71086032 曾诗仪 第二周作业

作业内容： 基于提供的微博数据进行基本的词频统计。数据文件为weibo.txt，一行为一条微博，分别是经纬度，文本，发布时间，用\t隔开。本次作业只使用文本内容。

1. 读取文件，用split进行分隔，并选出文本，一行视为一个文档。文档中可能会包含一些“噪声”（比如‘[’和‘]’等，可以删除）。

```
with open(filename, encoding="utf-8") as fp:
    text = fp.read()

text = re.sub('[^\u4e00-\u9fa5]+', '', text)
```

2. 使用jieba对所有文档进行分词，并统计词频

```
ls = jieba.lcut(text)  # 分词
# 统计词频
counts = {}
for i in ls:
    if len(i) > 1:
        counts[i] = counts.get(i, 0) + 1
```

3. 按词频进行排序。观察高频词和低频词。

```
ls1 = sorted(counts.items(), key=lambda x: x[1], reverse=True)  # 词频排序
```

4. 引入停用词表（上网搜索）进行停用词过滤，重新观察词频排序的结果。

```
words_1 = ''.join(counts.keys())

for word in words:  # 去掉停用词
    counts.pop(word, 0)
```

5. 用wordcloud对高频词进行可视化（词云）。
6. 对词性进行分析，观察不同词性的出现频率，并对特定词性的词进行可视化（词云）。

```
c = wordcloud.WordCloud(font_path="C:/Users/shiye/Desktop/simhei.ttf",
width=800, height=600, min_font_size=30,
                        max_font_size=300, max_words=100)
c.generate(" ".join(jieba.lcut(words_1)))
c.to_file("C:/Users/shiye/Desktop/python/pywordcloud.png")
```

7. （附加） 如果tuple来表示bigram，请统计所有的bigram的频率，并通过可视化观察高频的bigram。
8. （附加）可否利用词频来进行特征词的筛选？如果有了特征词，怎么通过其来对文本进行向量表示？如果有了向量表示，可否计算不同文本之间的距离（相似性）？

相关文档

jieba: https://github.com/fxsjy/jieba
wordcloud: https://amueller.github.io/word_cloud/references.html

作业附件：

weibo.txt

完整代码：

```
import re
import jieba
import zhon.hanzi
import wordcloud

filename = "C:/Users/shiye/Desktop/python/weibo.txt"  # 设置文件

punc = zhon.hanzi.punctuation  # 要去除的中文标点符号

with open('C:/Users/shiye/Desktop/python/cn_stopwords.txt', encoding="UTF-8") as
fp:
    words = fp.read()

# 读入文件
with open(filename, encoding="utf-8") as fp:
    text = fp.read()

text = re.sub('[^\u4e00-\u9fa5]+', '', text)

ls = jieba.lcut(text)  # 分词

# 统计词频
counts = {}
for i in ls:
    if len(i) > 1:
        counts[i] = counts.get(i, 0) + 1

words_1 = ''.join(counts.keys())
```

```
for word in words:  # 去掉停用词
    counts.pop(word, 0)

ls1 = sorted(counts.items(), key=lambda x: x[1], reverse=True)  # 词频排序

c = wordcloud.WordCloud(font_path="C:/Users/shiye/Desktop/simhei.ttf",
width=800, height=600, min_font_size=30,
                        max_font_size=300, max_words=100)
c.generate(" ".join(jieba.lcut(words_1)))
c.to_file("C:/Users/shiye/Desktop/python/pywordcloud.png")
```

输出结果：