jieba中文分词工具包

> 同济子豪兄B站视频专栏： https://space.bilibili.com/1900783
>
> 子豪兄Python交流QQ群：1077638784

概述

jieba是一个开源的中文分词库。广泛用于文本分析、词云绘制、关键词提取、自然语言处理等领域。

jieba非常容易上手学习。支持繁体中文分词、支持用户自定义词典。

项目Github地址： https://github.com/fxsjy/jieba

安装jieba

```
pip install jieba
```

jieba常用函数

精确模式：把最可能组成词语的词切开，没有冗余单词。

全模式：把所有可能组成词语的词切开，有冗余单词。

搜索引擎模式：在精确模式的基础上，对长词再次切分，适合用于搜索引擎分词。结果和全模式类似。

paddle模式：使用百度PaddlePaddle飞桨深度学习框架，调用双向GRU循环神经网络进行分词。

PaddlePaddle官网

```
import jieba

jieba.lcut('中国科学院大学的大学生')
# 精确模式：['中国科学院', '大学', '的', '大学生']

jieba.lcut('中国科学院大学的大学生',cut_all=True)
# 全模式：['中国', '中国科学院', '科学', '科学院', '学院', '大学', '的', '大学', '大学生', '学生']

jieba.lcut_for_search('中国科学院大学的大学生')
# 搜索引擎模式：['中国', '科学', '学院', '科学院', '中国科学院', '大学', '的', '大学', '学生', '大学生']

# 启动paddle模式，如果没有安装paddle会自动开始安装，耗时十秒左右
jieba.enable_paddle()

jieba.lcut('中国科学院大学的大学生',use_paddle=True)
# paddle模式：['中国科学院大学', '的', '大学生']

```

```
jieba.add_word('科学院大学') # 增加词语“科学院大学”
jieba.lcut('中国科学院大学',cut_all=True)
# 全模式：['中国', '中国科学院', '科学', '科学院', '科学院大学', '学院', '大学']
```

> 对于普通用户，使用精确模式即可。
>
> jieba.lcut('中国科学院大学的大学生')

jieba背后的“黑科技”算法原理

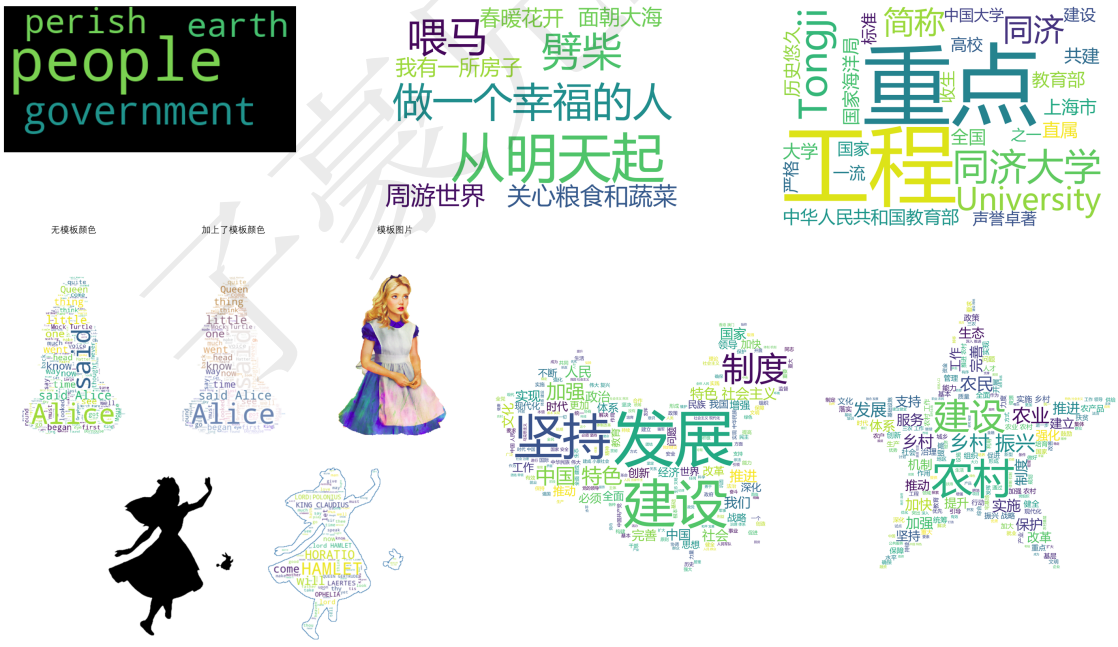
- 基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图 (DAG)
- 采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合
- 对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用 Viterbi 算法

> 使用隐马尔可夫、动态规划等概率模型，计算字符之间的关联概率，字符间概率大的就认为是一个词。
>
> 使用者不需要知道背后的算法原理，只需要会调用接口即可。

扩展案例

词云可视化：四行Python代码轻松上手到精通

https://www.bilibili.com/video/BV1i4411W76Z



词频统计

循环神经网络： https://www.bilibili.com/video/BV1K7411W7So?p=17