重要代码介绍

爬虫涉及四个部分：番剧列表链接数据爬取，番剧信息数据爬取，番剧评分数据爬取，b站给的真实推荐结果爬取。这几个部分的爬取流程基本是一致的，即先用selenium库由url通过模拟浏览器的方式获取页面源码，然后再用BeautifulSoup对源码进行解析，通过BeautifulSoup的一些用法获取所需要的信息。用其中主要的main函数和main函数中调用的函数见图1、2。最终得到的数据内容见表1。

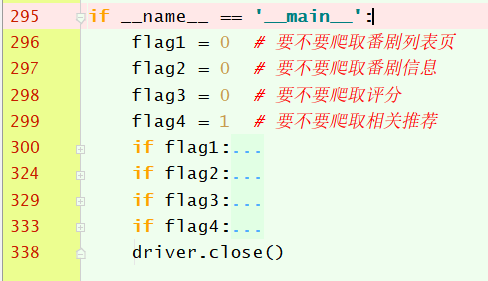


图 1 mian函数总体框架



图 2 mian中调用的函数

|  |  |
| --- | --- |
| **名称** | **数量** |
| **爬取番剧列表页数（/页）** | **20（每页有20个番剧链接）** |
| **爬取的番剧链接（/条）** | **480** |
| **去重后的番剧链接（/条）** | **400** |
| **爬虫爬取的番剧信息（/条）** | **380（无重复）** |
| **爬虫爬取的评分信息（/条）** | **77843** |
| **去重后的评分信息（/条）** | **77741** |
| **去重后的推荐数据信息（/条）** | **188** |

表 1数据内容

对数据进行去重、标准化、整合。代码见图3、4、5。

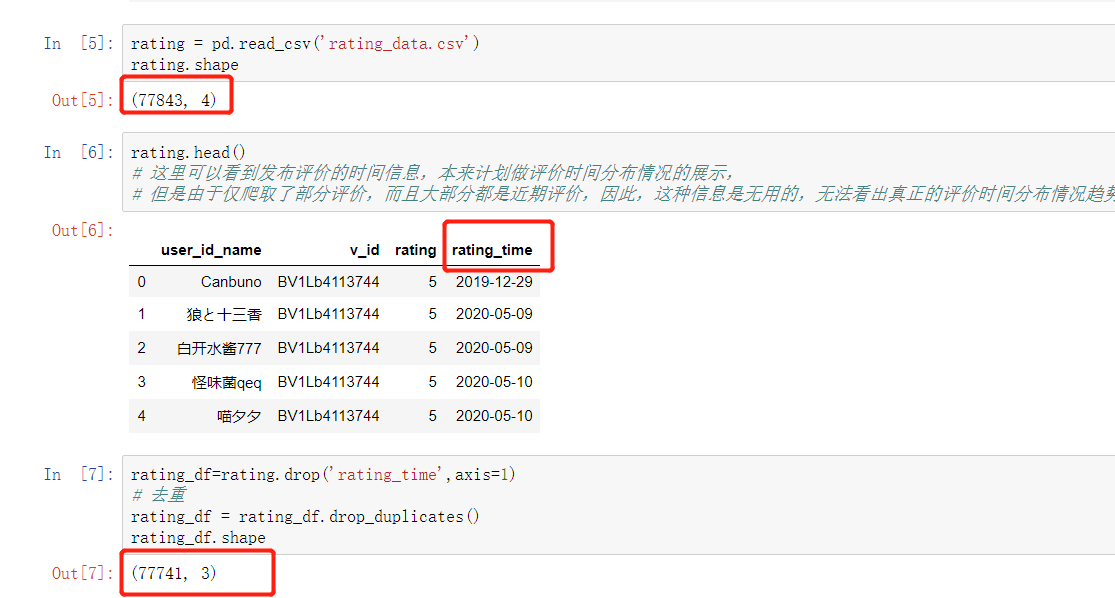


图 3 去重

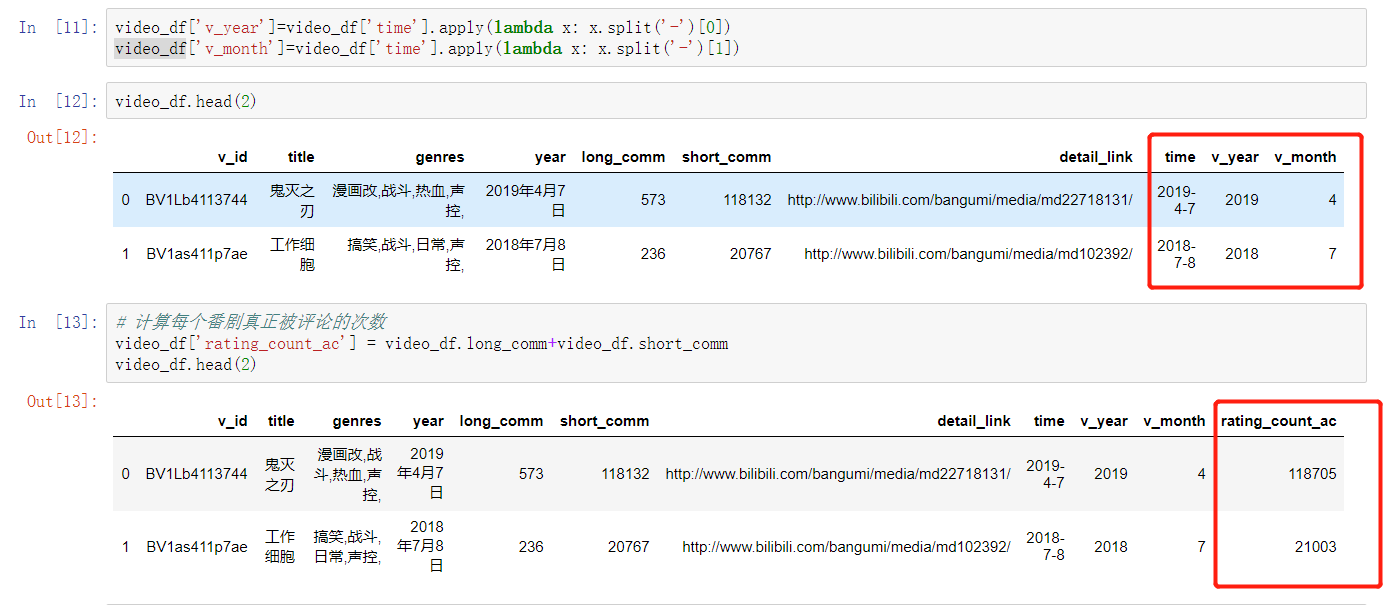


图 4 标准化



图 5 整合

设计推荐算法。主要涉及对新用户的基于内容的推荐，代码见图6；对老用户基于协同过滤的推荐（皮尔森、KNN、SVD），代码见图7，其中pivot是构造的评分矩阵（列名为用户id，行名为番剧id，值为评分），通过这个矩阵找到与目标番剧向量相似度高的向量，而这个向量的番剧id就是最终需要推荐的番剧id，其中这个相似度计算的方法有三种分布为皮尔森、KNN以及SVD。

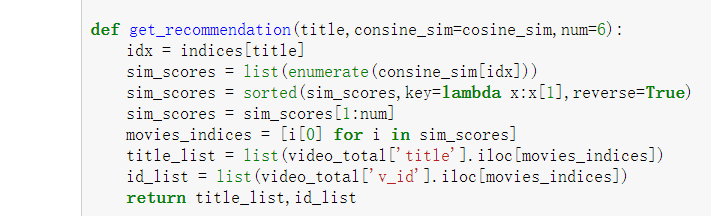
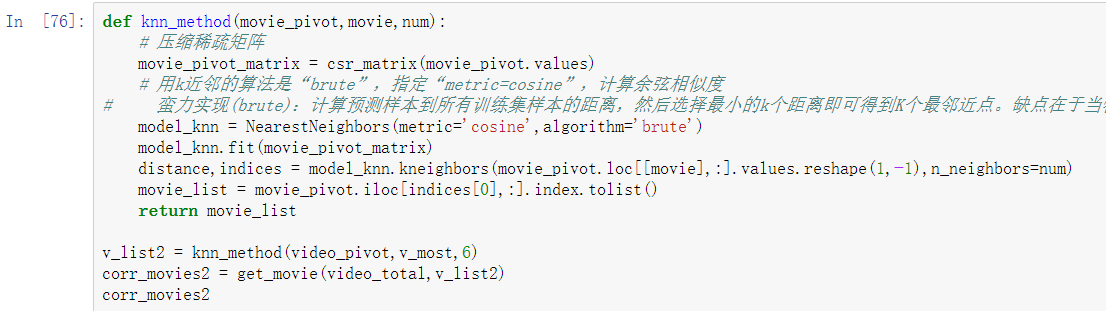


图 6 基于内容推荐





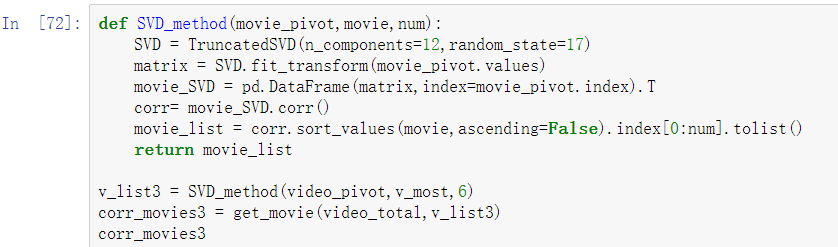


图 7 基于协同过滤的推荐

新增了结果合并方式是将基于协同过滤的推荐不同方法的结果合并并统计出现的次数，按照出现次数的高低取得前五个作为最终推荐。代码见图8。



图 8 合并方法

最后合并流程，对用户进行推荐，如果有用户就通过用户点评过的最高分数的番剧作为目标番剧，使用协同过滤的方法去找到这个番剧的相关番剧；如果无用户，即输入为空字符串，则会使用基于内容的推荐方法推荐热门的番剧。代码见图9。

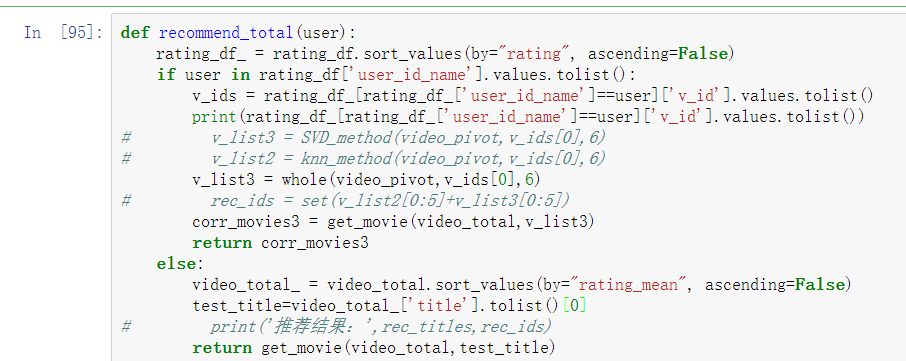


图 9 合并流程

**效果和结论**

通过基于年份和月份的可视化展示，见图10、11。可以看出大部分番剧都是在近些年播映，这也和这个视频平台的发展息息相关符合现实情况，另外对于月份的可以看出其实寒暑假期间番剧会多一些，这和番剧的受众也符合，番剧的受众大部分为学生，因此番剧的播映也与学生的假期相关。至于，评分的分布和番剧的数量分布很类似，符合常识，番剧多自然评论总数也会较多。

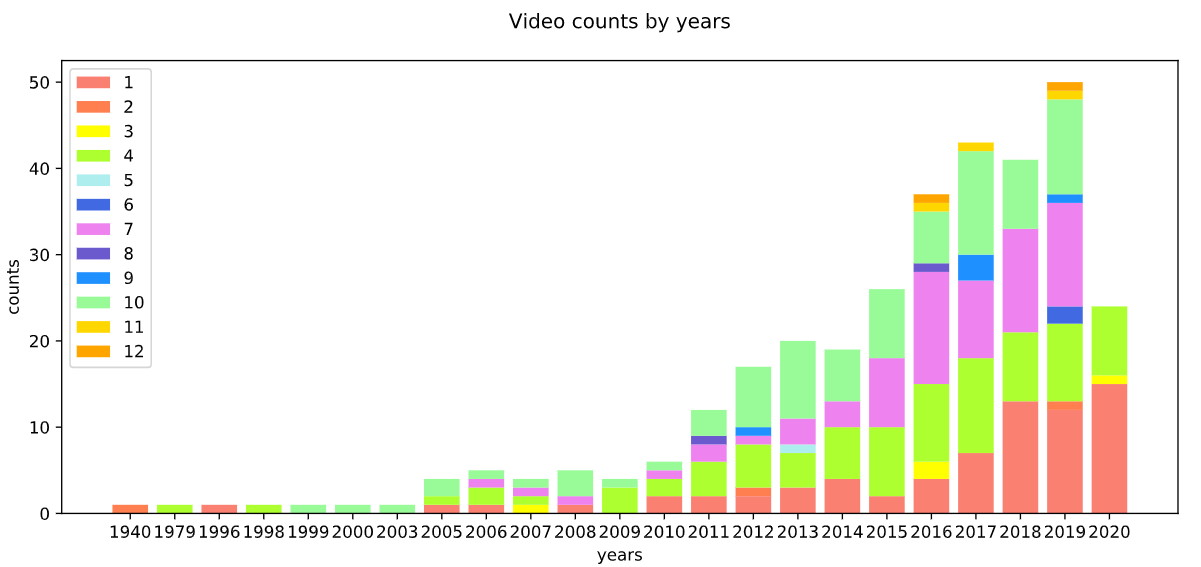


图 10每年和每月的番剧分布情况

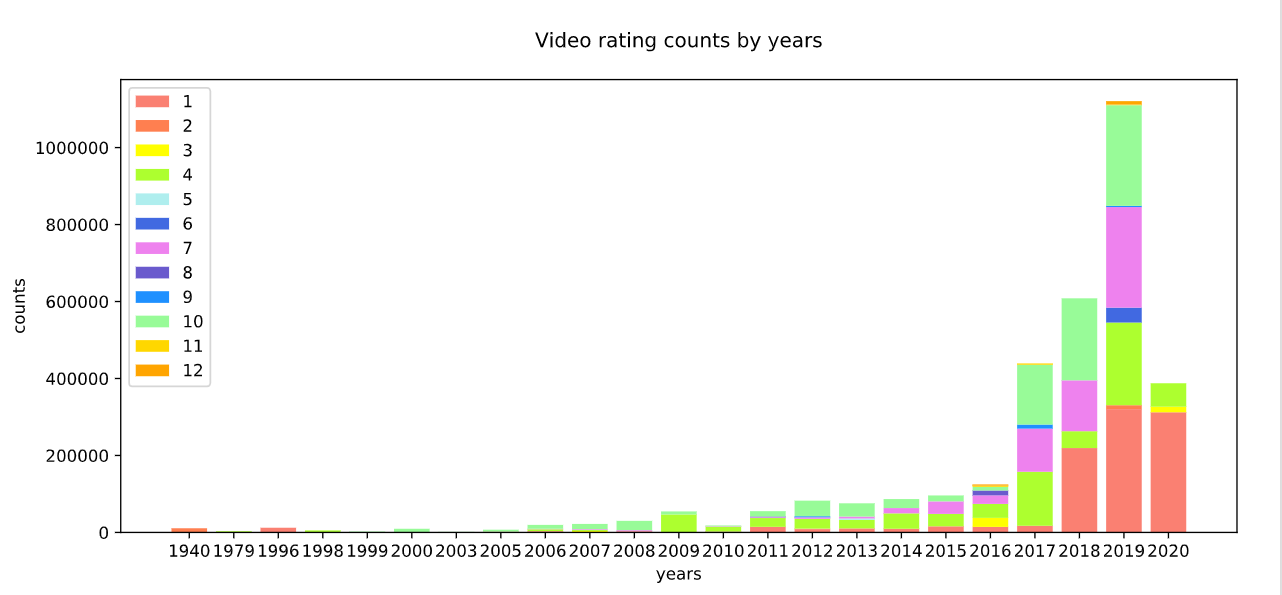


图 11 每年和每月的评分的番剧的评分分布情况

针对番剧推荐的结果，基于内容的推荐，通过结果的比对，发现预测结果较好，推荐番剧的类型几乎和目标番剧的类型是一致的，而真实的推荐，可能算法不同，看不出是通过什么方式进行的推荐，也看不出和目标番剧的相似性是偏向于什么。为此，如果想通过使用真实推荐作为标签作为基于内容的推荐算法的准确率计算，计算的结果是比较低的，也不应该把真实推荐作为基于内容的推荐的标签。见图12，相似的类型标签用红框圈出。



图 12 基于内容推荐与真实推荐对比

基于协同过滤的推荐，一共有三种方法分别为皮尔森、KNN以及SVD，结果见图13、14、15。可以发现皮尔森和KNN的结果是一样的，为此再次进行其他的测试，发现也是相同的，图16、17。会导致这样结果，猜测的原因是评分矩阵比较稀疏，用不同方法计算出来的相似性得到了类似的结果。相似的类型标签均用红框圈出。



图 13 基于协同过滤的推荐（皮尔森）



图 14 基于协同过滤的推荐（KNN）



图 15 基于协同过滤的推荐（SVD）



图 16 基于协同过滤的推荐（皮尔森）



图 17 基于协同过滤的推荐（KNN）

基于协同过滤的推荐，将协同过滤的方法结果合并，用计数的方式得到前五个推荐，图18。

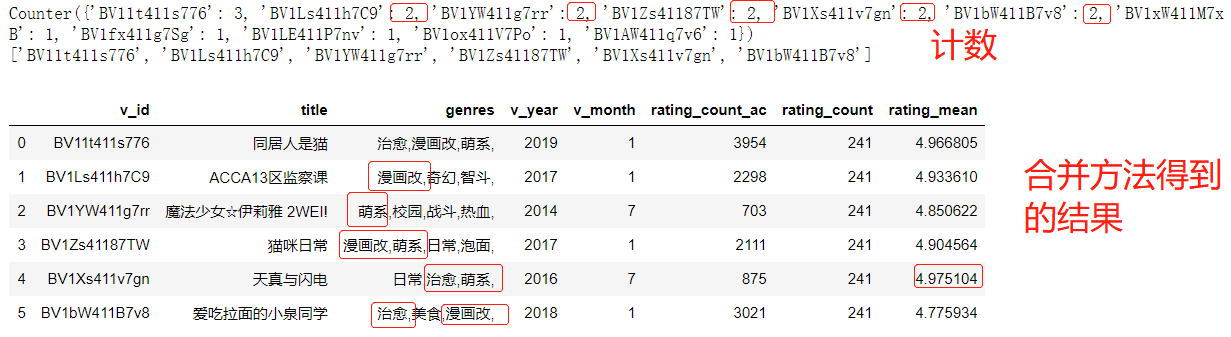


图 18 合并结果

将基于内容和基于协同过滤的方法合并，遇到老用户(图19)时，对用户评过分的番剧基于协同过滤的方法找相似的番剧，对于新用户(图20)，用基于内容的方式推荐热门番剧及其相似番剧。



图 19 合并流程的结果（老用户）



图 20 合并流程的结果（新用户）