BILIBILI 弹幕获取及分析

王琦 赵茂祥 王美心

背景

弹幕，用户与视频的即时互动

抓取弹幕，对弹幕特征进行统计分析与展示

理解用户的互动行为/内容，情感倾向

步骤

获取弹幕

- 番剧
- 分区热门

解析弹幕

解析/存储

分析展示

弹幕特征

统计图表 & 词云 & 情感得分

Input

- 视频地址
- 分区名称及时间跨度

Output

- html图表
- 标签 & 词云

获取

import requests

get()方法

import re

在网页源码中找到弹幕cid

re.compile() + .search()

番剧：https://www.bilibili.com/bangumi/play/ss32974/

普通投稿：https://www.bilibili.com/video/BV1R54y1i71G

获取

[番剧]

```
In [7]: r.encoding = 'utf-8'
        r.text    #HTTP响应内容的字符串形式,即网页内容
```


```
,"premiereInfo":{}},"epList":[{"loaded":true,"id":317441,"badge":"","badgeType":0,"badgeColor":"#FB7299","epStatus"
:2,"aid":667574781,"bvid":"BV1Ea4y1t74Q","cid":172398507,"from":"bangumi","cover":"\\u002F\\u002Fi0.hdslb.com\\u002
Fbfs\\u002Farchive\\u002F849d184f244c3026de2ae48598b83d24ba3dab9a.jpg","title":"1","titleFormat":"第1话","vid":"","
longTitle":"BALL","hasNext":true,"i":0,"sectionType":0,"releaseDate":"","skip":{},"hasSkip":false,"rights":{"allow_
demand":0,"allow_dm":1,"allow_download":1}},{"loaded":false,"id":317442,"badge":"限免","badgeType":0,"badgeColor":"
#FB7299","epStatus":2,"aid":455131795,"bvid":"BV1Q541147W4","cid":175514325,"from":"bangumi","cover":"\\u002F\\u002
Fi0.hdslb.com\\u002Fbfs\\u002Farchive\\u002F98103598aea1c00536d3d16ca69edbbe4590a300.jpg","title":"2","titleFormat"
:"第2话","vid":"","longTitle":"400分之3","hasNext":true,"i":1,"sectionType":0,"releaseDate":"","skip":{},"hasSkip":f
alse,"rights":{"allow_demand":0,"allow_dm":1,"allow_download":1}},{"loaded":false,"id":317443,"badge":"限免","badge
Type":0,"badgeColor":"#FB7299","epStatus":2,"aid":285297007,"bvid":"BV19f4y1U7P9","cid":178316299,"from":"bangumi",
"cover":"\\u002F\\u002Fi0.hdslb.com\\u002Fbfs\\u002Farchive\\u002Fa256aa92f4c106dd8f6d46cc86764f9535f38bcc.jpg","ti
tle":"3","titleFormat":"第3话","vid":"","longTitle":"正确的门","hasNext":true,"i":2,"sectionType":0,"releaseDate":"",
"skip":{},"hasSkip":false,"rights":{"allow_demand":0,"allow_dm":1,"allow_download":1}},{"loaded":false,"id":317444,
"badge":"限免","badgeType":0,"badgeColor":"#FB7299","epStatus":2,"aid":625265423,"bvid":"BV1Mt4y127TK","cid":181666
249,"from":"bangumi","cover":"\\u002F\\u002Fi0.hdslb.com\\u002Fbfs\\u002Farchive\\u002F234051ff7ff0df3f8250e25941d7
212dc701eb85.jpg","title":"4","titleFormat":"第4话","vid":"","longTitle":"绿色四月","hasNext":true,"i":3,"sectionType
":0,"releaseDate":"","skip":{},"hasSkip":false,"rights":{"allow_demand":0,"allow_dm":1,"allow_download":1}},{"loade
d":false,"id":317445,"badge":"限免","badgeType":0,"badgeColor":"#FB7299","epStatus":2,"aid":497983154,"bvid":"BV1zK
41157ko","cid":184632204,"from":"bangumi","cover":"\\u002F\\u002Fi0.hdslb.com\\u002Fbfs\\u002Farchive\\u002Fd07c859
a98cb13065e9ff7bba5a4d82d62163aa8.jpg","title":"5","titleFormat":"第5话","vid":"","longTitle":"王冠的去向","hasNext":
```

获取

[分区]

获得视频地址列表

https://s.search.bilibili.com/cate/search?main_ver=v3&search_type=video&view_type=hot_rank&order=dm&copy_right=-1&cate_id=21&page=1&pagesize=100&jsonp=jsonp&time_from=20200613&time_to=20200627

主要参数

view_type=hot_rank | order=dm | pagesize=100

time_from=20200613&time_to=20200627

import pandas

df[‘cid’] / df[‘tag’]

获取

```
In [26]: print(df.iloc[1])

senddate                                                         1592735414
rank_offset                                                               2
tag                                      VLOG,生活记录,生活,翔翔大作战,小翔哥,父亲节
duration                                                                642
id                                                                838550068
rank_score                                                            70242
badgepay                                                              False
pubdate                                                 2020-06-21 17:43:22
author                                                               拜托了小翔哥
review                                                                54698
mid                                                               353539995
is_union_video                                                            0
rank_index                                                                0
type                                                                  video
arcrank                                                                   0
play                                                                6724033
pic                     //i0.hdslb.com/bfs/archive/786dc5a10305c16ce4f...
description             大家好，我是小翔哥，未来在各个平台，我会通过@拜托了小翔哥 继续和大家分享好吃好玩好快乐的内...
video_review                                                          70242
is_pay                                                                    0
favorites                                                            117770
arcurl                          http://www.bilibili.com/video/av838550068
bvid                                                           BV1Wg4y1q7fk
title                                                          小翔哥回来了，我们重新开始
Name: 1, dtype: object
```

解析

构造请求的弹幕地址

danmaku = "https://api.bilibili.com/x/v1/dm/list.so?oid={}".format(cid)

import xml.etree.ElementTree as ET

从文件中解析xml文件

获取包含弹幕信息的节点

编码写入文件

解析

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


```
▼<i>
  <chatserver>chat.bilibili.com</chatserver>
  <chatid>136038300</chatid>
  <mission>0</mission>
  <maxlimit>500</maxlimit>
  <state>0</state>
  <real_name>0</real_name>
  <source>e-r</source>
  <d p="4.88700,1,25,16777215,1578716055,0,9fedd4ea,27039566709391364">开播了</d>
  <d p="11.76000,1,25,16777215,1578716176,0,89696f7f,27039629983088640">爷笑了</d>
  <d p="7.11100,1,25,16777215,1578716777,0,2fa90446,27039945055010818">啊啊</d>
  <d p="41.10700,1,25,16777215,1578720227,0,f29aae97,27041754081919040">夹击妹都</d>
  <d p="70.72800,5,25,15772458,1578722275,0,7a838972,27042827668029442">更多硬币 更强威力</d>
  <d p="5.06400,1,25,16777215,1578722376,0,a80d20d6,27042880592805892">激动啊</d>
  <d p="46.89000,1,25,16777215,1578722442,0,e124cf99,27042915101442048">初春！！！ </d>
  <d p="6.69800,1,25,16777215,1578722449,0,c8052046,27042919127973890">见证历史</d>
  <d p="96.41900,1,25,16777215,1578723041,0,57c86fd3,27043229325066244">没有教主吗</d>
  <d p="10.71900,1,25,16777215,1578723748,0,89eb0c58,27043600120938498">开播了，开播</d>
  <d p="31.32600,1,25,16777215,1578725238,0,8335d944,27044381101391874">姐姐大人！！！！！！ </d>
```

分析

词云关键词的准备

```
import jieba
import jieba.analyse
```

加载停用词，jieba.analyse.set_stop_words(‘.txt’)

tf-idf 提取关键词，jieba.analyse.extract_tags(content, topK=100, allowPOS=[])

分析

计算情感得分

分词，去除停用词

加载情感词/程度词/否定词

SCORE = 情感基础分数*程度倍数*(-1) if 包含否定词

展示

[番剧]

import pyecharts

.Pie() / .Line() / .Bar() / .WordCloud()

.html格式，可视化图表

展示

[分区]

```
import imageio
from wordcloud import WordCloud
import matplotlib.pyplot as plt
```

绘制词云jpg

情感得分分布

Time to run it

优化

嵌入到web应用中，部署到云端，用户通过互联网访问

数据存储管理，持久化

并行处理提高运行速度

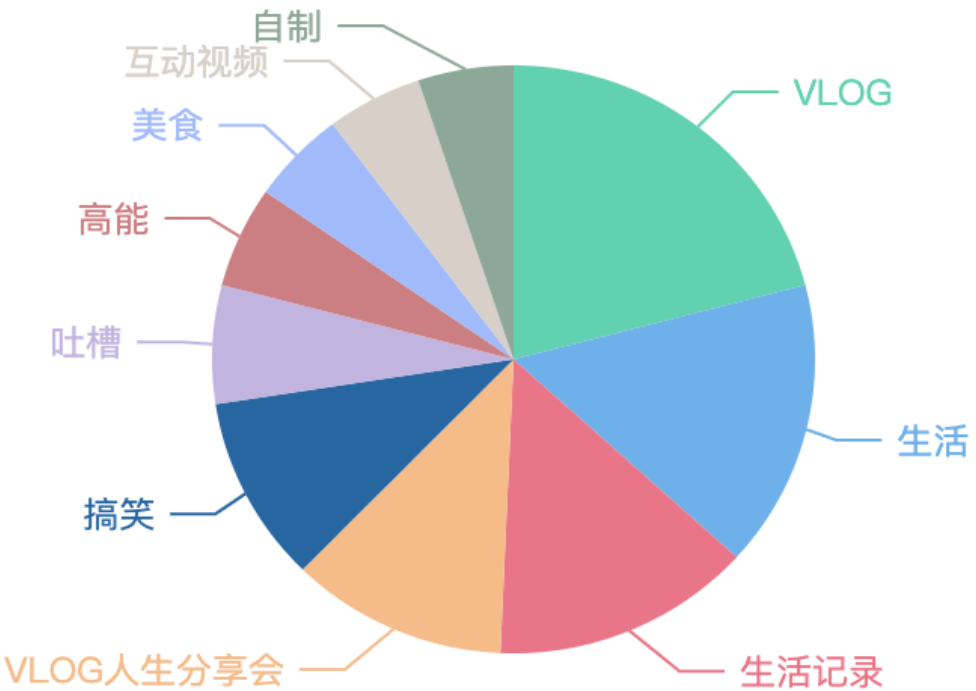
…

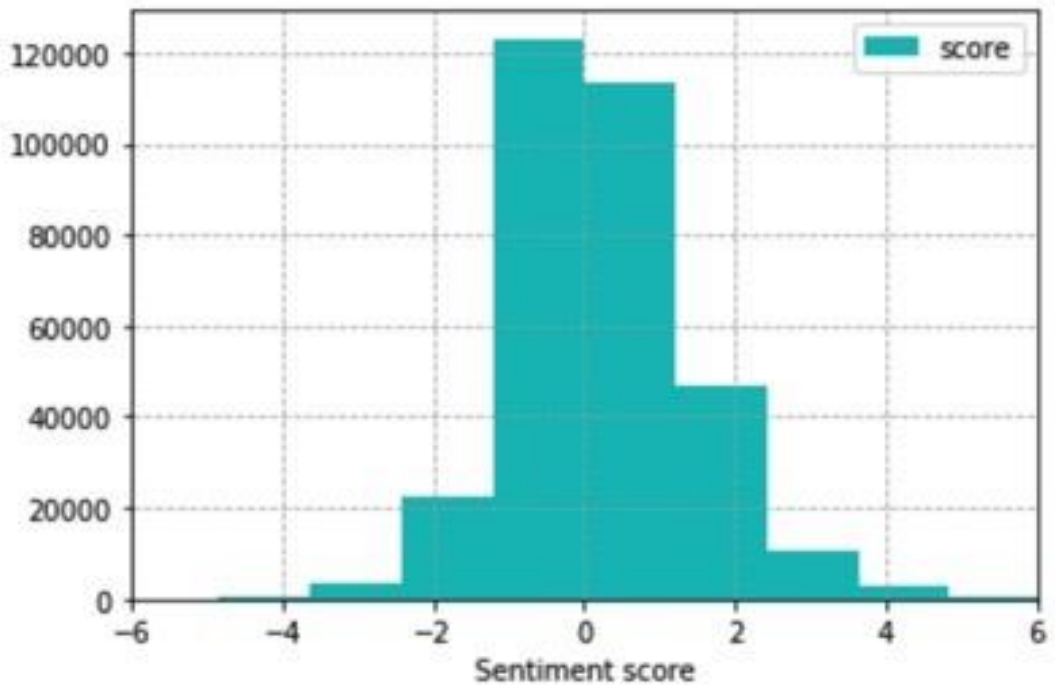
分工

王琦	番剧弹幕爬取、统计分析可视化
赵茂祥	分词词云、分区热门弹幕爬取
王美心	PPT、情感得分

Thank you!

++


自制
互动视频
美食
高能
吐槽
搞笑
VLOG人生分享会
VLOG
生活
生活记录


score
120000
100000
80000
60000
40000
20000
0
−6
−4
−2
0
2
4
6
Sentiment score

++

新婚
百年好合
甜甜
好好看
保护
快乐
好看
无悔
害怕
加油
裂开
领取
高考
可爱
喜欢
举手
猝不及防
拉低
恭喜
逍遥散
严谨
冲冲
好像
厉害
期许
六道轮回
正道
成功
好听
自愧不如
三生有幸
考试
白头偕老
吓死
分手
好评