



B站视频 热点追踪

目录

Content



1. 问题背景与介绍

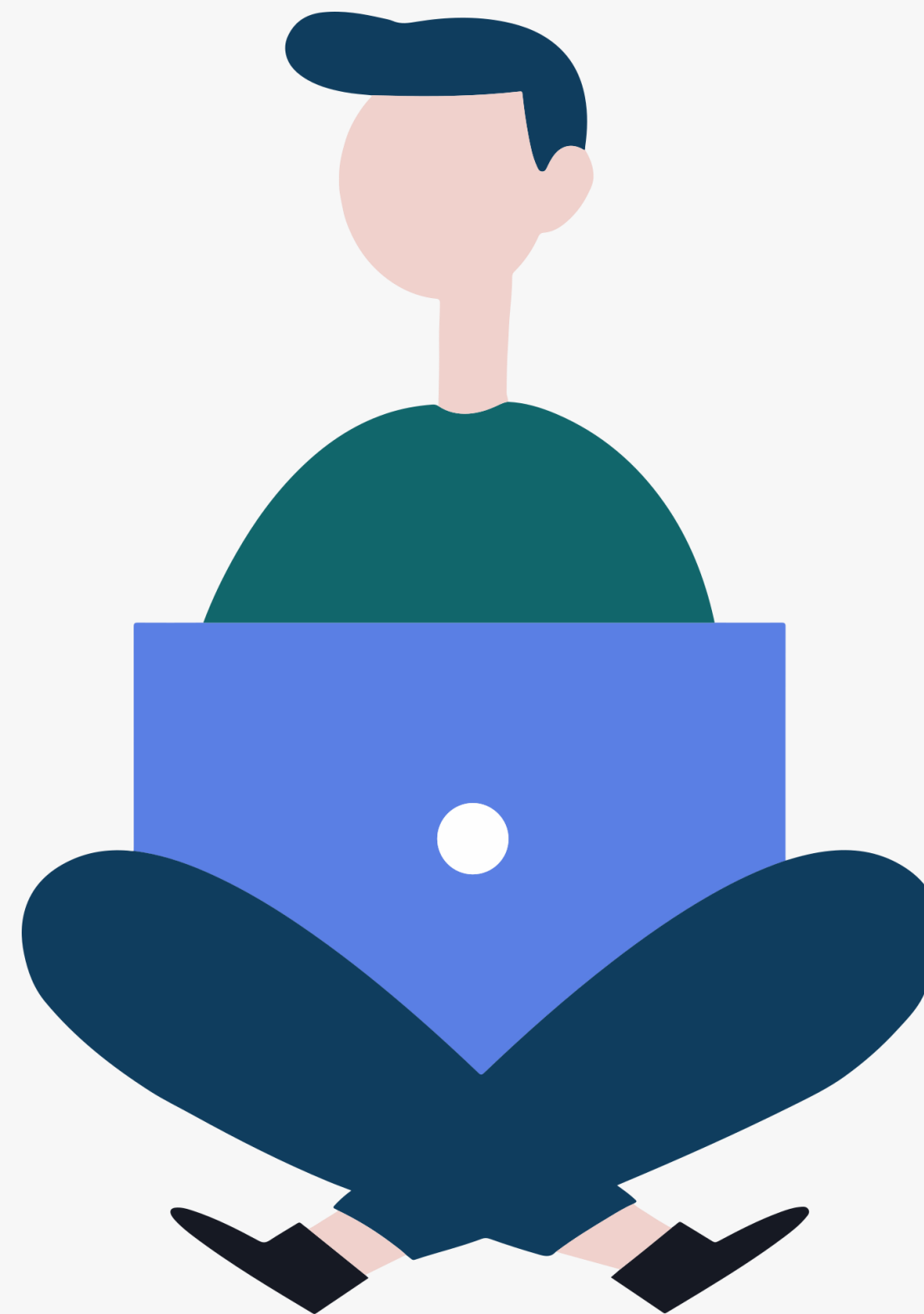
2. 数据采集

3. 可视化展示

4. 总结

01

问题背景与介绍



1.问题背景与介绍



1.B站创作激励

平台陷入成长放缓期
对高质量创作者的需求日益增长



2.新人UP主难以把握发展方向

普通人需要花费大量成本
学习提升视频质量、
摸索适合自己的发展方向



3.授人以鱼不如授人以渔

总结出“流量密码”
利于新人UP主快速融入平台
加速成长周期



02

数据采集

爬到网站和爬取视频标题和标签

爬取标题与标签

01 观察元素规律

- ◆使用检查来观察目标的XPath规律
- ◆记下文字内容记入列表

02 抓取视频标签

- ◆点击标题进入播放页面
- ◆检查视频下方标签的class
- ◆记下文字内容记入列表



爬取标题与标签

使用 `windows=browser.window_handles`

`browser.switch_to.window(windows[-1])`
`browser.close()`
完成每个视频页面的切换

利用for循环和 `try: ...except:`
.....

不断重复到每个视频里爬取的动作

```
for i in range(1,43):
    try:
        label_week.append([])

        windows=browser.window_handles
        browser.switch_to.window(windows[-1])
        elem_text=browser.find_element_by_xpath('//*[@id="app"]/div/div[2]/div[2]/div/div[%d]/div[2]/p'%i)
        print(elem_text.text)
        title_week.append(elem_text.text)
        elem_text.click()

        windows=browser.window_handles
        browser.switch_to.window(windows[-1])
        wait=WebDriverWait(browser,100)
        element=wait.until(EC.presence_of_all_elements_located)
        elems=browser.find_elements(by=By.CLASS_NAME, value="channel-name")
        for elem in elems:
            print(elem.text)
            label_week[i-1].append(elem.text)
        browser.close()
    except Exception as e:
        pass
    continue

windows=browser.window_handles
browser.switch_to.window(windows[-1])
browser.execute_script('var q=document.documentElement.scrollTop=0')
button=browser.find_element_by_xpath('//*[@id="app"]/div/div[1]/div[4]/div/span')
button.click()
windows=browser.window_handles
browser.switch_to.window(windows[-1])
try:
    wait=WebDriverWait(browser,100)
    element=wait.until(EC.presence_of_all_elements_located)
finally:
    print('加载完成')

time.sleep(2)
```

爬取标题与标签

```
csvFile=open('sample.csv','wt+',encoding='UTF-8-sig',newline='')
writer=csv.writer(csvFile)
writer.writerow(['标题','标签'])
```



标题	标签
约尔太太今天约会♥	穿搭'服饰'
丰收了 我真的很想画这幅画	绘画'
首次公开！91岁的袁隆平去医院路上歌唱祖国	
揉肉肉~~~~	萌宠'必剪创作'
黄焖鸡米饭真正的配方，15个字，一万块钱	家常菜'美食'
【医学博士】如何拯救草莓鼻？I 毛孔粗大还有救吗？	医学'护肤' '健康' '美妆' '搞笑'
废话连篇	鬼畜'
【花泽香菜自投稿】恋爱循环，再来一遍！	花泽香菜'恋爱循环' '声优'
永琪与紫菜蛋花兔的缺氧日记	
怎么办啊！我被求婚了！	vlog'
鉴定一下热门营销号谣言	科普'
【D×丁程鑫】少年主舞solo炸裂演绎神级舞台《D》	丁程鑫'时代少年团' '明星舞蹈'
卧槽，这也太酷了吧！	闪电侠' DC' '漫威' '超级英雄' '影视混剪' '影视剪辑'
【金坷垃】我是非洲哒！我要金坷垃~	金坷垃'万恶之源' '鬼畜' '搞笑'
我给僵尸做了一个自行车	
【4K超清】乘风破浪的姐姐3 主题曲《乘风》MV 上线！	王心凌'郑秀妍' '4K' '开口跪' '综艺'
【互动视频】沈阳之夏	互动视频'
得到了甲方的赞助	布偶猫'喵星人' '萌宠'
在召唤师峡谷，守护他们的热爱	LPL' '英雄联盟'
被眼前的这一幕震撼到了	摄影'
你们差点就打赢了RNG！	加勒比海盗' 搞笑配音' '英雄联盟' '搞笑'
“可惜你不看海贼，也不明白这个视频的分量.....”	索隆'路飞' '海贼王' 'AMV' 'MAD'
偏科天花板	说唱'鬼畜调教' '鬼畜' '考试' '搞笑'
主持人：我已经报警了③！！！！	朱广权'人类迷惑行为' '是在下输了' '搞笑'
让朋友发现新闻上正在通缉自己，他会是什么反应？【翼刀整蛊奇闻录】	搞笑'
我是看球的	国足'世界杯' '足球' '体育'

03

数据采集

弹幕与评论



视频弹幕爬虫

01 爬取排行榜中各视频 URL

◆利用Selenium访问B站各个排行榜

◆利用XPath找出每个视频的URL

02 透过 URL 找出每个视频 CID

◆利用正则表达式找到视频CID

03 利用 CID 抓取视频弹幕

◆将抓取到的CID丢入comment.bilibili.com

◆透过BeautifulSoup爬取所有弹幕，并存入CVS

```
<d p="176.46700,1,25,16777215,1653018797,0,458c519f,1056357251205324544,11">好可爱!!</d>
<d p="47.76000,1,25,16777215,1653007655,0,5feba680,1056263782348620032,11">感觉像开心果馅的五仁月饼??</d>
<d p="292.98700,1,25,16777215,1652992774,0,f6cd62a0,1056138952102122496,11">我看的时候已经点到了</d>
<d p="296.15100,1,25,16777215,1652922613,0,e62db043,1055550404764395008,11">助力每一个梦想</d>
<d p="320.90900,1,25,16777215,1652848686,0,74884bdd,1054930258639390976,11">投币助力每一个梦想</d>
<d p="310.13100,1,25,16777215,1652800340,0,a7af9727,1054524703886021120,11">助力每一个梦想</d>
<d p="401.02200,1,25,16777215,1652799865,0,8172a2c8,1054520716864721664,11">现在的小孩子真有礼貌 (doge)</d>
<d p="294.22500,1,25,16707842,1652700008,0,972d4b,1053683055887201536,11">已经点到了哟 (ღ>ღ<*)</d>
<d p="178.14900,1,25,16777215,1652588737,0,3db1aebe,1052749643546501888,11">哇同一个纪录片</d>
```

```
1 def get_cid(url):
2     driver.get(url)
3     WebDriverWait(driver, 100).until(EC.presence_of_all_elements_located)
4     html = driver.page_source
5     cid = re.findall('https://upos-hz-mirrorakam.akamaized.net/upgcxcode/\d{2}/\d{2}/\d{9}', html)
6     cid_num = cid.pop() #将所有找到的放在列表并取出最后一个
7     cid_num = cid_num[-9:] #保留最后9位cid
8     return cid_num
9
10 def get_danmu(cid):
11     danmu = []
12     url = "https://comment.bilibili.com/" + str(cid) + ".xml" #输入视频cid能取得所有弹幕的url
13     request = requests.get(url)
14     request.encoding='utf8'
15     bs = BeautifulSoup(request.text, 'html.parser')
16     results = bs.find_all('d')
17     for result in results:
18         danmu.append(result.text)
19     return danmu
20
21 driver = webdriver.Chrome('chromedriver', options=chrome_options)
22 driver.get("https://www.bilibili.com/v/popular/weekly?num=164")
23 WebDriverWait(driver, 100).until(EC.presence_of_element_located((By.CLASS_NAME, 'no-more'))))
24
25 urlList = []
26 danmu_weekly = []
27 for k in range(1, 45):
28     names = driver.find_elements(by=By.XPATH, value='//*[@id="app"]/div/div[2]/div[2]/div/div[' +
29     for name in names:
30         url = name.get_attribute("href")
31         urlList.append(url)
32
33 for i in urlList:
34     cid = get_cid(url)
35     danmu = get_danmu(cid)
36     danmu_weekly.append(danmu)
```

视频评论爬虫

01 点击进入排行榜中各视频

- ◆利用Selenium访问B站各个排行榜
- ◆利用XPath找到每个视频标题并进入

02 抓取视频评论

- ◆利用XPath找到视频所有评论
- ◆抓取视频前4个评论，并存入CSV



```
1 browser=webdriver.Chrome()
2 browser.get('https://www.bilibili.com/v/popular/weekly?num=164')
3 wait=WebDriverWait(browser,100)
4 element=wait.until(EC.presence_of_all_elements_located)
5
6 windows=browser.window_handles
7 browser.switch_to.window(windows[-1])
8 browser.execute_script('var q=document.documentElement.scrollTop=0')
9
10 comments = []
11 title_week = [] # 每周必看视频的标题
12 for i in range(1,43):
13     try:
14         comments.append([])
15         windows=browser.window_handles
16         browser.switch_to.window(windows[-1])
17         elem_text=browser.find_element_by_xpath(f'//*[@id="app"]/div/div[2]/div[2]/div/div[{i}]/div[2]/p')
18         elem_text.click()
19         title_week.append(elem_text.text)
20         time.sleep(2)
21         windows=browser.window_handles
22         browser.switch_to.window(windows[-1])
23         browser.execute_script('window.scrollTo(0,document.body.scrollHeight)')
24         time.sleep(10)
25         browser.execute_script('window.scrollTo(0,document.body.scrollHeight)')
26         time.sleep(5)
27
28         for j in range(1,5):
29             elem_text=browser.find_element_by_xpath(f'//*[@id="comment"]/div/div[2]/div/div[4]/div[{j}]/div[2]/p')
30             print(elem_text.text)
31             comments[i-1].append(elem_text.text)
32         browser.close()
33     except Exception as e:
34         pass
35     continue
```



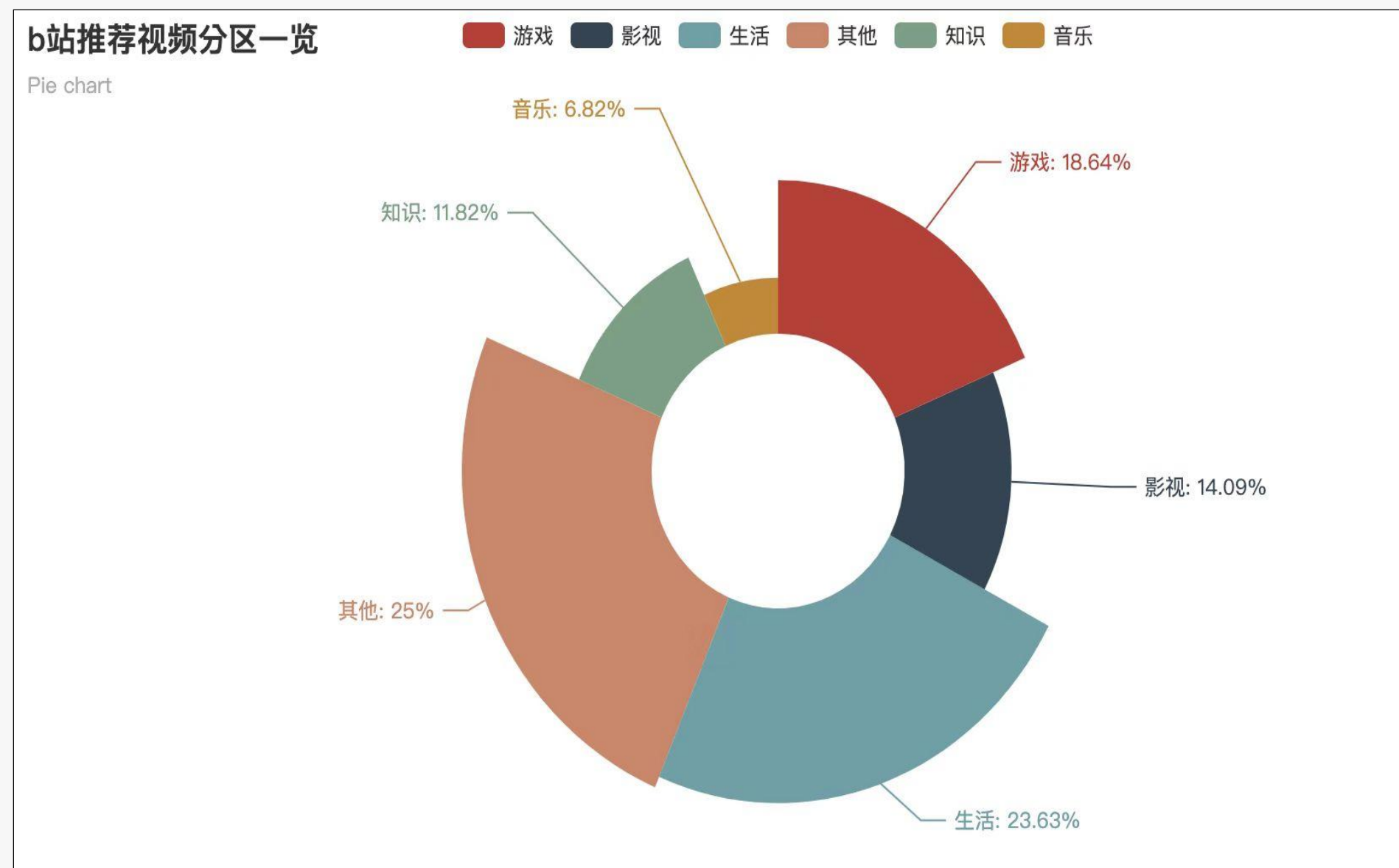
04

数据分析

可视化展示

1. 热门视频标签一览

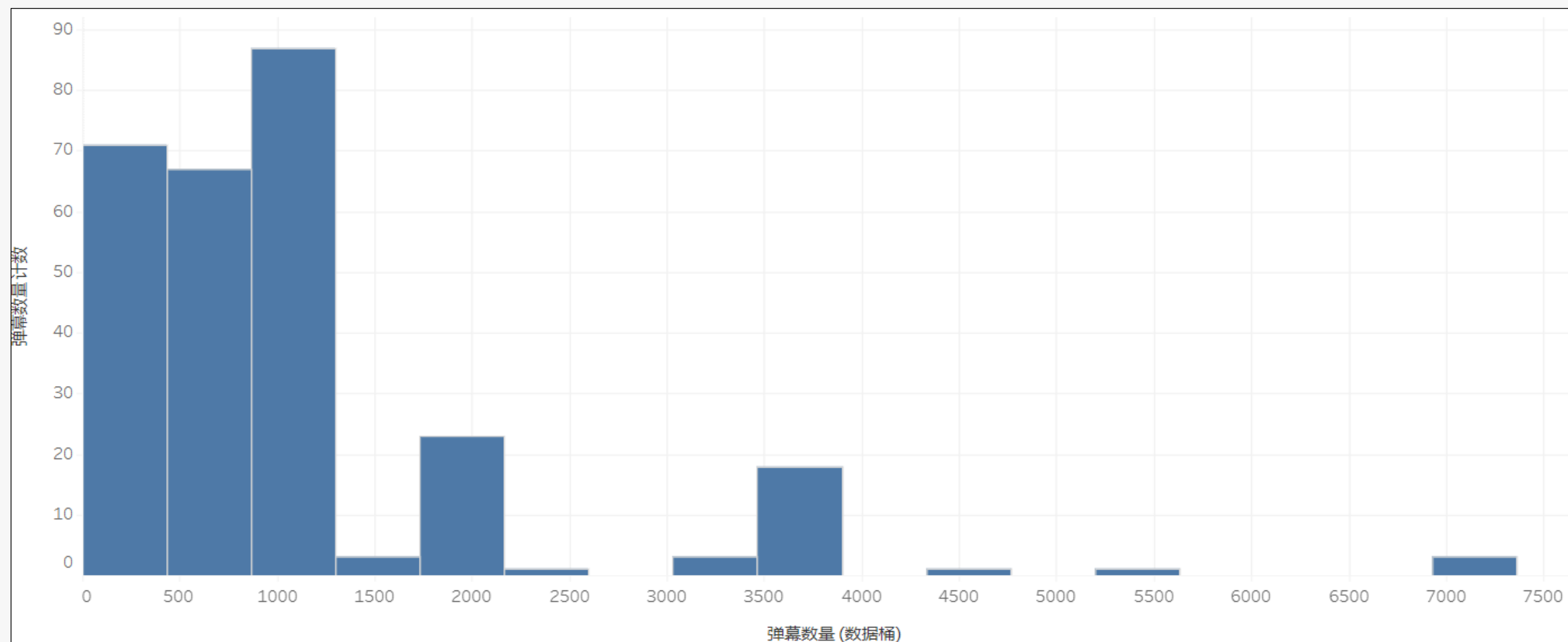
按照分区上看生活区和游戏区最多，此外影视区和知识区也不少。
类型上比较轻松的如带“搞笑”标签会比较受欢迎。



2.视频弹幕性质一览——弹幕数量分布

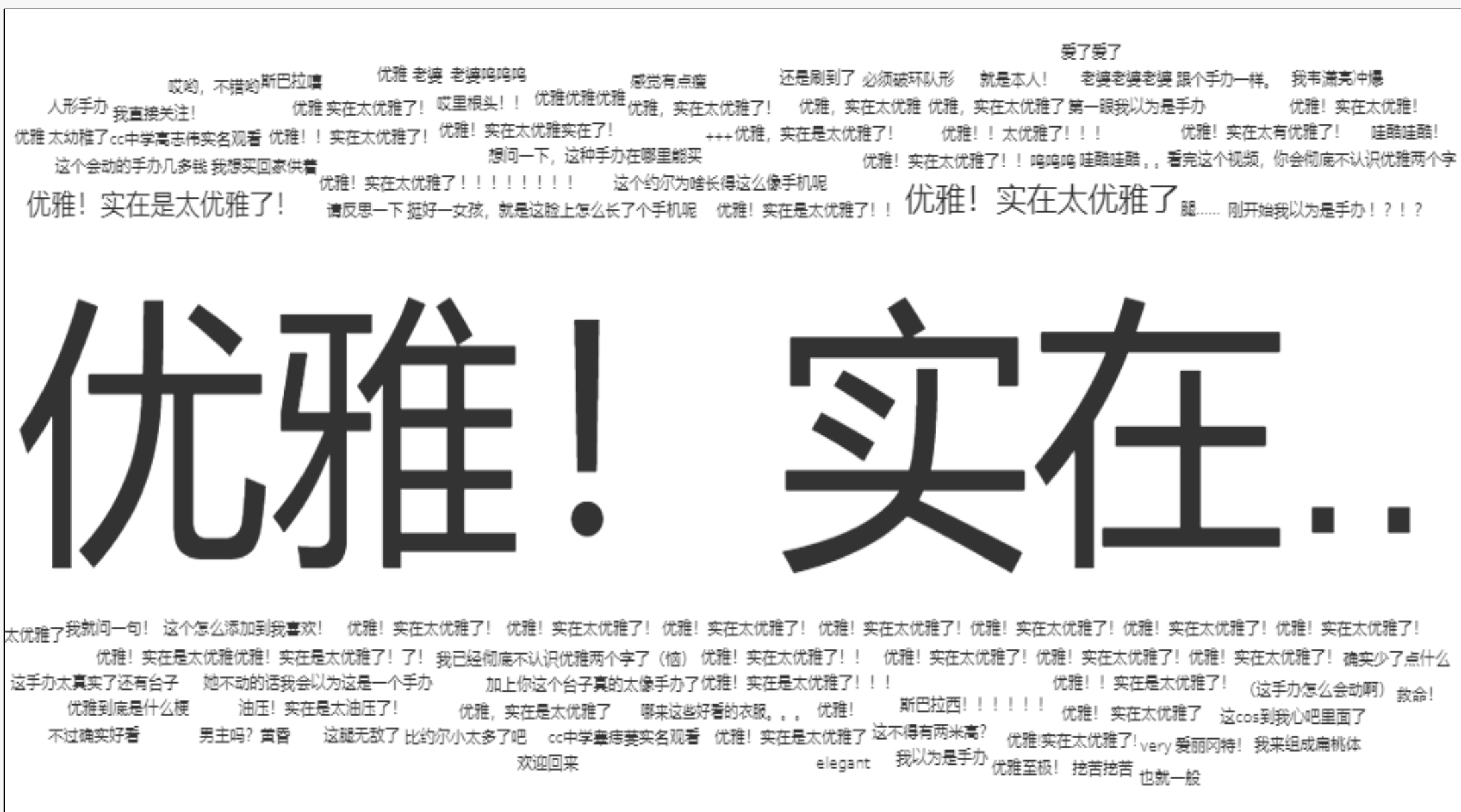
我们统计了共277条视频的弹幕数据。在直方图外，我们统计得出不少视频具有相同弹幕数（120，360，1200等），可能是由于弹幕数量过多而选择性显示。

热门弹幕视频弹幕数量主要在1500条以下大致均匀分布，最低弹幕数甚至只有71条。
这说明视频上热门并不需要大量弹幕，热门条件可能是其他因素。



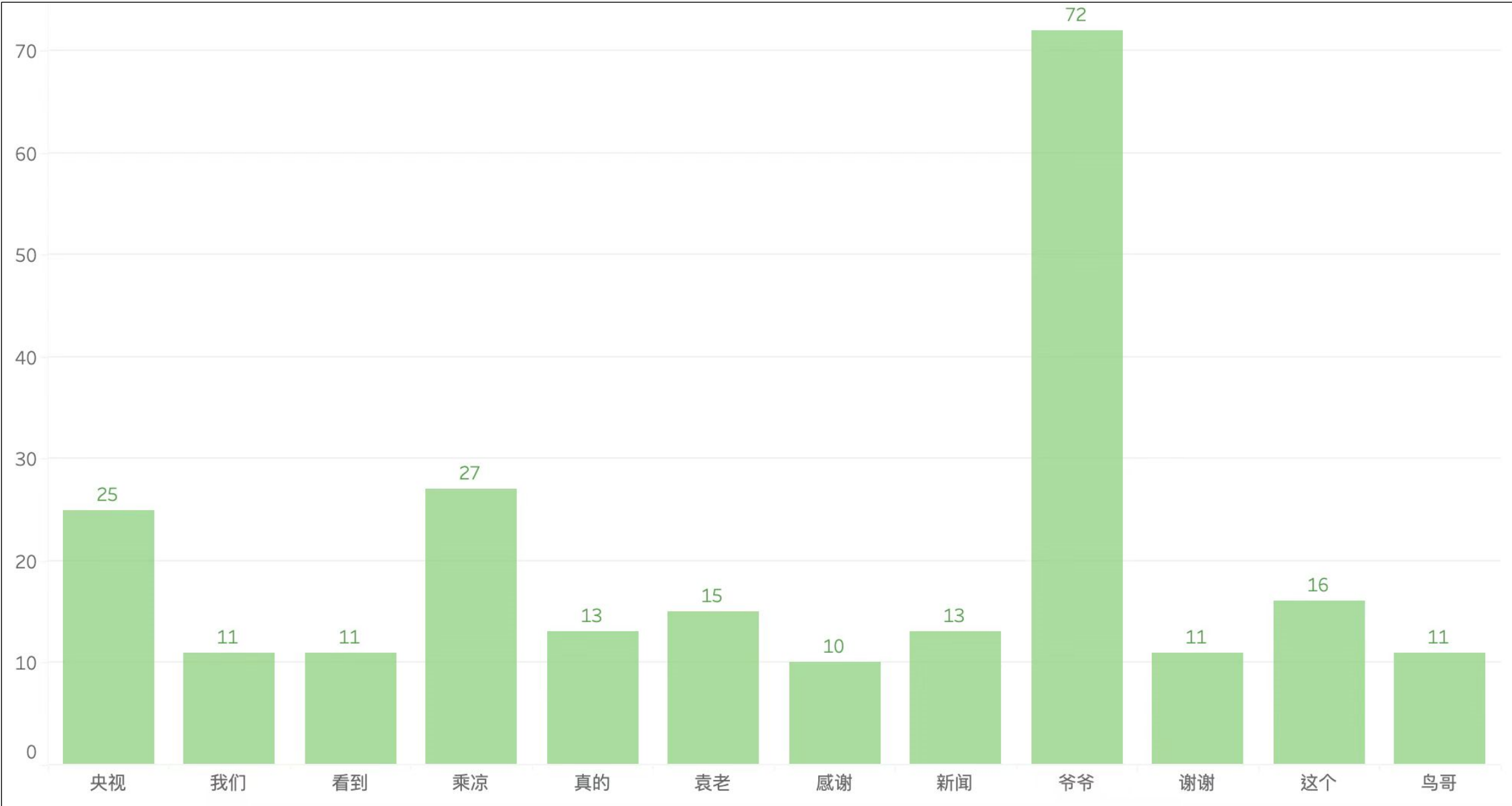
2.视频弹幕性质一览——从弹幕热词看视频内容

从视频弹幕的热门词汇中可以大致推断出视频内容。我们将以《约尔太太今天约会♥》
《【4K60FPS】王心凌《爱你》经典现场！她太可爱了》和《丰收了 我真的很想画这幅画》
为例分析视频内容与弹幕之间关系。



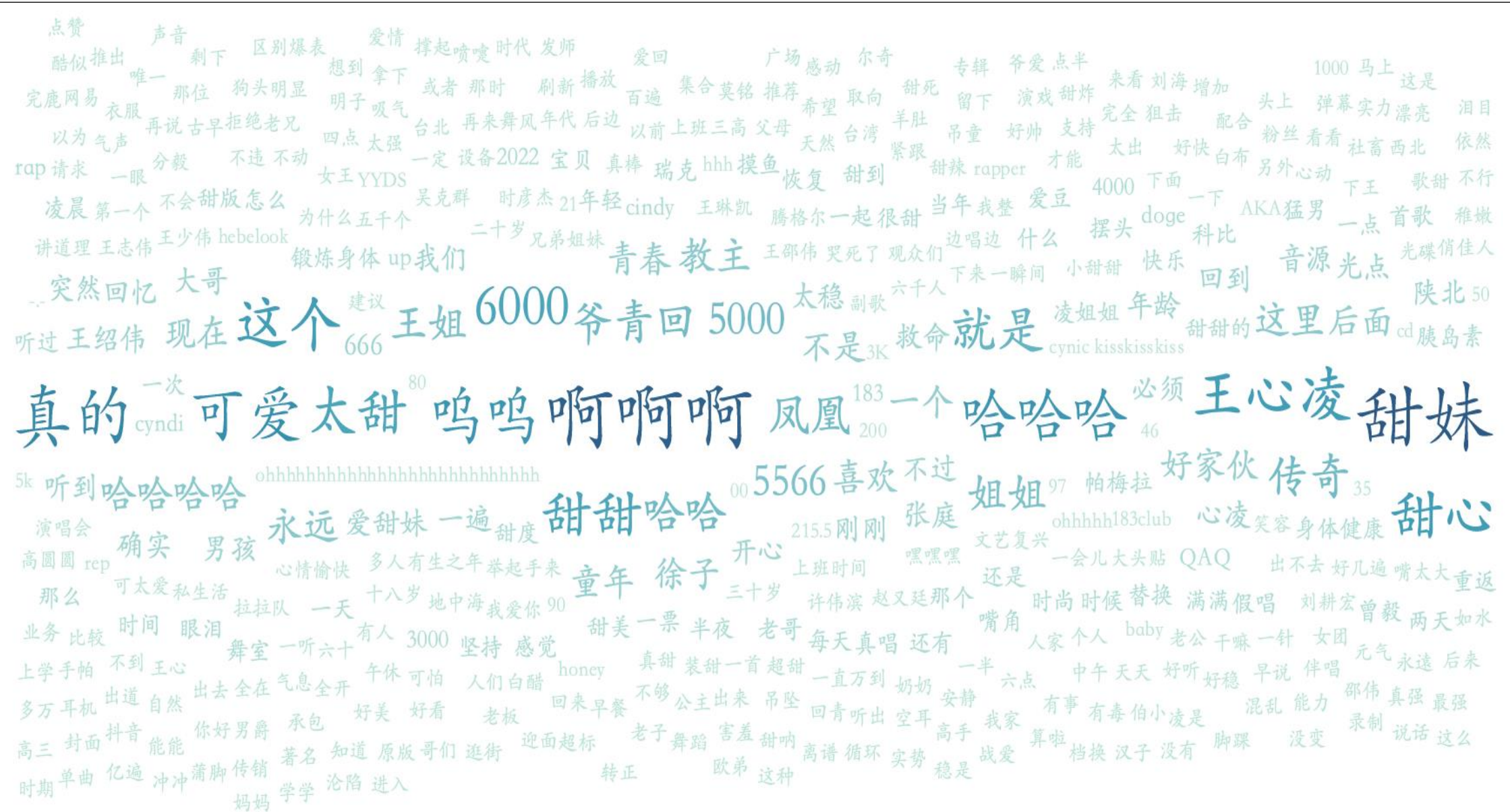
2.视频弹幕性质一览——从弹幕热词看视频内容

从视频弹幕的热门词汇中可以大致推断出视频内容。我们将以《约尔太太今天约会♥》
《【4K60FPS】王心凌《爱你》经典现场！她太可爱了》和《丰收了 我真的很想画这幅画》
为例分析视频内容与弹幕之间关系。



2.视频弹幕性质一览——从弹幕热词看视频内容

从视频弹幕的热门词汇中可以大致推断出视频内容。我们将以《约尔太太今天约会♥》
《【4K60FPS】王心凌《爱你》经典现场！她太可爱了》和《丰收了 我真的很想画这幅画》
为例分析视频内容与弹幕之间关系。



总结

01 投制视频时需要选择热门相关分区

如果up主希望自己的视频获得更多热度，则应当将视频尽量投至热门分区如生活区、游戏区等。这样视频获得推荐的可能性会大大提高。

02 视频内容需要“紧跟时事”

通过对弹幕数量和弹幕内容进行分析可得，就算由于时长原因弹幕数量少，只要视频内容为当前热点，上热门的概率也会相应提升。





Thank You