# jieba

## 分词模式

- 精确模式 — 最精确地切开，适合文本分析
- 全模式 — 句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义
- 搜索引擎模式 — 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
- paddle模式
  - 利用PaddlePaddle深度学习框架，训练序列标注（双向GRU）网络模型实现分词。同时支持词性标注
  - paddle模式使用需安装paddlepaddle-tiny，pip install paddlepaddle-tiny==1.6.1
  - 目前paddle模式支持jieba v0.40及以上版本
- 其他
  - 支持繁体分词
  - 支持自定义词典

## 安装

- pip方式 — pip install jieba / pip3 install jieba
- 下载安装 — 先下载 http://pypi.python.org/pypi/jieba/ ，解压后运行 python setup.py install
- 手动安装 — 将 jieba 目录放置于当前目录或者 site-packages 目录

## 代码

- **jieba.cut** 参数
  - 要切分的字符串
  - cut_all — 控制是否使用全模式
  - HMM — 控制是否使用HMM模型
  - use_paddle — 控制是否使用paddle模式下的分词模式
    - paddle模式采用延迟加载的方式，通过enable_paddle接口安装 paddlepaddle-tiny，并且import相关代码

  *待分词的字符串可以是 unicode 或 UTF-8 字符串、GBK 字符串。返回的结构都是一个可迭代的 generator，可以使用 for 循环来获得分词后得到的每一个词语(unicode)*
- **jieba.cut_for_search**
  - 参数
    - 需要切词的字符串
    - HMM — 是否使用HMM模型
  - 该方法适合用于搜索引擎构建倒排索引的分词，粒度比较细
- jieba.lcut
- jieba.lcut_for_search

  *直接返回 list*
- jieba.Tokenizer(dictionary=DEFAULT_DICT) — 新建自定义分词器，可用于同时使用不同词典 jieba.dt 为默认分词器，所有全局分词相关函数都是该分词器的映射

```
# encoding=utf-8
import jieba

jieba.enable_paddle()# 启动paddle模式。 0.40版之后开始支持，早期版本不支持
strs=["我来到北京清华大学","乒乓球拍卖完了","中国科学技术大学"]
for str in strs:
 seg_list = jieba.cut(str,use_paddle=True) # 使用paddle模式
 print("Paddle Mode: " + '/'.join(list(seg_list)))

seg_list = jieba.cut("我来到北京清华大学", cut_all=True)
print("Full Mode: " + "/ ".join(seg_list))  # 全模式

seg_list = jieba.cut("我来到北京清华大学", cut_all=False)
print("Default Mode: " + "/ ".join(seg_list))  # 精确模式

seg_list = jieba.cut("他来到了网易杭研大厦")  # 默认是精确模式
print(", ".join(seg_list))

seg_list = jieba.cut_for_search("小明硕士毕业于中国科学院计算所，后在日本京都大学深造")  # 搜索引擎模式
print(", ".join(seg_list))
```

## 词典

- 载入词典
  - 开发者可以指定自己自定义的词典，以便包含 jieba 词库里没有的词。虽然 jieba 有新词识别能力，但是自行添加新词可以保证更高的正确率
  - **jieba.load_userdict(file_name) # file_name 为文件类对象或自定义词典的路径**
  - **词典格式和 Dict.Txt 一样，一个词占一行；每一行分三部分：词语、词频（可省略）、词性（可省略），用空格隔开，顺序不可颠倒。File_Name 若为路径或二进制方式打开的文件，则文件必须为 Utf-8 编码**
  - 词频省略时使用自动计算的能保证分出该词的词频
- 调整词典
  - 使用 add_word(word, freq=None, tag=None) 和 del_word(word) 可在程序中动态修改词典
  - 使用 suggest_freq(segment, tune=True) 可调节单个词语的词频，使其能（或不能）被分出来
  - 注意：自动计算的词频在使用 HMM 新词发现功能时可能无效

## 关键词抽取

- 基于 TF-IDF
  - 关键词的抽取：其实就是用文章中切分好的词来代替文章的准确程度
  - 导包 — import jieba.analyse
  - **jieba.analyse.extract_tags(sentence, topK=20, withWeight=False, allowPOS=())** 参数
    - sentence 为待提取的文本
    - topK 为返回几个 TF/IDF 权重最大的关键词，默认值为 20
    - withWeight 为是否一并返回关键词权重值，默认值为 False
    - allowPOS 仅包括指定词性的词，默认值为空，即不筛选
  - **jieba.analyse.TFIDF(idf_path=None) 新建 TFIDF 实例，idf_path 为 IDF 频率文件**
- 基于 TextRank
  - **jieba.analyse.textrank(sentence, topK=20, withWeight=False, allowPOS=('ns', 'n', 'vn', 'v'))** 直接使用，接口相同，注意默认过滤词性。
  - jieba.analyse.TextRank() 新建自定义 TextRank 实例

## 词性标注

- jieba.posseg.POSTokenizer(tokenizer=None) — 新建自定义分词器，tokenizer 参数可指定内部使用的 jieba.Tokenizer 分词器。jieba.posseg.dt 为默认词性标注分词器
- 标注句子分词后每个词的词性，采用和 ictclas 兼容的标记法
- 除了jieba默认分词模式，提供paddle模式下的词性标注功能。paddle模式采用延迟加载方式，通过enable_paddle()安装paddlepaddle-tiny，并且import相关代码

| 标签 | 含义 | 标签 | 含义 | 标签 | 含义 | 标签 | 含义 |
|---|---|---|---|---|---|---|---|
| n | 普通名词 | f | 方位名词 | s | 处所名词 | t | 时间 |
| nr | 人名 | ns | 地名 | nt | 机构名 | nw | 作品名 |
| nz | 其他专名 | v | 普通动词 | vd | 动副词 | vn | 名动词 |
| a | 形容词 | ad | 副形词 | an | 名形词 | d | 副词 |
| m | 数量词 | q | 量词 | r | 代词 | p | 介词 |
| c | 连词 | u | 助词 | xc | 其他虚词 | w | 标点符号 |
| PER | 人名 | LOC | 地名 | ORG | 机构名 | TIME | 时间 |

## 并行分词

- 将目标文本按行分隔后，把各行文本分配到多个 Python 进程并行分词，然后归并结果，从而获得分词速度的可观提升
- 基于 python 自带的 multiprocessing 模块，目前暂不支持 Windows
- 使用方法
  - jieba.enable_parallel(4) # 开启并行分词模式，参数为并行进程数
  - jieba.disable_parallel() # 关闭并行分词模式
- 注意：并行分词仅支持默认分词器 jieba.dt 和 jieba.posseg.dt

## Tokenize

- Tokenize：返回词语在原文的起止位置
- 注意，输入参数只接受 unicode

## 命令行分词

```
使用: python -m jieba [options] filename

结巴命令行界面。

固定参数:
 filename              输入文件

可选参数:
 -h, --help            显示此帮助信息并退出
 -d [DELIM], --delimiter [DELIM]
           使用 DELIM 分隔词语，而不是用默认的' / '。
           若不指定 DELIM，则使用一个空格分隔。
 -p [DELIM], --pos [DELIM]
           启用词性标注；如果指定 DELIM，词语和词性之间
           用它分隔，否则用 _ 分隔
 -D DICT, --dict DICT  使用 DICT 代替默认词典
 -u USER_DICT, --user-dict USER_DICT
           使用 USER_DICT 作为附加词典，与默认词典或自定义词典配合使用
 -a, --cut-all         全模式分词（不支持词性标注）
 -n, --no-hmm          不使用隐含马尔可夫模型
 -q, --quiet           不输出载入信息到 STDERR
 -V, --version         显示版本信息并退出
```