大作业2设计文档

洪一宁 2020011022

0. 目录结构

/code

- data.json 为爬虫存储的数据
- get_data.py 为爬虫程序1（具体两个程序的区别见后文）
- 大作业2设计文档.pdf
- /creator 为爬虫爬取的up主头像
- /data_analyzation 为数据分析相关代码 具体文件对应内容见后文
- /my_website 为django文件所在的文件夹
  - data.json 复制自上级目录
  - db.sqlite3 自动创建文件
  - manage.py 自动创建文件
  - /my_website部分自动创建文件忽略不写
    - urls.py url匹配
    - settings.py 设置
    - views.py 数据处理、前往网站
  - /static
    - /creator 复制自上上级目录
    - /pic 复制自上上级目录
  - /templates
    - creators.html up主列表页
    - homepage.html 主页-视频列表页
    - search.html 搜索页
    - search_result_up.html up主搜索结果页
    - search_result_video.html 视频搜索结果页
    - up_information.html up主信息页
    - video_information.html 视频信息页
- /new_data 为爬虫程序2
- /pic 为爬虫爬取的视频封面

1. 项目设计

1.1 系统功能

1.1.1 主页-视频列表页

此页展示了所有爬取视频的列表，点击列表中的项目可前往对应视频信息页。页面可进行翻页、跳转操作，其右侧有导航栏，可以前往up主列表页和检索页。

1.1.2 up主列表页

此页展示了所有爬取up的列表，点击列表中的项目可前往对应up主信息页。页面可进行翻页、跳转操作，其右侧有导航栏，可以前往视频列表页和检索页。

1.1.3 搜索页

此页有一关键词输入框，有一对单选按钮，有一个搜索按钮。输入单关键词，选择是对up或者视频进行搜索后单击搜索按钮或按enter键可以前往对应搜索结果页。页面右侧有导航栏，可以前往up主列表页和视频列表页。

1.1.4 搜索结果页

1.1.4.1 视频搜索结果页

此页展示了所有标题或简介能够严格匹配关键字的视频的列表，点击列表中的项目可前往对应视频信息页。页面可进行翻页、跳转操作，其右侧有导航栏，可以前往视频列表页、up主列表页和检索页。

1.1.4.2 up主搜索结果页

此页展示了所有用户名或个人简介能够严格匹配关键字的up主的列表，点击列表中的项目可前往对应up主信息页。页面可进行翻页、跳转操作，其右侧有导航栏，可以前往视频列表页、up主列表页和检索页。

1.1.5 视频信息页

此页展示了单个视频的详细信息，包括标题、封面、简介、播放数、评论数、分享数、投币数、点赞数、收藏数、前五条评论、投稿时间、弹幕数。点击封面可前往B站原视频。页面右侧有对应up主头像、用户名、个人简介、uid、粉丝数，点击头像或用户名可前往对应up主信息页。后侧下方有导航栏，可以前往u视频列表页、up主列表页和检索页。

1.1.6 up主信息页

此页展示了单个up主的详细信息，左侧列有up主的所有投稿视频封面及对应标题，可进行翻页、跳转操作，点击视频可以进入对应视频信息页。页面右侧有导航栏，可以前往视频列表页、up主列表页和检索页。

1.2 数据量

5000个视频 3857个up主

1.3 使用算法

1.3.1 爬虫

由于不清楚考察要求，因此文件夹中有两个爬虫程序，其中第一个使用遍历av号的方法获得访问的地址，另一个通过爬取B站VOCALOID·UTAU区的视频列表获得地址。由于在得知不可以提前准备视频编号时爬虫工作已经基本完成，因此爬虫数据仍由第一个程序生成。第二个程序也可以正常工作，程序位于new_data文件夹下。除获取被爬取视频地址方式和爬取数据的存储位置不同外，两个程序没有区别。

爬虫爬取的信息包含视频标题、视频简介、视频播放页url、封面图片、播放量、弹幕数、上传时间、点赞数、投币数、收藏数、转发数、评论数、前5条评论、作者ID、作者简介、作者头像图片、粉丝数、是否为联合投稿、作者uid。

其中前5条评论通过selenium访问视频页面实现。粉丝数通过selenium访问up主个人空间实现。视频标题、视频简介、视频播放页url、封面图片、播放量、弹幕数、上传时间、点赞数、投币数、收藏数、转发数、评论数、是否为联合投稿、作者uid、作者ID通过bs4抓取视频页源代码实现。作者简介、作者头像图片通过bs4抓取up主个人主页源代码实现。代码中通过注释可以看出每部分如何实现。

爬取的数据中，封面图片存储至pic文件夹，up主头像存储至creator文件夹，其余信息按格式存储至data.json中

1.3.2 网页

网页的界面布局及格式设计主要由css、javascript共同完成，通过div标签分区实现。

本地数据的读取与处理全部位于view.py文件中。html主要完成全部显示工作。

所有数据都从json中以字典的格式读入python，并通过列表、字典的方式进行较快的处理。

搜索部分主要应用了python的字典、列表、str的find()函数完成。

在所有数据处理完毕后，view.py将数据以django变量的形式发送给html。

html全部分为左区和右区。左区为主体，右区为少量信息和导航栏。通过django的for和if语法实现动态网页。部分内容的显示通过js脚本实现。

网络架构、地址参数传输及正则表达式的匹配通过django实现。

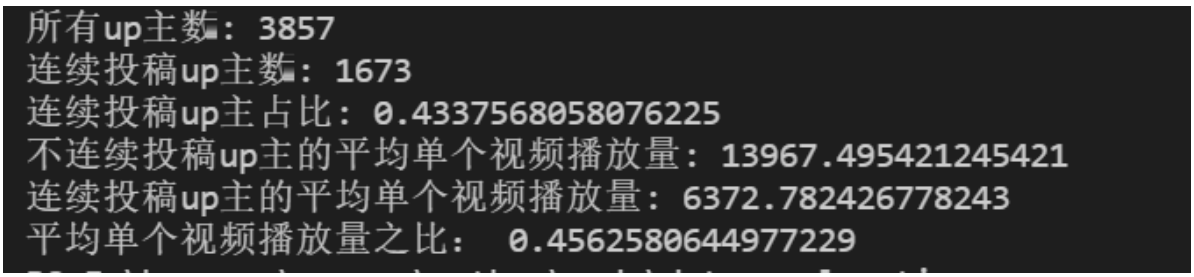
2. 数据分析

2.1 up主连续投稿行为分析

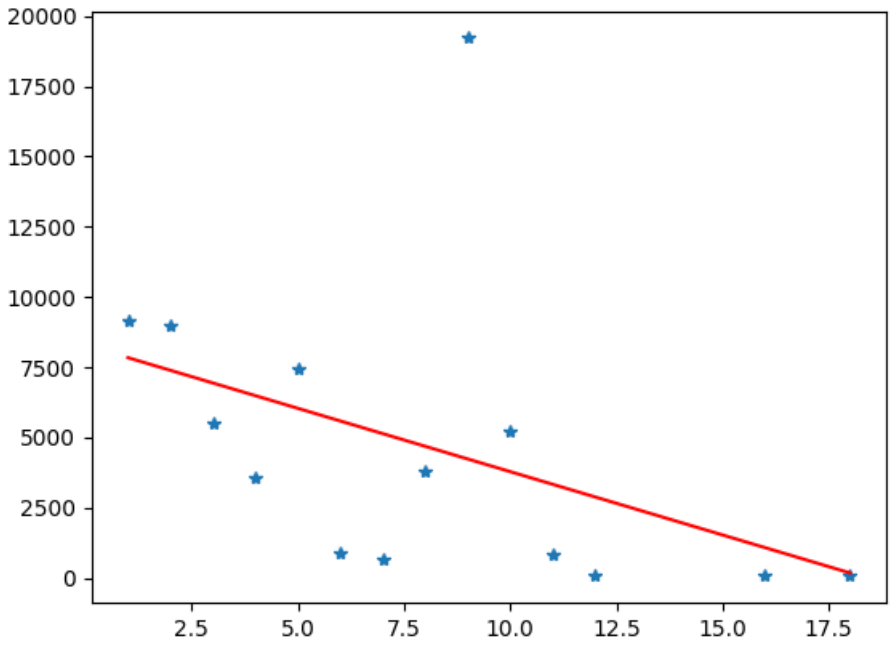
对应代码creator_behavior.py

由于爬取的是2018年8月22日约十二点至约十五点半这一段时间内所有的视频投稿，故如果爬取数据中，一个up主有多个稿件，则定义这个up主有连续投稿行为。

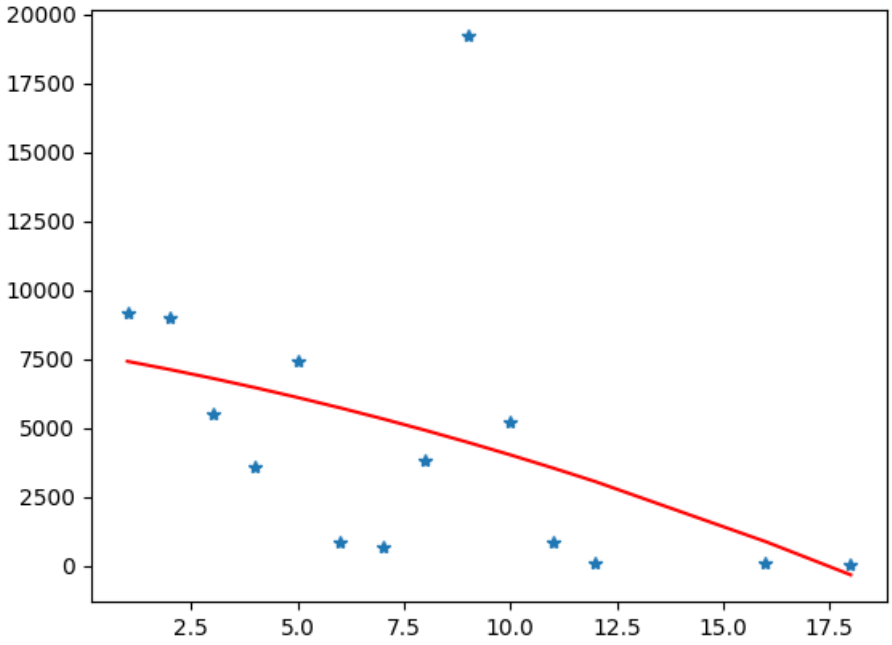
根据数据，共有3857位up主，其中1673位有连续投稿行为，占比0.4338。有连续投稿行为的up主的单个视频平均播放量为6372.78。没有连续投稿行为的up主的单个视频平均播放量为13967.50。二者相差0.456倍，差距较大，说明有连续投稿行为的up主可能没有连续投稿行为的up主那样关心视频质量。



对up主短时间内投稿的数量与单个视频平均播放量做线性拟合，得到表达式 $y = -450.5x + 8291$，其中y为平均播放量，x为投稿数量，二者负相关，绘图如下



二次拟合表达式为 $y = -9.822x^2 - 269x + 7702$，绘图如下



两种拟合都呈现负相关关系

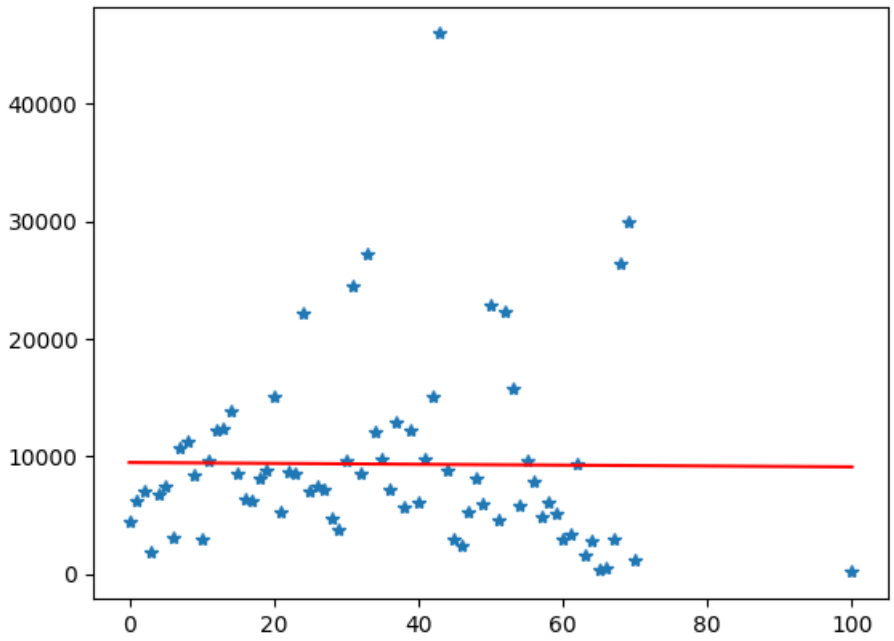
2.2 up主个人简介长度与播放量关系

对应代码introduce_and_view_cnt.py

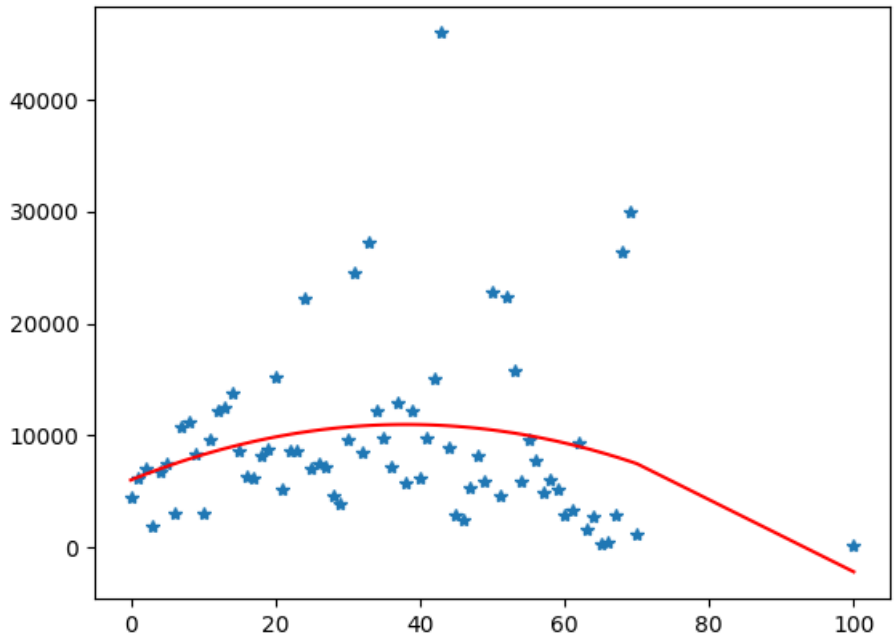
根据常识，个人简介如果长度超过10，则更有可能是一段提供了关于up主有效信息的文字。以10为界，个人简介长度大于10的视频为2752，单个视频平均播放量为10415.568；个人简介长度小于等于10的视频数为2248，单个视频平均播放量为5561.84653。二者差相差1.873倍，说明愿意通过个人简介更加详细地介绍自己的up主更有可能提供高质量的视频。

```
个人简介长度大于10的视频数：2752
个人简介长度小于等于10的视频数：2248
个人简介长度大于10的up主的平均单个视频播放量：10415.568313953489
个人简介长度小于等于10的up主的平均单个视频播放量：5561.84653024911
相差倍数：1.872681717718483
```

然而，线性拟合却表明，up主的个人简介与播放量关系很小，其拟合式为 $y = -3.719x + 9493$，其中y为播放量，x为个人简介长度



通过二次拟合，发现个人简介长度适中的up主播放量较多。拟合式为
$y = -3.432x^2 + 261.1x + 602$。

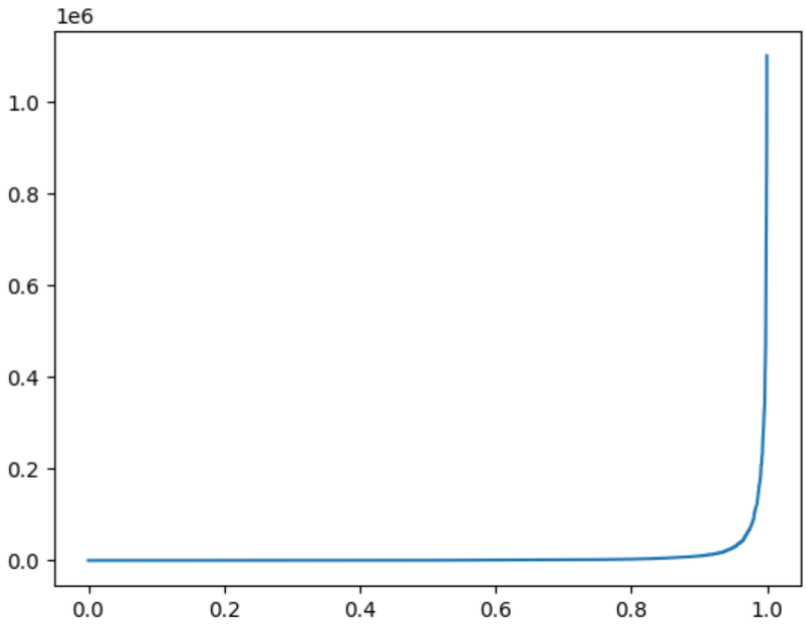


更高次的拟合也说明了这个结论，即个人简介长度适中的up主播放量较多。

2.3 B站播放量中的“二八定律”

对应代码play_times.py

将所有视频的播放量进行排序并画图如下。纵轴为播放量，横轴为比例。



根据分析，播放量前10.8%的up主拥有全部播放的89.297%。这一数据基本符合经济学二八定律，即20%的人拥有社会80%的财富，但此处的数据更加极端。推测其原因是网络内容具有便于传播的特点，一个优质视频一旦收到关注，将获得巨大的播放量。