/现如今/，/机器/学习/和/深度/学习/带动/人工/智能/人工智能/飞速/
的/发展。'
```

自定义词典

- 特定应用领域专业名词、人名等。
- 分词工具包训练数据的局限性、滞后性，新词、网络热词等。
- 解决方法：利用模型的新词发现功能、自行训练（不鼓励）、自定义词典。
- 构建自定义词典的步骤：数据分析——初步分词结果分析——关键词抽取（如TF-IDF）——调整词典/载入词典

自定义词典

```
>>> text = "自然语言处理“
>>> "/".join(jieba.cut(text, cut_all=False))
'自然语言/处理'

>>> # 调整词典
>>> jieba.add_word("自然语言处理")
>>> "/".join(jieba.cut(text, cut_all=False))
'自然语言处理'

>>> # 加载自定义词典
>>> jieba.load_userdict('user_dict.txt')
>>> "/".join(jieba.cut(text, cut_all=False))
'自然语言处理'
```

去除停止词

- 下载停用词表

- 根据具体应用调整停止词表。如：情感分析任务中，“不”、“非常”可能会很重要。

感谢聆听