BTH004-歌词分词与哈希存储

姓名：赵水 学号：201806150329

1 问题：

分别对汪峰和郑钧的几首歌词进行分词处理，并利用哈希存储（字典）来统计每个词语出现的次数

2 算法：

利用分词工具将分好的词，出现次数以键值对的形式存入字典，并输出词语和次数

3 代码：

```
import jieba

txt = open("汪峰歌词.txt","r",encoding='utf-8').read()

for ch in '，\n!"#$%&()*+,-./:;<=>?@[\\]^_‘{|}~': #将特殊符号替换为空格
    txt = txt.replace(ch, " ")

words = jieba.lcut(txt) #分词
counts = {}
for word in words: #将分好的词，出现次数以键值对的形式存入字典
    if word != " ":
        counts[word] = counts.get(word,0) + 1

items1 = list(counts.items()) #生成链表方便处理
items1.sort(key=lambda x: x[1], reverse=True)     # 根据词语出现的次数进行从大到小排序

txt = open("郑钧歌词.txt","r",encoding='utf-8').read()

for ch in '，\n!"#$%&()*+,-./:;<=>?@[\\]^_‘{|}~':
    txt = txt.replace(ch, " ")

words = jieba.lcut(txt)
counts = {}
for word in words:
    if word != " ":
        counts[word] = counts.get(word,0) + 1

items2 = list(counts.items())
items2.sort(key=lambda x: x[1], reverse=True)

print('%-10s%-10s%-10s%-10s%-10s' %("rank","wangfeng", "num","zhengjun","num"))
for i in range(10):
    word1, count1 = items1[i]
    word2, count2 = items2[i]
    print("%-10s%-10s%-10s%-10s%-10s" %(i+1,word1, count1,word2,count2))
```

4 输出

rank	wangfeng	num	zhengjun	num
1	的	165	我	101
2	我	101	你	69
3	你	59	的	59
4	在	46	在	23
5	更	25	想	15
6	高	23	让	15
7	飞	22	了	14
8	知道	20	说	14
9	为	20	是	12
10	相信	19	这	11