爬虫与数据可视化作业报告

|  |  |
| --- | --- |
| 系别 | 软件工程 |
| 组长 | 杜上海 |
| 学号 | 201992009 |
| 班级 | 软测1901 |
| 成绩 |  |

2022 年 1 月 3 日

1. **项目背景及意义**

随着网络的迅速发展，万维网成为大量信息的载体，如何有效地提取并利用这些信息成为一个巨大的挑战。搜索引擎(SearchEngine)，例如传统的通用搜索引擎百度，搜狐和Google等，作为一个辅助人们检索信息的工具成为用户访问万维网的入口和指南。但是，这些通用性搜索引擎也存在着一定的局限性。因此，人们对爬虫的设计出现。我们小组通过尝试利用爬虫技术来实现资源的高效获取。

1. **项目创新点**

利用爬虫技术，结合可视化方法，实现资源的高效获取以及直观反映。

1. **项目设计**

本次大作业主要利用python爬虫对B站视频信息进行统计并分析，其中统计的信息储存在csv文件中，通过python自带的第三方库生成简单的柱状图，可以直观看出不同分区中的视频信息。

1. **项目实现**
2. 数据采集部分

本部分中主要使用了re，requests 2个模块的功能，其中用requests的get方法来获取网页的html文件，并通过分析网页的内容与结构，用BeautifulSoup进行整理





1. 可视化部分

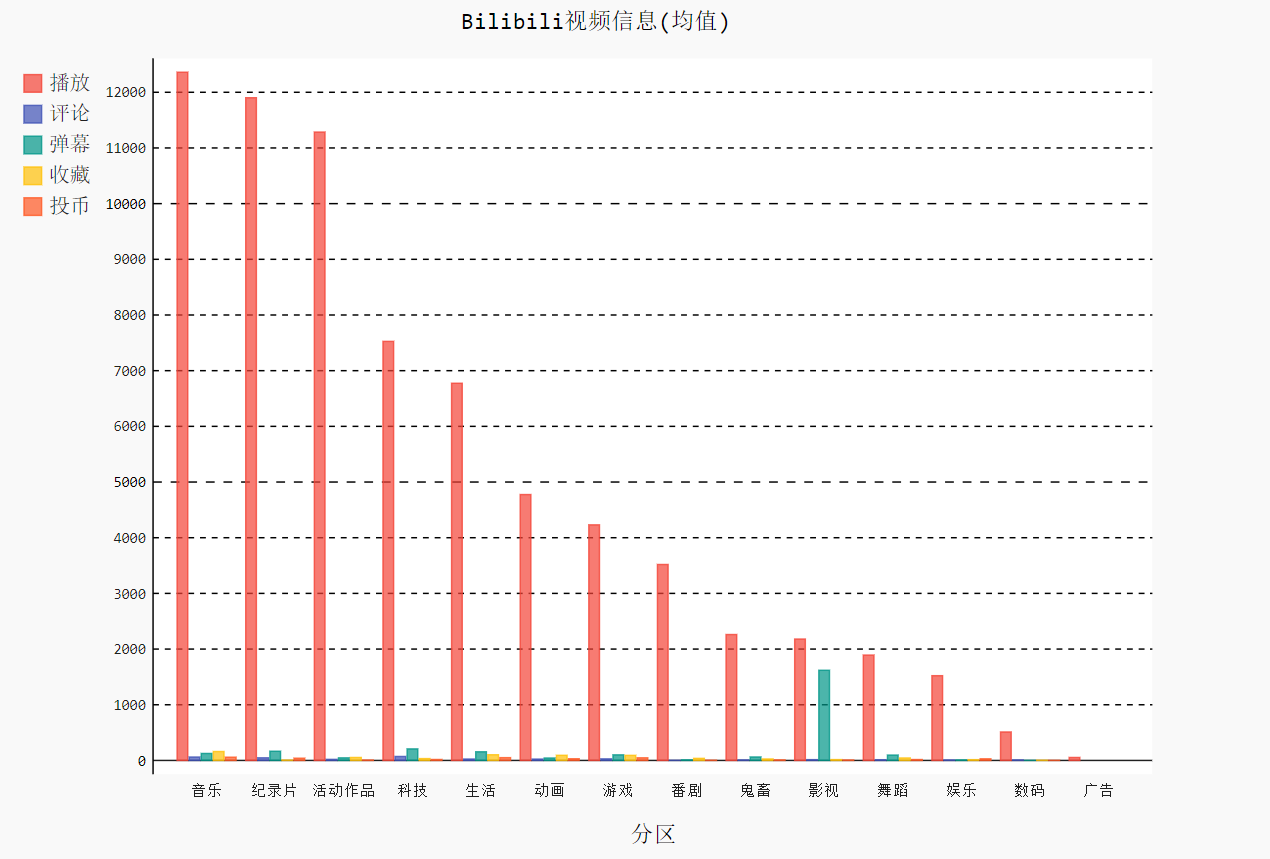
通过利用python自带的第三方库pandas，对csv中的信息简单的进行条形图的生成



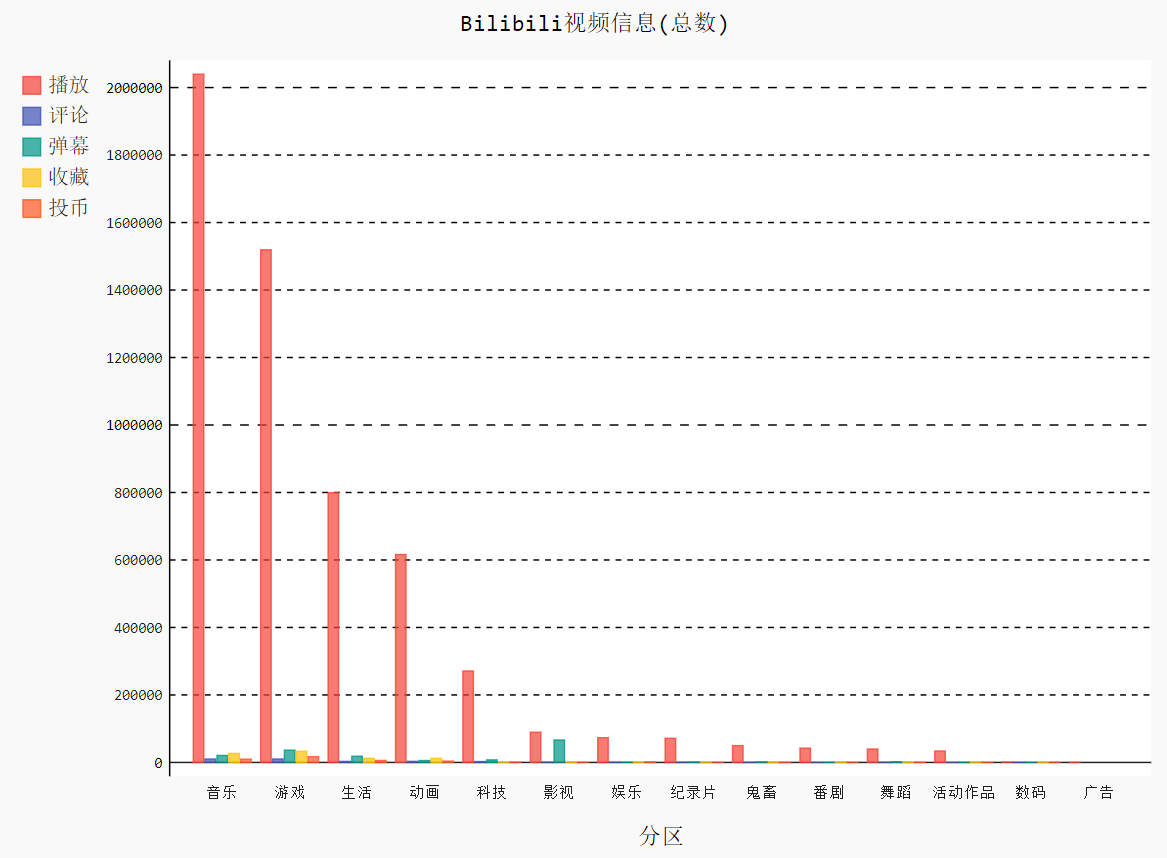
1. **项目测试与反馈**

通过对b站视频信息的实际爬取，计算和实际可视化过程，生成的测试结果如下所示：

1. 视频相关数据均值



1. 视频相关数据总数



1. **项目分析与总结**

此次项目利用了当今流行的爬虫技术来爬取需求信息，通过python中的re，requests 2个模块来获取所需网页信息，进而转存为本地csv文件后，通过可视化pandas模块来实现相关信息的图表展示，实现可视化目的。

通过这个项目，让我们所有小组成员更深入的去接触了爬虫技术，并且在实际运用中对这项技术有了深刻体会。同时对于python这门强大语言的使用也有了一定的了解，通过它实现了最初制定的项目需求，收获颇丰。

1. **项目感想**

此次项目很好的锻炼了我们学习技术、使用技术的能力的能力。从项目分析，项目详细设计，到项目编写，项目测试，每个人都功不可没。这次项目深刻让我们体会到了团队开发的乐趣和重要性。

开源软件基础这门课程带着我们领略了许许多多优秀的开源项目与软件。我们在这门课程中深深体会到了学习技术、分享技术的乐趣与重要性。在今后的工作学习中，我们会带着这份经验与兴趣，朝着技术与软件的方向不断迈进。

1. **小组分工**

杜上海---爬虫代码编写

陈英奇--数据分析

郭宣麟--可视化代码编写

潘宇豪--结构设计

倪学颖--测试