

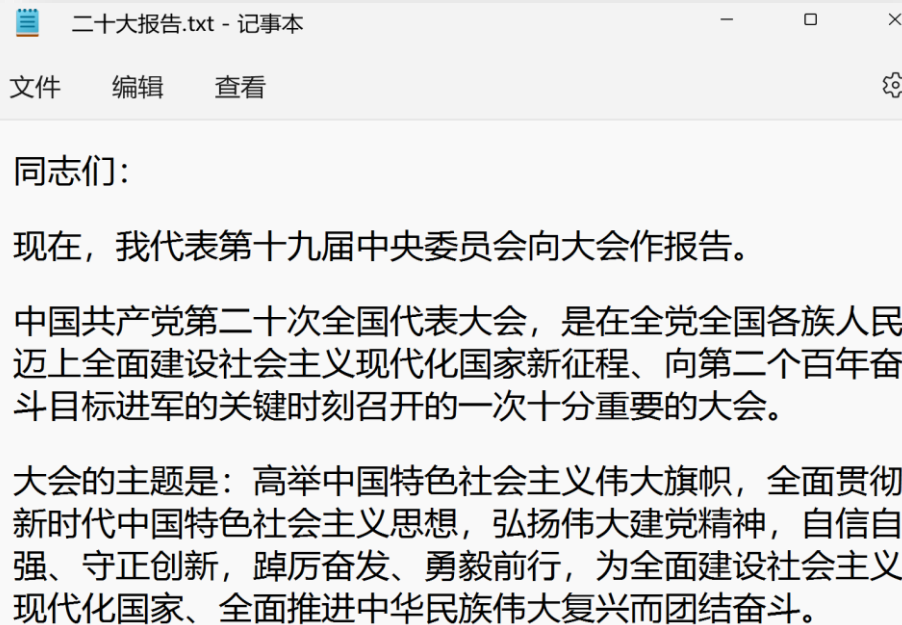


Python与金融数据挖掘(4)

文欣秀

wenxinxiu@ecust.edu.cn

案例分析一



文件读操作

read(): 读取整个文件到字符串中

```
fobj=open("二十大报告.txt", "r", encoding="utf-8")  
paper=fobj. read()  
print(paper)  
fobj. close()
```

问题: 总是忘记关闭文件怎么办?

文件读操作

with语句： 化简代码、处理异常

```
with open("二十大报告.txt", "r", encoding ='utf-8') as fobj:
```

```
    paper=fobj. read()
```

```
print(paper)
```

文件写操作

write() : 将一个字符串写入文件中

```
with open("XX财富网简介.txt","r",encoding="utf-8") as fobj:  
    paper=fobj.read()  
    before=input("被替代字: ")  
    after=input("替代为字: ")  
    result=paper.replace(before,after)  
with open("result.txt", "w") as f2:  
    f2.write(result)
```

案例分析二

| | A | B |
|----|----------|-----|
| 1 | 学号 | 姓名 |
| 2 | 20002518 | 邹轩敏 |
| 3 | 21012973 | 李一凡 |
| 4 | 21012974 | 王绘雯 |
| 5 | 21012975 | 黄佳妮 |
| 6 | 21012976 | 李雨杉 |
| 7 | 21012978 | 胡凯 |
| 8 | 21012979 | 党嘉懿 |
| 9 | 21012980 | 马睿 |
| 10 | 21012981 | 赵骏飞 |
| 11 | 21012982 | 袁洲力 |

随机点名

点一名学生

点三名学生

点五名学生

文件读操作

readlines(): 读取整个文件并创建列表

| | A | B |
|----|----------|-----|
| 1 | 学号 | 姓名 |
| 2 | 20002518 | 邹轩敏 |
| 3 | 21012973 | 李一凡 |
| 4 | 21012974 | 王绘雯 |
| 5 | 21012975 | 黄佳妮 |
| 6 | 21012976 | 李雨杉 |
| 7 | 21012978 | 胡凯 |
| 8 | 21012979 | 党嘉懿 |
| 9 | 21012980 | 马睿 |
| 10 | 21012981 | 赵骏飞 |
| 11 | 21012982 | 袁洲力 |

```
fobj= open("student.csv", 'r')
aList=fobj. readlines()
print (aList)
for i in aList:
    print(i. strip())
fobj. close()
```

文件读操作

化简方法：直接在文件对象上循环读取内容

```
with open("student.csv", 'r') as fobj:  
    for i in fobj:  
        i=i. strip()  
        code, name=i. split(",")  
        print(name)
```

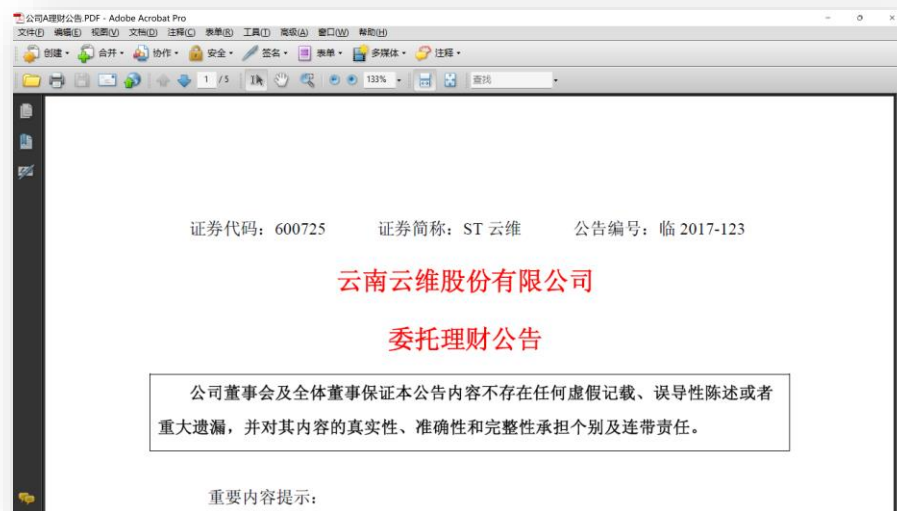

文件写操作

writelines(): 字符串列表按行写入文件

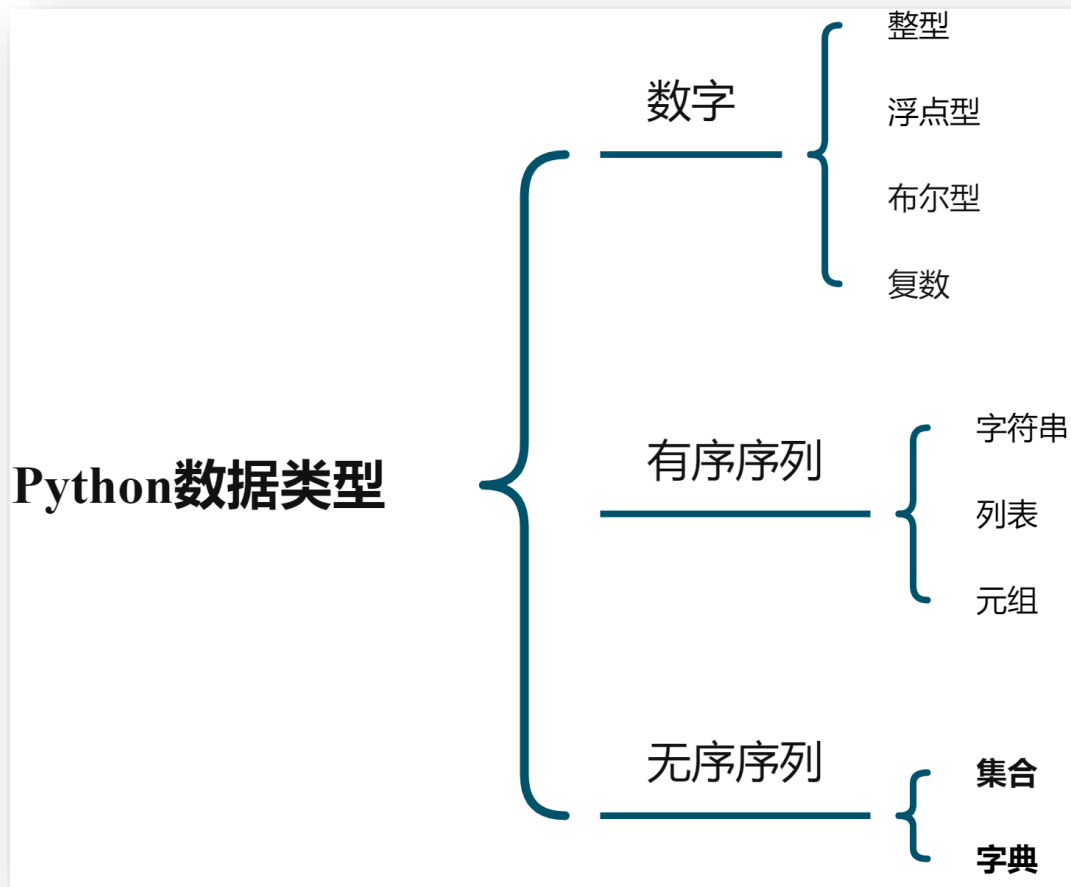
```
aList=[]  
with open("name.csv", 'r', encoding="utf-8") as f1:  
    for line in f1:  
        line=line. strip()  
        new=line[0]+len(line[1:])*"*"  
        aList. append(new + "\n")  
with open("新名字.csv", 'w', encoding="utf-8") as f2:  
    f2. writelines(aList)
```



案例分析



数据类型



| 代码 | 名称 |
|--------|-------|
| 000001 | 上证指数 |
| 399001 | 深证成指 |
| 899050 | 北证50 |
| 000300 | 沪深300 |
| 399005 | 中小100 |
| 399006 | 创业板指 |

常用字典方法

di.keys(): 返回包含字典所有**键**的列表

di.values(): 返回包含字典所有**值**的列表

di.items(): 返回包含所有(**键**、**值**)项列表

di.get(key,[default]): 返回**键key**对应的**值**，若
key不存在，则返回default

di.update(a): 将字典a中的键值对添加到di中

课堂练习

若dic={"000001":"上证指数","899050":"北证50"},
则执行 dic.get('000300')的结果是 ()

A、 "北证50"

B、 False

C、 报错

D、 空值

词云相关库

matplotlib: 用于绘图的第三方库

wordcloud: 用于词云展示的第三方库

imageio: 读取和写入各种图像的第三方库

爱心词云

| | A | B |
|----|----------|-----|
| 1 | 学号 | 姓名 |
| 2 | 20002518 | 邹轩敏 |
| 3 | 21012973 | 李一凡 |
| 4 | 21012974 | 王绘雯 |
| 5 | 21012975 | 黄佳妮 |
| 6 | 21012976 | 李雨杉 |
| 7 | 21012978 | 胡凯 |
| 8 | 21012979 | 党嘉懿 |
| 9 | 21012980 | 马睿 |
| 10 | 21012981 | 赵骏飞 |
| 11 | 21012982 | 袁洲力 |



```

from random import *
counts={ }#创建一个空字典
with open("student.csv", 'r') as fobj:
    for i in fobj:
        if i[:2]=="学号":
            continue
        i=i. strip()
        code, name=i. split(",")
        counts[name]=randint(30,100)
import matplotlib. pyplot as plt
from wordcloud import WordCloud
from imageio.v2 import imread

```


爱心词云

```
pic = imread('love.png')
wc=WordCloud(mask=pic,font_path='msyh.ttc', #中文字体
              repeat=False, #内容是否可以重复
              background_color='white', #设置背景颜色
              max_words=100, #设置最大词数
              max_font_size=120, #设置字体最大值
              min_font_size=10, #设置字体最小值
              random_state=50, #设置有配色方案
              scale=1) #按照比例进行放大画布

wc.generate_from_frequencies(counts)
plt.imshow(wc)
plt.show()
```

如何实现分词？

二十大报告.txt - 记事本

文件 编辑 查看

同志们：

现在，我代表第十九届中央委员会向大会作报告。

中国共产党第二十次全国代表大会，是在全党全国各族人民迈上全面建设社会主义现代化国家新征程、向第二个百年奋斗目标进军的关键时刻召开的一次十分重要的大会。

大会的主题是：高举中国特色社会主义伟大旗帜，全面贯彻新时代中国特色社会主义思想，弘扬伟大建党精神，自信自强、守正创新，踔厉奋发、勇毅前行，为全面建设社会主义现代化国家、全面推进中华民族伟大复兴而团结奋斗。



关于文本词频统计

词频统计的内涵：累加问题，即对文档中的每个词设计一个计数器，词语出现一次，计算器加1，词和次数是一对出现，构成

<单词>：<出现次数>

键值对：字典

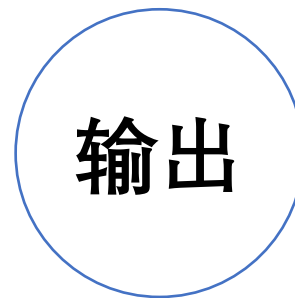
词频统计问题的IPO描述



从文件中读取一
篇待分析的文章



采用字典数据结构
统计词语出现的频率



根据词频进行图形
绘制或统计高频词语

jieba库分词原理

- ◆ 提供中文词库 `pip install jieba`
- ◆ 将待分词的内容与分词词库进行比对
- ◆ 通过图结构和动态规划方法找到最大概率词组
- ◆ 增加自定义中文单词的功能

jieba库三种分词模式

精确模式：将句子最精确地切开，适合文本分析

```
>>>import jieba
```

```
>>>jieba.lcut("中华人民共和国是一个伟大的国家")
```

jieba库三种分词模式

全 模 式： 把句子中所有可以成词的词语都扫描出来，
速度非常快，但不能消除歧义

```
>>>import jieba
```

```
>>>jieba.lcut("中华人民共和国是一个伟大的国家",cut_all=True)
```

jieba库三种分词模式

搜索引擎模式： 在精确模式基础上， 对长词再次切分，
提高召回率， 适合用于搜索引擎分词

```
>>>import jieba
```

```
>>>jieba.lcut_for_search("中华人民共和国是一个伟大的国家")
```


添加专属名词

```
>>>import jieba
```

```
>>>jieba.lcut("习大大希望中国的老百姓有更好的生活")
```

```
>>>jieba.add_word("习大大")
```

```
>>>jieba.lcut("习大大希望中国的老百姓有更好的生活")
```

二十大报告词云案例 (1)

```
import jieba
import matplotlib.pyplot as plt
from wordcloud import WordCloud
from imageio.v2 import imread
fobj=open("二十大报告.txt","r",encoding="utf-8")
txt=fobj.read()
words=jieba.lcut(txt)
```

单词计数方法一

```
aList=["上海","北京","上海","云南","北京","上海"]  
counts={ }  
for word in aList:  
    if word not in counts:  
        counts[word]=1  
    else:  
        counts[word]=counts[word]+1  
print(counts)
```

单词计数方法二

```
aList=["上海","北京","上海","云南","北京","上海"]
```

```
counts={ }
```

```
for word in aList:
```

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

```
print(counts)
```


二十大报告词云案例 (3)

```
wc=WordCloud(mask=pic,font_path='msyh.ttc', #中文字体
              repeat=False, #内容可以重复
              background_color='white', #设置背景颜色
              max_words=110, #设置最大词数
              max_font_size=120, #设置字体最大值
              min_font_size=10, #设置字体最小值
              random_state=50, #设置配色方案
              scale=10)

wc.generate_from_frequencies(counts)
plt.imshow(wc) #将数值以图片形式显示出来
plt.show()
```



集合定义

- ◆ 集合使用大括号 $\{ \}$ 来包裹
- ◆ 集合相当于只有键没有值的字典
- ◆ 集合内的元素不可重复出现
- ◆ 集合内的元素是不可变的
- ◆ 集合内的元素没有先后关系

集合运算一示例

```
>>> a={"江西铜业","神州长城","中集集团","古井贡酒"}
>>> h={"中集集团","江西铜业","小米集团","阿里影业"}
>>> a & h          {'中集集团','江西铜业'}
>>> a | h          {'中集集团','古井贡酒','小米集团','阿里影业','神州长城','江西铜业'}
>>> a - h          {'神州长城','古井贡酒'}
>>> a ^ h          {'小米集团','阿里影业','神州长城','古井贡酒'}
>>> "小米集团" not in a  True
```

集合运算二示例

```
>>> a={"江西铜业","神州长城","中集集团","古井贡酒"}
```

```
>>> h={"中集集团","江西铜业","小米集团","阿里影业"}
```

```
>>> s={"江西铜业","中集集团"}
```

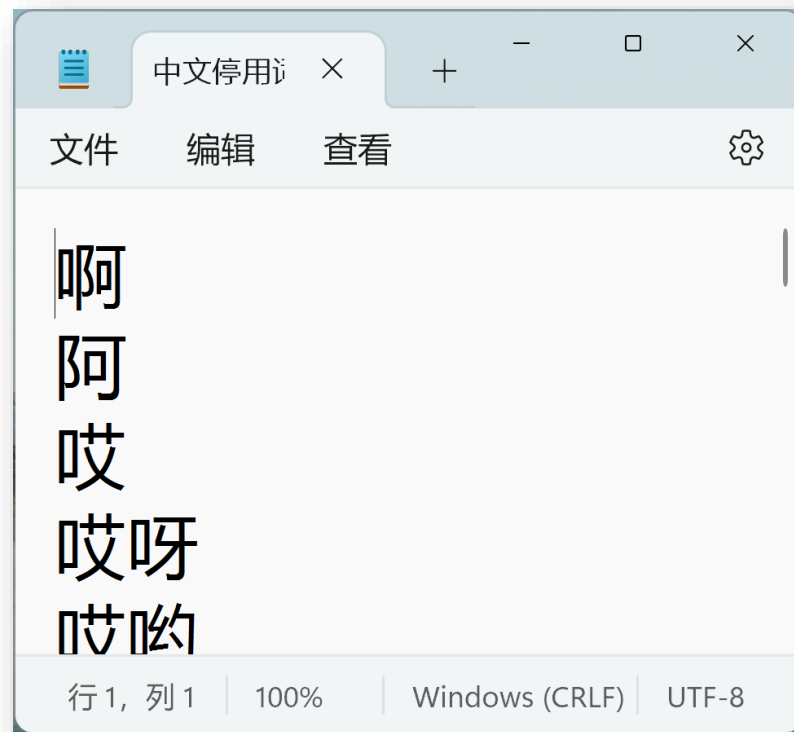
```
>>> s<=a True
```

```
>>> s > h False
```

```
>>> s< a True
```

```
>>> a==h False
```


拓展问题



思考：如何从文件中读取所有停用词并进行判断？

从文件获取排除词

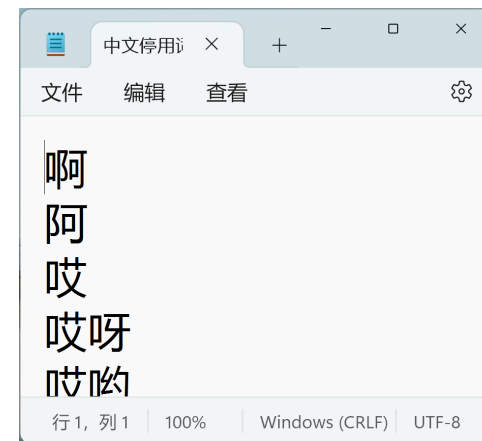
```
excludes=set()
```

```
with open("中文停用词.txt", "r") as handle:
```

```
    for i in handle:
```

```
        i=i.strip()
```

```
        excludes.add(i)
```



二十大报告词云案例（修改3）

```
counts={}
```

excludes=set()

```
with open('中文停用词.txt','r',encoding='utf-8') as fobj:
```

```
for i in fobj:
```

i=i.strip()

excludes. add(i)

for word in words:

```
if len(word)==1:
```

continue

elif word in excludes:

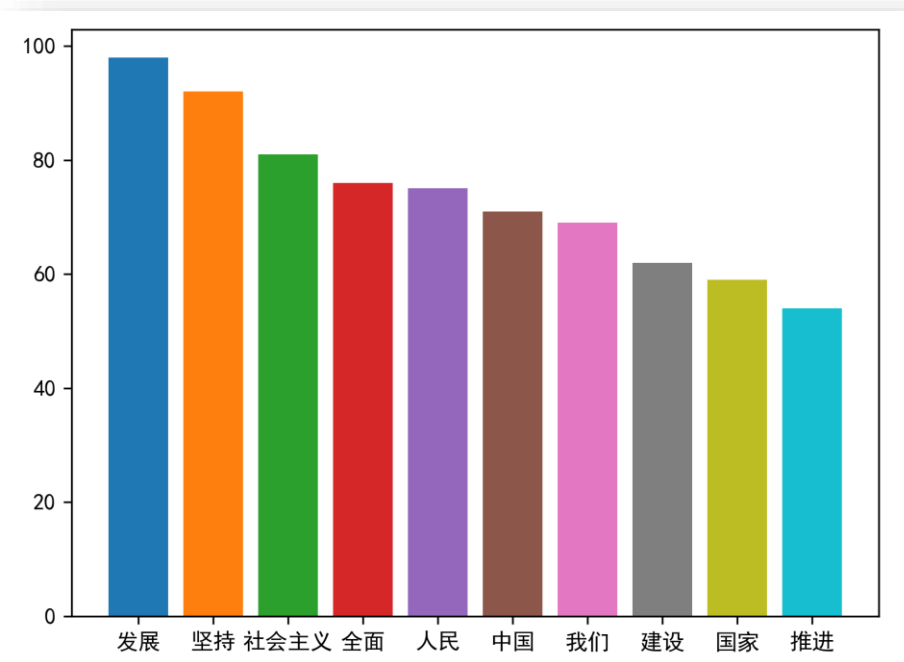
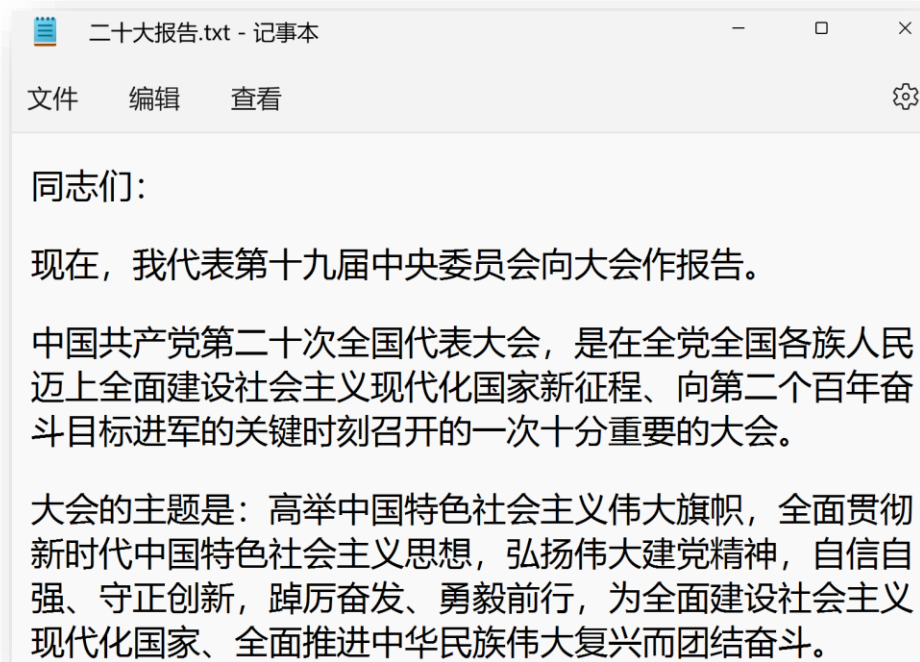
continue

else:

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



拓展问题



问题： 如何根据词频画出柱状图？

二十大报告词频统计案例

```
import jieba
import matplotlib.pyplot as plt
from wordcloud import WordCloud
from imageio.v2 import imread
fobj=open("二十大报告.txt","r",encoding="utf-8")
txt=fobj.read()
words=jieba.lcut(txt)
```


二十大报告词频统计案例

```
counts={ }
excludes=set()
with open("中文停用词.txt","r",encoding="utf-8") as fobj:
    for i in fobj:
        i=i.strip()
        excludes.add(i)
for word in words:
    if len(word)==1:
        continue
    elif word in excludes:
        continue
    else:
        counts[word]=counts.get(word,0)+1
```



二十大报告词频统计案例

```
items=list(counts.items())
```

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

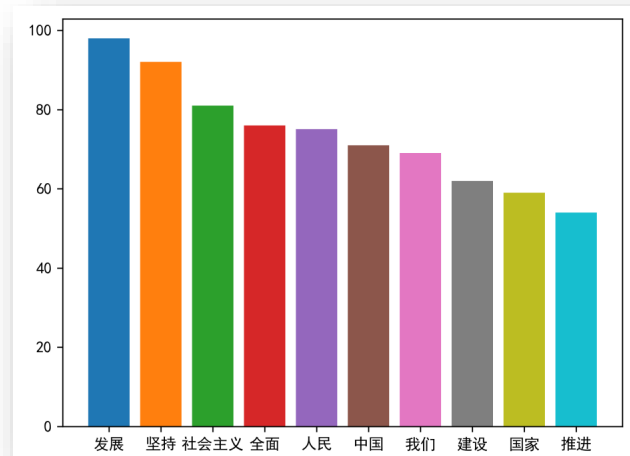
```
plt.rcParams['font.sans-serif']=['SimHei'] #用来正常显示中文标签
```

```
for i in range(10):
```

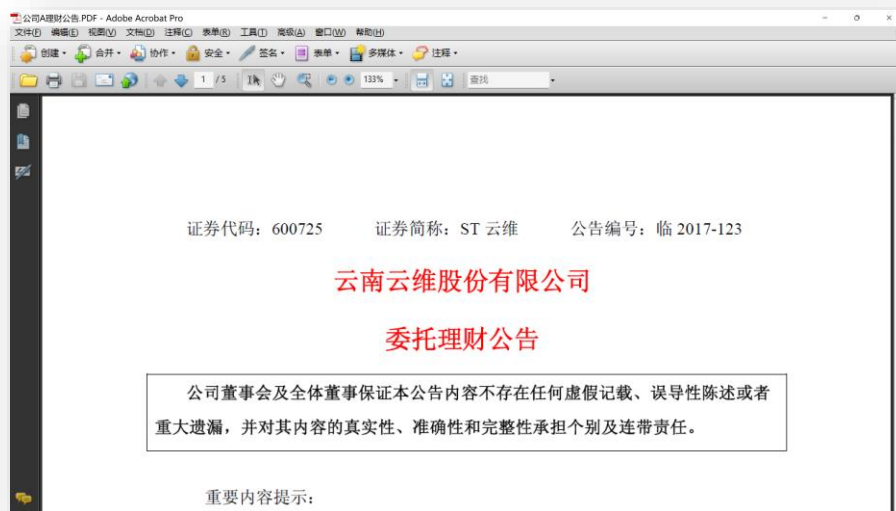
```
    word,count=items[i]
```

```
    plt.bar(word,count)
```

```
plt.show()
```



拓展问题



思考： 如何从pdf文件中读取数据进行词频统计？

PDF文件读取

```
import pdfplumber
```

```
pdf = pdfplumber.open('公司A理财公告.PDF')
```

```
pages = pdf.pages
```

```
text_all = []
```

```
for page in pages: # 遍历pages中每一页的信息
```

```
    text = page.extract_text() # 提取当页的文本内容
```

```
    text_all.append(text) # 汇总每一页内容
```

```
text_all = "".join(text_all) # 把列表转换成字符串
```

```
print(text_all) # 打印全部文本内容
```

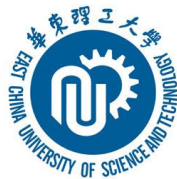
```
pdf.close()
```

生成词典

```
import jieba
import matplotlib.pyplot as plt
from wordcloud import WordCloud
from imageio.v2 import imread
words=jieba.lcut(text_all)
counts={ }
for word in words:
    if len(word)==1:
        continue
    else:
        counts[word]=counts. get(word,0)+1
pic = imread('cloud.jpg')
```

绘制词云

```
wc=WordCloud(mask=pic,font_path='msyh.ttc', #中文字体
              repeat=False, #内容可以重复
              background_color='white', #设置背景颜色
              max_words=110,          #设置最大词数
              max_font_size=120,      #设置字体最大值
              min_font_size=10,       #设置字体最小值
              random_state=50,        #设置配色方案
              scale=10)
wc.generate_from_frequencies(counts)
plt.imshow(wc) #将数值以图片形式显示出来
plt.show()
```

谢 谢