上 海 财 经 大 学

**Shanghai University of Finance and Economics**

数据科学导论期末项目报告

**基于哔哩哔哩每周必看对视频热门度相关性研究**

**项目GitHub链接：**

|  |  |
| --- | --- |
| **院（系）：** | **信息管理与工程学院** |
| **班 级：** | **大数据1班** |
| **学 号：** | **2020111627** |
| **姓 名：** | **欧阳逸** |
| **日 期：** | **2022年6月20日** |

摘 要

在短视频充斥的时代，拥有高质量、原创度高的长视频无疑使得b站具有吸引更多年轻人的差异化竞争力。根据数据调查，b站平均年龄为为21岁，新注册用户不到20岁，大量Z世代年轻人在b站上观看自己喜爱UP主的视频，观看的同时留下许多交互活动产生的用户数据如弹幕、评论、点赞等。由于b站注册需要答题的门槛，这些视频的点赞量、播放量、收藏量等数据往往具有较强的真实性与可分析性。通过分析这些数据并将其可视化，可以很好的展示当代年轻人喜爱潮流与b站每周必看选取相关算法。

为了进一步了解b站每周必看选取视频的具体标准与当代年轻人长视频喜爱倾向，笔者爬取了近五十周的b站每周必看视频榜单并对相关元素分析统计，采用相关机器学习的学习方法来提取选取视频特征。

**关键词：**哔哩哔哩 机器学习 数据可视化

目 录

[**一、绪论 4**](#_Toc106024879)

[第一节、 项目背景 4](#_Toc106024880)

[第二节、项目内容 4](#_Toc106024881)

[第三节、项目组织框架 5](#_Toc106024882)

[**二、数据采集 5**](#_Toc106024883)

[第一节、主题页面结构特征分析及数据结构分析 5](#_Toc106024884)

[第二节、爬虫代码示例 8](#_Toc106024885)

[**三、数据集介绍、预处理及储存 9**](#_Toc106024886)

[第一节、B站数据集介绍 10](#_Toc106024887)

[第二节、数据预处理 10](#_Toc106024888)

[第三节 数据集储存 12](#_Toc106024889)

[**四、数据分析与可视化 14**](#_Toc106024890)

[（一）根据数据集统计分析 14](#_Toc106024891)

[（二）根据数据集拟合分析 18](#_Toc106024892)

[（三）结果分析 20](#_Toc106024893)

[**五、项目总结 20**](#_Toc106024894)

# 绪论

## 项目背景

近十年来B站凭借其独特的二次元文化环境在Z世代网络上扎根，从一开始的小众二次元原创内容，逐渐发展为泛二次元圈层；然后从这些广泛对二次元有兴趣的青年群体泛化为更多的自然属性的年轻人群体。B站之所以能吸引这么多用户并拥有高留存率，不仅仅与B站本身的多元性原创内容有关，还与B站的相关推荐算法、排行榜、每周必看推荐有关，让每一个用户都能看见全新且适合的内容，同时鼓励更多的新人UP主参与更加丰富的内容创作之中。在大数据时代下，对于哔哩哔哩相关数据的分析有着重大意义，对用户来说，根据分析的数据可以选择更加优质的UP主与潜在高质量视频，提高用户相关体验；对于B站来说，可以改进推荐算法，增加更多创作激励与活动，促进内容创作进一步发展。因而，基于哔哩哔哩的数据分析是有重要意义的。

可以观察到，相对优质的视频将实时选入排行榜中，其中每周最为典型、最具代表性的高质量视频将被选入每周必看榜单。这个榜单既可以充分展示B站用户所喜爱视频的类型、所喜爱的UP主等等，也能够反应出B站相关推荐算法机制和榜单选取标准。本项目旨在利用哔哩哔哩每周必看相关数据，分析榜单选取标准以及用户视频、UP主喜爱度。

## 第二节、项目内容

本项目基于哔哩哔哩，中国年轻世代高度聚集的文化社区和视频平台展开研究,首先使用网络爬虫从视频平台上抓取项目实验所需要的相关网络数据，并将其通过数据清晰整合为哔哩哔哩数据集，然后根据所需数据得到各项数据（播放量、点赞、投币、收藏）相关关系、排行前十的视频与UP主、每周必看在时间维度上视频数量的变化、热门视频标题词云展示，最后将利用机器学习方法构建相关度模型，结合播放量占比图等数据分析，最终得到喜爱度分析结果与榜单选取标准。

## 第三节、项目组织框架

本报告内容将按如下顺序进行安排：

第一章为绪论，介绍本项目的项目背景以及研究内容；

第二章为数据采集。详细描述本项目数据的采集过程与部分爬虫代码示例；

第三章将对本项目使用的数据集以及储存方式进行相关介绍，展示数据的预处理过程；

第四章将展示本项目数据分析过程，并将实验结果可视化后进行解释；

第五章，总结项目。

# 二、数据采集

## 第一节、主题页面结构特征分析及数据结构分析

1.打开源码

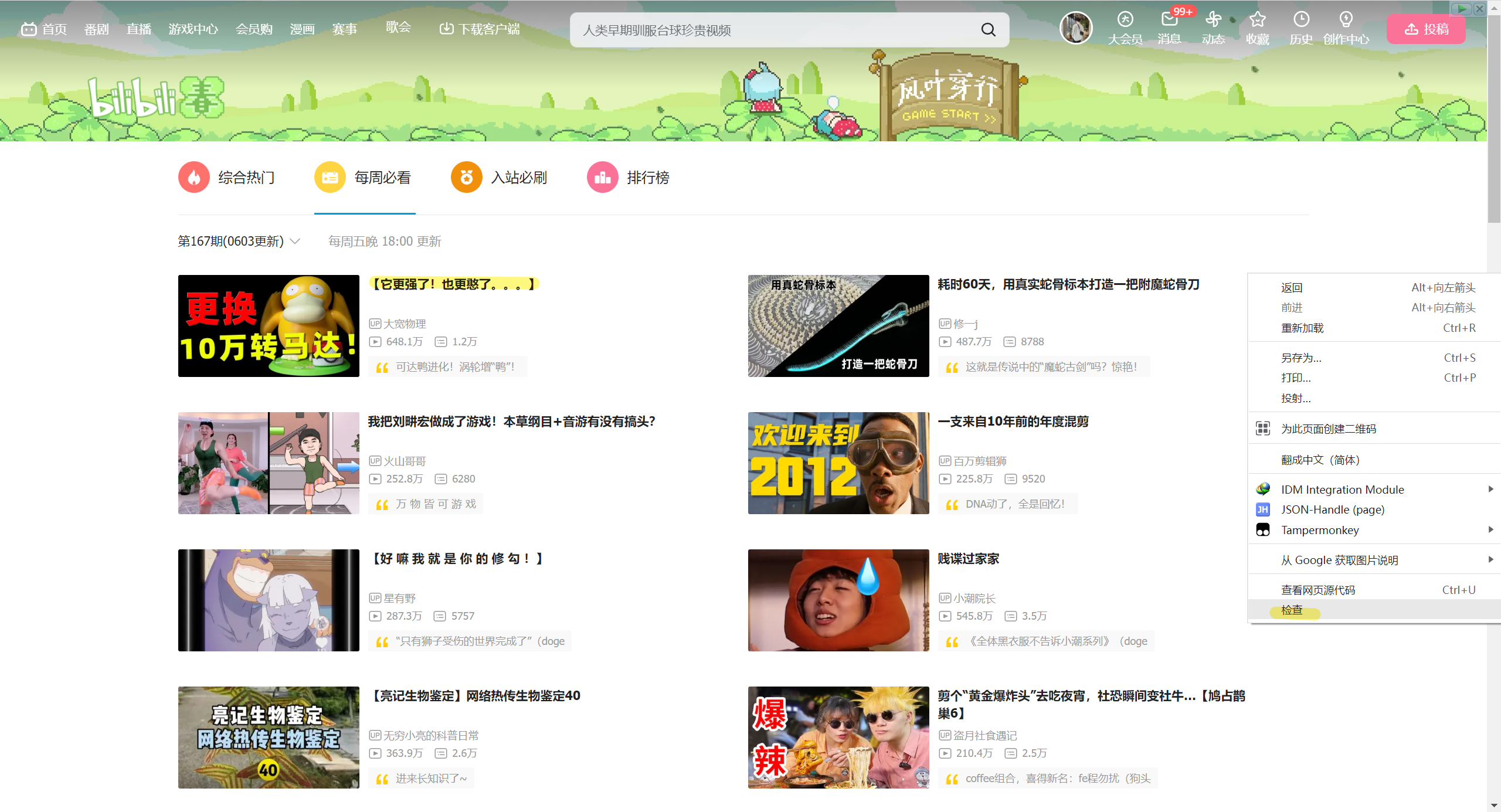


图2.1.1 每周必看页面

2. 不进行html网页的解析，而是对接口数据进行JSON数据提取。

复制数据，在开发者模式的network中进行搜索，刷新页面可获取到JSON数据连接，点击预览即可看到JSON数据内容，其中包含每周必看的所有具体信息。

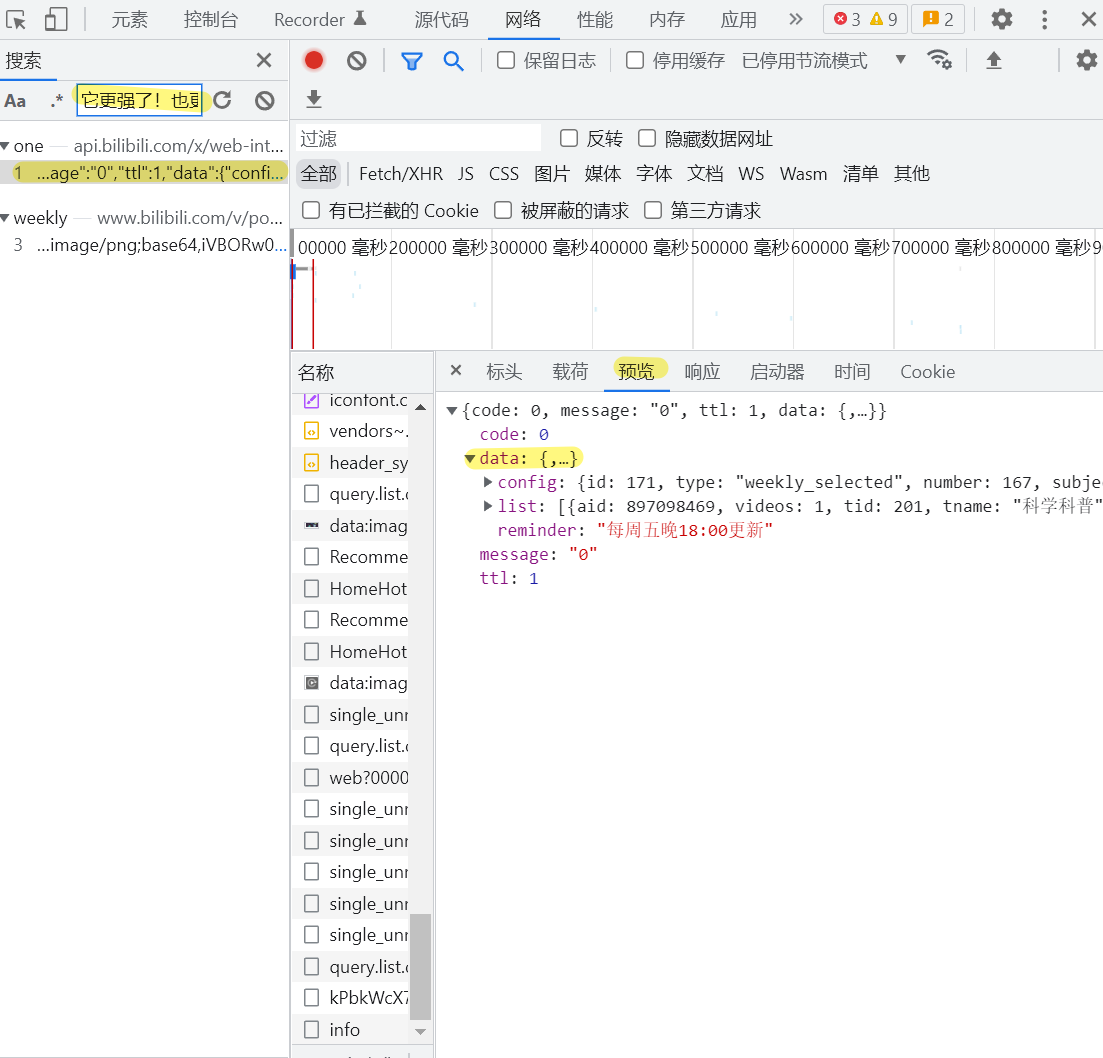


图2.1.2 获取目标链接

3. 获取请求链接以及请求时的headers模拟浏览器访问服务器

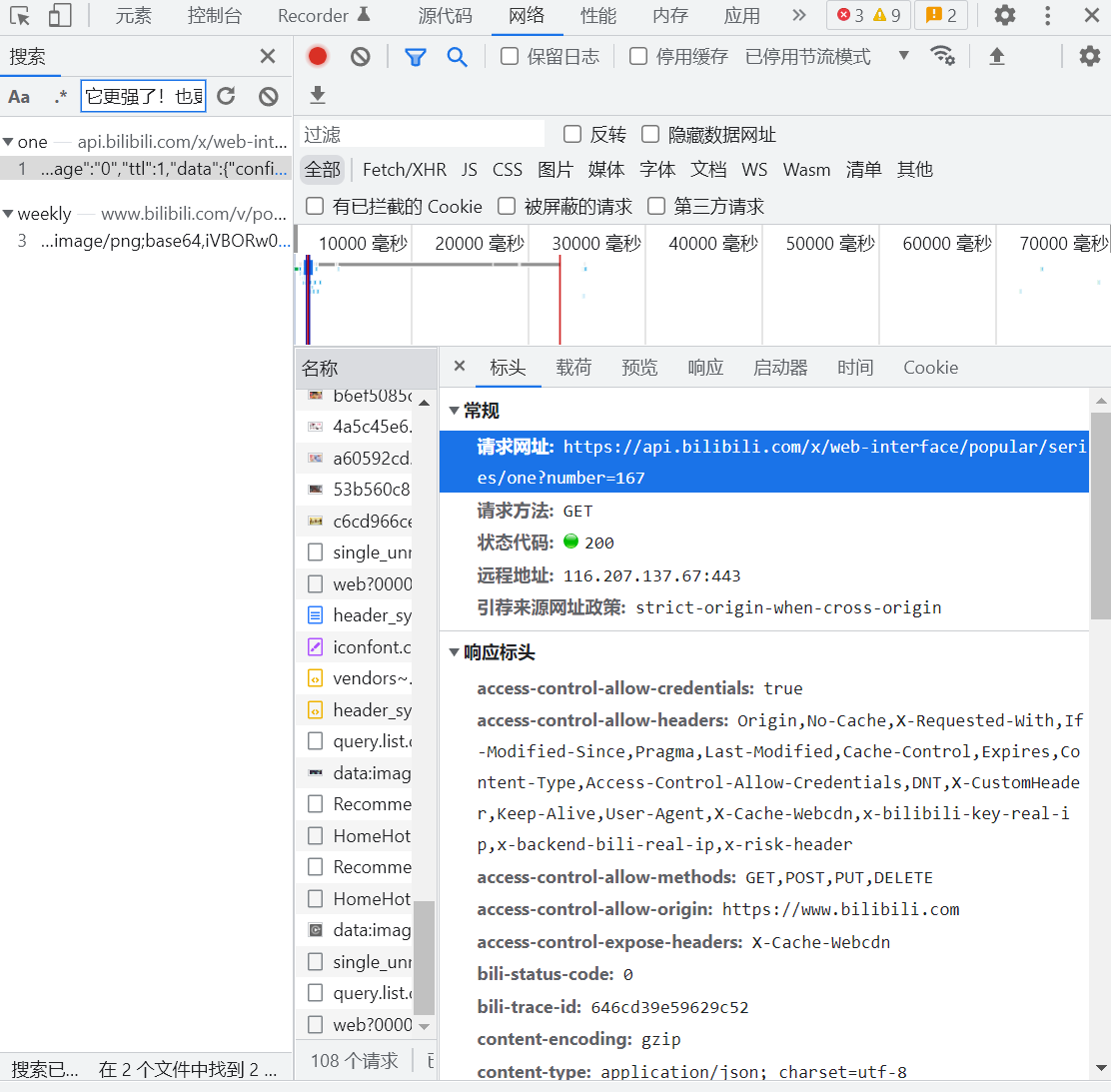


图2.1.3 复制请求网址

4.复制链接后粘贴在新页面中即可预览JSON文件



图2.1.4 浏览器打开JSON文件

5.分析数据结构

Data下面的list包含了本周所有视频的相关数据，以list[0]为例，可以找到可能需要的数据有：aid, tname, title, owner下的mid, name, stat下的view, reply, favorite, coin, like



图2.1.5 分析JSON数据结构

6.更改链接中number的周数即可获得所需周数的数据



图2.1.6 更改链接参数

## 第二节、爬虫代码示例

本项目使用python中快速、高层次的web抓取框架Scrapy进行多网页数据抓取，根据上述网页分析和数据结构分析编写了bilispider.py和items.py,并且对settings进行修改以应对反爬虫机制。一下相关代码示例分别是主要的bilispider.py和items.py。

项目中首先将参数start\_urls起始网址设置为<https://api.bilibili.com/x/web-interface/popular/series/one?number=167>，通过循环遍历过去五十周的每周必看数据页面链接，然后用Scrapy Request发送请求访问页面数据。得到数据后，根据上述结构分析，逐一将每个节点下的like、view等数据保存在items中，然后返回至items.py。

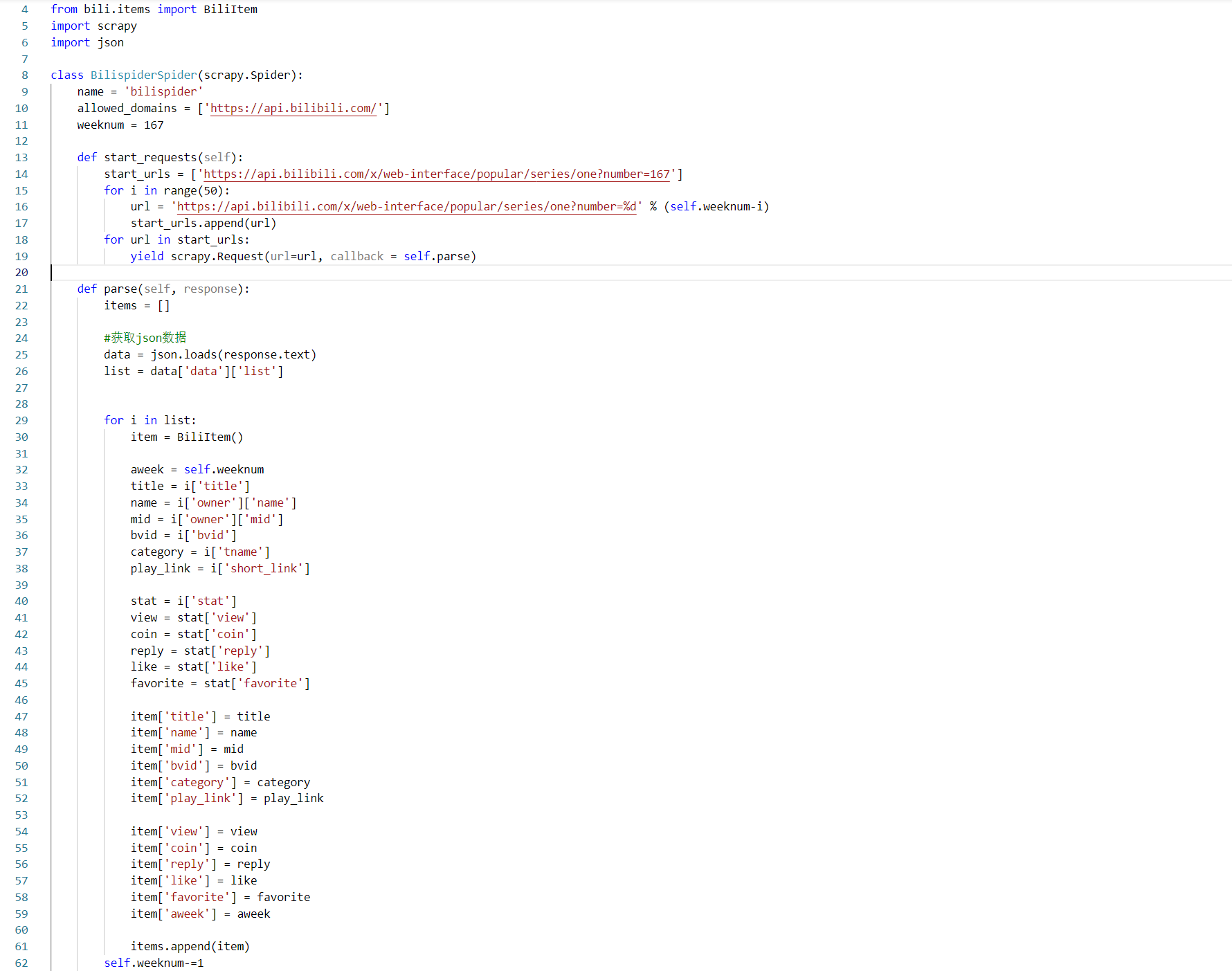


图2.2.1 bilispider.py

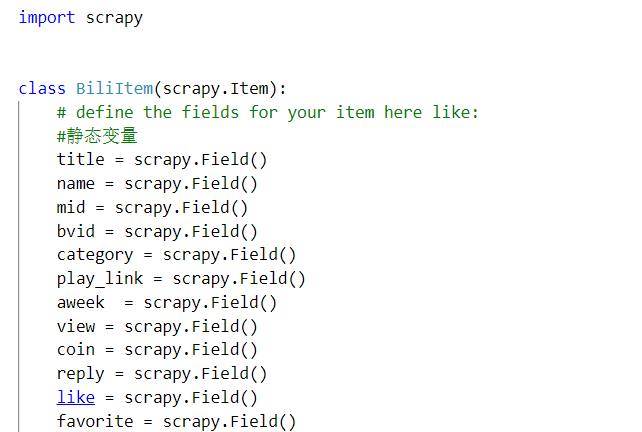


图2.2.2 items.py

相关代码及设置编写好后，在终端进入bili文件夹，输入scrapy crawl bilispider -o bili.csv可得到爬取后的csv文件。

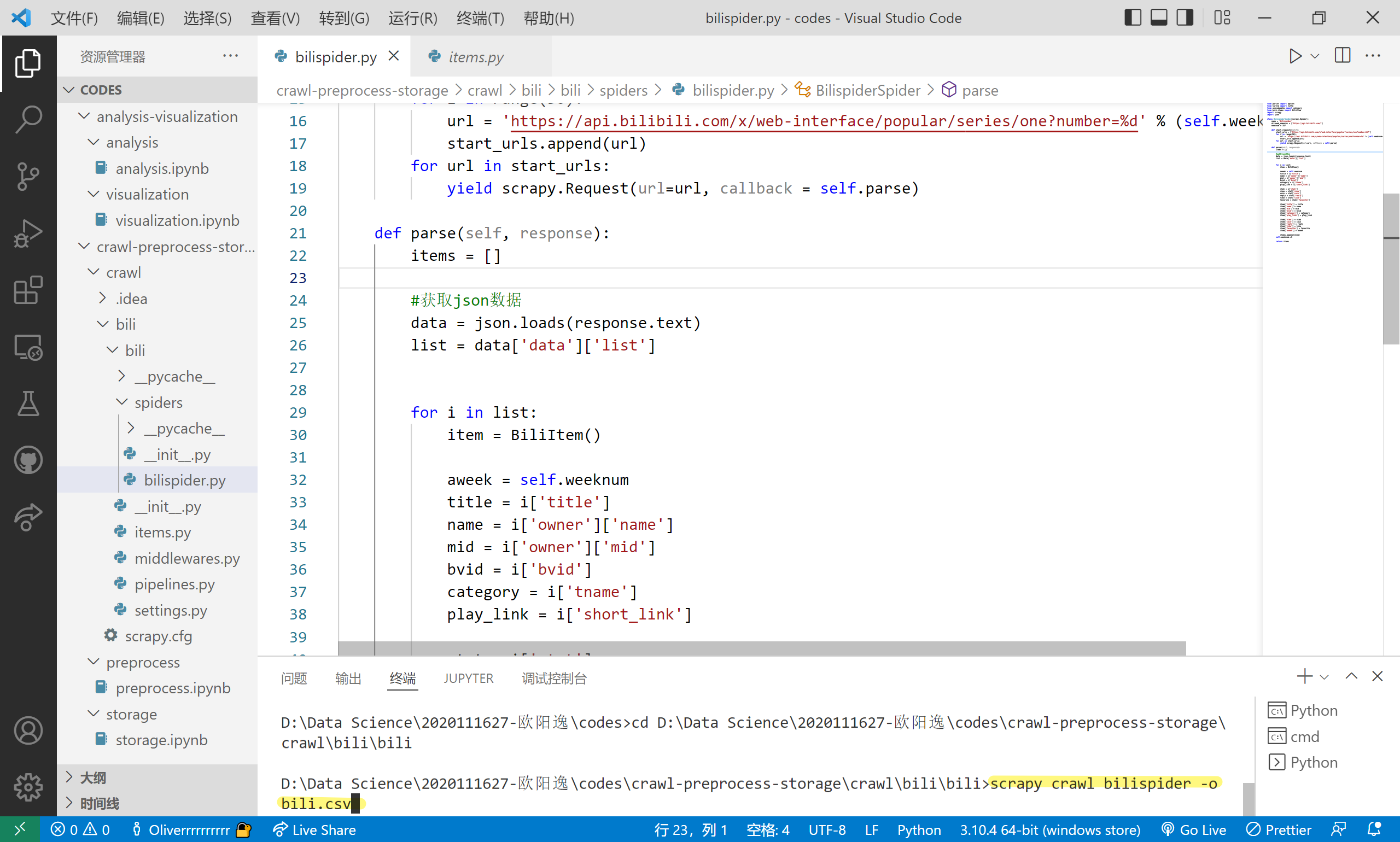


图2.2.3 获取csv文件

# 三、数据集介绍、预处理及储存

本项目按照第二招所介绍的数据采集方法从B站平台获取了本项目所需的数据集，记为B站数据集。本章将对该数据集进行介绍，并说明数据的预处理过程。

## 第一节、B站数据集介绍

在爬取的B站数据集中共包含2092条视频，对于每一个视频，我们收集了视频相关数据如view、like等，最终得到25104条B站视频数据，其中各字段的说明如表3-1所示：

|  |  |
| --- | --- |
| **字段名** | **说明** |
| aweek  bvid  category  coin  favorite  like  mid  name  play\_link  reply  title  view | 周数  视频编号  视频类别  投币数  收藏量  点赞数  UP主编号  UP主名字  播放链接  评论数  标题  播放量 |

表3-1 B站视频数据集字段说明

## 第二节、数据预处理

首先，我们从收集的数据集中提出不需要的字段，保留本次项目所需要的字段不为空的记录aweek, category, coin, favorite, like, name, title, view.接着我们单独将每个视频标题数据拿出来，为了找到热门视频标题中出现频率较高的词语，首先用jieba的lcut方法进行截取，用collection的Counter方法对词频进行统计，发现停用词较多，分析效果较差。之后我们尝试用extract\_tags方法对title进行分析截取，再用wordcloud对词云进行绘制。在经过上述预处理步骤后得到的实验数据集上绘制了词云，展示结果如图3.2所示：

从图3.2可以看出，在过去五十周的每周必看视频标题的热词主要有：一个、原神、我们、游戏、动画、视频、什么、这么等等，其中语气词或者疑问词如这么、什么、如何包含较多，可以推断出标题具有一定暗示性、神秘性可以增加用户观看兴趣，提高视频播放量。而像视频标签词如动画、游戏、原神、中国等可以通过标题展示视频相关联的主题，因而能吸引到更多感兴趣的目标人群进行观看。

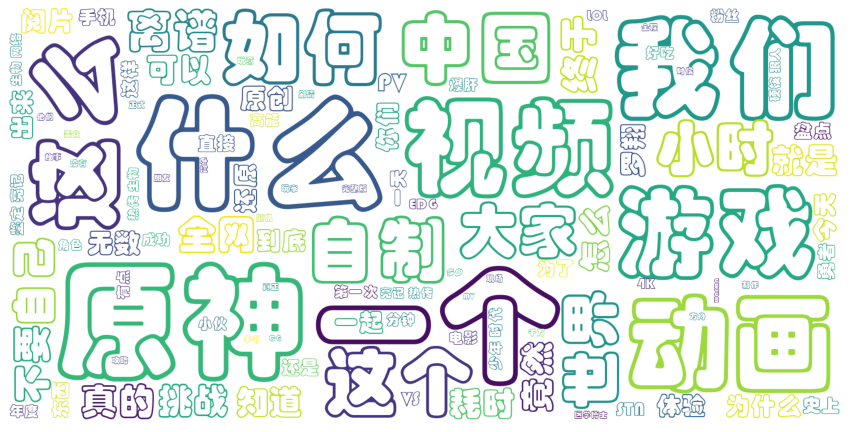
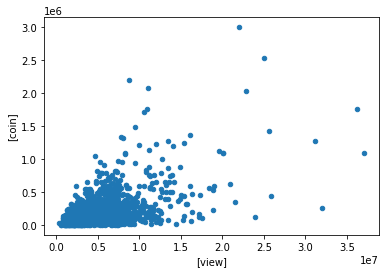
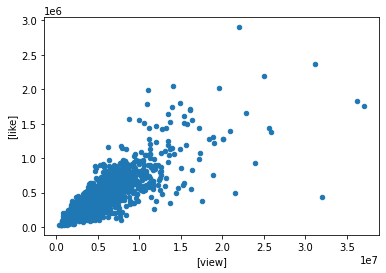


图3.2.1 B站视频标题实验数据集词云

除此之外，为了探究影响播放量的相关因素，我们对播放量与点赞、投币、收藏、回复的之间的关系通过散点图进行研究。从图3.3中我们可以看出，播放量与点赞数关联性很强，和投币数与收藏量都离散度都不高，具有正相关度，与回复数相关性很弱，由此可以大致确定播放量与点赞、投币、收藏存在关联性。



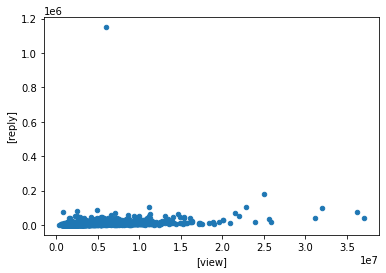
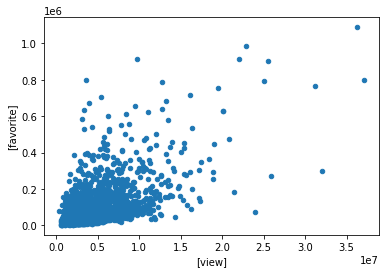


图3.2.2 播放量与点赞、投币、收藏、回复关系散点图

## 第三节 数据集储存

对数据进行预处理之后，还需要将处理后的数据集进行储存和持久化。考虑到数据集体量较小，我们选择用较为轻量且交互性较强的MySQL来储存数据集，并且将数据存储在xlsx文件中以持久化。操作步骤如下：

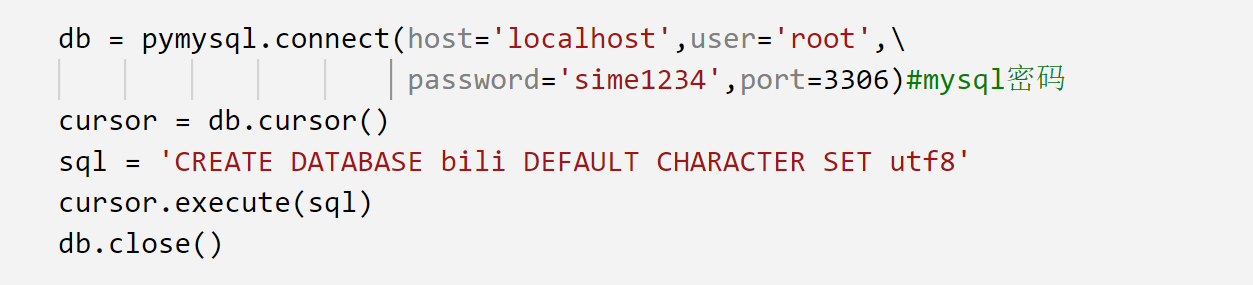


图3.3.1 连接MySQL，创建bibi数据库



图3.3.2 创建表biliweek



图3.3.3 打开并读取bili.csv

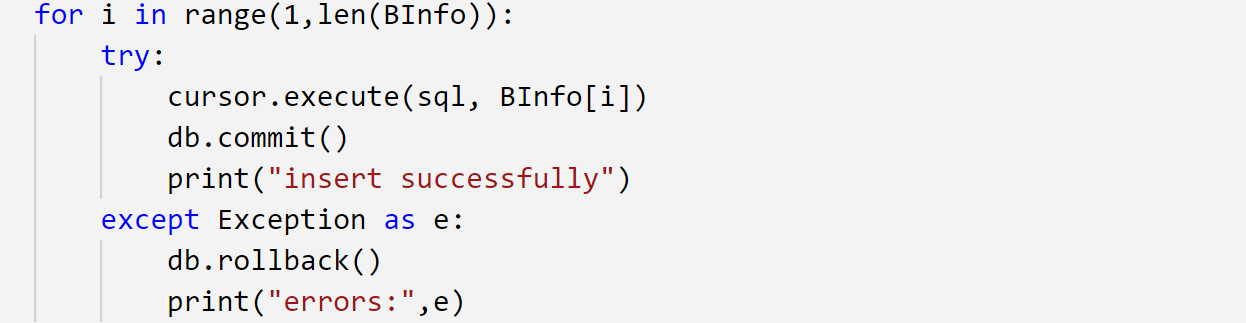


图3.3.4 插入数据

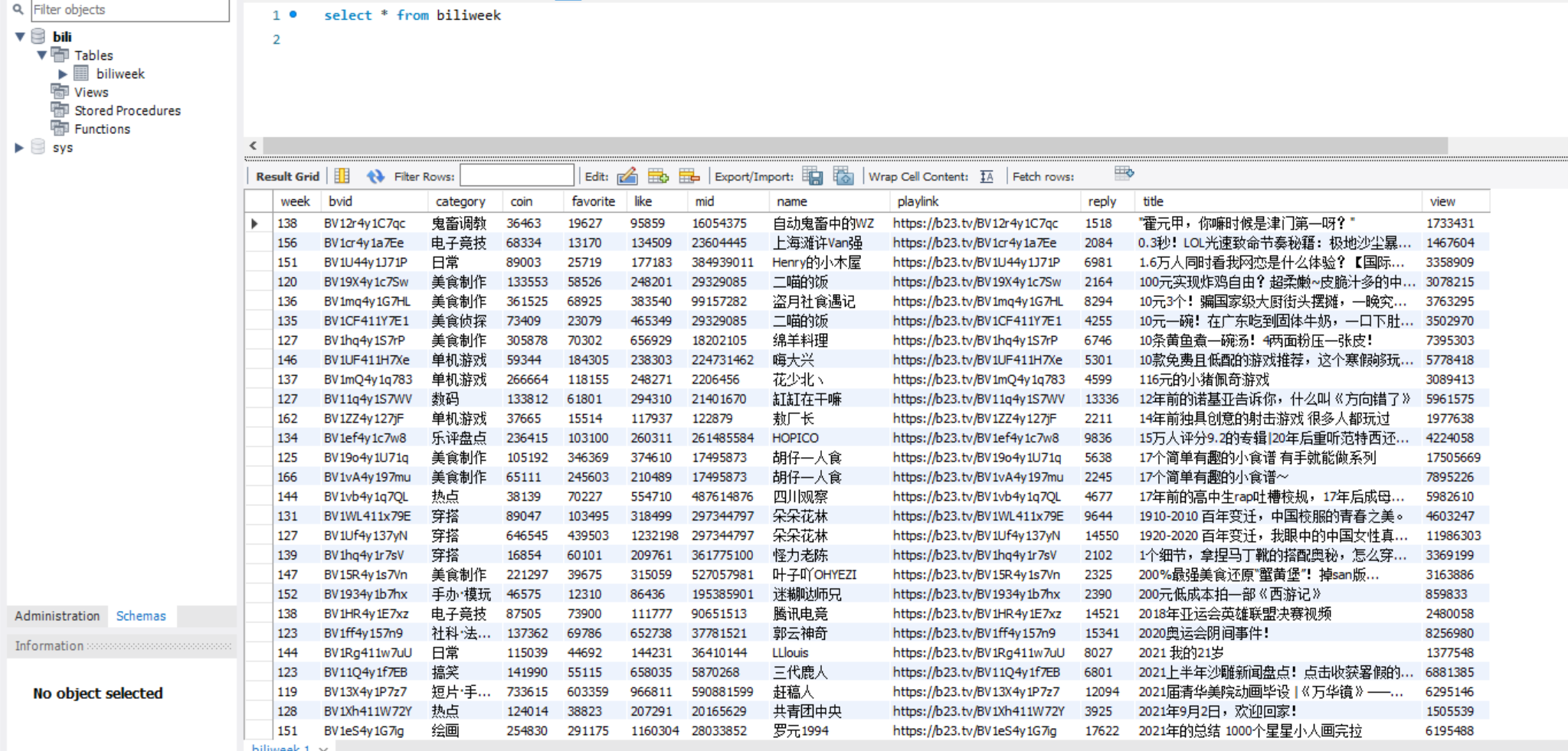


图3.3.5 MySQL储存数据查询结果

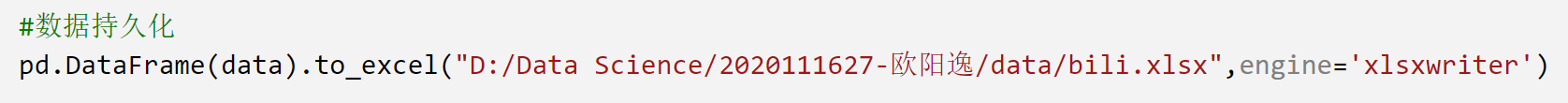


图3.3.6 数据持久化

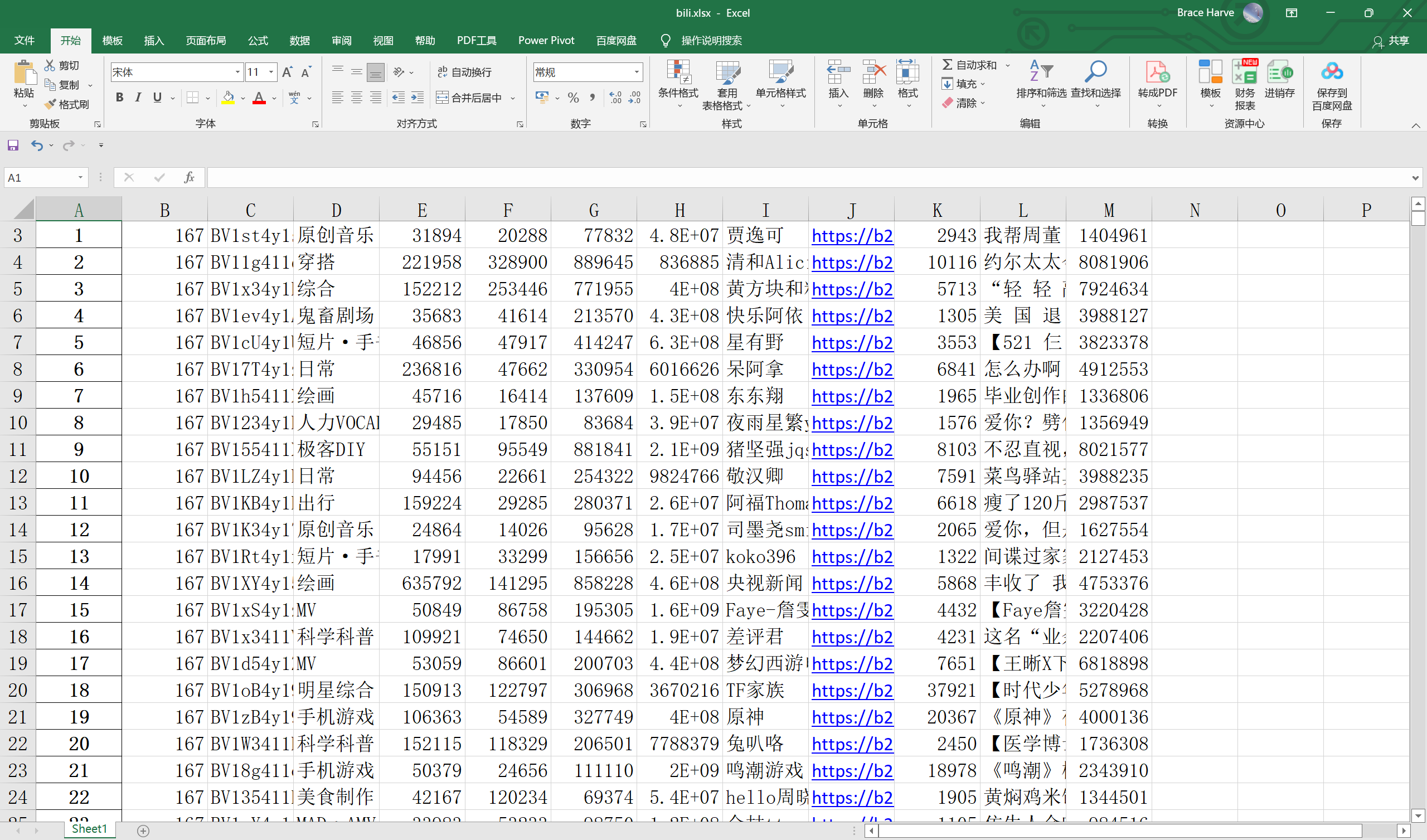


图3.3.7 bili.xlsx储存数据

# 四、数据分析与可视化

## （一）根据数据集统计分析

在对数据进行预处理和筛选之后，为了探究视频选取的相关标准，我们对数据各个维度进行相关统计分析。

根据B站数据集统计得到五十周每周必看榜单中播放量前十的视频如图3.1所示，其中播放量最高的视频为“⚡萨 日 朗！！！⚡”，达到了37056387次播放。前十名的视频类别为：'人力VOCALOID', '电子竞技', '明星综合', '舞蹈综合', '手工', '翻唱', '单机游戏', '翻唱','手机游戏', '极客DIY'，可以看出有关唱歌、游戏的视频容易用户循环播放，提高播放量，从而被选入榜单。

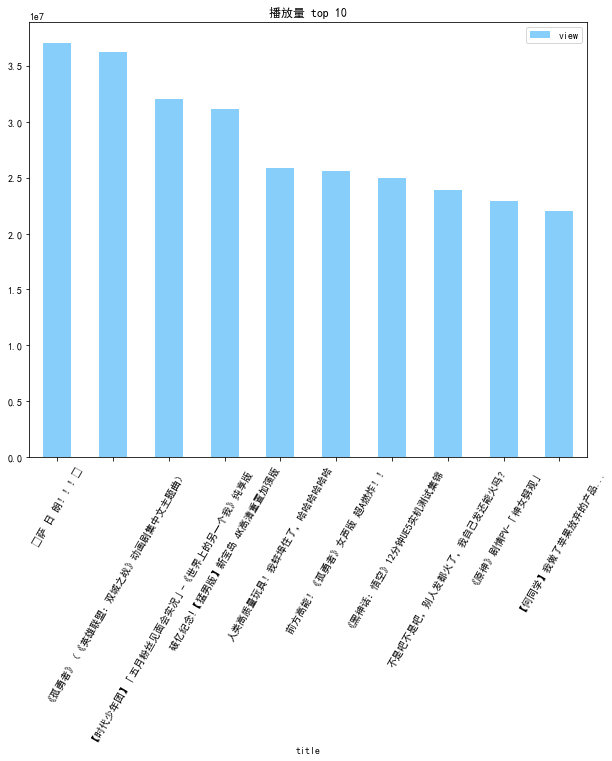


图4.1 播放量前十视频

在图4.2中可以看到五十周内上榜次数最多的前十名UP主，分别为：'盗月社食遇记', '凉风Kaze', '老番茄', '绵羊料理', '小约翰可汗', '原神', '小潮院长', 'TF家族', '无穷小亮的科普日常', '大漠叔叔'，上榜次数均高于18次，其中盗月社食遇记五十周内上榜了三十五次，可以看出其视频质量高，为过去五十周内最热门的UP主之一。

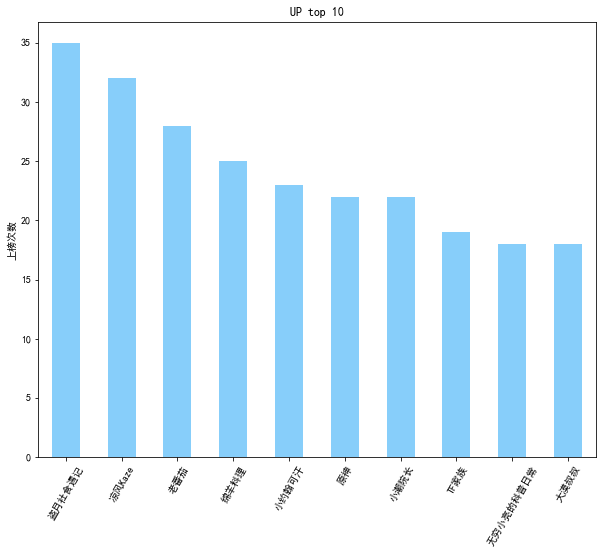


图4.2 上榜次数前十的UP主

之后我们对上榜视频各个类别进行数量统计，用来分析哪种视频更受B站用户欢迎，更容易被选入每周必看榜单。由图4.3可以明显看出搞笑类视频上榜数量为224，呈现断崖式高居榜首，其占比达到了10.70%；其后的单机游戏、日常、美食制作、手书配音上榜数量均超过100。由此可知B站当前制作搞笑日常、美食配音的视频相对来说质量高、用户基数大，拥有一定粉丝数的相关领域UP主更容易上榜，同时竞争激烈，新人UP所分得的流量资源可能相对较少。同时由此图可以观察到基本每个领域都会有所涉及，相对小众且高质量的视频也会被选入每周必看。

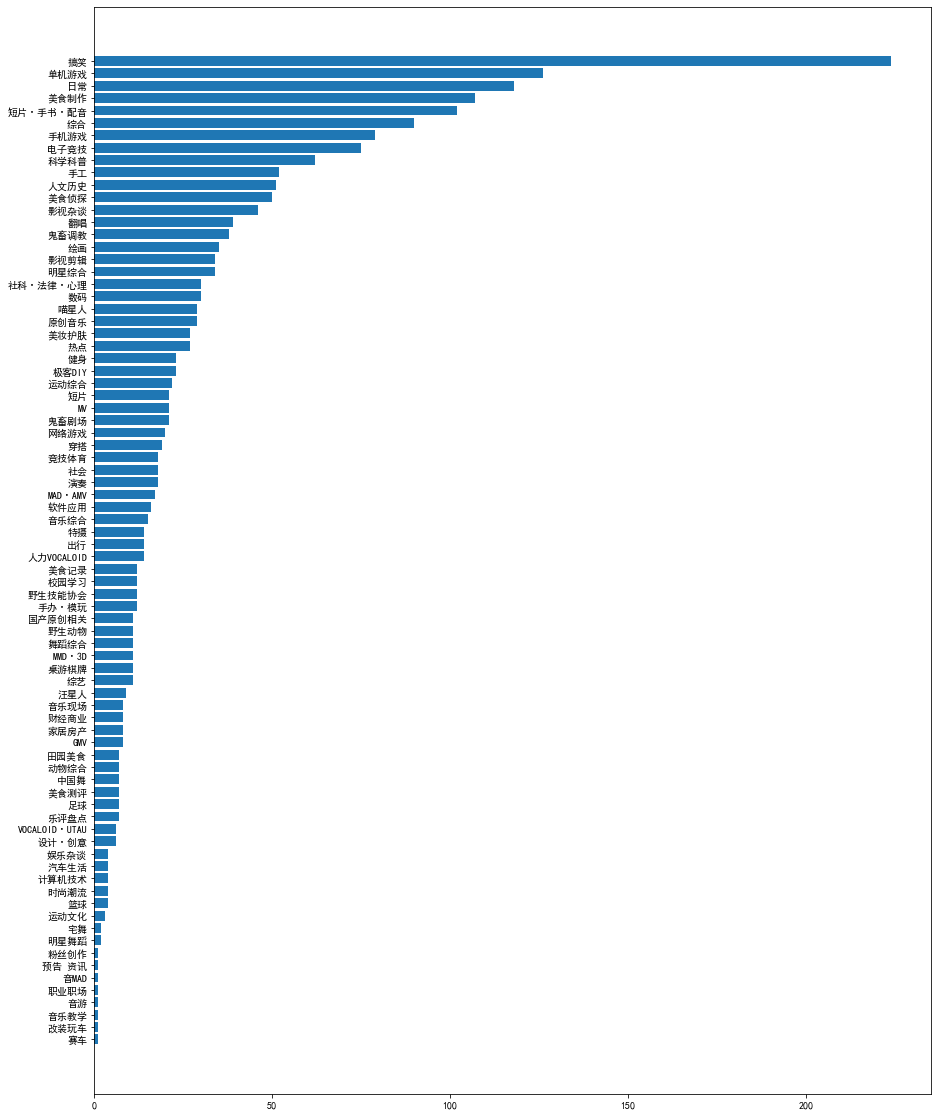


图4.3 各类别视频上榜数量

在播放量标准上，我们制作了图4.4的播放量比例饼图，从图中可以得知大部分视频的播放量都在一百到五百万之间，占64.53%；同时五百万到一千万之间的视频占27.25%，一千万以上的视频有6.02%，小于一百万的只有2.20%。可知要上每周必看，播放量至少要在百万播放以上。

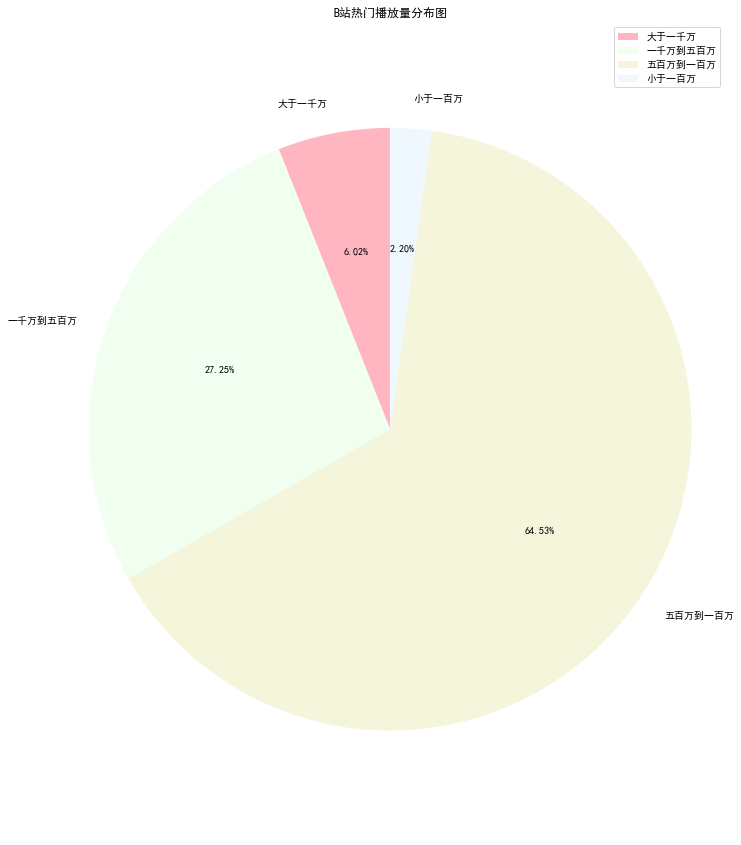


图4.4 各层次播放量占比

对于B站推荐机制来说，点赞、投币、收藏对于视频播放量影响较大，为此我们分析上榜视频的三者对于播放量的比例，即每一次播放用户进行这三项操作的比率。由图4.5可知，点赞比例最高，达到了0.0845，即每一百次播放中会有九个人对视频点赞；同时投币和收藏的比例为0.0408，0.0227。由于点赞不需要消耗用户硬币，也不会占用收藏夹，加上UP主视频对点赞需求高，用户点赞相对比例较高，而投币和收藏较少。

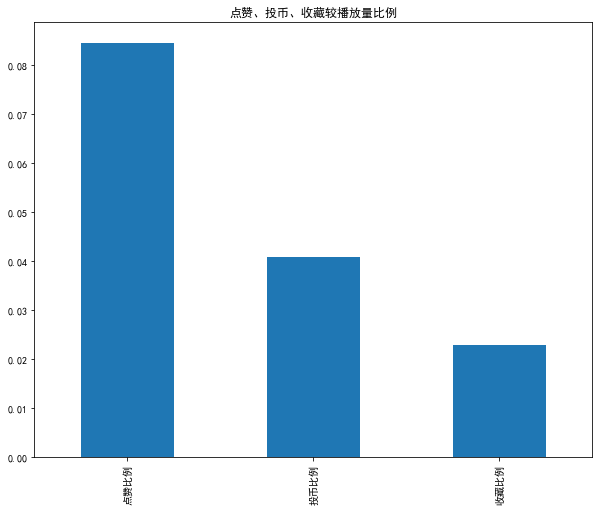


图4.5 点赞、投币、收藏较播放量比例

从图4.6中，我们对每周上榜视频数量进行统计去，可以看到五十周内视频数量呈现上升的趋势，推测与视频质量变好，优质UP主增多有关，同时在后十周内反复波动。根据数据统计得到，五十周内上榜视频数量平均数为41.84，方差为20.6144，其中峰值周数为156周，达到了51条视频，最低为161周仅有34条视频。由此说明榜单名额并不是固定数量的，而是随着平台与视频发展变化的，整体大趋势为不断增加优质视频。

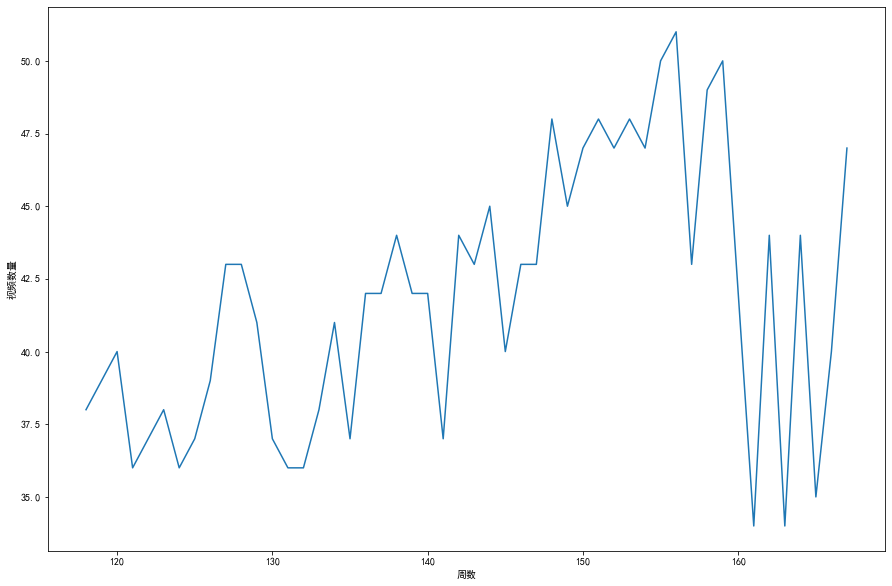


图4.6 上榜视频数量随周数变化

## （二）根据数据集拟合分析

从数据预处理过程中，我们大致了解了点赞、投币、收藏、播放量之间的关系，为了更好预测其之间变化的相关性，我们使用机器学习的线性回归对相关性进行预测模拟。

图4.7，图4.8，图4.9中，分别是播放量与点赞量相关性拟合、收藏量与播放量相关性拟合、收藏量与播放量相关性拟合。可以看出这三对数据相关性都比较强，离散程度不高，其中播放量和点赞量相关程度最高，拟合效果最好，而收藏量和播放量拟合效果相对较差，由此可见播放量和点赞呈现强正相关性，优质的视频点赞量多将促进播放量正向增加，同样随着播放量增加，新用户的观看将带来更多点赞，投币和收藏同理。

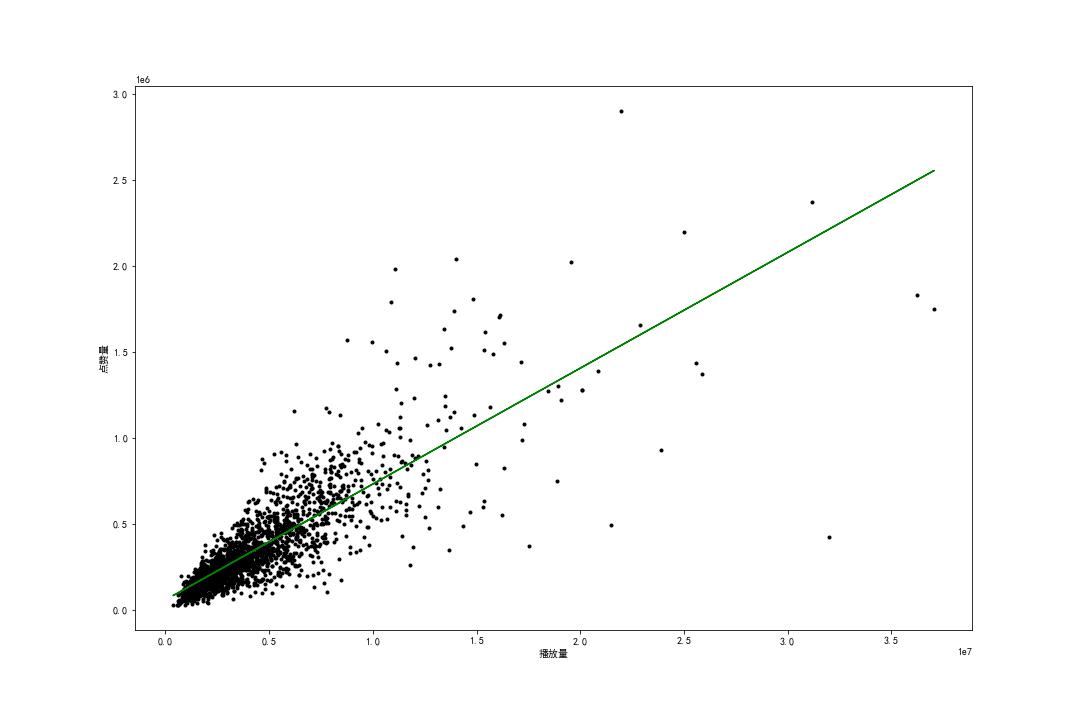


图4.7 播放量与点赞量相关性拟合

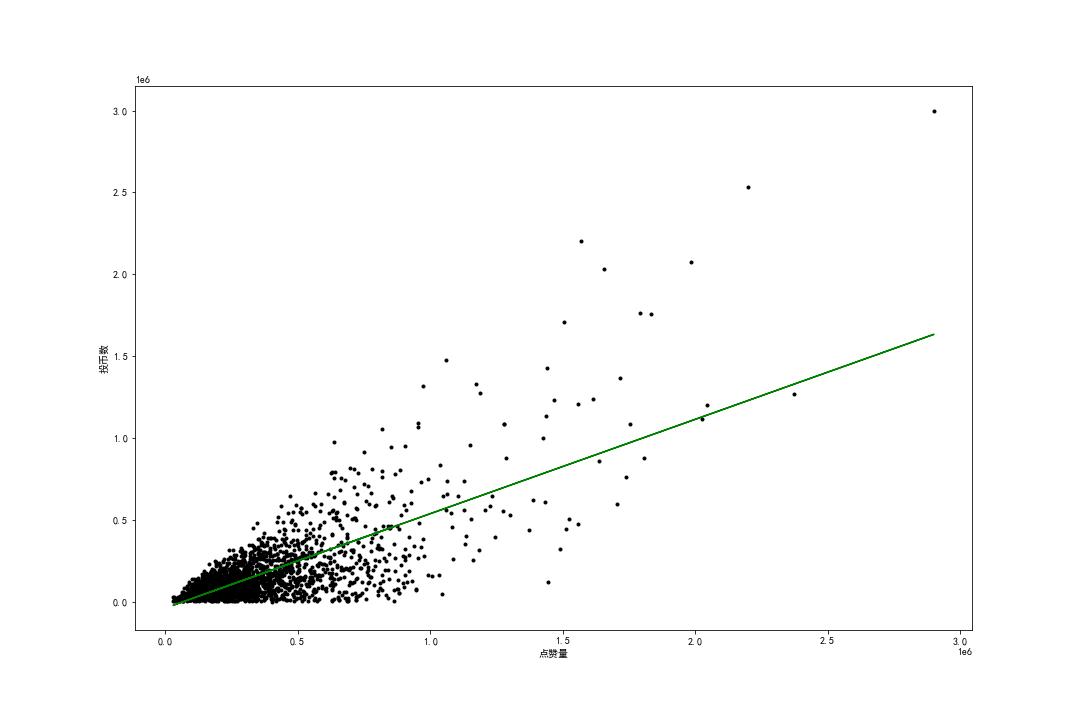


图4.8 点赞量与投币数相关性拟合

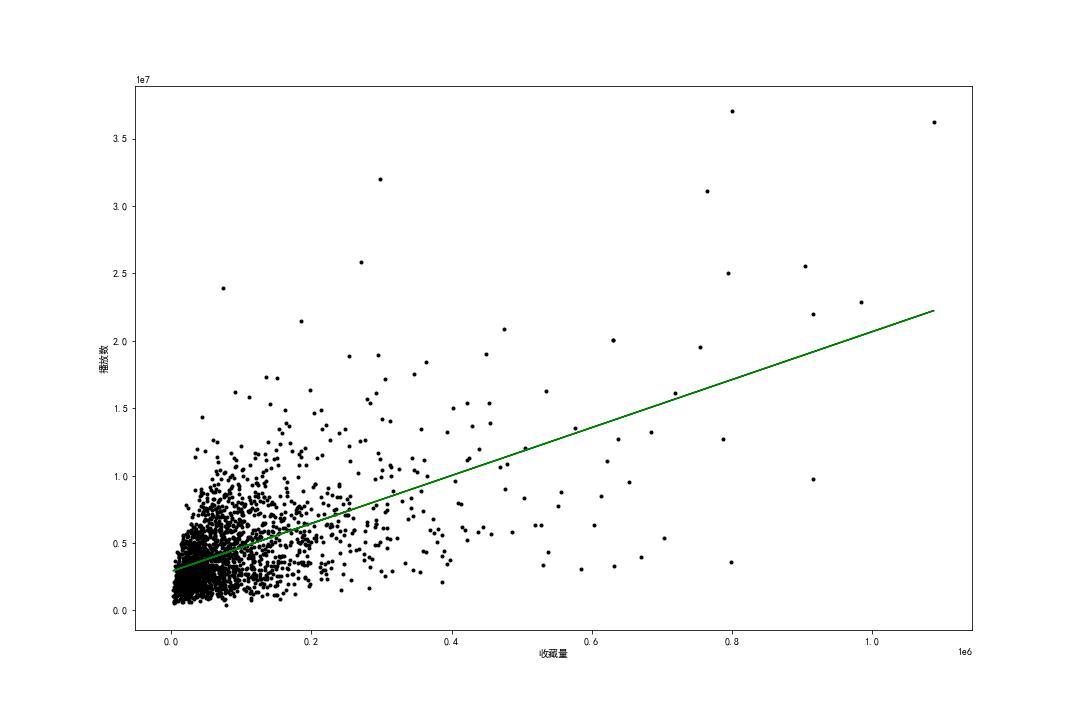


图4.9 收藏量与播放量相关性拟合

## （三）结果分析

从上述统计分析和机器学习拟合分析来说，总体可以得到以下结论：通过对数据的分析和可视化，可以得到词云，方便分析文本。可以得到视频类型以搞笑、单机游戏和日常为主要输出分布。可以看到从前五十周到现在每周必看视频大体上呈上升趋势。可以看到观看和点赞、点赞和硬币、收藏和播放中点赞和播放的相关性比较强，而且离散性不强。可以得到上榜视频标准为至少百万播放，同时我们得到排名前十的UP主和播放量前十的视频。

# 五、项目总结

本项目基于哔哩哔哩每周必看榜单，对B站用户喜爱倾向和上榜视频标准展开研究。首先，本项目根据每周必看数据特点设计了相应的网络爬虫来获取实验所需的数据；其次，本项目对爬取的数据进行了清洗和预处理，并用词云，散点图的形式对数据进行可视化展示；接着，本项目对数据进行进一步数据分析和可视化，包括统计分析和机器学习拟合分析，在数据集上获取了上榜标准以及播放量相关性预测拟合；最后，本项目对实验结果做出了分析，得到改善用户体验和B站推荐机制的结论。