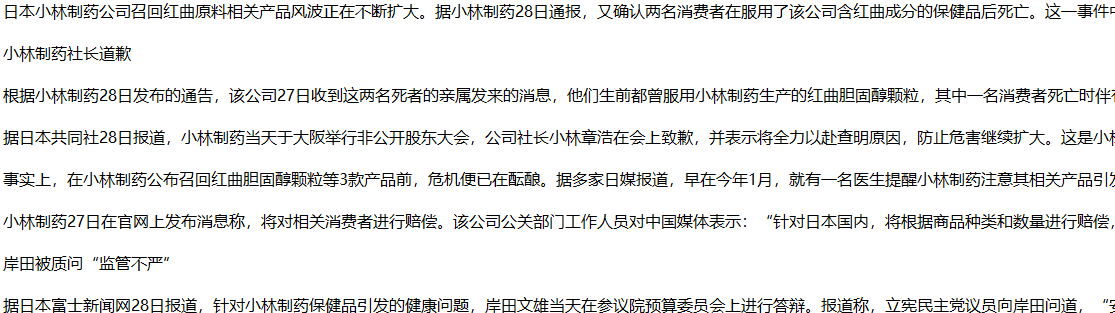
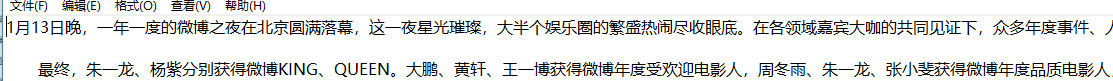
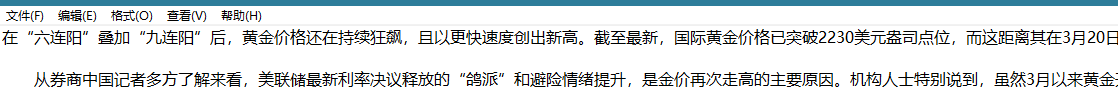
我在新浪新闻选择了三个文本

一、小林制药

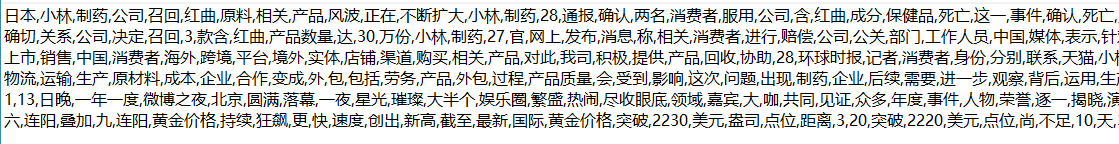


二、朱一龙娱乐新闻

三、黄金价格

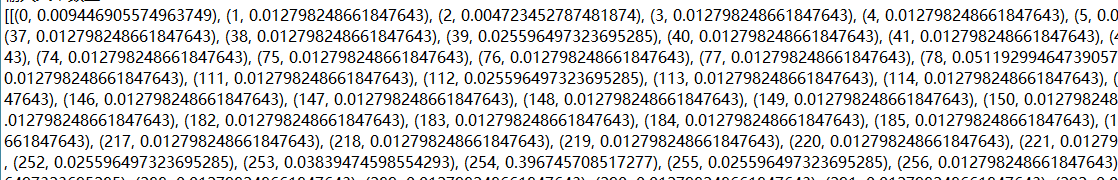


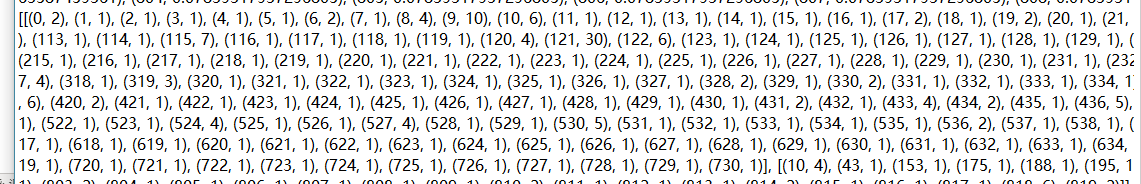
将三个文本进行数据清洗包括：标点符号，停用词清洗，分词



然后进行TFIDF权重计算

以下是词的权重和文本向量空间

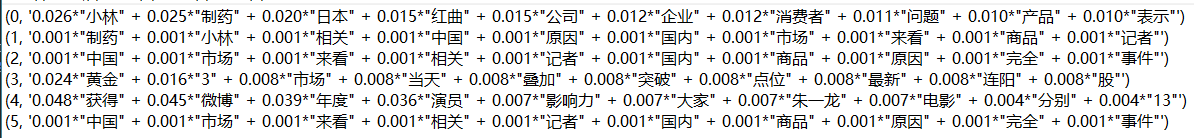




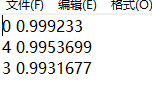
然后进行LDA主题模型的训练

我设定LDA分了6个主题

训练后：



再对文本进行模型验证



上面的数据说明

第一个文档是属于第0个主题后面的0.999是概率

再看第0个主题的词，小林，制药，日本说明很符合

第二个文档属于第4个主题

年度，朱一龙，电影说明很符合

第三个文档属于第3个主题

黄金，市场，突破说明很符合

以下是我的代码

import os  
import jieba  
import codecs  
from gensim import corpora  
from gensim.models import LdaModel  
from gensim import models  
import matplotlib.pyplot as plt  
import numpy as np  
# 获取当前目录  
current\_dir = os.getcwd()  
output\_file\_path = os.path.join(current\_dir, 'output.txt')  
if os.path.exists(output\_file\_path):  
 with open(output\_file\_path, 'w', encoding='utf-8'):  
 pass # 创建或清空文件  
else:  
 open(output\_file\_path, 'w', encoding='utf-8').close() # 创建新文件  
# 指定输入文件路径  
files\_to\_clean = [  
 os.path.join(current\_dir, '新浪新闻', 'test1.txt'),  
 os.path.join(current\_dir, '新浪新闻', 'test2.txt'),  
 os.path.join(current\_dir, '新浪新闻', 'test3.txt'),  
]  
  
# 加载停用词  
stopwords\_path = os.path.join(current\_dir, 'hit\_stopwords.txt')  
with open(stopwords\_path, 'r', encoding='utf-8') as f:  
 stop\_words = [line.strip() for line in f.readlines()]  
  
  
def dataclean(file\_path):  
 # 读取文本文件内容  
 with open(file\_path, 'r', encoding='utf-8') as file:  
 text = file.read()  
  
 # 定义一个包含常见中文标点的集合  
 chinese\_punctuation = set('，、。！？；：“”‘’《》【】「」『』…—·～')  
  
 # 删除标点符号  
 text\_without\_punctuation = ''.join(c for c in text if c not in chinese\_punctuation)  
  
 # 对文本进行分词  
 words = jieba.lcut(text\_without\_punctuation)  
  
 # 删除停用词  
 filtered\_words = [word for word in words if word not in stop\_words]  
 # 移除空格，并将换行符视为分隔符进行处理  
 filtered\_words\_no\_space\_and\_newline = [word for word in filtered\_words if word.strip()]  
 # 将分词结果以逗号分隔的形式写入到新的txt文件中，每个文档结果占一行  
 with open(output\_file\_path, 'a', encoding='utf-8') as output\_file:  
 output\_file.write(','.join(filtered\_words\_no\_space\_and\_newline) + '\n') # 添加换行符  
  
 print(f"文档 '{os.path.basename(file\_path)}' 的分词结果已保存至 '{output\_file\_path}' 文件中。")  
  
  
# 清洗所有指定的文档  
for file in files\_to\_clean:  
 dataclean(file)  
  
  
te = []  
with open(output\_file\_path, 'r',encoding='utf-8') as fp:  
 for line in fp:  
 line=line.split(',')  
 te.append([w for w in line ])  
print('输人文本数量:',len(te))  
dictionary=corpora.Dictionary(te)  
corpus=[dictionary.doc2bow(text) for text in te]  
tfidf = models.TfidfModel(corpus)  
corpus\_tfidf = tfidf[corpus]  
# print(list(corpus\_tfidf))  
# print (list(corpus))  
lda=LdaModel(corpus=corpus,id2word=dictionary,num\_topics=6,passes=100,random\_state=42)  
doc\_topic=[a for a in lda[corpus]]  
topics\_r = lda.print\_topics(num\_topics=6, num\_words=10)  
topic\_name=codecs.open('topies\_result3.txt','w')  
for v in topics\_r:  
 topic\_name.write(str(v)+'\n')  
fp2=codecs.open('documents\_result.txt','w')  
for t in doc\_topic:  
 c = []  
 c.append([a[1] for a in t])  
 print(t)  
 m =max(c[0])  
 for i in range(0,len(t)):  
 if m in t[i]:  
 fp2.write(str(t[i][0]) + ' ' + str(t[i][1]) + '\n')  
 break