**一、背景与挖掘目标**

知乎是国内知名的知识分享平台

知乎问题“疫情尚未结束，多国禁止粮食出口，老百姓需要囤积粮食吗？”，截止本文写作时，该问题已有5022个关注者，6186748次浏览，回答数1469个，最高赞1.7万。本文针对该问题，爬取了全部1469个回答，然后运用python3对这些回答采用聚类的方法进行了分析。

图1:问题的知乎页面

**二、分析方法与过程**

**（一）数据获取**

本文通过对<https://www.zhihu.com/question/383352936>的知乎问题进行爬虫，获取了number,author,voteup\_count,content,headline,gender,created\_time,updated\_time,comment\_count,thanks\_count,excerpt 的数据。

图2：爬虫结果文件

爬虫代码展示如下：其中黄色部分替换为自己的知乎手机号和验证码即可

from zhihu\_oauth import ZhihuClient

from zhihu\_oauth.exception import NeedCaptchaException

import time

import random

import pandas as pd

client = ZhihuClient()

try:

client.login('phone number', 'code')

except NeedCaptchaException:

# 保存验证码并提示输入，重新登录

with open('a.gif', 'wb') as f:

f.write(client.get\_captcha())

captcha = input('please input captcha:') # 验证码在文件夹中，手动输入

client.login(' phone number ', ' code ', captcha)

client.save\_token('token.pkl')

id=383352936

question=client.question(id)

answers=question.answers

rows=[]

num=1

for ans in answers:

if num<=question.answer\_count:

print("No.%d"%num,ans.author.name,ans.voteup\_count,ans.author.headline,ans.author.gender,ans.created\_time,ans.comment\_count)

dic={

'No.':num,

'author':ans.author.name,

'voteup\_count':ans.voteup\_count,

'content':ans.content,

'headline':ans.author.headline,

'gender':ans.author.gender,

'created\_time':ans.created\_time,

'updated\_time':ans.updated\_time,

'comment\_count':ans.comment\_count,

'thanks\_count':ans.thanks\_count,

'excerpt':ans.excerpt

}

rows.append(dic)

num+=1

df=pd.DataFrame(rows)

#print(df)

df.to\_csv("疫情尚未结束，多国禁止粮食出口，老百姓需要囤积粮食吗？.csv",encoding='utf-8',index=False)#sep='\t',

**（二）聚类分析**

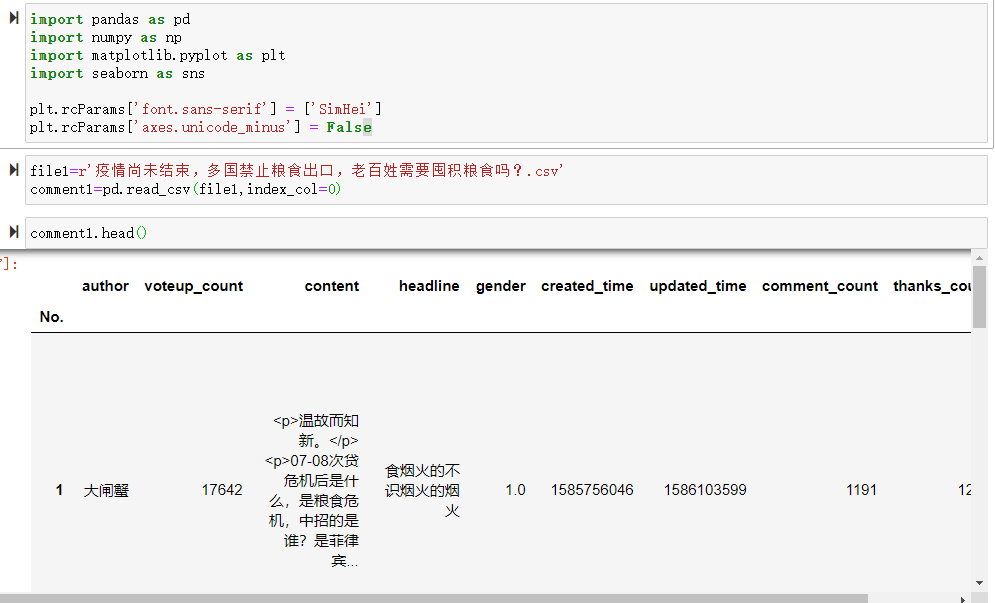
**1.引入相关的包，导入文件**

图3：文件和包的导入

2.**数据处理**

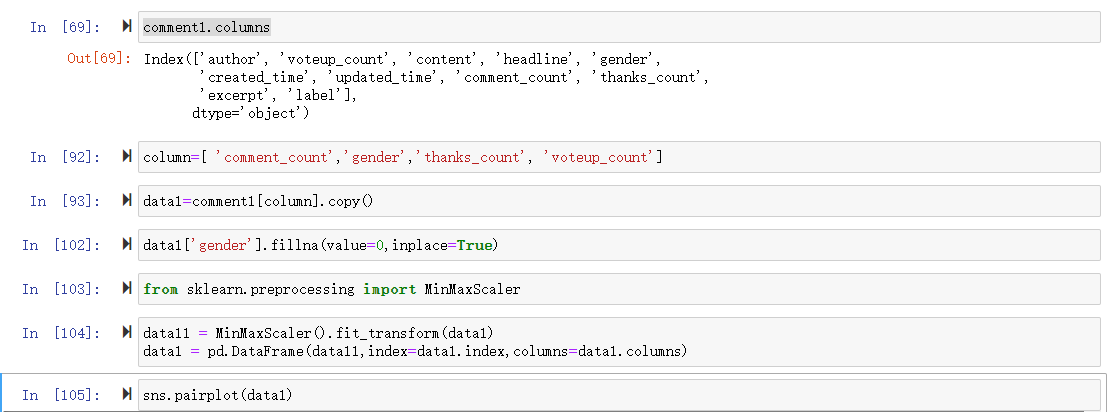
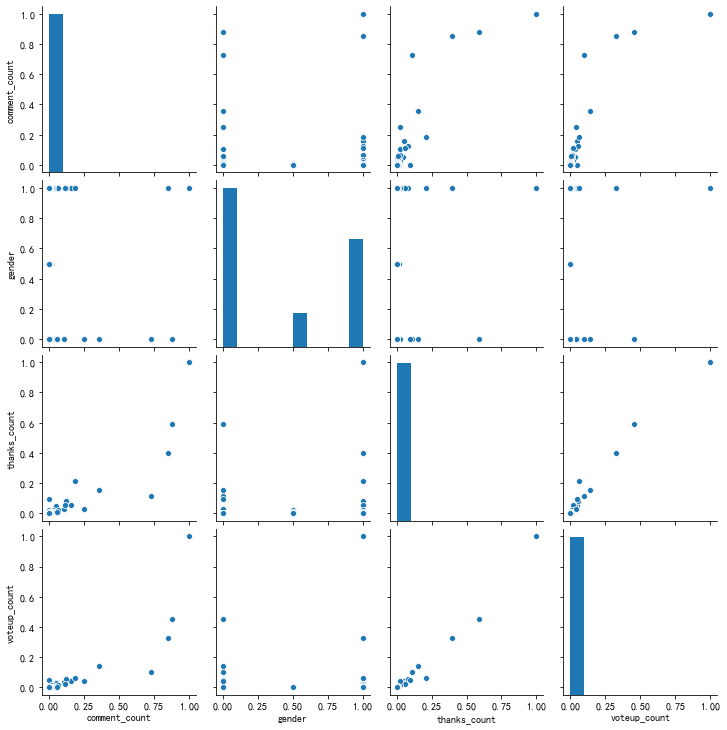
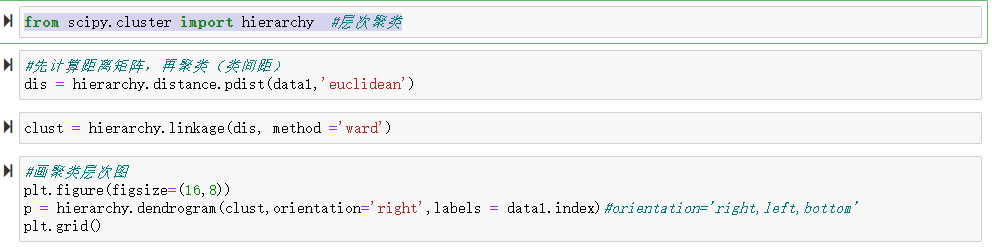
选取主要维度comment\_count，gender，thanks\_count，voteup\_count，对性别进行缺失值处理，整体数据进行标准化处理，将处理后的数据转化为DataFram格式，输出pairpolt图。对角线上是各个属性的直方图（分布图），而非对角线上是两个不同属性之间的相关图。

图4：数据处理

图5：pairpolt图

**3.层次聚类**

由树状聚类图可知，分4类比较合适

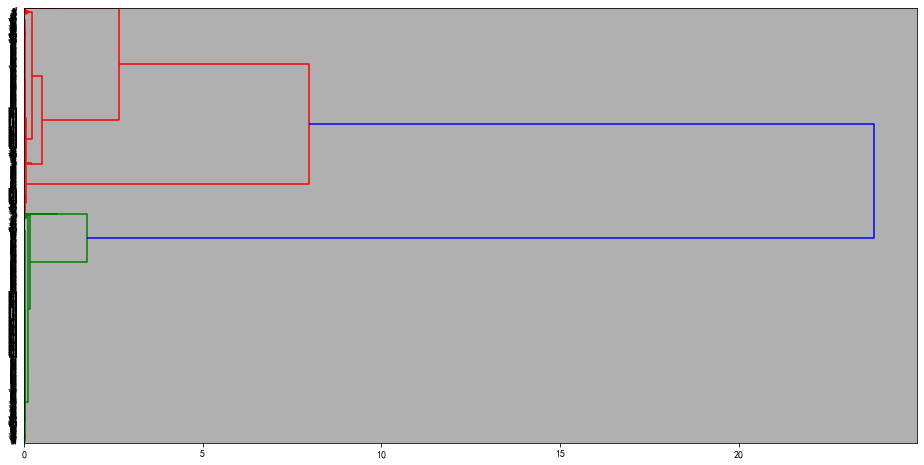
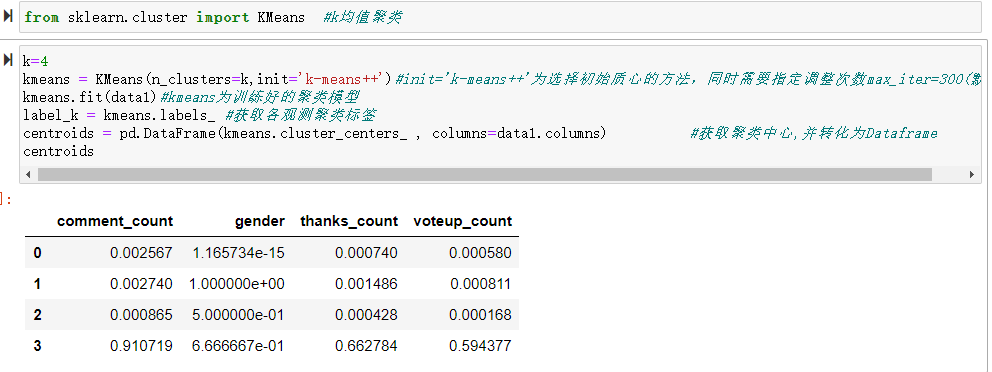
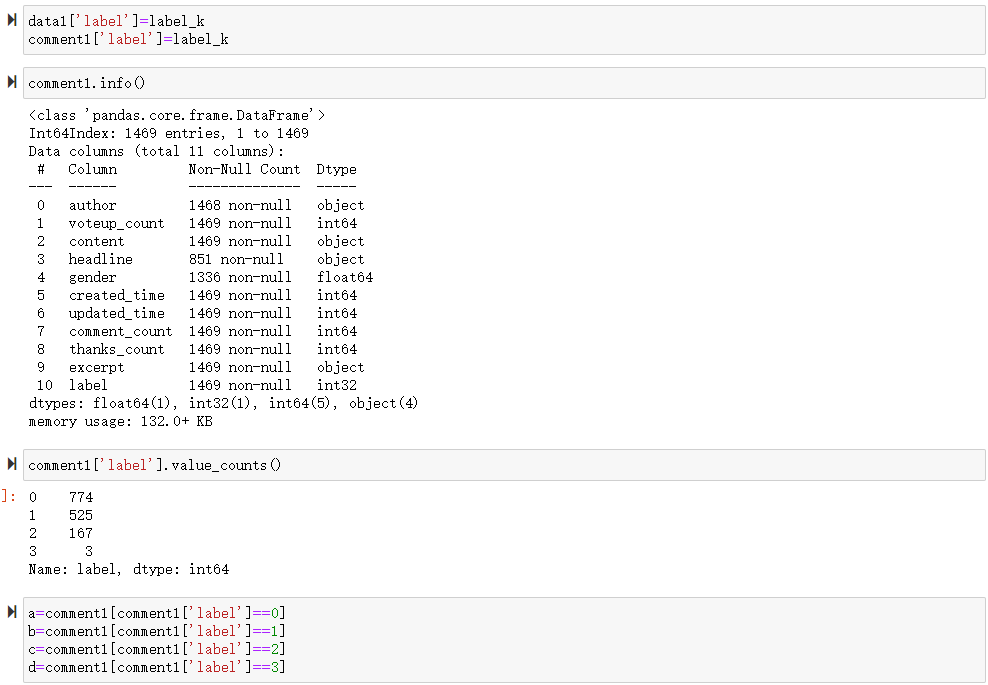
图6：层次聚类代码

图7：树状聚类图

**4.K均值聚类**

用kmeans聚类获取各观测聚类标签，并可以看出3类是点赞量极高的评论，0类，1类次之，最少的是2类，将其分别输出为四个文本文件，图9展示其中极高的三个

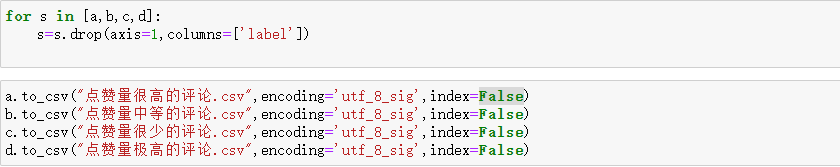
图8：K均值聚类代码

图9：点赞量极高的评论

**（三）文本特征分析**

主要调用jieba和wordcloud，去除掉无关列信息，设置禁用词，只保留中文字符。以点赞量很高.text为例，保存文本为a1.txt,最终输出词云图片点赞很高.png

代码如下：

import wordcloud

import jieba

import imageio

import matplotlib.pyplot as plt # plt 用于显示图片

import matplotlib.image as mpimg # mpimg 用于读取图片

import numpy as np

import pandas as pd

import re

a1=pd.read\_csv('点赞量很高的评论.csv',encoding='utf\_8\_sig')

del a1['voteup\_count']

del a1['headline']

del a1['created\_time']

del a1['updated\_time']

del a1['comment\_count']

del a1['thanks\_count']

del a1['excerpt']

del a1['label']

del a1['gender']

a1 = pd.DataFrame(a1, columns=['content'])

a1.to\_csv("a1.txt",encoding='utf\_8\_sig')

filename = r"a1.txt"

inf = pd.read\_csv(filename)

inf\_need = inf['content']

inf\_need = inf\_need.values

inf\_txt = ''

for i in range(0, 774):

inf\_txt = inf\_txt + str(inf\_need[i])

inf\_txt\_CN = re.findall(r'[\u4e00-\u9fa5]', inf\_txt)

string = ''

for i in range(0, 110000):

string = string + inf\_txt\_CN[i]

with open(r"a1.txt", "w", encoding='utf-8') as f:

f.write(string)

f.close()

## 设置禁用词

stopwords=set()

stopwords.add('我们')

stopwords.add('已经')

stopwords.add('还是')

stopwords.add('但是')

stopwords.add('应该')

stopwords.add('真的')

stopwords.add('什么')

stopwords.add('那么')

stopwords.add('其实')

stopwords.add('就是')

stopwords.add('没有')

stopwords.add('自己')

stopwords.add('一个')

stopwords.add('这些')

stopwords.add('因为')

stopwords.add('只是')

stopwords.add('这种')

stopwords.add('如果')

stopwords.add('不是')

stopwords.add('他们')

stopwords.add('这个')

stopwords.add('这么')

stopwords.add('怎么')

stopwords.add('比如')

stopwords.add('当然')

stopwords.add('怎么')

stopwords.add('而是')

stopwords.add('或者')

stopwords.add('比如')

stopwords.add('你们')

stopwords.add('一下')

stopwords.add('是不是')

stopwords.add('等等')

stopwords.add('多少')

stopwords.add('如何')

stopwords.add('一天')

stopwords.add('一点')

stopwords.add('的话')

stopwords.add('也许')

mk = imageio.imread('1.jpg')

w = wordcloud.WordCloud(width=1000, ## scale 调整清晰度

height=700,

background\_color='white',

mask=mk,

stopwords=stopwords,

font\_path='msyh.ttc',

scale=10)

f = open('a1.txt',encoding='utf-8')

txt = f.read()

article\_contents = ""

#使用jieba进行分词

words = jieba.cut(txt,cut\_all=False)

for word in words:

#使用空格来分割词

article\_contents += word+" "

string=""

string = "".join(article\_contents)

w.generate(string)

w.to\_file('点赞很高.png')

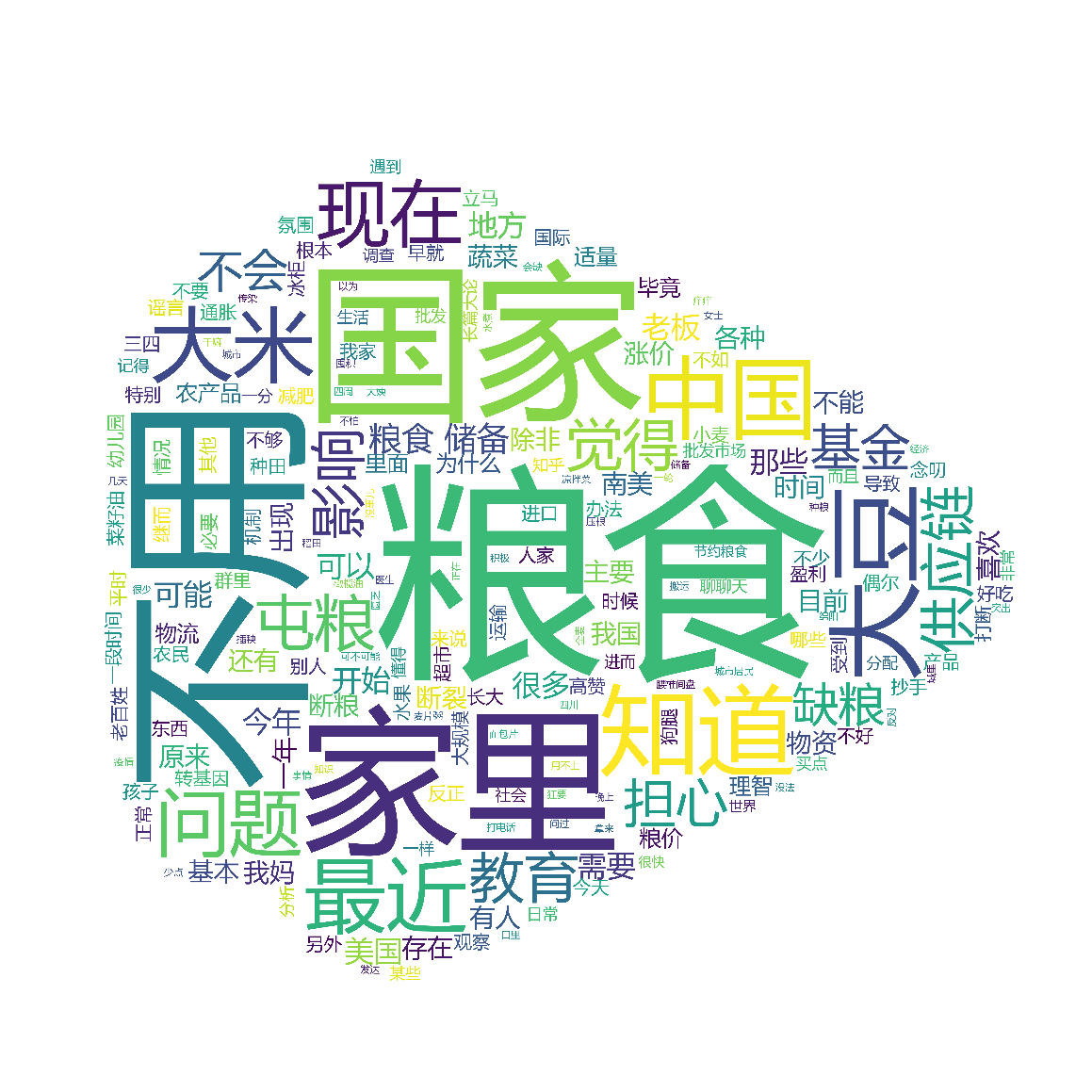
f.close()

替换掉保存文本和图片名称，可以得到其他三个点赞极高，点赞中等，点赞很少的词云图片

点赞很高的词云：图10

点赞极高的词云：图11

点赞中等的词云：图12

点赞很少的词云：图13

**三、结果分析**

通过Kmeans聚类将所有1469个回答划分为四类，对应的数量分别为极高赞3个，很高赞774个，中等赞525个，很少赞167个，输出对应的CSV文件，并针对每个文本进行词频过滤从而得到词云。

通过词云我们可以看见在知乎问题“疫情尚未结束，多国禁止粮食出口，老百姓需要囤积粮食吗？”下，

（1）四类用户都提及了关键词：国家(中国，我国)，粮食，大米，

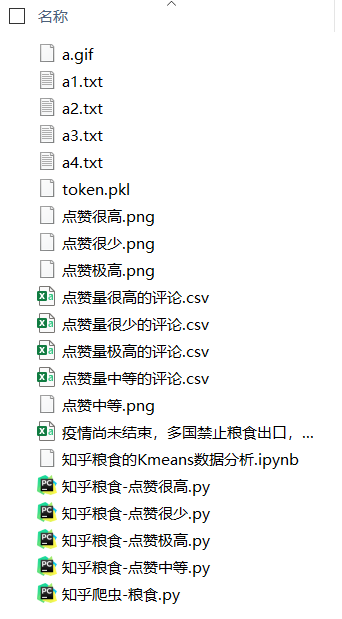
（2）极高赞用户评论的出发点在于我国粮食、大米、小麦、稻谷的国内的库存、消费，谈及了国际形势下的进口问题。该类用户知识面较广，且属于知乎上被关注数较多的大V

（3）很高赞用户评论集中在疫情对于我国大米、小米、大口的进口影响，并且很高频率的家里囤粮。该类用户推测属于文化教育水平较高的城市白领中产阶级，才会比较关注风险。

（4）中等赞用户评论聚焦于国内粮食、大豆的进口、储备情况。推测该类用户对袁隆平和耕地红线的提及，没有像很高赞用户和很少赞用户提及家里，而是用了大家，属于单身青年和在校学生群体。

（5）很少赞用户评论认为不太用屯粮，提及了大口和粮食。该类对应用户的评论字数普遍较少。

附：

本次论文产生的数据和代码汇总截图：图14