jieba库——中文分词处理

1.特点

- 支持三种分词模式：
  - 精确模式：试图将句子最精确地切开，不存在冗余单词，适合文本分析；
  - 全模式：把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：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。
- 支持繁体分词
- 支持自定义词典
- MIT 授权协议

2.安装与使用

- 全自动安装：`easy_install jieba` 或者 `pip install jieba` / `pip3 install jieba`
- 半自动安装：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
- 通过 `import jieba` 来引用

3.原理

- 基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图 (DAG)
- 采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合
- 对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法

4.用法

- **分词**
  - `jieba.cut` 方法接受三个输入参数: 需要分词的字符串；cut_all 参数用来控制是否采用全模式；HMM 参数用来控制是否使用 HMM 模型
  - `jieba.cut_for_search` 方法接受两个参数：需要分词的字符串；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
  - 待分词的字符串可以是 unicode 或 UTF-8 字符串、GBK 字符串。注意：不建议直接输入 GBK 字符串，可能无法预料地错误解码成 UTF-8
  - `jieba.cut` 以及 `jieba.cut_for_search` 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)，或者用
  - `lcut` **和** `lcut_for_search` **直接返回 list（cut返回的是generator，可以通过for获取）**
  - `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。

```python
import jieba

str="奋斗创造历史，实干成就未来。我们要更加紧密地团结在以习近平同志为核心的党中央周围，高举中国特色社会主义伟大旗帜，以习近平新时代中国特色社会主义思想为指导，迎难而上，开拓进取，以经济社会发展的优异成绩迎接中华人民共和国成立70周年，为决胜全面建成小康社会、夺取新时代中国特色社会主义伟大胜利，为把我国建设成为富强民主文明和谐美丽的社会主义现代化强国、实现中华民族伟大复兴的中国梦不懈奋斗！"
print("精确模式：",jieba.lcut(str))
# 精确模式： ['奋斗', '创造', '历史', '，', '实干', '成就', '未来', '。', '我们', '要', '更加', '紧密', '地', '团结', '在', '以', '习近平', '同志', '为', '核心', '的', '党中央', '周围', '，', '高举', '中国', '特色', '社会主义', '伟大旗帜', '，', '以', '习近平', '新', '时代', '中国', '特色', '社会主义', '思想', '为', '指导', '，', '迎难而上', '，', '开拓进取', '，', '以', '经济社会', '发展', '的', '优异成绩', '迎接', '中华人民共和国', '成立', '70', '周年', '，', '为', '决胜', '全面', '建成', '小康社会', '、', '夺取', '新', '时代', '中国', '特色', '社会主义', '伟大胜利', '，', '为', '把', '我国', '建设', '成为', '富强', '民主', '文明', '和谐', '美丽', '的', '社会主义', '现代化', '强国', '、', '实现', '中华民族', '伟大', '复兴', '的', '中国', '梦', '不懈', '奋斗', '！']
print("全模式：",jieba.lcut(str,cut_all=True))
# 全模式： ['奋斗', '创造', '历史', '', '', '实干', '成就', '未来', '', '', '我们', '要', '更加', '加紧', '紧密', '地', '团结', '结在', '以', '习近平', '同志', '为', '核心', '的', '党中央', '中央', '周围', '', '', '高举', '中国', '国特', '特色', '社会', '社会主义', '会主', '主义', '伟大', '伟大旗帜', '大旗', '旗帜', '', '', '以', '习近平', '新', '时代', '中国', '国特', '特色', '社会', '社会主义', '会主', '主义', '思想', '想为', '指导', '', '', '迎难而上', '', '', '开拓', '开拓进取', '进取', '', '', '以', '经济', '经济社会', '社会', '发展', '的', '优异', '优异成绩', '成绩', '迎接', '中华', '中华人民', '中华人民共和国', '华人', '人民', '人民共和国', '共和', '共和国', '成立', '70', '周年', '', '', '为', '决胜', '全面', '建成', '小康', '小康社会', '社会', '', '', '夺取', '新', '时代', '中国', '国特', '特色', '社会', '社会主义', '会主', '主义', '伟大', '伟大胜利', '大胜', '胜利', '', '', '为', '把', '我国', '国建', '建设', '设成', '成为', '富强', '民主', '主文', '文明', '和谐', '谐美', '美丽', '的', '社会', '社会主义', '会主', '主义', '现代', '现代化', '强国', '', '', '实现', '中华', '中华民族', '民族', '伟大', '复兴', '的', '中国', '梦', '不懈', '奋斗', '', '']
print("搜索引擎模式：",jieba.lcut_for_search(str))
# 搜索引擎模式： ['奋斗', '创造', '历史', '，', '实干', '成就', '未来', '。', '我们', '要', '更加', '紧密', '地', '团结', '在', '以', '习近平', '同志', '为', '核心', '的', '中央', '党中央', '周围', '，', '高举', '中国', '特色', '社会', '会主', '主义', '社会主义', '伟大', '大旗', '旗帜', '伟大旗帜', '，', '以', '习近平', '新', '时代', '中国', '特色', '社会', '会主', '主义', '社会主义', '思想', '为', '指导', '，', '迎难而上', '，', '开拓', '进取', '开拓进取', '，', '以', '经济', '社会', '经济社会', '发展', '的', '优异', '成绩', '优异成绩', '迎接', '中华', '华人', '人民', '共和', '共和国', '中华人民共和国', '成立', '70', '周年', '，', '为', '决胜', '全面', '建成', '小康', '社会', '小康社会', '、', '夺取', '新', '时代', '中国', '特色', '社会', '会主', '主义', '社会主义', '伟大', '大胜', '胜利', '伟大胜利', '，', '为', '把', '我国', '建设', '成为', '富强', '民主', '文明', '和谐', '美丽', '的', '社会', '会主', '主义', '社会主义', '现代', '现代化', '强国', '、', '实现', '中华', '民族', '中华民族', '伟大', '复兴', '的', '中国', '梦', '不懈', '奋斗', '！']

print("cut方式：","/".join(jieba.cut(str)))
```

```
# cut方式：  奋斗/创造/历史/，/实干/成就/未来/。/我们/要/更加/紧密/地/团结/在/以/习近平/同志/为/核心/的/党中央/周围/，/高举/中国/特色/社会主义/伟大旗帜/，/以/习近平/新/时代/中国/特色/社会主义/思想/为/指导/，/迎难而上/，/开拓进取/，/以/经济社会/发展/的/优异成绩/迎接/中华人民共和国/成立/70/周年/，/为/决胜/全面/建成/小康社会/、/夺取/新/时代/中国/特色/社会主义/伟大胜利/，/为/把/我国/建设/成为/富强/民主/文明/和谐/美丽/的/社会主义/现代化/强国/、/实现/中华民族/伟大/复兴/的/中国/梦/不懈/奋斗/！
```

- 自定义词典
  - 导入词典
    - 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
    - 用法： 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。
  - 修改词典
    - 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。
- **关键词提取**
  - 基于TF-IDF算法的关键词抽取

    `jieba.analyse.extract_tags(sentence, topK=20, withWeight=False, allowPOS=())`

    - 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

```
import jieba
import jieba.analyse

topK = 10
str = "奋斗创造历史，实干成就未来。我们要更加紧密地团结在以习近平同志为核心的党中央周围，高举中国特色社会主义伟大旗帜，以习近平新时代中国特色社会主义思想为指导，迎难而上，开拓进取，以经济社会发展的优异成绩迎接中华人民共和国成立70周年，为决胜全面建成小康社会、夺取新时代中国特色社会主义伟大胜利，为把我国建设成为富强民主文明和谐美丽的社会主义现代化强国、实现中华民族伟大复兴的中国梦不懈奋斗！"
tags = jieba.analyse.extract_tags(str, topK=topK)
print("关键词："."/".join(tags))
#关键词：  社会主义/习近平/特色/奋斗/中国/70/实干/迎难而上/时代/开拓进取
```

  - 基于TextRank算法的关键词抽取

    `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
- 参数同TF-IDF
- 原理

  1. 将待抽取关键词的文本进行分词
  2. 以固定窗口大小(默认为5，通过span属性调整)，词之间的共现关系，构建图
  3. 计算图中节点的PageRank，注意是无向带权图

```
import jieba
import jieba.analyse

topK = 10
str = "奋斗创造历史，实干成就未来。我们要更加紧密地团结在以习近平同志为核心的党中央周围，高举中国特色社会主义伟大旗帜，以习近平新时代中国特色社会主义思想为指导，迎难而上，开拓进取，以经济社会发展的优异成绩迎接中华人民共和国成立70周年，为决胜全面建成小康社会、夺取新时代中国特色社会主义伟大胜利，为把我国建设成为富强民主文明和谐美丽的社会主义现代化强国、实现中华民族伟大复兴的中国梦不懈奋斗！"
rank_tags=jieba.analyse.textrank(str)
print("关键词:","/".join(rank_tags))
#关键词：社会主义/中国/特色/迎接/时代/发展/美丽/中华人民共和国/夺取/核心/民主/同志/富强/实干/历史/创造/小康社会/奋斗/建成/全面

print("返回词频:",end='')
for x, w in jieba.analyse.textrank(str, withWeight=True):
    print('%s %s' % (x, w))

'''
社会主义 1.0
中国 0.7655752163879971
特色 0.7606848907467376
迎接 0.5918618250218707
时代 0.5801818990697842
发展 0.4452834347499381
美丽 0.44357047004598554
中华人民共和国 0.4432116549802776
夺取 0.4390284892330628
核心 0.4208688126723878
民主 0.41912165478978064
同志 0.41780833853185173
富强 0.4160181916644136
实干 0.4103928347070348
历史 0.4083222628557034
创造 0.40715519996779415
小康社会 0.40624859449243583
奋斗 0.37586251593731507
建成 0.34595342553273856
全面 0.3420793697316446
'''
```

- 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。