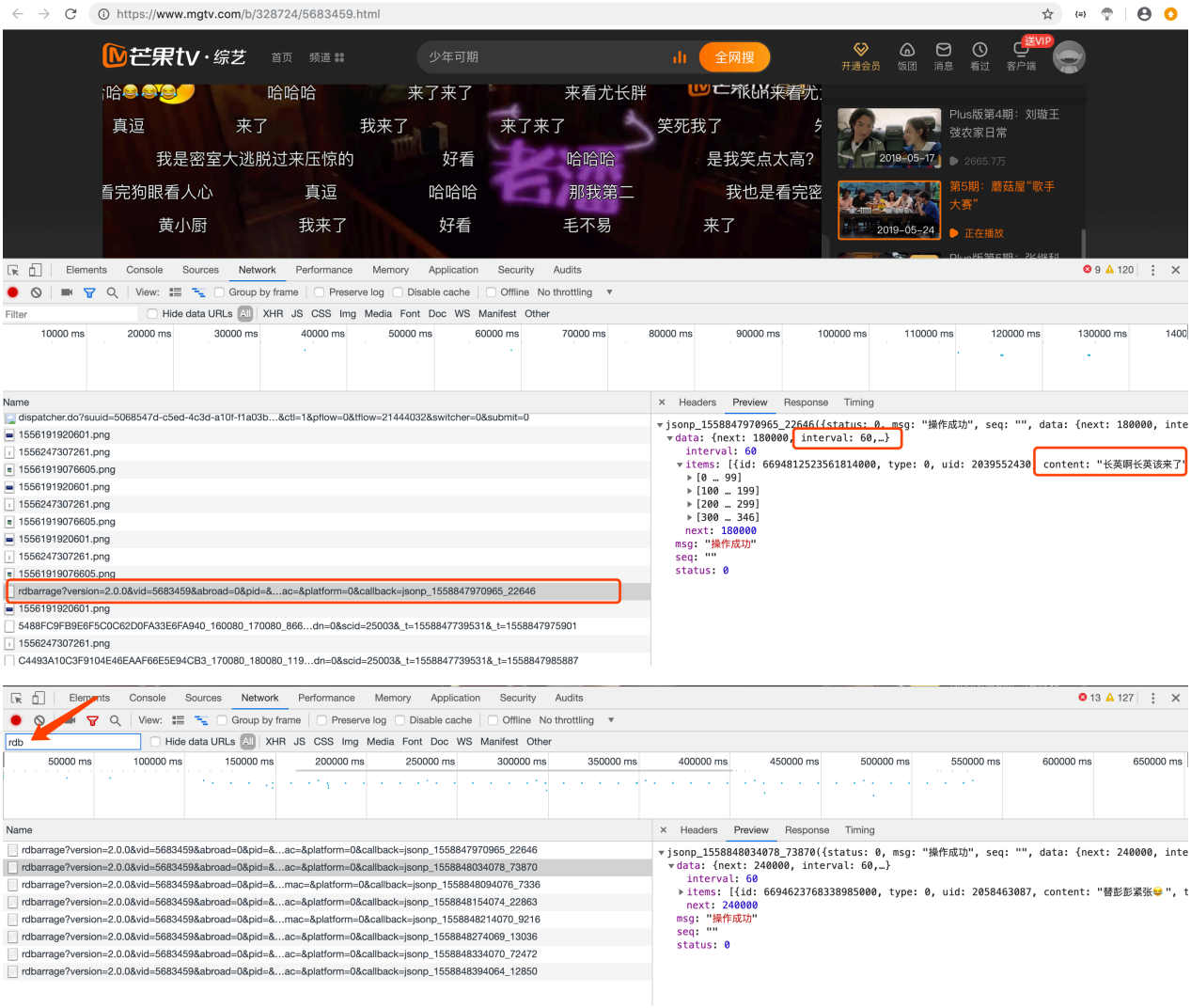
《向往的生活》弹幕抓取与分析

《向往的生活》是湖南卫视一档十分温馨的生活类真人秀综艺节目，目前第三季正在更新中，常驻嘉宾加入了张子枫，深受广大观众的喜欢。而该节目的豆瓣评分也达到了7.9。这档综艺以明星艺人到村寨里体验生活为主线，融入了美食，劳动，幽默的元素，让人边看边有身临其境的感觉，仿佛自身也真正进入了“向往的生活”。

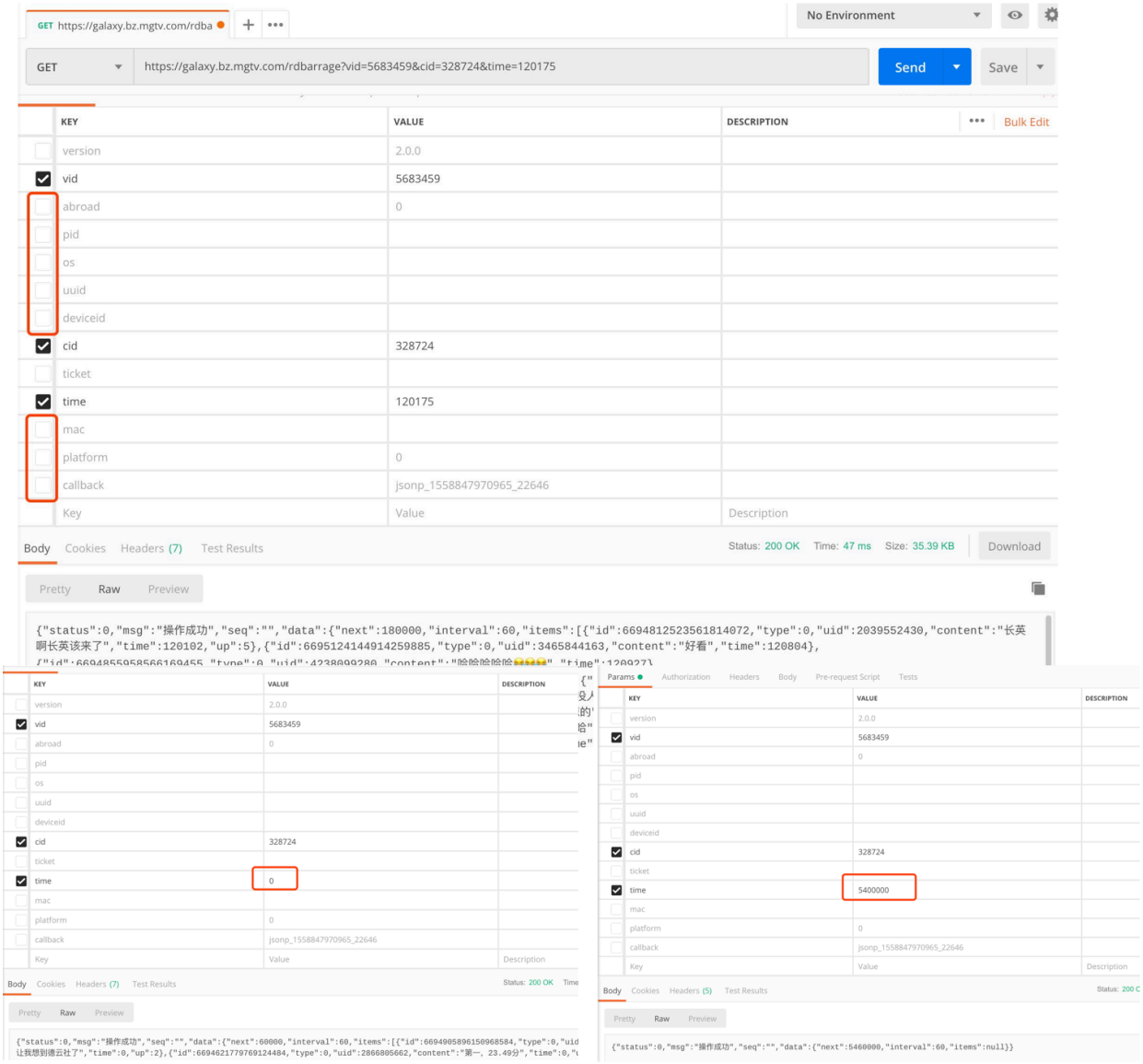
近些天在看节目的时候，看到弹幕上大家讨论的特别热闹，突发奇想能不能把所有的弹幕爬下来做一下分析呢。一方面探究一下弹幕数据抓取有没有特别之处，另一方面通过弹幕对这个节目的口碑一探究竟。接下来以上周五刚更新的第5期为例，进行弹幕数据抓取。代码主要使用requests库，抓取结果存储在csv文件中。

网页分析

在芒果TV网页版打开第5期节目，等待广告加载完毕，同时打开chrome开发者工具的network选项卡。由于请求很多，而且随着时间推移，会越来越多。所以我采取了先清空再等待的方式。发现前面大多加载的都是图片，自然这不是我们的目标。过了一会儿之后，发现一条可疑的请求，见下图所示，点击一看，真的出现了弹幕内容。interval是60，猜测可能是表示一个间隔，每60s会有一个新的请求。于是使用filter过滤了以“rdb”开头的请求，发现这些都是弹幕，而且next都是60000的倍数，猜测表示的是60000毫秒，也就是60秒。



接下来我们需要确认弹幕的翻页逻辑，也就是这些弹幕链接的统一规律。这里推荐一个很好用的网页请求分析工具postman。它不仅可以用来分析网页的请求参数，还能够提供不同语言的请求代码，稍加修改就可以使用。把刚刚我们找到的链接贴到postman中。如图所示，可以看到请求的参数，点击send按钮之后能看到请求的结果。由于参数很多，可以考虑去掉一些无用的参数。最终发现，只需要保留vid，cid，time三个参数即可。猜测vid表示节目id，cid表示视频id，time应该是请求时刻，是一个相对值。并且请求结果中，而每一条弹幕的时间，都要比time数值大。结合上文的分析逻辑，可以得出每一个请求结果都是请求时间60s内的弹幕。如果我们要获取所有的弹幕，就可以通过改变time的值来实现。最小的time取值应该是0，最大的应该就是和视频时长最接近的60000倍数的毫秒数。这里的节目时长为89:49。经过验证，果然如此，接下来我们就可以用代码来实现了。



代码实现

使用requests构造网络请求，并用一个循环控制翻页，爬取全部的弹幕。解析返回的json数据并使用pandas存储到Excel中。详细代码如下所示，一共45行。

```
import requests
import pandas as pd
import time
import datetime
from fake_useragent import UserAgent

ua = UserAgent()
url = "https://galaxy.bz.mgtv.com/rdbarrage"

rdb_content = {'id': [], 'type': [], 'uid': [], 'content': [], 'add_time':
[], 'ups': []}
count = 0

print("爬取开始时间: {}".format(datetime.datetime.now().strftime('%Y-%m-%d
%H:%M:%S')))
```

```
for i in range(0, 91):

    querystring = {"version": "2.0.0", "vid": "5683459", "cid": "328724",
"time": i*60000}

    headers = {
        'User-Agent': ua.random
        }

    try:
        response = requests.request("GET", url, headers=headers,
params=querystring).json()
        items = response['data']['items']
        if items is None:
            print("爬取完毕！弹幕数量{}".format(count))
            break
        else:
            for item in items:
                rdb_content['id'].append(item.get('id')) #弹幕id
                rdb_content['type'].append(item.get('type')) #弹幕类型
                rdb_content['uid'].append(item.get('uid')) #用户id
                rdb_content['content'].append(item.get('content')) #弹幕内容
                rdb_content['add_time'].append(item.get('time')) #弹幕时间
                rdb_content['ups'].append(item.get('up', 0)) #d弹幕点赞数
                count = count + 1

            print("爬取第{}分钟的弹幕..., 当前弹幕数量{}".format(i + 1, count))
        time.sleep(5)
    except:
        print("第{}分钟弹幕爬取失败!当前弹幕数量{}".format(i + 1, count))
        continue
    rdb_df = pd.DataFrame(rdb_content)
    rdb_df.to_csv('rdb.csv', index=None)
```

运行效果截图：

可以看出，在本次爬取时，弹幕数量已经将近3w条，而此时节目更新还不到2天，在一定程度可以反映出该节目的火爆程度。接下来我们对弹幕数据做一些深入的分析，从数据的角度看这期节目。

数据可视化

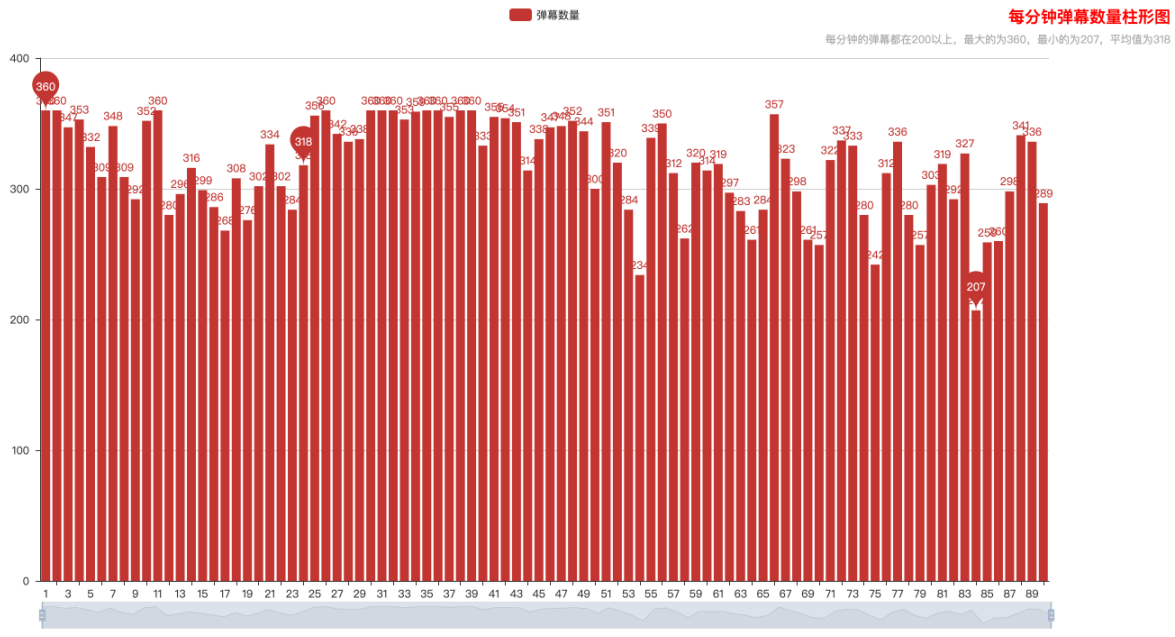
以上爬取的数据，有一些字段存在缺失，但是占比极小，因此采取删除的方式处理，最终剩余28602条有效数据。

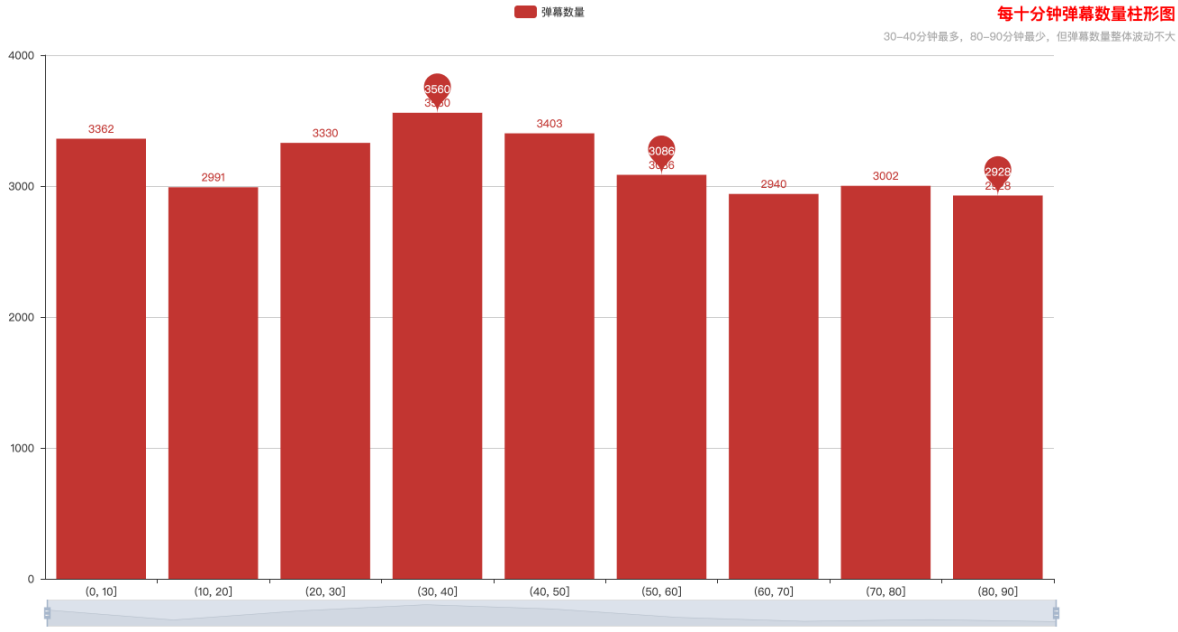
```
data2 = data.dropna()
data2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 28602 entries, 0 to 28605
Data columns (total 6 columns):
id           28602 non-null object
type         28602 non-null int64
uid          28602 non-null int64
content      28602 non-null object
add_time     28602 non-null float64
ups          28602 non-null float64
dtypes: float64(2), int64(2), object(2)
memory usage: 1.5+ MB
```

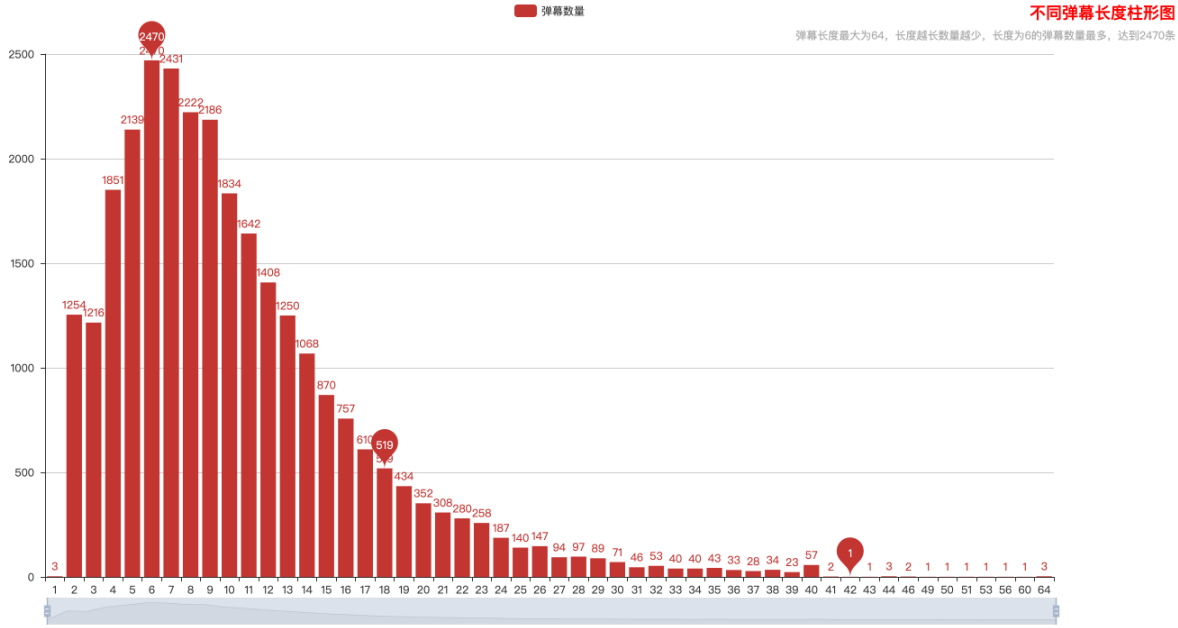
01 不同时间段弹幕数量的分布

节目时长大约90分钟，我们分别以1分钟和10分钟为单位，看一下弹幕数量。每分钟弹幕都在200以上，最小的为207，最大的为360，平均值为318。30-40分钟弹幕最多，一般此时是节目的高潮时段，80-90分钟弹幕数量最少，此时节目已经接近尾声。可以看出，虽然随着时间推移，弹幕数量有所波动，但整体来讲，在各个时间，弹幕波动不剧烈，也反映出节目能够持续保持较高的热度，可谓“分分钟都是精彩”。





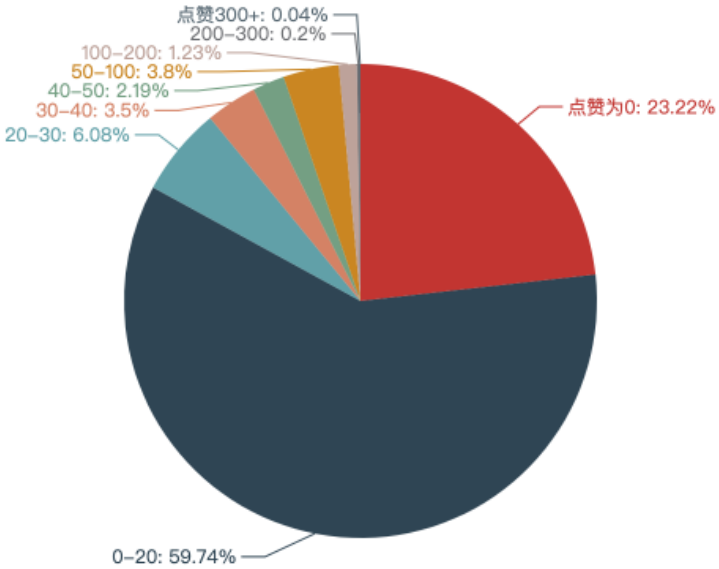
02 不同长度的弹幕数量分布



可以看出，弹幕长度最大为64，长度越长，弹幕数量越少，长度为6的弹幕数量最多，达到2470条。大多数弹幕的长度都集中于10个字上下，趋向于口语化。这也符合我们的认知，10字左右已经足以表达用户看剧的心情和观点。当然也有不嫌麻烦的用户，弹幕数量达到了30字以上，也有极少量的弹幕长度达到了50以上。出于好奇，我们可以看一下长度超过50的弹幕都说了啥，见下图所示，多少能够感受到观众十分走心地在享受节目。

```
9454          都没认真看吗他两一开始没看见后来反应了一下才晃了下身子，结果那时候彭彭手也伸回去了，我觉得这样反而避免了尴尬，谁没事在节目直播上
18108         我们的幸福生活在我们的身边有很多人都在一起的时候就会觉得自己很好的人都是这样的人都会有自己的生活方式的生活生活中的的一一个回复问
19031                  我黄儿一直看着爆米花 估计很想试试吧 但没人cue 不好意思上前 要大华在 一定会带着他玩吧 我想梅溪湖了
23272      常英和她的胖儿子啥时候来啊？盼望着  盼望着\n常英和她的胖儿子啥时候来啊？盼望着  盼望着\n常英和她的胖儿子啥时候来啊？盼望
24682                  我会劈柴，会做饭，还会掰玉米，最主要是我爱何老师，彭彭，黄老师。什么时候节目组才会邀请我去向往的生活。。。。。
26181              黃老師不說了嘛，因為彭彭去種田了，沒人生火，想念以往有大華幫他生火。以前兩兄弟主外主內兩不誤。又不是否認彭彭，亂說啥呢。
27856                  感谢何老师黄老师彭彭妹妹对黄子人工的照顾，感谢向往的生活邀请黄子人工来，感谢幕后所有的工作人员，辛苦啦
Name: content, dtype: object
```

03 弹幕点赞数分布



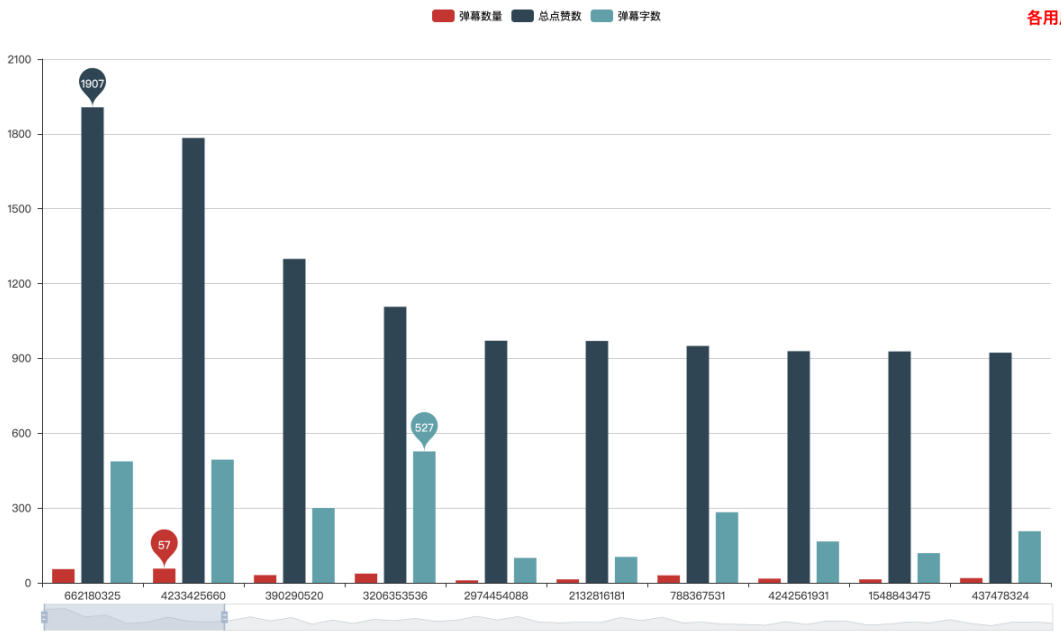
点赞数量区间

可以看出有接近四分之一分弹幕没有获得点赞。近6成的弹幕点赞量在20以下，点赞量20以上的弹幕不到20%。我们同样可以看一下点赞大于300的弹幕都说了啥，但从弹幕就能感受到节目整体的欢乐气氛。

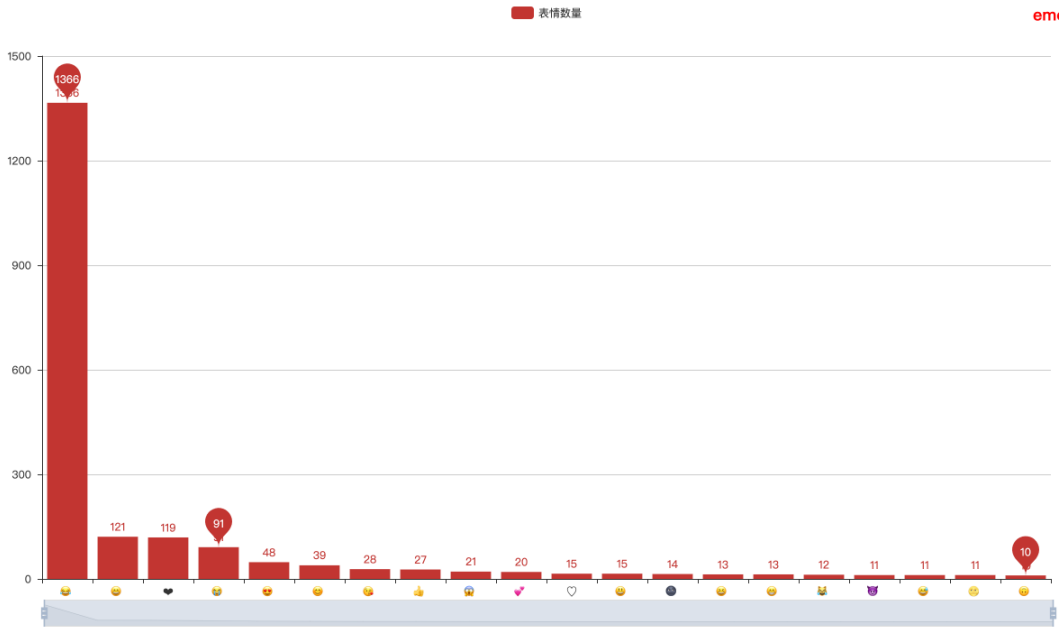
```
9069          快放下哈哈哈哈哈哈哈哈
10033         哈哈哈  好尴尬的笑容
11009          黄老师今日标准假笑
13058        池子好皮哈哈哈哈哈哈哈哈😄
13447         哈哈哈哈哈哈对比明显
17659      池子：忽然好像意识到了什么.....
24887     把我挤得这么边上哈哈哈哈哈哈哈
25461        兄妹两一起惶恐  哈哈哈哈
25473       使不得使不得  妹妹惊慌了😄
25544     哈哈哈，来自长辈乎如其来的关爱
25550          狼叔怕妹妹吃不到好的
25713          老父亲，爱的注视。
25802    黄老师喜欢跟大华开玩笑，哈哈哈哈
Name: content, dtype: object
```

04 用户发布的弹幕数量，点赞数，弹幕总字数对比

我们的数据中共有17268名用户发布了28602条弹幕，人均发布弹幕1.66条。按照点赞数降序排列取前10，观察弹幕数量，点赞数，弹幕总字数。可以看出，点赞数高的用户，发布的弹幕数量也多，字数相应也很多。下图只展示了前10的情况，也可以调整下面的区间，看更多用户的弹幕表现。需要指出的是，只能取到用户ID，无法分析用户的偏好情况。



05 弹幕使用emoji表情情况



我们的数据中，使用表情的弹幕数量为1430，共使用了166种，2439次表情，"笑哭"这个表情使用数量最多，远远超过其他表情。一定程度上说明了这个表情受欢迎的程度，也侧面反映出节目本身的幽默效果很多，让网友们哭笑不得。

06 词云图

通过对弹幕进行分词处理，绘制出以下的词云图

看着这个词云图，瞬间感觉有溢出屏幕的欢乐，好像耳朵也能音乐听见断断续续的“哈哈哈哈”声，群众的眼睛是雪亮的，能让人如此开心的节目，火起来自然也就不足为奇了。

至此，我们基本完成了《向往的生活》第5期节目弹幕的抓取与简单的可视化分析工作。更多有趣的点大家可以自己去分析和发现。本来我还调用了百度的情感分析API，想对弹幕的情感倾向进行分析，但效果好像不是太好，结果就没有贴出来。感兴趣的盆友可以看代码。在后台回复“向往的生活”获取取源代码以及本文的pdf版本。祝大家生活愉快！