**词频统计分析**

**学号：20161152136**

**姓名：浦杰**

**1.什么是“词频分析”？**

：词频分析，就是对某一或某些给定的词语在某文件中出现的次数进行统计分析。

**2.那么它能做哪些事情？**

eg：分析你喜欢歌手的表达习惯是怎样的？

判断一段歌词是周杰伦写的还是吴亦凡写的？

电影是那些人主演的，那些是导演？

某小说的人物出场顺序是怎样的？

老师课堂中强调最多的是什么？

……等等。

**这个时候我们需要使用python的jieba库。**

jieba库：优秀的中文分词组件。支持繁体分词、自定义词典。可以对文档进行三种模式的分词：

1、精确模式：试图将句子最精确地切开,适合文本分析;

2、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;

3、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。

**jieba库中文分词原理**

1、利用一个中文词库，确定汉字之间的关联概率。

2、汉字间概率大的组成词组，形成分词结果。

3、除了分词，用户还可以添加自定义词组。

**jieba库安装**

打开cmd，进入D:\python\Scripts>键入：pip install jieba

待安装成功即可。

**jieba常用方法：**

jieba.cut(str)   :  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。返回生成器。

jieba.lcut(str) ：精确模式，返回一个列表类型的分词结果。参数同上。

>>> import jieba

>>> jieba.lcut("西南林业大学以树木树人至真至善为校训")

Building prefix dict from the default dictionary ...

Dumping model to file cache C:\Users\9527\AppData\Local\Temp\jieba.cache

Loading model cost 1.021 seconds.

Prefix dict has been built succesfully.

['西南', '林业大学', '以', '树木', '树人', '至', '真', '至', '善为', '校训']

jieba.lcut(str)  : 全模式，返回一个列表类型的分词结果，有冗余。

>>> jieba.lcut("西南林业大学以树木树人至真至善为校训",cut\_all=True)

['西南', '林业', '林业大学', '业大', '大学', '以', '树木', '树人', '至', '真', '至', '善为', '校训']

jieba.lcut\_for\_search(str) : 搜索引擎模式，返回一个列表类型的分词结果，有冗余。

>>> jieba.lcut\_for\_search("西南林业大学以树木树人至真至善为校训")

['西南', '林业', '业大', '大学', '林业大学', '以', '树木', '树人', '至', '真', '至', '善为', '校训']

>>>

Jieba.add\_word(str) : 向分词词典增加新词str。

**案例实践：**

 分析[美]塞林格小说《麦田里的守望者》（前3节）出现次数最多的词语。

首先下载好《麦田里的守望者》以txt格式、utf-8编码。导入jieba库：

import jieba

打开文件：

txt = open("mtldswz.txt", encoding="utf-8").read()

words = jieba.lcut(txt)

#定义空集合，并借此进行进行统计：

counts = {}

for word in words:

counts[word] = counts.get(word,0) + 1

#dict\_items转换为列表,并以第二个元素排序：

items = list(counts.items())

items.sort(key=lambda x:x[1], reverse=True)

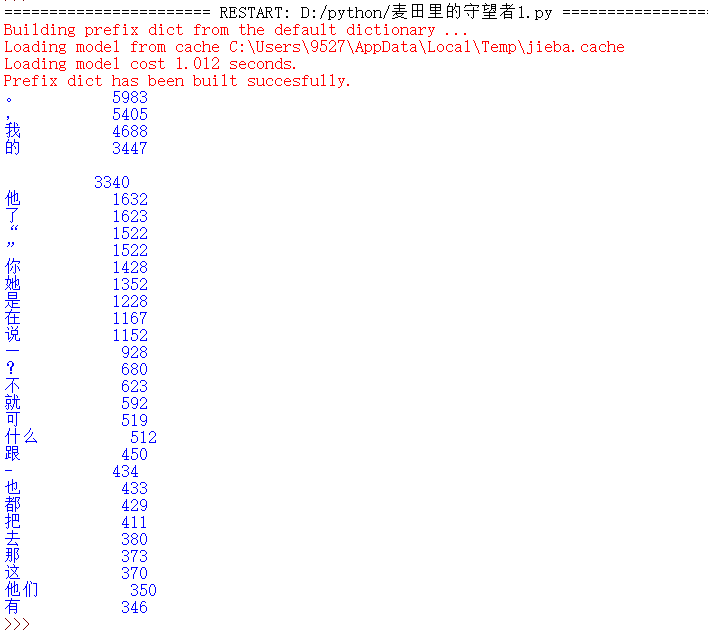
#以格式化打印前30名：

for i in range(30):

word, count = items[i]

print ("{0:<10}{1:>5}".format(word, count))

结果：



可以看到，存在非常多的垃圾数据。因为把文档中的标点、空格、没有意义的字、词语全部进行了统计。

这并不是我们想要的数据。这个时候我们需要使用停用词表。

**停用词表**

停用词：停用词是指在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据（或文本）之前或之后会自动过滤掉某些字或词，这些字或词即被称为Stop Words（停用词）。

停用词表便是存储了这些停用词的文件。在网上下载停用词表，命名Stop.txt。

下面只需要修改原代码即可：

import jieba

txt = open("mtldswz.txt", encoding="utf-8").read()

#加载停用词表

stopwords = [line.strip() for line in open("Stop.txt",encoding="utf-8").readlines()]

words = jieba.lcut(txt)

counts = {}

for word in words:

#不在停用词表中

if word not in stopwords:

#不统计字数为一的词

if len(word) == 1:

continue

else:

counts[word] = counts.get(word,0) + 1

items = list(counts.items())

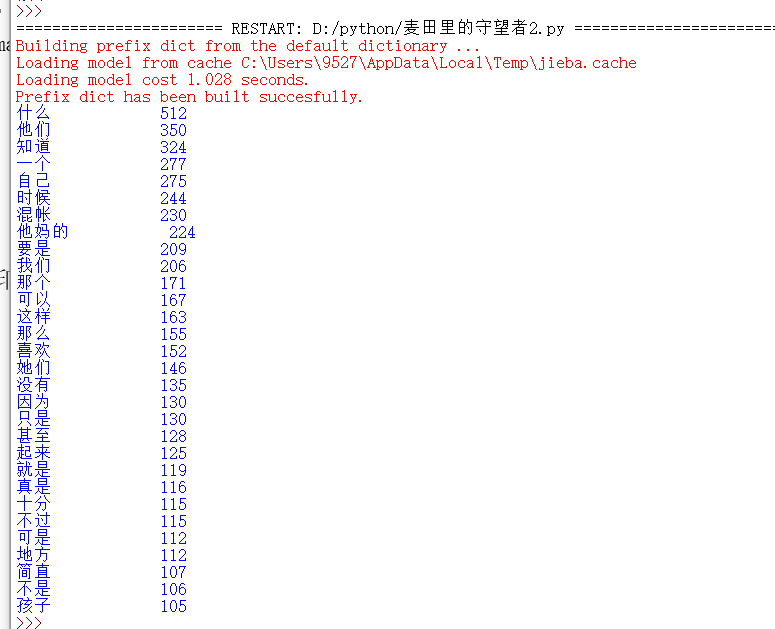
items.sort(key=lambda x:x[1], reverse=True)

for i in range(30):

word, count = items[i]

print ("{:<10}{:>7}".format(word, count))

得到结果：



最后得到了我们想要的结果。打印的即为出现最多的词语和出现的次数。