**课程名称：**Python语言与数据分析

课 程 报 告

项目名称 基于新浪新闻网的要闻标题爬取与分析

班 级 金融科技实验班196

学 号 0193829

姓 名 於心怡

任课教师 丁菊玲

开课学期： 2020 至 2021 学年 第 2 学期

完成时间： 2021 年 7 月 7 日

**《**项目名称**》数据分析报告**

**目 录**

目录

[1 概述 1](#_Toc76625277)

[2 数据描述 1](#_Toc76625278)

[3 数据分析内容 1](#_Toc76625279)

[4 数据分析图表 1](#_Toc76625280)

[5 数据分析结果 3](#_Toc76625281)

[6 总结 4](#_Toc76625282)

[附录-数据分析代码 5](#_Toc76625283)

# 1 概述

新浪新闻是新浪集团旗下的新闻资讯客户端，致力于利用大数据和人工智能技术，为用户提供更丰富、场景化和个性化的内容阅读体验。相较于代表国家的央视新闻，新浪新闻网更能体现民间热点，舆情趋势。故对新浪新闻网每日标题的提取，生成年度词云进行分析，可以把握时事热点，统揽年度要闻，了解群众关心的大事。

# 2 数据描述

通过编写爬虫程序访问[新浪新闻网](file:///C%3A%5CUsers%5CNapstablook%5CDownloads%5C%E6%96%B0%E6%B5%AA%E6%96%B0%E9%97%BB%E7%BD%91)（https://news.sina.com.cn/）对要闻标题进行爬取，然后将获取的数据存入文件中用jieba库进行分词以便生成词云。由于新闻标题均为字符串格式，数据不需要进行清洗和格式化，但爬取新闻标题是主要的难题，主要使用BeautifulSoup解析html网页，get\_text函数提取标题。详细方法叙述将在代码中注释。

# 3 数据分析内容

2018，2019，2020，2021年年度热点——以词云形式。

# 4 数据分析图表



图表 1：2018年新闻标题词云



图表 2：2019年新闻标题词云



图表 3：2020年新闻标题词云



图表 4：2021年新闻标题词云

# 5 数据分析结果

历年的标题词云都较好地总结了当年的时事热点。

**2018年：**关键词为“中国”，“外交部”，“习近平”，“特朗普”，“贸易战”，“大陆”，“台湾”，“蔡英文”等，说明2018年我国新闻的重点在于国际关系与政治问题，关键词点明了当年中美贸易战的热点话题，特朗普上任后频频刁难中国，其自身也落为众人的笑柄。两岸关系依然紧张，台独势力仍旧顽固。“周年”体现的是2019年是建国50周年。较小的关键字如“朝鲜”，“金正恩”，“俄”，“普京”，“世界杯”等，也体现出了一定的国际关系与时事热点。

**2019年：**政治上特朗普统治下的美国与中国的关系进一步恶化。香港特别行政区爆发暴乱，部分香港青年化身黑衣暴徒袭击港警，破坏公共交通，损坏公共财物，我国政治形势严峻。这样突如其来的暴乱，很难不让人联想到境外势力的插手。经济上贸易战还在持续，但影响已逐渐变小，华为引领行业开发5G技术。

**2020年：**新冠肺炎爆发，我国新闻的重点也从国际关系上转到了国内情势。“疫情”，“肺炎”是出现频率最高的关键词，另外“湖北”，“武汉”体现了疫情的重灾区，“实时”，“地图”表现了民众对全国疫情情况的关心与担忧。2020年是我们面对疫情严峻挑战的一年，也是全民族团结一致共同抗疫的一年，更是体现中国实力的一年。

**2021年：**虽然2021年才过半，根据生成的词云我们还是可以对国内形势窥得一二，虽然新冠肺炎已在我国得到有效控制，但是疫情的防控仍然是重中之重。同时频率较低的关键词还有“疫苗”，“输入”，“台湾”，“月”，“棉花”之类，反映了我国在新冠肺炎疫苗上的研究成果，目前感染者主要为境外输入的情况，两岸关系的重点，我国航空技术的发展，“新疆棉花”事件。

# 6 总结

本次实验我主要碰到的难题是数据获取。本想通过tushare的接口获取cctv央视新闻网标题，但由于tusharepro的积分制，初始的120积分仅有一分钟访问两次的权限，遂作罢。只好自己编写爬虫程序对网页内容进行爬取。

在网上查阅了很多相关文章，大多都很松散，用的方法也各异，好的文章并不多，综合信息后有了一些思路。本来想还是抓取CCTV的新闻，尝试了许久之后未果，CCTV新闻网的网页不愧是国家的网页，对我而言过于复杂，不能抓取出每日的url以进行循环，每日新闻url不变，好像只是内部嵌套，修行尚浅的我只好抱憾放弃。然后转战新浪新闻网，新浪新闻网每日新闻主页的url格式固定，只变动时间，这让爬取的工作简洁了不少，后面的工作就是对网页内容进行解析，抓取标题部分，然后分词生成词云，具体的步骤将在代码中注释。

经过这次数据爬取与分析，我感受到了python的强大，也认识到了自己的不足，虽然代码写了很久，但是回过头来看，也没有多么的复杂，实现的分析也没有特别的细致有效。想要熟练地掌握这门语言需要长此以往的学习与实践练习，在编写代码的过程中，本门课程在学习通上的课件也帮助了我很多，课件上的基础知识很丰富，对于各个方面的介绍也比较详尽。

这也是我第一次见到python运行一次需要这么久，打破了我对python“秒出结果”的印象。我也感受到学好一门编程语言是非常实用的，基础不牢地动山摇，相比于网上搜集文章而言，我认为更好的方式还是挑一本好书，从0重新开始，认认真真地学好python，这也是我暑假准备完成的目标。

# 附录-数据分析代码

import requests

from bs4 import BeautifulSoup

import datetime

import numpy as np

from pandas import Series

import jieba

import wordcloud

from PIL import Image

#爬取数据部分：

lst=[]

begin=datetime.date(2021,1,1)

end=datetime.date(2021,7,8)

#以2021年为例，确定起止日期

for i in range((end-begin).days+1):

 day=begin+datetime.timedelta(days=i)

DATE=day.strftime('%Y%m%d')

#实现起止日期内的每日循环，并以YYYYMMDD格式输出

 url='https://news.sina.com.cn/head/news'+DATE+'am.shtml'

 #经过对网页url的分析，可知每日网页的构成为：

 https://news.sina.com.cn/head/news+当日日期+am.shtml

strhtml = requests.get(url)

#Get方式获取网页数据，存下字符串形式的html网页代码

 strhtml.encoding='utf-8'

 soup=BeautifulSoup(strhtml.text,'lxml')

 data = soup.select('#syncad\_1 > h1> a')

 #进入开发者模式后，对想要提取的标题进行检查，copy selector 路径，删掉h1后面的子项后用soup.select进行提取

 for item in data:

 news=item.get\_text() #对数据进行清洗，上一步得到的数据是带所有信息（a标签，标题，herf，链接）的总数据。使用get.text函数获取<a>标签内的标题

 lst.append(news)

s=Series(lst)

s.to\_excel('d:/news1.xlsx')

f = open('d://title.txt','w',encoding='utf-8')

for item in lst:

 f.write(item)

f.close() #将爬取的数据存入文件

#分词制图部分：

txt = open("D://title.txt", "r",encoding='utf-8').read()

words = jieba.lcut(txt)

txt\_1 = " ".join(words)

STOPWORDS=['了','在','来','的','中','我','大','被','或','都','还','已','个',

 '将','这个','这','又','他','对','是','要','为','年','与','有','向',

 '人','不','和','同','再','你','图','就','后','新','副','把','最','说',

 '这些','这样'] #经过前几次生成的词云，总结出一些无意义词

image = Image.open('d://china.jpg') # 作为背景轮廓图

graph = np.array(image) #把背景图化为二进制矩阵

w = wordcloud.WordCloud(font\_path="msyh.ttc",

 mask=graph,

 width=1000,

 height=700,

 background\_color="white",

 stopwords=STOPWORDS, #避免无意义词

 collocations=False, #避免重复

 )

w.generate(txt\_1)

w.to\_file("d://2021.png")