

基于 b 站弹幕的 数据挖掘与分析

唐欣妍 10225501455

华东师范大学

2022 级 数据科学与大数据技术

目录

- 1、引言.....3
 - 1.1 研究背景.....3
 - 1.2 研究方法.....3
 - 1.3 研究难点.....3
- 2、数据获取.....4
 - 2.1 目标数据.....4
 - 2.2 代码实现.....5
 - 2.3 数据处理.....6
- 3、数据分析与可视化.....7
 - 3.1 分析 b 站排行榜前十的视频.....7
 - 3.1.1 弹幕情感分析.....8
 - 3.1.2 弹幕词云图.....10
 - 3.1.3 弹幕在视频的出现时间分析.....11
 - 3.1.4 弹幕发送日期分析.....11
 - 3.2 分析比较标题党与非标题党视频的弹幕情况.....12
 - 3.2.1 对“标题党”进行操作性定义.....12
 - 3.2.2 获取标题党与非标题党视频弹幕数据.....13
 - 3.2.3 两种视频的弹幕数量比较.....14
 - 3.2.4 对两种视频的弹幕进行情感分析.....14
 - 3.3 数据可视化整体看板图.....15
- 4、总结与展望.....15
 - 4.1 结论.....15
 - 4.2 建议.....16
 - 4.3 不足与改进之处.....16
 - 4.4 展望.....16

1、引言

1.1 研究背景

近年来，互联网行业不断繁荣发展，作为其热门产品之一的视频网站亦是吸引了一大批用户。其中，哔哩哔哩(www.bilibili.com，以下简称“b站”)是一个成功的例子，它以极具互动性的弹幕为特色，鼓励用户们积极参与视频互动。很多用户会为视频发送弹幕，在视频引起共鸣时刻发送的尤为多。

因此，笔者猜测弹幕能在一定程度上反映用户的情感，而据此也可以探究视频的一系列特征等等。同时，从弹幕的发送时间或许可以看出视频的热点时段、以及用户登录b站的喜好时间等信息。

为此，笔者将挖掘b站部分弹幕信息，对相关数据进行分析及可视化操作。

1.2 研究方法

在Windows操作系统下，利用Visual Studio Code环境编写Python代码爬取b站特定弹幕数据——弹幕内容、弹幕在视频中的时间、弹幕发送时间，并同时利用Python的多个库及DataEase网站进行数据可视化与分析。

1.3 研究难点

①一开始尝试在类似以下界面获取红笔圈出来的弹幕信息，然而得到下面第二张图的结果：




```
2023-12-09 11:50:28 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.bilibili.com/video/BV1A94y1E7Gt/?vd_source=0b78dd91e96cec5d5636dee0a2bc41af> (referer: None)
<!DOCTYPE html><html><head><title>验证码_哔哩哔哩</title><meta name="viewport" content="width=device-width,user-scalable=no,initial-scale=1,maximum-scale=1,minimum-scale=1,viewport-fit=cover"><meta name="spm_prefix" content="333.1291"><script>window._BiliGreyResult={"method":"base","grayVersion":"76"}</script><script type="text/javascript" src="//www.bilibili.com/gentleman/polyfill.js?features=Promise%2CObject.assign%2CString.prototype.includes%2CNumber.isNaN"></script>
```

这说明利用爬虫爬取该网站会被网页验证码拦截，无法从中获取数据。
因此笔者在网上寻找解决方法，得知可以尝试别的网站——在原网址的“bilibili”前加“i”，跳转到如下网址，并进一步进到下图红框圈起来的网址。

ibilibili.com/video/BV1A94y1E7Gt/?spm_id_from=333.880.my_history.page.click&vd_source=0b78dd91e96cec5d563ddee0a2bc41af

“蜡笔小新从来不是幼稚片”-“蜡笔小新从来不是幼稚片”



0:00

视频图片 作者头像 弹幕 高清视频下载(不支持版权的视频下载) 爆鱼人的网站 无魔法GPT

AID: 366511988

CID: 1348525291

视频图片: http://i2.hdslb.com/bfs/archive/ed6f10d8bb9118c5e9e7d0c6ecbb4d4dd8f73059.jpg

作者头像: https://i1.hdslb.com/bfs/face/6f8de9082e9d1290d2e2a732444917974b4a9c03.jpg

弹幕地址: https://api.bilibili.com/x/v1/dm/list.so?oid=1348525291

视频描述: “蜡笔小新从来不是幼稚片”-小新, 你是否有许多问号

【一刻talks】郭明: 精准定制 正在到来的未来旅行趋势
旅行的意义到底是什么, 是放松、开阔眼界又或是寻找本真呢? 相信每个人都有自己的需

3分钟看懂插管、呼吸机 and 价格
3分钟看懂插管、呼吸机 and 价格

FINK! ALL COUPLE MV《小船》(韩版自制)
土豆 转自ICFINK!, 近来很想李利姐, 很想FINK!啊, 哈哈, 太暴露年龄了,

[双语]分析是魔法: 闪电到底该不该烧?
youtube 原名 Analyzing Is Magic: What's Wr

【美国广告】麦当劳从来不让别人失望 拍广告主要目的是为了
忍饥受渴吧
https://mobile.r.ty/mission/#/share/vid

2、数据获取

2.1 目标数据

在上一步中, 笔者找到了能够获取弹幕信息的网址 (且可以实现爬虫), 跳转至此, 我们可以看到如下界面:

api.bilibili.com/x/v1/dm/list.so?oid=1348525291

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<?xml version="1.0"?>
<chatserver chat.bilibili.com/>
<chatid 1348525291/>
<mission 0/>
<maxlimit 1000/>
<state 0/>
<real_name 0/>
<source k v/>
<p 63.29900,5.25,16777215,1701438585,0,44eebf0,1462531871903671808,10>好犀利的文字</p>
<p 59.71400,5.25,16777215,1701268515,0,ff25670f,1461259115765484289,10>火钳刘明</p>
<p 91.86900,5.25,15138834,1701437879,0,2ab9363b,146252595169423344,10>动感光波! biubiubiu</p>
<p 10.31900,1.25,15138834,1701437131,0,69216880,1462519679321968128,10>我们心理课的材料</p>
<p 80.48400,5.25,41194,1701432427,0,1c3547cc,1462480217993018624,10>成长是个艰难的过程</p>
<p 114.47300,5.25,16707842,1701430627,0,a08593c2,1462465118490207232,10>蒙古.D.黄狼</p>
<p 49.27200,5.25,16707842,1701425945,0,346a363a,146242584372411104,10>不许骂人! <doge></p>
<p 139.52300,1.25,16777215,1701418185,0,4f441b1b,1462360748150944512,10>我和天空的距离少了两米左右, 我长大了啊……</p>
<p 70.18700,5.25,16777215,1701406131,0,e307f265,1462259626199452672,10>阿梅! </p>
<p 120.98900,1.25,16777215,1701405628,0,8e952fc,1462255411796483328,10>海军大将</p>
<p 40.01200,1.25,16777215,1701390927,0,bf0b0544,1462132089536302080,10>压ao</p>
<p 30.06300,5.25,15138834,1701357224,0,a39e4d3c,1461849370403516416,10>这里是fire老师离开的时候, 妮妮偷偷哭</p>
<p 132.28300,1.25,15138834,1701327364,0,194bdf9,1461598884915723776,10>年少以为不在乎人言可畏, 长大却发现最怕文字间的触动……</p>
<p 105.80900,1.25,16777215,1701322134,0,a871539f,1461555014274407680,10>我去, 破防了。超人已被击倒</p>
<p 19.92500,1.25,16777215,1701307156,0,33c21e6,1461429369292768768,10>龙的填词绝</p>
<p 5.19400,1.25,16777215,1701268458,0,f4e6f42e,1461259115924867840,10>最壮观激动了情</p>
<p 99.45000,1.25,16777215,1701262515,0,1d76ea13,1461259115480271616,10>大人为什么觉得动画无聊 难道他们的超人已被击倒</p>
<p 23.43300,5.25,16765698,1701255805,0,341cec72,1461259115278945027,10>百 万 填 词</p>
<p 4.98000,1.25,16777215,1701517356,0,901f5d8e,1463192651280133376,10>什么b音乐……</p>
<p 103.24100,4.25,16777215,1701270888,0,2184ef26,1461259115773872896,10>会火, 放一个超人</p>
<p 101.98700,5.25,16776960,1701258248,0,ea4c5fbd,1461259115413162752,10>难道他们的超人已被击倒 (百万填词) </p>
<p 105.42100,5.25,16646914,1701257308,0,fffe8014,1461259115278945026,10>我们的超人有没有被击倒呢? </p>
<p 2.33800,1.25,16777215,1701255659,0,b20324ee,1461259115278945024,10>Dada la </p>
<p 106.76900,1.25,16777215,1702336727,0,3650c64b,1470066033581811968,10>必须三连</p>
<p 108.25100,1.25,16777215,1702306703,0,e58a28f,1469814172413009664,10>牛逼</p>
<p 44.62300,1.25,16777215,1702306621,0,3f25d11c,1469813489488044288,10>呆神yyds</p>
<p 32.46500,1.25,16777215,1702306603,0,3f25d11c,1469813335666144256,10>妮妮和正男是一对哇……</p>
<p 146.92900,1.25,15138834,1702304327,0,e33f839b,1469794243286783744,10>666666666</p>
<p 149.89600,5.25,16765698,1702290737,0,457566b3,1469680239042313472,10>我取消了点赞, 但是长按</p>
<p 37.48000,1.25,16777215,1702283774,0,d757berb,1469621832201264640,10>压压压压压压</p>
<p 100.28700,1.25,16777215,1702282976,0,fad93c64,1469615136288319488,10>点开3秒我就三连的视频真不多</p>
<p 56.46400,1.25,15138834,1702281912,0,bcec4adf,1469606215230572288,10>大爱小新</p>
<p 26.03500,1.25,16777215,1702227269,0,3cb68e5,1469147830517358080,10>1.54听</p>
<p 112.70000,5.25,15138834,1702222806,0,9a840301,1469110391681416192,10>泪目了</p>
<p 37.46600,1.25,16777215,1702217638,0,d872e6e3,1469067042844963328,10>是不是担心 变成一只野兽</p>
```

在这个界面中, 每一行 p 之后的第一个数字表示这条弹幕在视频中出现的 (如视频的第 1 分 22 秒), 第五个数字表示弹幕的发送时间戳 (代表了日期及时间, 如 2023/12/10 10: 00), 而最后的黑色字表示了弹幕内容。

为了详细分析弹幕内容和弹幕时间的信息, 笔者选取这三类数据为所要获取的目标数据。

2.2 代码实现

对于网站数据，我们需要用爬虫代码去获取。在本次项目里，笔者调用了 Python 的 requests 库，并基于这个库来爬取 b 站弹幕数据。

```
import requests
import xml.etree.ElementTree as ET
import pandas as pd
```

调用相关的库

```
#获取弹幕
def get_bilibili_danmaku(cid):
    url = f"https://comment.bilibili.com/{cid}.xml" #f前缀代表格式化字符串，里面可以包含{cid}这种占位符，使其之后能被替换
    response = requests.get(url)

    if response.status_code == 200:
        return response.content
    else:
        print(f"Error accessing the API. Status Code: {response.status_code}")
    return None
```

输入视频的 cid 号，从对应网站上获取弹幕

```
#解析弹幕
def parse_danmaku(xml_content):
    root = ET.fromstring(xml_content) #解析xml格式的字符串
    danmaku_list = []

    for d in root.iter('d'): #遍历xml树中所有标签为'd'的元素
        danmaku_text = d.text
        danmaku_attr = d.attrib #返回d的属性字典
        time_info = danmaku_attr.get('p', '').split(',')

        if len(time_info) > 0:
            video_time = int(float(time_info[0])) # #弹幕在视频中的出现时间（秒），转为整数秒
            hours = video_time // 3600
            minutes = (video_time % 3600) // 60
            seconds = video_time % 60
            real_time = f"{hours}:{minutes}:{seconds}" # 格式化时间，取整到小时、分钟、秒'''

            date = time_info[4]

            danmaku_list.append({
                '弹幕': danmaku_text,
                '视频中时间': video_time,
                '发送时间戳': date
            })
    return danmaku_list
```

对弹幕进行解析

```
def save_danmaku_to_excel(danmaku_list, filename): #将获取的弹幕信息保存到excel中
    df = pd.DataFrame(danmaku_list)
    df.to_excel(filename, index=False, engine='openpyxl')
    print(f"保存了 {len(danmaku_list)} 条弹幕到 {filename}")
```

为了方便后续操作，将获得的弹幕数据以 excel 文件存储

```
if __name__ == '__main__':
    cid = input('输入视频的cid:')
    danmaku_xml = get_bilibili_danmaku(cid)

    if danmaku_xml:
        danmaku_list = parse_danmaku(danmaku_xml)
        filename = f"{cid}的弹幕.xlsx"
        save_danmaku_to_excel(danmaku_list, filename)
    else:
        print("无法获取弹幕。")
```

主函数实现

2.3 数据处理

运行以上的代码，我们可以获取这样形式的 excel 文件：

#	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	弹幕	视频中时间戳	发送时间戳												
2	好犀利的	0h 1min	1701438585												
3	火钳刘明	0h 0min	1701268515												
4	动感光波!	0h 1min	1701437879												
5	我们心理!	0h 0min	1701437131												
6	成长是个	0h 1min	1701432427												
7	蒙奇.D.黄	0h 1min	1701430627												
8	不许骂人	0h 0min	1701425945												
9	我和天空!	0h 2min	1701418185												
10	阿梅!	0h 1min	1701406131												
11	海军大将	0h 2min	1701405628												
12	压ao	0h 0min	1701390927												
13	这里是fi	0h 0min	1701357224												
14	年少以为	0h 2min	1701327364												
15	我去，破	0h 1min	1701322134												
16	龙的填词	0h 0min	1701307156												
17	最怕鬼畜	0h 0min	1701268458												
18	大人为什	0h 1min	1701262515												
19	百万填	0h 0min	1701255805												
20	什么b音乐	0h 0min	1701517356												

首先，笔者对获取的数据进行了基本的去重、清洗等处理工作。并把数据存到了 MySQL 中（利用 create 语句创造了相应的表格）。

然而，上图的“发送时间戳”是对时间的一种特殊表示，对数据分析有一定的难度与阻碍。而笔者希望能得到日期格式的发送时间（例如“xx 年 xx 月 xx 日 xx 时”），便于后续的直观分析。因此笔者在 excel 里编写函数并调整数字格式，将“发送时间戳”转为所需要的形式，如下所示：

LOGIO		✖ ✓ fx		=(C2+8*3600)/86400+70*365+19	
A	B	C	D		
1 弹幕	视频中时间	发送时间戳	时间戳对应的发送时间		
2 估计再临要献出自己的牛仔帽了	0h 14min	1702130456	=(C2+8*3600)/86400+70*365+19		

excel 中函数实现

弹幕	视频中时间	发送时间戳	时间戳对应的发送时间
好犀利的文字	0h 1min 3s	1701438585	2023/12/1 21:49
火钳刘明	0h 0min 59s	1701268515	2023/11/29 22:35
动感光波! biubiubiu~	0h 1min 31s	1701437879	2023/12/1 21:37
我们心理课的素材	0h 0min 10s	1701437131	2023/12/1 21:25
成长是个艰难的过程	0h 1min 20s	1701432427	2023/12/1 20:07
蒙奇.D.黄猿	0h 1min 54s	1701430627	2023/12/1 19:37
不许骂人! (doge	0h 0min 49s	1701425945	2023/12/1 18:19
我和天空的距离少了两米左右,我长	0h 2min 19s	1701418185	2023/12/1 16:09
河梅!	0h 1min 10s	1701406131	2023/12/1 12:48
每军大将	0h 2min 0s	1701405628	2023/12/1 12:40
玉ao	0h 0min 40s	1701390927	2023/12/1 8:35
这里是fire老师离开的时候,妮妮偷	0h 0min 30s	1701357224	2023/11/30 23:13
羊少以为不在乎人言可畏,长大却发	0h 2min 12s	1701327364	2023/11/30 14:56
我去,破防了.超人已被击倒	0h 1min 45s	1701322134	2023/11/30 13:28
龙的填词绝	0h 0min 19s	1701307156	2023/11/30 9:19
最怕鬼畜动了情	0h 0min 5s	1701268458	2023/11/29 22:34
大人为什么觉得动画无聊 难道他们	0h 1min 39s	1701262515	2023/11/29 20:55
百万填词	0h 0min 23s	1701255805	2023/11/29 19:03
什么b音乐.....	0h 0min 4s	1701517356	2023/12/2 19:42
会火,放一个超人	0h 1min 43s	1701270888	2023/11/29 23:14
难道他们的超人已被击倒~ (百万填	0h 1min 41s	1701258248	2023/11/29 19:44
我们的超人有没有被击倒呢?	0h 1min 45s	1701257308	2023/11/29 19:28
lada la	0h 0min 2s	1701255659	2023/11/29 19:00
好棒的填词啊,他们的超人是否已被	0h 2min 20s	1702132444	2023/12/9 22:34
个了,认可	0h 2min 29s	1702126979	2023/12/9 21:02
卓眼泪因为你们不玩扮家家酒	0h 0min 31s	1702126487	2023/12/9 20:54
唉,德朗医生好像在非洲遇到抢劫的	0h 1min 19s	1702125770	2023/12/9 20:42
火焰!!!	0h 0min 36s	1702125711	2023/12/9 20:41
刀我为什么要用童年啊?	0h 0min 18s	1702125679	2023/12/9 20:41
雅,实在是雅	0h 0min 4s	1702104017	2023/12/9 14:40

经处理过的数据形式

此外,为了便于对这些数据利用后续的情感分析代码进行分析,笔者将 excel 文件转化为了 csv (utf-8) 格式文件,这样后续代码可以按 csv 文件的列读取特定弹幕数据,且 utf-8 性质可以让其完整地保留弹幕的中文内容。

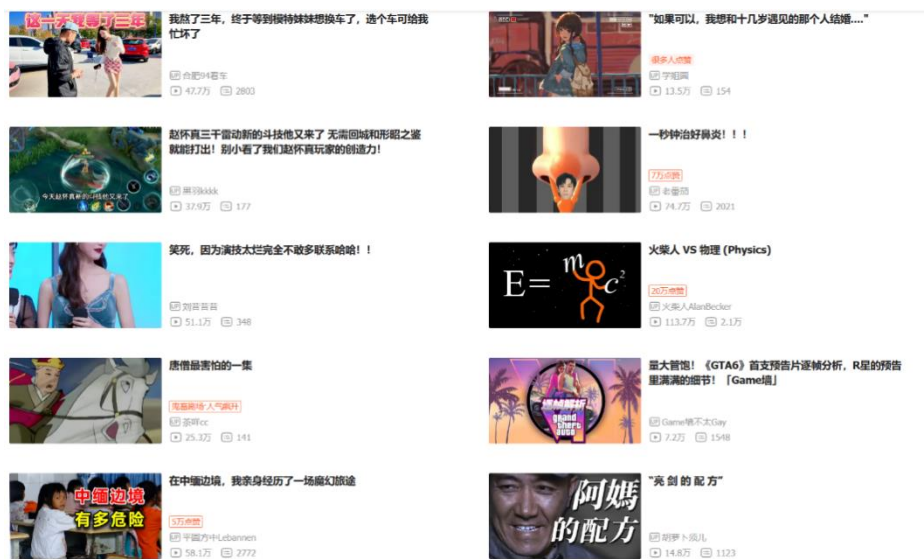


将 excel 文件转化为 csv (utf-8) 格式文件

3、数据分析与可视化

3.1 分析 b 站排行榜前十的视频

首先,笔者希望能探索 b 站热门视频的性质特征,因此选取当日 (2023 年 12 月 10 日) b 站排行榜前十的视频为代表,获取其弹幕信息。



当日 b 站排行榜前十的视频



爬虫获取的弹幕数据信息（excel 文件及对应 csv 文件）

由于这些热门视频弹幕数量普遍较多，少则几百条，多则几万条，因此此次爬虫累计获得了上万条数据信息，内容充实丰富。

3.1.1 弹幕情感分析

首先，由于笔者猜测弹幕能在一定程度上反映用户的情感，因此在 python 中调用 snownlp 这一情感分析的库，对每一个视频的弹幕进行情感分析，并计算出每个视频弹幕中积极、消极或中立情绪的占比，利用 matplotlib 的子库 pyplot 绘制出相应的饼图。


```

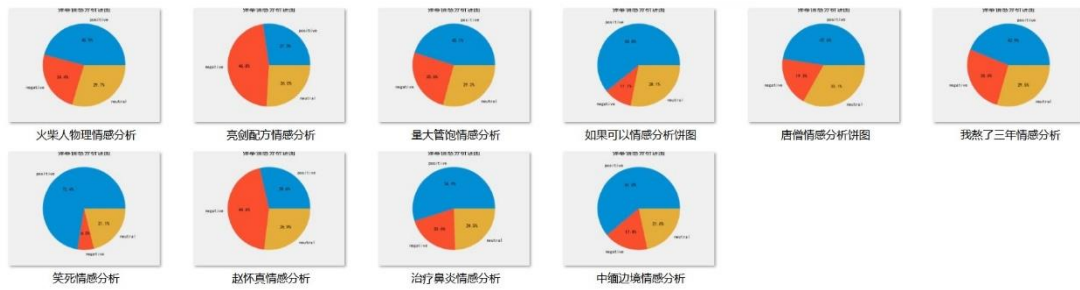
file = open('火柴人物理_弹幕.csv', encoding = 'utf-8')
text = []
for line in file.readlines():      #依次读入弹幕文件的每一行
    content = line.split(',')
    text.append(content[0])

emotions = {'positive':0,
            'negative':0,
            'neutral':0}

for item in text:
    s = SnowNLP(item)
    if s.sentiments > 0.6:
        emotions['positive'] += 1
    elif s.sentiments < 0.4:
        emotions['negative'] += 1
    else:
        emotions['neutral'] += 1
    #print(item)

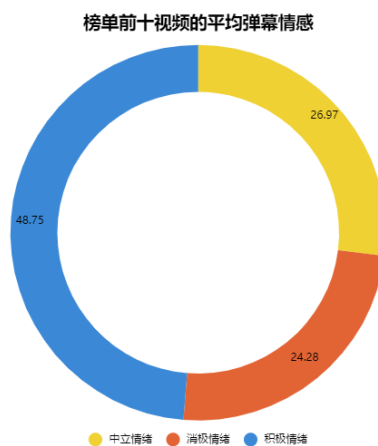
```

利用 snownlp 对每个视频弹幕进行情感分析



绘制出弹幕的情绪占比图（蓝色：积极，黄色：中立，红色：消极）

为了便于看这些视频的共同特征，我们将这些数据汇总，计算出榜单前十视频的平均情感：



可以看出，在这些视频中，积极情绪的弹幕普遍占了将近一半，而中立情绪弹幕数与消极情绪弹幕数接近，消极情绪的弹幕略少。

因此得出结论：热门视频往往是能调动用户们积极情绪的。也由此引发对视频创作者的建议：预测能够引发观众积极情绪的内容并围绕其重点展开。

3.1.2 弹幕词云图

同时，为探究包含不同情绪的弹幕对应的弹幕具体文字内容，笔者选取在上述情感分析中积极情绪占比最高的视频和最低的视频，分别绘制词云图。这需要在 python 中调用 wordcloud 库实现。

```
img = imageio.imread('皮卡丘.webp')

wc = WordCloud(width = 10000, height = 8000,
               background_color = 'white',
               font_path = 'C:\Windows\Fonts\msyh',
               mask = img)
wc.generate(txt)

wc.to_file('testcloud2.jpg')
```

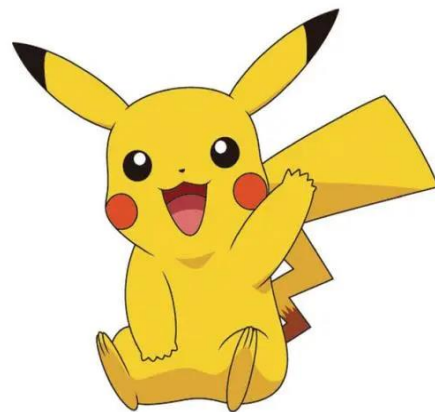
利用 wordcloud 库实现词云图



积极情绪占比最高的视频 弹幕词云



积极情绪占比最低的视频 弹幕词云



©Nintendo·Creatures·GAME FREAK·TV Tokyo·ShoPro·JR Kikaku ©Pokémon

分别选取了蜡笔小新、皮卡丘的白底图片作为词云模板

词云图中，字体越大的词句是出现频数越多的。在上左图中，可见“哈哈哈哈哈”等具有明显积极情绪的词出现了多次，而上右图中，出现多次的是具有反讽意味的“天才”，对应了占比较高的消极情绪。

由此可见，词云图与上述情感分析的结果相匹配，两者均较好地反映出了 b 站视频弹幕的情感倾向。

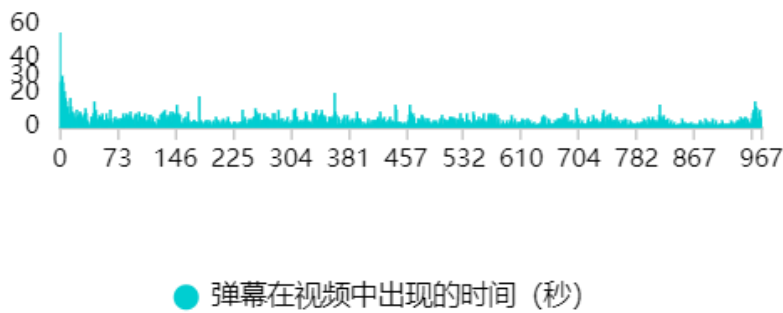
同时，由于 b 站弹幕往往包含了网络热词，这对于研究网络流行语有不可或缺的研究价值，例如平台方可以根据这些弹幕内容信息，制作相应的营销广告等，从而增加吸引力。

3.1.3 弹幕在视频的出现时间分析

由于用户们往往会在视频中某些精彩的时刻发送弹幕，因此视频某个时刻出现的弹幕数量能一定程度上反映此刻视频吸引观众的程度。

笔者选取了榜单前十中弹幕数量最多的视频，以此为例，分析弹幕在视频出现时间的变化，从而探究该视频的高光之处。

弹幕在视频中出现时间的变化图（以弹幕数最多的视频为例）



绘制的相应折线图

我们从上图可以看出，在视频的开头时刻弹幕数量达到了最大值；之后逐渐下降并在一定范围内不断波动，在约 47 秒、180 秒、360 秒、816 秒迎来小范围的峰值；在视频结尾处弹幕数量又逐渐增加。

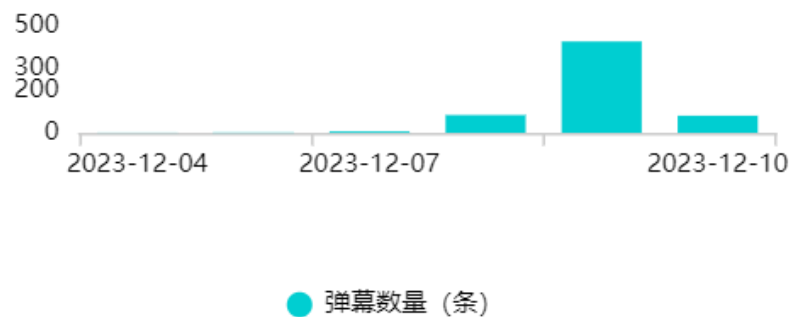
由此推测，用户们最倾向于在视频开头发送弹幕，在视频结尾处也会有较高的意愿去发送弹幕。而在这条视频中，约 47 秒、180 秒、360 秒、816 秒的时刻可能是视频的“高光时刻”，能引发用户们共鸣，吸引观众们发送弹幕。

因此，视频创作者可以根据这样的趋势，结合视频内容研究视频热度的走向，从而在之后更有针对性地创作。

3.1.4 弹幕发送日期分析

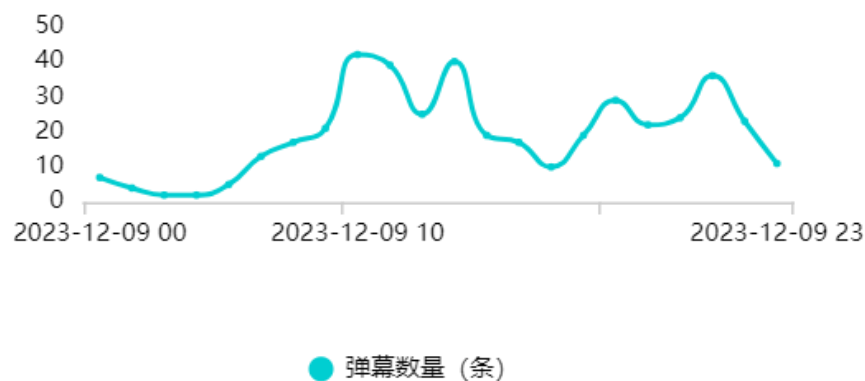
笔者还希望能探究用户们往往在视频发出之后多久发送弹幕，以及每天何时发送弹幕。因此，选取这个榜单中投送时间最早的视频（2023-12-04 发出）为例，进行分析。

弹幕数量随日期（年月）的变化



首先，我们比较从12月4日开始的每一天弹幕数量情况（由于统计时间为12月10日上午，因此此日弹幕数还不全）。可以看出弹幕数在视频发出的几天内逐日增多，在12月9日达到峰值，因此一个视频“变火”往往也需要几天的积累。这对于视频创作者把握视频热度走势有一定的帮助。

弹幕数在一天内随小时的变化（以12/9为例）



接着，笔者选取了12月9日的弹幕数据，观察一日内弹幕数量随时间（小时）的变化。由上图可见，用户们在10-13点、18-21点普遍发送弹幕数量较多，推测这两个时间段都是用户们休息的时间，因此会有较多视频的收看和互动量。

因此，笔者建议创作者可尝试在中午或晚上的时间发布视频；同时，平台方也可以在这些时间段推送特定广告或宣传等，便于获取流量。

3.2 分析比较标题党与非标题党视频的弹幕情况

在b站这种视频平台，除了热门榜单的视频具有较高的研究价值，还有一类不可忽视的现象——创作者经常会使用夸张的标题用词（也称“标题党”）来吸引用户的标题，希望获取流量。那么，这种标题是否真正能有所效果呢？笔者猜测，这客观上确实容易增加播放量，但也容易让观众感到内容与实际不符，带来较差的评价。

3.2.1 对“标题党”进行操作性定义

为了方便寻找数据，我们首先列出了“标题党”的一些定义，在此文中，其总体含义为：为了吸引观众而给视频冠以“名不副实”的标题。

其中分为标题党 I 类型和标题党 II 类型——

类型 I：名不副实的地方在于声明自己某种独特性

代表例子：最 XXXX 的 XX，你一定没有看过的 XX，震惊！XXXXXXXX.....

这些视频往往利用观众猎奇心理，将视频的某种特色元素过度放大，吸引点击。但很多时候，当观众带着高预期点击这些视频的时候，会感到视频内容与预期不符，发出“就这”，“很一般嘛”之类的弹幕。

类型 II：名不副实的地方在于将视频中 1% 的噱头内容拿出来做标题

代表例子：在 XX 游戏里面感受二次元的魅力.....

但视频只有几帧和二次元有关，其余内容只是 XX 游戏普通内容，被吸引的二次元观众纷纷感到失落

3.2.2 获取标题党与非标题党视频弹幕数据

由于 b 站游戏区包含的视频类别较为丰富，也往往能见到“标题党”视频，因此我们以游戏区为目标，在其中找到了十组标题党与非标题党的视频，每组内的两个视频主题接近，但标题用词有明显差异：

【这操作放眼整个王者也是非常炸裂的!!! -
哔哩哔哩】 <https://b23.tv/npX9zmZ>

第一组

【2023 上半《年度集锦》总有一波操作能惊
艳到你-哔哩哔哩】 <https://b23.tv/WXe3pce>

【【老 E】这款游戏里的人竟敢不穿裤子! -哔
哩哔哩】 <https://b23.tv/rM1fjW9>

第二组

【「科技美学直播」黑鲨游戏手机 3 开箱上手
体验 LPDDR5 | UFS3.0 | 升降游戏按键 3499
元起售-哔哩哔哩】 <https://b23.tv/BaLsjG>

选取的视频示例图



用爬虫代码获取的弹幕数据

同样地，由于一共爬取了 20 个视频的弹幕，且这些视频有一定的热度，因此能够获得

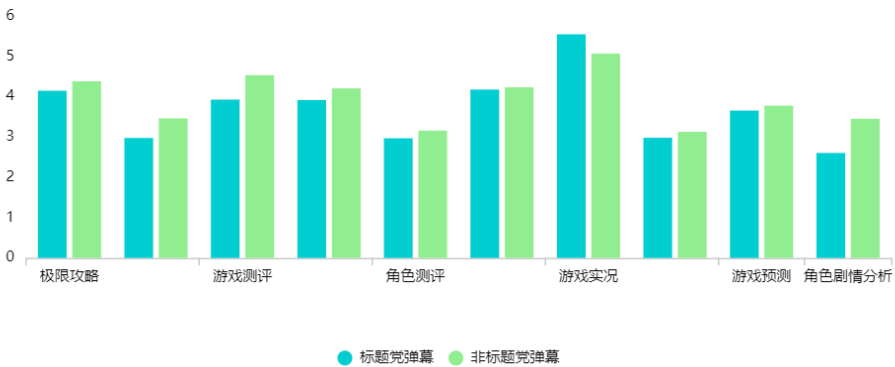
上万条的弹幕数据信息，数据量十分可观。

3.2.3 两种视频的弹幕数量比较

弹幕数量一定程度上能反映视频的热度。因此，笔者统计了两种视频的弹幕数，并通过图表对其进行比较（由于数据之间差异比较大，故对所有数取对数来绘图表示）：

	A	B	C
1	视频类型	标题党弹幕数	非标题党弹幕数
2	精彩操作	955	1340
3	游戏测评	8436	34000
4	游戏实况	352000	117000
5	游戏抽卡	8229	16000
6	游戏预测	4516	5954
7	极限攻略	14000	24000
8	角色测评	926	1433
9	角色剧情分析	401	2832
10	沙雕二创	944	2886
11	游戏速通	15000	17000

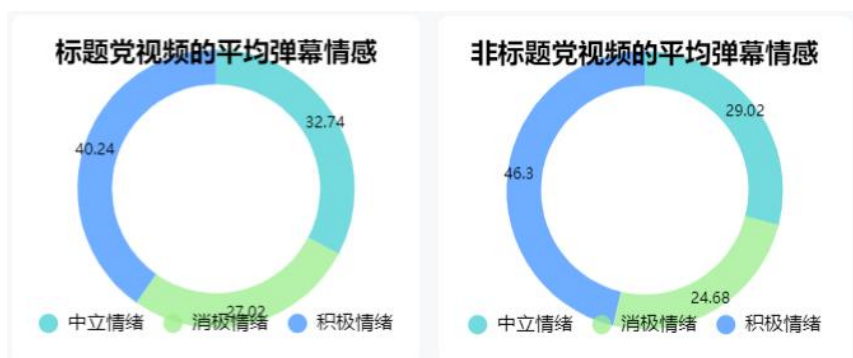
游戏视频“标题党” vs “非标题党” 弹幕数（分别对二者取了对数）



由上图我们可以看出，标题党视频的弹幕数量在大多数情况下都小于非标题党视频。可以得知：标题党视频并没有更为吸引人，相反还可能带来反效果，降低热度与互动性。

3.2.4 对两种视频的弹幕进行情感分析

那么，标题党与是否会影响观众们对于视频的互动情绪呢？笔者采用 3.1.1 一样的方法，对两种视频计算出平均弹幕情感，结果如下图所示：

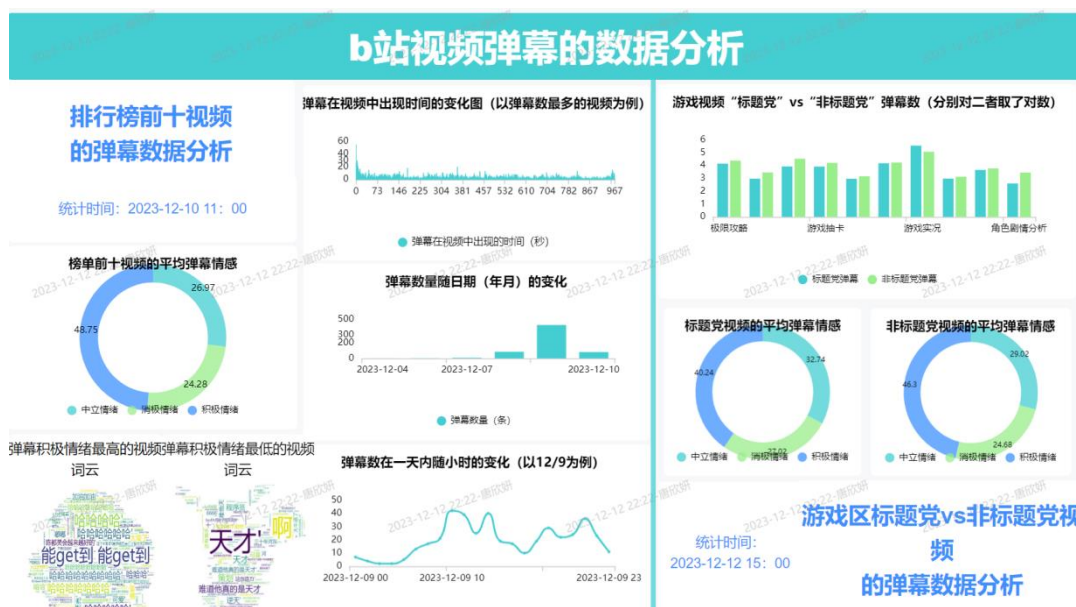


因此，非标题党视频中的弹幕普遍比标题党有更多的积极情绪、更少的消极情绪。

这验证了笔者的猜想：标题党视频容易让观众感到内容与实际不符，带来较差的评价。

故笔者建议视频创作者在关注视频热度的同时，还要重视视频内容与标题的相关性，谨慎斟酌标题用词；而平台也应当严格管控虚假夸大的“标题党”行为，建设平台良好风气环境。

3.3 数据可视化整体看板图



使用 DataEase 网站制作而成

4、总结与展望

4.1 结论

通过对 b 站视频弹幕的数据分析，我们认识到：

(一) 关于热门视频：

往往是能调动观众们积极情绪的、引发情感共鸣；而一个视频变火也不是一蹴而就的，往往需要几天的积累；同时，用户们普遍喜欢在中午和晚上的时间收看视频、发送弹幕。

(二) 关于标题党视频：

虽然它能因夺人眼球的标题获得一定观看量，但因其夸大事实等容易引起观众的反感，从而弹幕数量和弹幕积极情绪都常常不如非标题党视频。

4.2 建议

（一）对于视频创作者而言：

首先，由于热门视频往往是能引起观众普遍的情感共鸣的，因此创作视频时需洞察观众情感需求。

其次，创作者可尝试在中午或晚上的时间发布视频，由于观众倾向于在这两个时间段收看视频，因此较有可能获得更多的播放量。

此外，创作者也应当谨慎用词，避免与视频内容不符的“标题党”行为，引发观众反感。

（二）对于平台管理者而言：

为了让平台有较多且长期的受众，视频质量对于平台而言至关重要。

因此，平台需严格管控视频内容，对视频创作者进行宣传教育（参照“给视频创作者的建议”），携手创作者、观众们共同营造良好的平台环境。

此外，平台也可以参考“弹幕词云图”内的词句，获悉实时的网络热词并加入到日后的广告营销中，用热点吸引观众们的注意，获取流量。

4.3 不足与改进之处

不足之处：

①首先，在对弹幕进行情感分析时，笔者发现 snownlp 库可能存在对语义理解有误的情况，例如某个视频很多弹幕是有关“致敬”的，但由其分析出的积极情绪只占了约 27%，明显是不符合常理的。

②弹幕由于其流行词属性，且字数简短，往往会包含复杂的含义，也可能会造成情感分析的偏差。

③在选取标题党与非标题党视频时，由于不同视频会有多种存在差异的因素，因此无法精准地控制变量。

改进之处：

对于弹幕的情感分析，我们日后会寻找更针对中文语言的情感分析库，从而更准确地界定词句的情感，提升结果可靠性。

4.4 展望

笔者希望能在数据分析这条道路上继续探索更多有趣的现象、分析背后的原理——或许可以进一步研究 b 站弹幕的更多信息，如字体大小、颜色等，或是对其他网站的数据进行挖掘……在如今的数字时代，我们将积极学习新兴知识，并把所学应用到实际中去，为数字社会贡献自己的力量！