1.词性分析

1.1直接输入数据

```from nltk import word_tokenize, pos_tag
text = "I am learning Natural Language Processing on Analytic Vidhya"
tokens = word_tokenize(text)
print(pos_tag(tokens))

[('I', 'PRP'), ('am', 'VBP'), ('learning', 'VBG'), ('Natural', 'NNP'), ('Language', 'NNP'), ('Pr
```python
```

1.2读取txt文件数据

```
    f = open("speech1.txt","r")   #设置文件对象
    text = f.read()
      #将txt文件的所有内容读入到字符串str中
print(text)

I believe we have made great steps

tokens = word_tokenize(text)
print(pos_tag(tokens))

[('I', 'PRP'), ('believe', 'VBP'), ('we', 'PRP'), ('have', 'VBP'), ('made', 'VBN'), ('great', 'J
```

2.英文词频统计

2.1直接输入数据
```

```word="I'm a boby, I'm a girl. When it is true, it is ture. thit are cats, the red is red."
word=word.replace(',','').replace('.','')
word=word.split()
print(word)
setword=set(word)
for i in setword:
    count=word.count(i)
    print(i,'出现次数: ',count)

["I'm", 'a', 'boby', "I'm", 'a', 'girl', 'When', 'it', 'is', 'true', 'it', 'is', 'ture', 'thit',
girl 出现次数:  1
true 出现次数:  1
When 出现次数:  1
it 出现次数:  2
is 出现次数:  3
ture 出现次数:  1
boby 出现次数:  1
thit 出现次数:  1
are 出现次数:  1
red 出现次数:  2
the 出现次数:  1
I'm 出现次数:  2
cats 出现次数:  1
a 出现次数:  2
```

2.2读取txt文件数据

```
        f=open('speech2.txt')
        readline=f.readlines()
        word=[]#存储单词

        #得到文章的单词并且存入列表中:
        for line in readline:
                #因为原文中每个单词都是用空格 或者逗号加空格分开的,
                line=line.replace(',','').replace('.','')#除去逗号,句号只要空格来分开单词
                line=line.strip()
                wo=line.split(' ')
                word.extend(wo)
        return word  #返回单词列表
def clear_account(list1):
   #创建字典
        wokey={}
        wokey=wokey.fromkeys(list1)
        word_1=list(wokey.keys())
        #然后统计单词出现的次数,并将它存入一个字典中
        for i in word_1:
                wokey[i]=list1.count(i)
        return wokey  #返回单词统计
def sort_1(wokey):
        #排序,按values进行排序，如果是按key进行排序用sorted(wokey.items(),key=lambda d:d[0],reverse=True)
        wokey_1={}
        wokey_1=sorted(wokey.items(),key=lambda d:d[1],reverse=True)
        return wokey_1
def main(wokey_1):
# 输出前10个
<p class="mume-header " id="输出前10个"></p>

    for i, (x, y) in enumerate(wokey_1):
        if i in range(0, 10):
            print('%s 出现次数: %d'%(x,y))
main(sort_1(clear_account(read_file())))

have 出现次数: 2
the 出现次数: 2
we 出现次数: 2
are 出现次数: 2
is 出现次数: 2
We 出现次数: 1
not 出现次数: 1
only 出现次数: 1
a 出现次数: 1
parliamentary 出现次数: 1
```

3.中文词频统计

```
import jieba

article = open('speech3.txt','r').read()
dele = {'。','! ','? ','的','“','”',' (',') ',' ','》','《',', ','...','.....',':'}
jieba.add_word('大数据')
words = list(jieba.cut(article))
articleDict = {}
articleSet = set(words)-dele
for w in articleSet:
    if len(w)>1:
        articleDict[w] = words.count(w)

articlelist = sorted(articleDict.items(),key = lambda x:x[1], reverse = True)

for i in range(10):
    print(articlelist[i])
```

```
('复工', 8)
('办公厅', 3)
('工人', 3)
('复产', 3)
('新招', 2)
('甚至', 2)
('疫情', 2)
('补贴', 2)
('企业', 2)
('措施', 2)
```