# 第二组作业报告

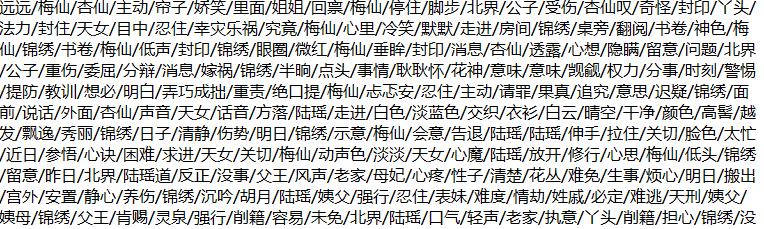
# 文本处理

### 代码：chuangxingshijian\jie\one.py

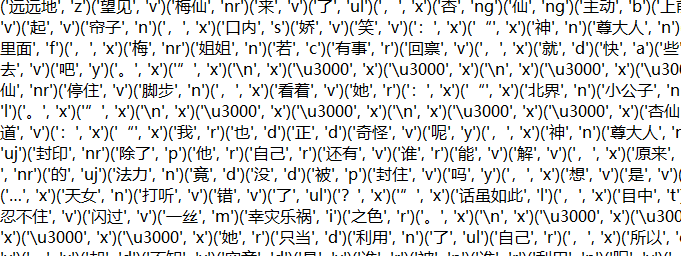
准备好的文本：chuangxinshijian\jie\xs.txt

准备好的停词表：chuangxinshijian\jie\stopwords.txt

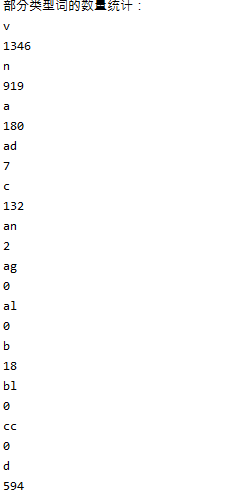
### 去停词处理后的结果预览：



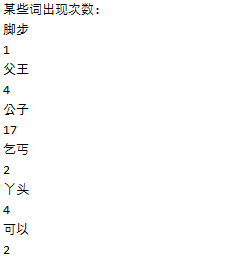
### 词性标注预览：



### 词类数量统计预览：



### 部分词出现次数统计：



## 文本特征提取

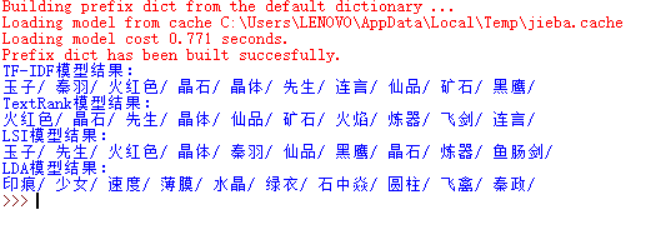
这里是利用TF-IDF模型和TextRank模型对文本进行的特征提取

### 将去停词处理后的文本进行文本特征提取：

### 代码：chuangxinshijian\zande\(alys.py/kt.py)

### 文本：chuangxinshijian\zande\jiebats.txt

### 运行结果：



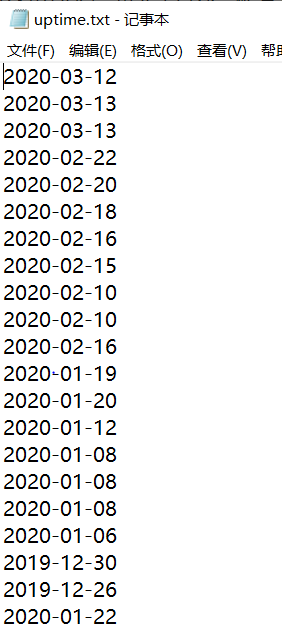
## Github项目主题提取

主要进行对GitHub上项目api上的内容、时间的提取：

#### 提取内容的预览：

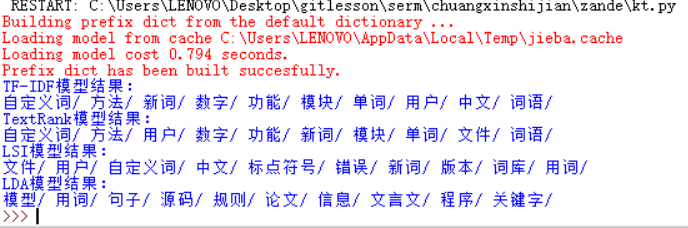
### 

### 时间的预览：

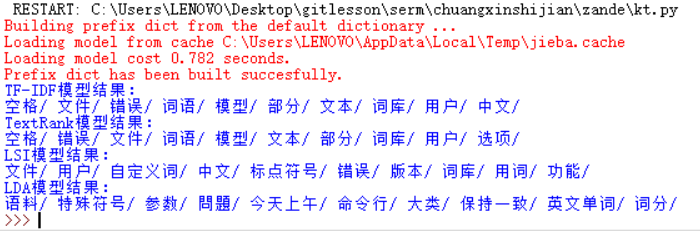


### 之后对于提取出来的数据进行版本的匹配归档，然后根据不同的版本提取出不同的关键词：

0.39版本关键词提取结果：



0.36版本关键词提取结果：



### 最后是关于根据数据集个性化的扩展停用词表的想法：

原思路为通过关键词交集找出高频无用词，之后转为通过统计频率判断高频无用词。

运行结果：

