Python与金融数据挖掘(4)

文欣秀

wenxinxiu@ecust.edu.cn

案例分析一


中国共产党第二十次全国代表大会


社会主义
发展
中国
人民
坚持
全面
我们
现代化
建设
国家
推进
实现
加强
推动
时代
体系
战略
特色
伟大
必须
中国共产党
中华民族
实施
复兴
创新
促进
历史
思想
贯彻
斗争
法治
强国
人才
群众
全党
坚定
确保
历史性
更加
我国
人类
完善
精神
党中央
政治
文化
经济
治理
科技
军事
事业
增强
工作
奋斗
维护
一系列
坚决
教育
道路
创造
提出
提升
团结
青年
改革
革命
党的建设
能力
解决
生态
党和国家
统筹
长期
保护
着力
深入
不断
社会
提高
巩固
国际
弘扬
建成
落实
保障
全球
高质量
构建
自信
问题
领域
健全
领导
基本
各族人民
稳定
重大
制度
马克思主义
共同
反对

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)
```

案例分析二

	A	B
1	学号	姓名
2	20002370	权泽睿
3	20002512	张宸煜
4	20002513	费诚成
5	20002514	李凌瑶
6	20002515	刘宇晨
7	20002516	邓庚麒
8	20002517	王飞扬
9	20002519	黄志鹏
10	20002520	周学勤
11	20002521	诸建飞
12	20002522	徐盛
13	20002523	余铠
14	20002524	朱晟
15	20002525	董旭
16	20002526	赵逸帆

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	20002370	权泽睿
3	20002512	张宸煜
4	20002513	费诚成
5	20002514	李凌瑶
6	20002515	刘宇晨
7	20002516	邓庚麒
8	20002517	王飞扬
9	20002519	黄志鹏
10	20002520	周学勤
11	20002521	诸建飞
12	20002522	徐盛
13	20002523	余铠
14	20002524	朱晟
15	20002525	董旭
16	20002526	赵逸帆

```
fobj= open("student.csv", 'r', encoding ='utf-8')
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

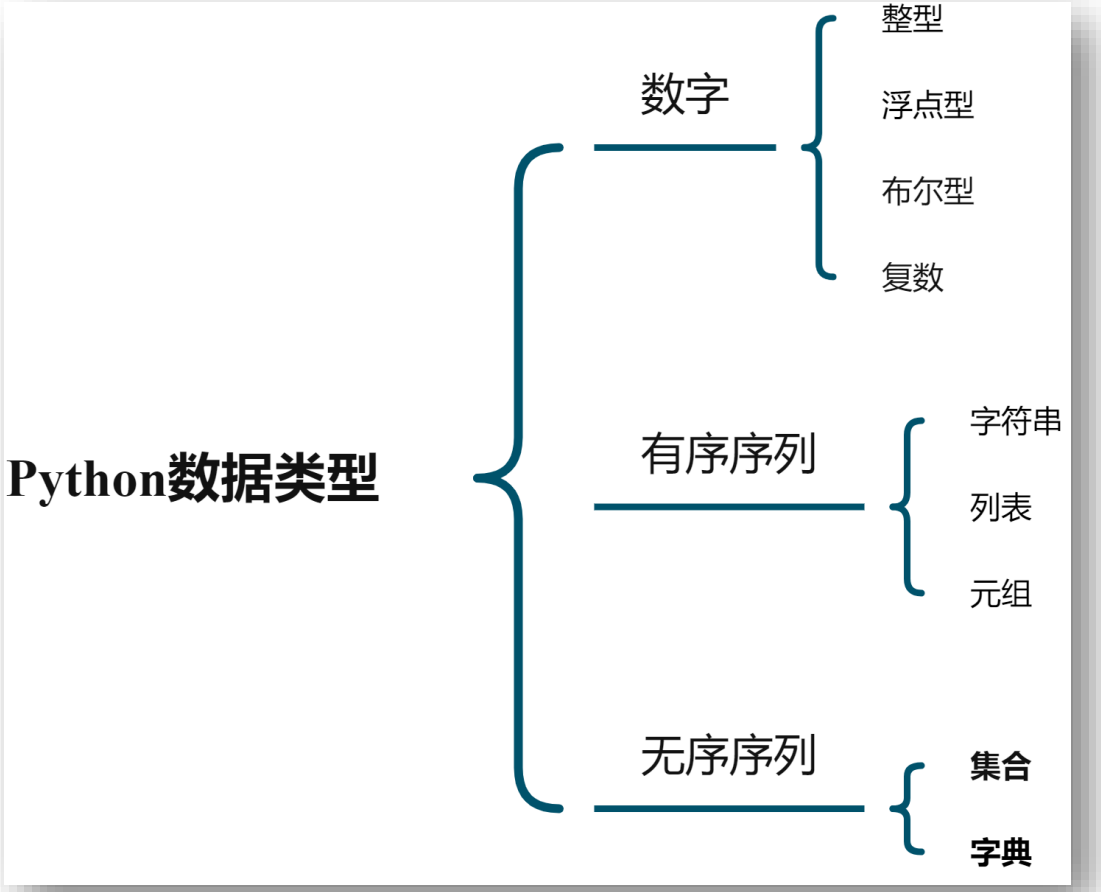
```
with open("student.csv", 'r', encoding="utf-8") as fobj:
    for i in fobj:
        i=i. strip()
        code, name=i. split(",")
        print(name)
```

如何制作词云？

社会主义
发展
坚持
人民
中国
我们
全面
推进
建设
国家
现代化
体系
时代
制度
重大
战略
特色
加强
推动
安全
伟大
社会
实现
文化
政治
加快
实施
保障
复兴
贯彻
必须
完善
精神
世界
民主
健全
领导
思想
历史
中国共产党
中华民族
马克思主义

赵逸帆
许力权
陈奕伟
崔艺耀
吴昕阳
苗志鹏
方子恒
陈弘悦
张宸煜
鲍骏伟
周天奕
游子缘
孙一宁
俞文涛
卢永飞
李响
单陈龙
应天翼
余铠
刘旭
姓名
余梓涵
梁桂锐
邓庚麒
杭业晟
阿拉帕提·牙力昆
朱思宇
颜荣岗
薛李丹
宋佳薇
王昌震
张境文
黄嘉庆

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

字典定义

- 是Python 中的映射数据类型，用**{ }**包裹
- 由**键-值**对构成，键值对使用冒号**:**分隔
- 键必须唯一，必须是不可变数据类型
- 一般以**数字、字符串**等不可变对象作为键
- 值可以是**任意类型**的Python 对象

字典示例

```
>>> test = {}                                  #创建一个空字典
>>> info= {'000001': '上证数', '399001' : '深证成指'}
>>> info['000300']= '沪深300'                  #新增元素
>>> info['000001']= '上证指数'                 #修改元素值
>>> del info['399001']                         #删除元素值
```

课堂练习

正确定义一个字典的是（ ）

A、 a=["A": 10, "B": 20, "C": 30]

B、 a=("A": 10, "B": 20, "C": 30)

C、 a={A:10, B: 20, C: 30}

D、 a={"A":10, "B":20, "C": 30}

常用字典方法

di.keys()： 返回包含字典所有**健**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])： 返回健**key对应的值**，若key不存在，则返回default

课堂练习

若dic1 = {'甲' :3, '乙' :1, '丙':5 , '丁' :8}，则执行print(dic1.get('乙', '未找到'))的结果是（ ）

A、未找到

B、1

C、报错

D、输出空值

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

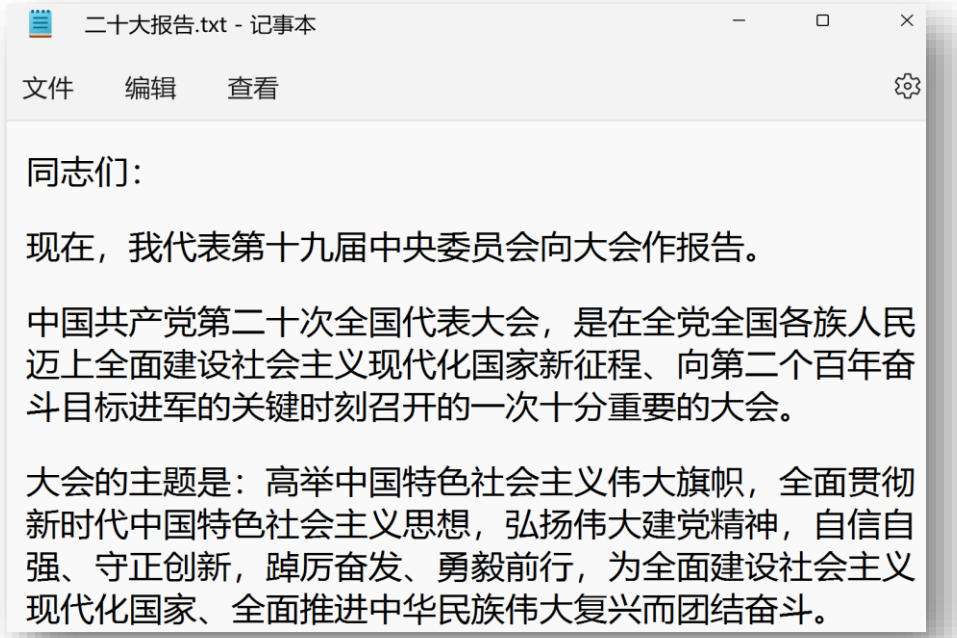
	A	B
1	学号	姓名
2	20002370	权泽睿
3	20002512	张宸煜
4	20002513	费诚成
5	20002514	李凌瑶
6	20002515	刘宇晨
7	20002516	邓庚麒
8	20002517	王飞扬
9	20002519	黄志鹏
10	20002520	周学勤
11	20002521	诸建飞
12	20002522	徐盛
13	20002523	余铠
14	20002524	朱晟
15	20002525	董旭
16	20002526	赵逸帆

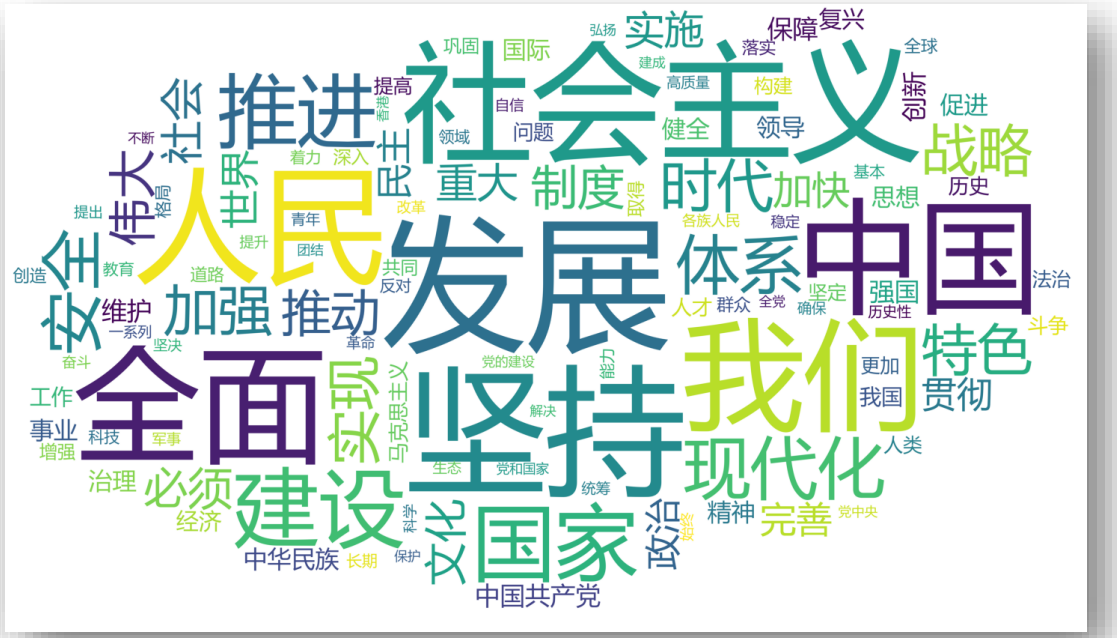

```
from random import *
counts={}#创建一个空字典
with open("student.csv", 'r', encoding="utf-8") as fobj:
    for i in fobj:
        if i[:2]=="学号":
            continue
        i=i. strip()
        code, name=i. split(",")
        counts[name]=randint(30,100)
import matplotlib. pyplot as plt
from wordcloud 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。


社会主义
发展
中国
人民
坚持
我们
全面
建设
国家
现代化
推进
战略
特色
体系
时代
制度
加强
推动
实施
中华民族
中国共产党

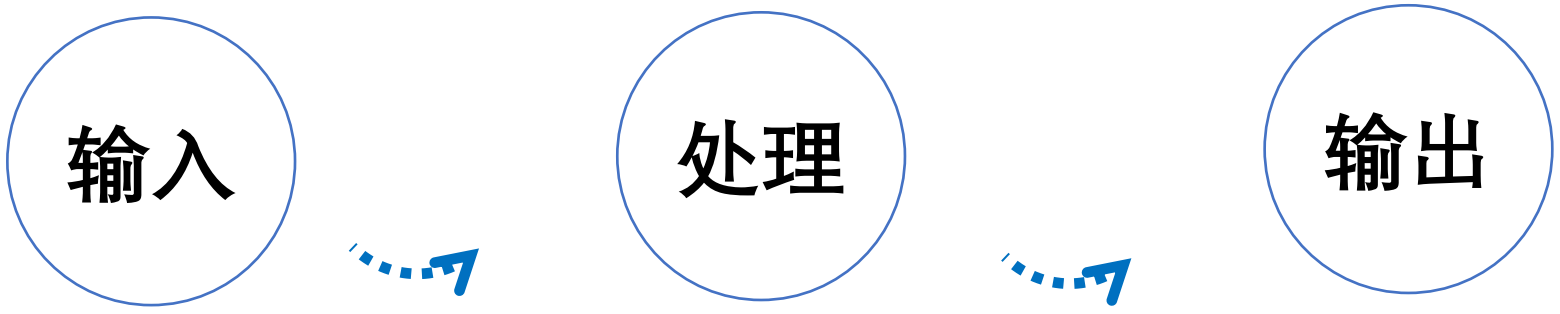
关于文本词频统计

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;

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;

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应用实例分析

```
>>>import jieba

>>>jieba.lcut("中华人民共和国是一个伟大的国家")

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

>>>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（2）

```
counts={}
for word in words:
        if len(word)==1:
            continue
        else:
            counts[word]=counts. get(word,0)+1
pic = imread('cloud.jpg')
```

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()
```

二十大报告词云案例

问题：如何去掉部分词语？

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—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
```

二十大报告词云案例（修改2）

```
counts={}
excludes={"不断","一系列","基本"}
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

二十大报告.txt - 记事本

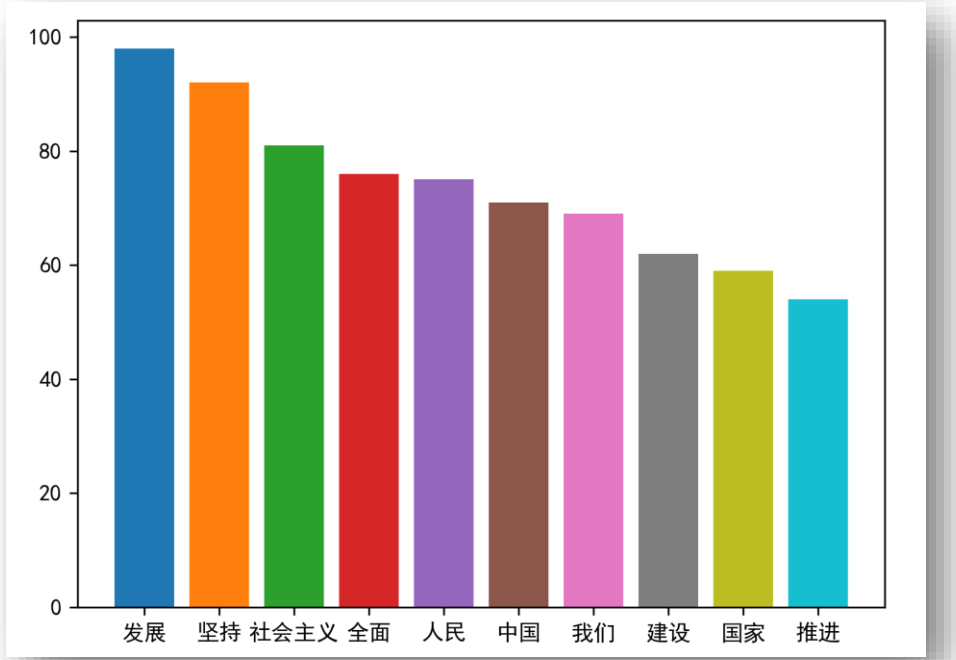
文件　编辑　查看

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。



二十大报告词频统计案例

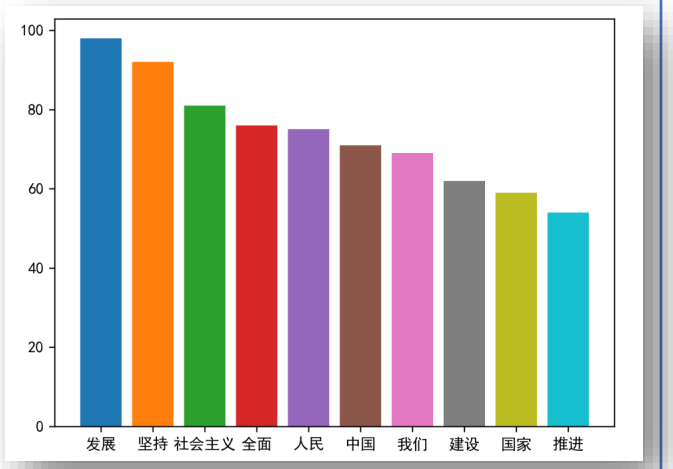
```
import jieba
import matplotlib.pyplot as plt
from wordcloud import WordCloud
from imageio.v2 import imread
fobj=open("二十大报告.txt","r",encoding="utf-8")
txt=fobj.read()
words=jieba.lcut(txt)
```

二十大报告词频统计案例

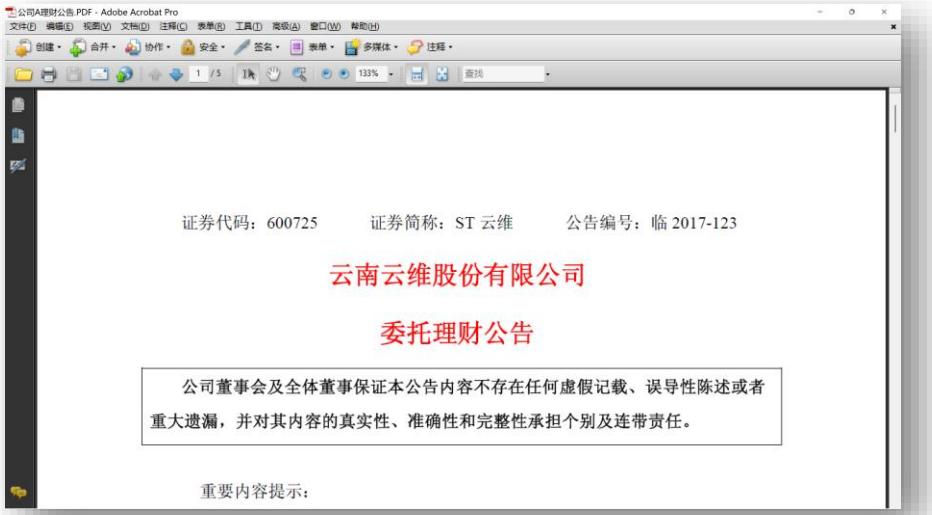
```
counts={}
excludes={"不断","一系列","基本"}
for word in words:
        if len(word)==1:
            continue
        elif word in excludes:
            continue
        else:
            counts[word]=counts. get(word,0)+1
```

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


证券代码：600725　　证券简称：ST 云维　　公告编号：临 2017-123

云南云维股份有限公司

委托理财公告

公司董事会及全体董事保证本公告内容不存在任何虚假记载、误导性陈述或者重大遗漏，并对其内容的真实性、准确性和完整性承担个别及连带责任。

重要内容提示：

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)  # 通过列表.append()方法汇总每一页内容
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')
```

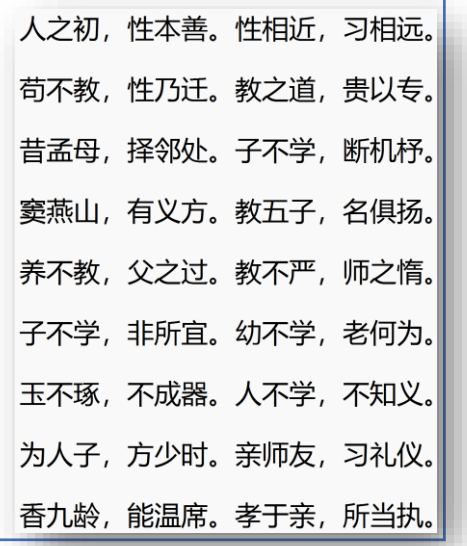
绘制词云

```
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()
```

文件写操作

write()： 将一个**字符串**写入文件中

```
with open("三字经.txt","r",encoding ='utf-8') as f1:
    paper=f1. read()
with open("新三字经.txt", "w") as f2:
    new=paper[::-1]
    new=new.replace("。","")
    f2.write(new)
```



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)
```

谢　谢