# 弹幕和评论区的信息特征研究(Barrage and comment)

### 研究对象：

本实验进行弹幕和评论区信息特征研究，所选择的视频网站实证对象为BiliBili平台，它是较早使用弹幕功能的视频在线播放平台之一。具有用户评论互动、热门评论，即时发送弹幕、弹幕点赞、弹幕字体颜色等功能。

### 研究的背景：

随着BiliBili弹幕功能的推出，视频的交互性和及时沟通性显著得以提高，视频的距离感更加拉近。弹幕的受众和影响力不断扩大，弹幕和评论信息数量也呈指数级增长，这同时也带来了大量无用信息的弹幕和评论。弹幕和评论的信息有用性检测变得重要，这也能凸显信息有用性检测的学术价值和商业价值。

### **意义：**

本课程实验主要进行三个方面实际操作：一是探索用户自身特征与信息有用性关系。二是探索文本特征，包括文本长度等特性。三是探索信息特征，实验将弹幕关键词作为主要衡量指标。本实验以检测弹幕信息有用性作为目标，从用户特征、表达形式，弹幕和评论的信息效用特征三个角度得到各等级信息有用性的检测结果，概括相关特征作为具有决定性的影响变量。对弹幕文本内容进行挖掘与审核，能正确引导视频观看用户合理讨论，发送文明弹幕，有助于营造和谐稳定的视频观看环境。

### 文件说明：

* data：存取在不同视频爬取到的数据的文件夹。
* Barrage.csv：弹幕信息文件。
* Comment.csv：评论信息文件。
* Barrage.png：弹幕云图文件。
* Comment.png：评论云图文件。
* emoji.png：表情包云图文件。

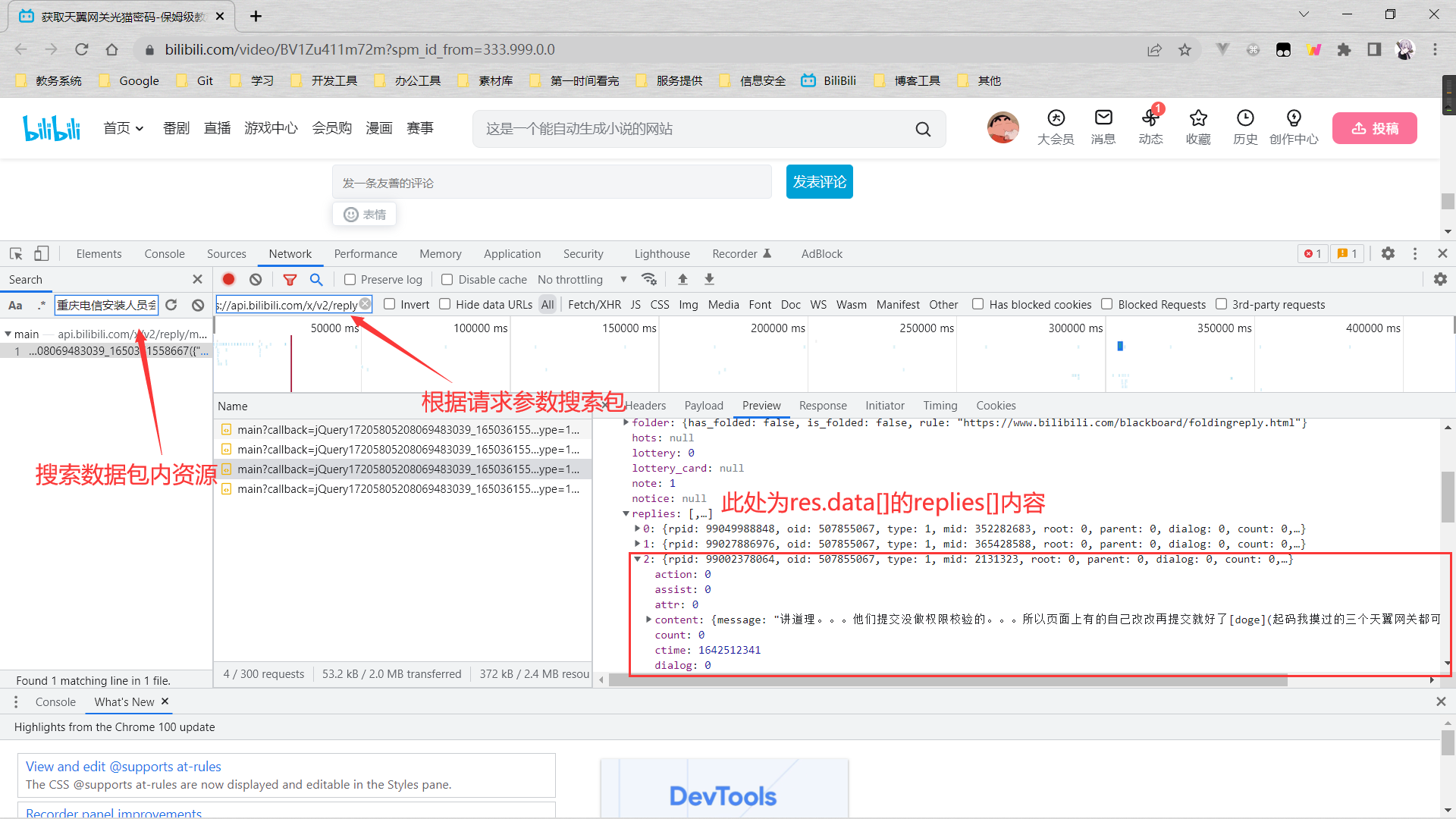
### Python环境

* Python3.9
* pandas
* numpy

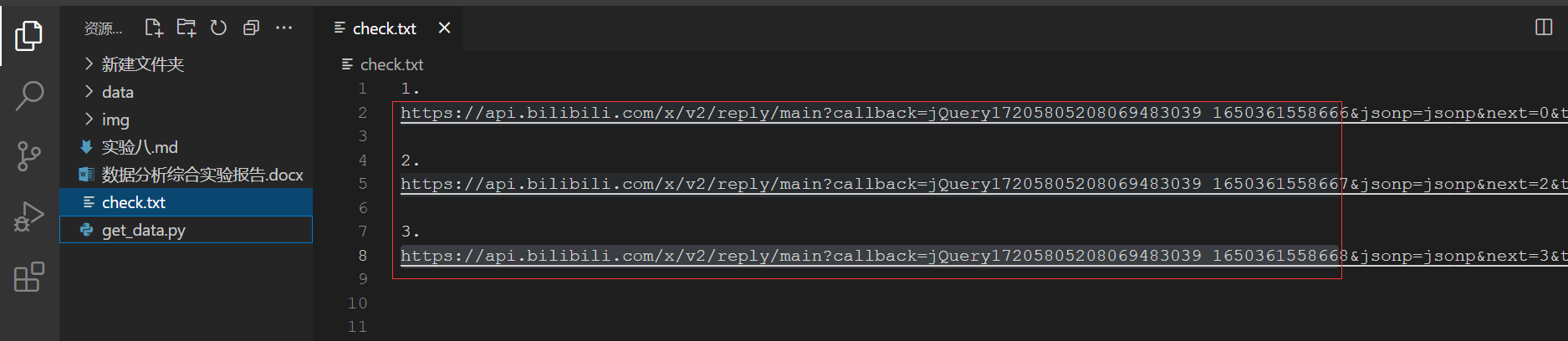
## 数据获取

### 数据获取方式

通过爬虫爬取获取数据。



**相关理论：**浏览器进入哔哩哔哩官网，打开一个视频。进入开发者工具，清除数据包后，向下拖动评论区可以发现哔哩哔哩的评论区采用的是懒加载模式。浏览器发送GET请求，服务器收到请求后返回数据一次返回固定的长度。当页面显示的长度大于浏览器返回包的长度时进行懒加载一次，加载一次浏览器发送一次GET请求，服务器返回一次数据。返回数据里的 res.data包括了评论区里面的真实reply。找到这些回答的数据包以后，删除相应包的参数，可以发现请求的浏览器发出的请求的url，将url复制到筛选框内，筛选对应的包，可以找到所有请求评论评论的所有url。通过抓包分析到三个url即可得到请求评论区数据的请求包规律。



**爬取评论区确定到的请求网址**

start\_url = https://api.bilibili.com/x/v2/reply/main?jsonp=jsonp&next={}&type=1&oid=507855067&mode=3&plat=1&\_=1650361573280

**爬取弹幕确定到的请求网址**

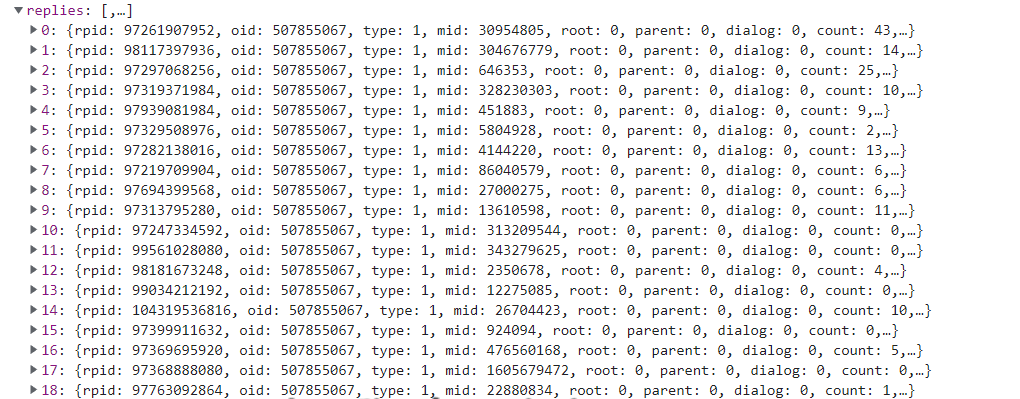
https://api.bilibili.com/x/v2/dm/web/seg.so?type=1&oid=474033384&pid=507855067&segment\_index=2

## 数据探索

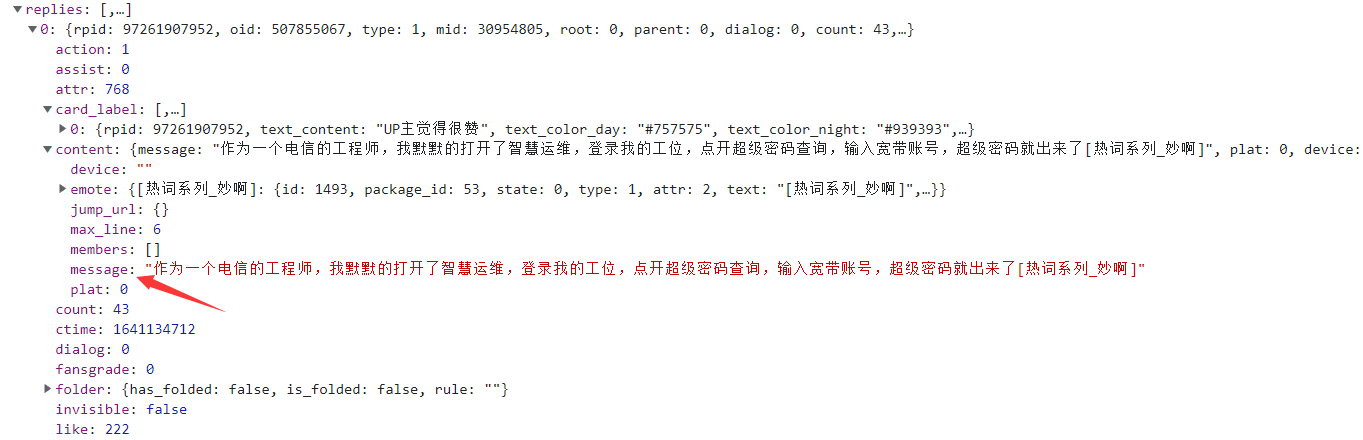
### 数据基本信息

（简要介绍数据的基本信息，如数据文件类型，包含的内容等等）

replies为有逻辑编号的列表



需要的数据在replies[0].content.message



一级评论：https://api.bilibili.com/x/v2/reply/main?jsonp=jsonp&next=0&type=1&oid=208143004&mode=3&plat=1

next：翻页  
 oid：视频编号（aid）  
 mode：1、2表示按热度、时间排序; 0、3表示按热度排序，并显示评论和用户信息

二级评论：https://api.bilibili.com/x/v2/reply/reply?jsonp=jsonp&pn=1&type=1&oid=208143004&ps=10&root=5453611704

pn：翻页  
 oid：视频oid  
 root：楼主的回复的rpid  
 ps: 单页显示数量（最大为20）

### 数据读取

读取数据使用了requests.get函数。通过requests.get(url, headers)构造一个向服务器请求资源的url对象。这个对象是手动生成的。这时候的返回的是一个包含服务器资源的Response对象。包含从服务器返回的所有的相关资源。

**获取数据代码：**

def main():  
 print('爬取懒加载节点{}的数据!'.format(i))  
 start\_url = f'https://api.bilibili.com/x/v2/reply/main?jsonp=jsonp&next=1&type=1&oid=507855067&mode=3&plat=1&\_=1650361573280'  
 headers = {  
 'user-agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.4896.88 Safari/537.36',  
 'referer': 'https://www.bilibili.com/video/BV1Zu411m72m?spm\_id\_from=333.999.0.0'  
 }  
 response = requests.get(start\_url, headers = headers).json()  
 ic(response)

由于本机的cmd的下载会调入Acaconda模块，而vscode的控制台会调用此处C:\Python310\python.exe，需要手动加Acaconda路径，或者将Acaconda的环境变量暴露出去，就可以解决该问题

&D:\Data\Acaconda\python.exe D:\Study\course\Python\NO8\get\_data.py

## 数据清洗与预处理

**评论信息数据清洗**

提取replies.message内容，该内容即为评论信息

replies = jsonpath(response,'$..replies')  
message\_list = jsonpath(replies,'$..message')  
name = message['member']['uname'] # 昵称  
sex = message['member']['sex'] # 性别  
sign = message['member']['sign'] # 签名  
rcount = message['rcount'] # 回复数  
like = message['like'] # 点赞数  
content = message['content']['message'] # 评论内容

**弹幕信息数据清洗**

提取response.text内容，该内容即为弹幕信息

在线正则表达式测试 https://tool.oschina.net/regex#

要提取中文，测得正则式[\u4e00-\u9fa5]+，但是空格会造成输出换行，使用.\*?([\u4e00-\u9fa5]+).\*

res = re.findall('.\*?([\u4e00-\u9fa5]+).\*',response.text)

## 数据分析及可视化

def parse\_img(big\_list):  
 print('----------词云生成中------------')  
 data = ''.join(big\_list)  
 stylecloud.gen\_stylecloud(data,font\_path="C:/Windows/Fonts/simfang.ttf")  
 img = Image.open("D:/Study/course/Python/NO8/stylecloud.png")  
 img.show()  
 print('----------词云已生成------------')

### 个体差异性

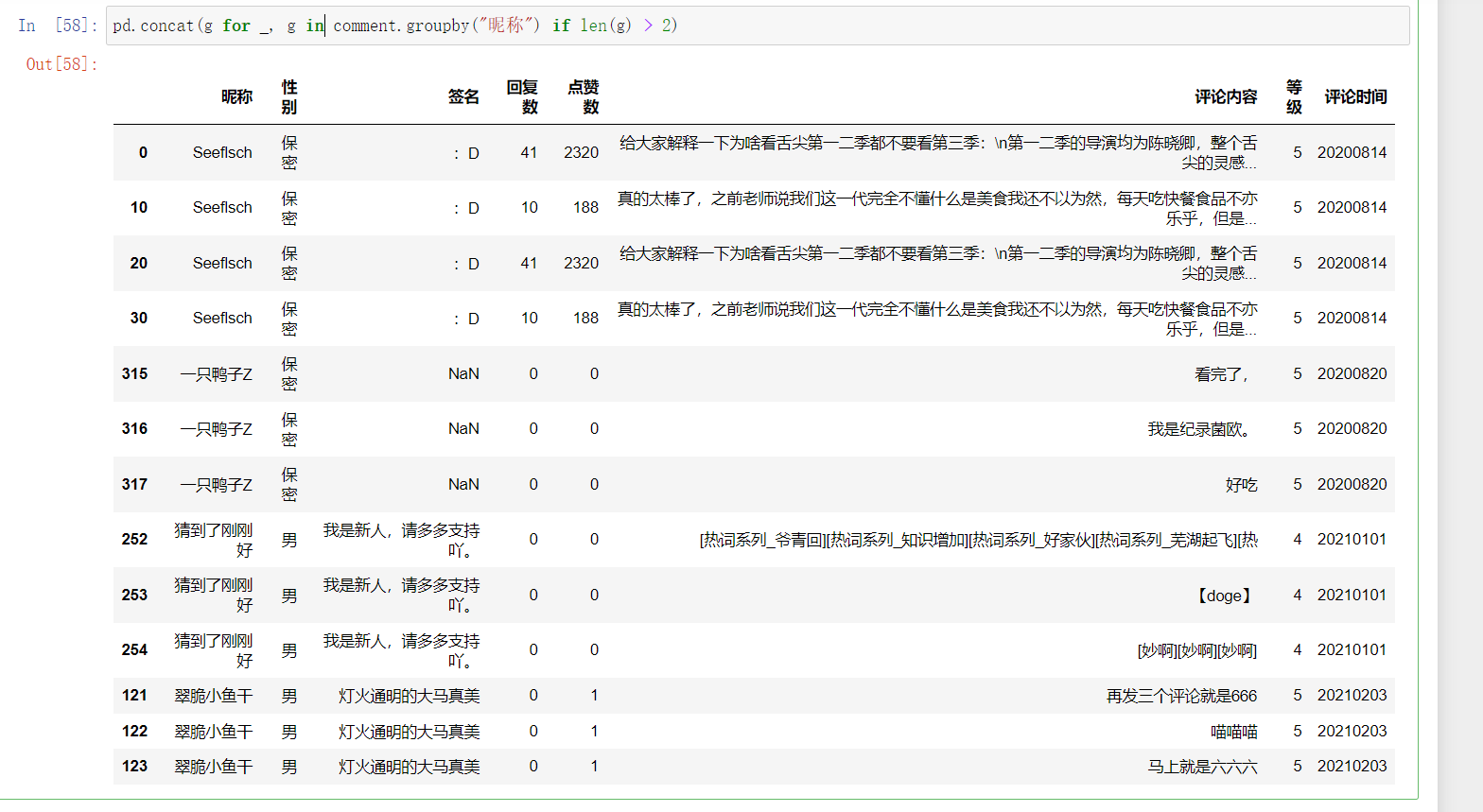
男生: 平均值为: 4.99，方差为: 0.63  
女生: 平均值为: 4.84，方差为: 0.75

个体用户特征里的用户会员等级差异，如上方数据所示。大量用户会员等级普遍较高，均值为4.8级左右，且差异性较为不明显。

11111111111

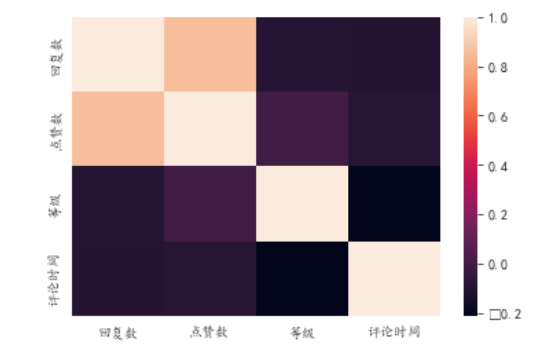
用户发送会员等级，与回复度和点赞数关系如上方表格所示。会员等级反映了对视频平台的时间或金钱投入，个体用户特征里的高级会员，UP主能够有更多的效果或文字，会员等级高的用户具有更强的依赖性。因等级的升级需要花费大量时间和用户活跃度，所以达到六级是较为难的事。不同的会员等级的用户，其弹幕信息使用经验和倾向可能有所差异，对弹幕信息有用性有一定影响。

### 针对多次评论的异常账户分析



这里仅仅从同一昵称发送的评论数进行分析，并不是严谨的。部分人会多次评论与回复互动，或一段话发送出去后进行补充回复。但可以看到检测到某一昵称是否存在恶意刷评论的行为，和某一昵称在一个视频里的互动过程。

### 评论区热图



我们可以很快看到一些相关性；

1. 点赞数与回复数呈强正相关。
2. 回复数与等级 和评论时间之间存在弱负相关。
3. 点赞数与评论时间之间存在轻微负相关；

有了这些信息，可以做一些观察。

1. 回复数和评论时间弱负相关，bilibili平台似乎不会根据评论时间来选择热门回复评论，这是一个有待探索的假设。
2. 点赞数与评论时间之间存在轻微负相关，哔哩哔哩平台似乎不会根据评论时间来选择热门评论。因此数据是按热门度添加的，先添加热门评论，最后添加一般评论。如果打算使用这些数据来建立一个模型，那么最好在将其分解或者对其进行随机化。这是我们可以探索的另一个假设。

### 评论内容字符数大小

11111111111

《舌尖上的中国第一季第七集》分析收集视频评论区数据发现，字符数小于20的有：占比多于50%，大部分评论文本长度为10个字左右，可见一半的用户倾向于即时发表感受或想法，不到1/10的用户会考虑根据热点主题进行针对性长篇评论。

### 评论关键字频率云图



从传播学角度研究，网络媒体有丰富的符号化表达方式，评论大致可分为文字类和非文字类符号，《巅峰时期的周杰伦有多恐怖》收集分析视频评论区数据后，发现关键词大部分为文字类符号，而《舌尖上的中国第一季第七集》分析收集视频评论区数据发现大部分表现为观众当时的观看心情。

### 评论区表情包数据频次分析

11111111111

### 评论区表情包云图



评论区的表情包有其独特语境，在语音、文字和语义方面不同于其他网络语言。可见弹幕信息时效性较高且表现稳定，主题相似度偏低。其中最受大家喜爱的是doge表情。



### 弹幕字数分析

11111111111

可见绝大部分弹幕文本长度为10个字符以内。弹幕文字较短，但是具有丰富的表现形式。用户可能更注重信息内容，对于弹幕表达的规范性和完整度不太关注。弹幕也改变了用户的视觉体验，增强了视频的亲近感。

### 弹幕云图



《舌尖上的中国第一季第七集》分析收集视频弹幕数据发现，信息有用性在不同程度上取决于个人的主观认知。弹幕信息中传递的大都是用户当时情绪。

## 心得体会

相关性有助于探索新的数据集。例如使用seaborn的热图，在很短时间就可以看到输入数据的相关性，并得到想法来探索问题，比较容易看到最强的相关性在哪里。

本实验尚有许多值得改进的地方。例如评论区的分析中，文本字数占据重要贡献度，但是其他带有感情的特征项无法做到有用性检测，未能涵盖文本语言上的特征。在全面性上也有进一步提高的空间，算法分类精度效果有限，时间允许情况下应使用机器学习算法进行模型训练，对情感文本与含义进行更深层次挖掘，进行深度分析。如果要使用这些数据来建立模型，应该再将原始数据分解或者对其进行随机化。

学生签字：

提交日期：2022年04月23日