

**课程报告**

**课程名称：大数据导论**

**专业班级： CS1706**

**学 号： U201714762 U201714763**

**姓 名： 梁一飞 贺子杰**

**指导教师： 王蔚**

**报告日期： 2019.12.31**

**计算机科学与技术学院**

**目 录**

[1 报告题目：bilibili视频分析 3](#_Toc28887643)

[1.1 测试环境说明 3](#_Toc28887644)

[1.2 数据集特征说明 3](#_Toc28887645)

[1.3 测试应用说明 3](#_Toc28887646)

[1.4 研究目的及意义 4](#_Toc28887647)

[1.5 问题挑战 4](#_Toc28887648)

[1.6 测试结果 4](#_Toc28887649)

[1.7 角色分工 6](#_Toc28887650)

[1.8 心得体会与总结 6](#_Toc28887651)

1 报告题目：bilibili视频分析

* 1. 测试环境说明
* CPU：AMD Ryzen 5 3500U 4核 8逻辑处理器
* 内存：8G DDR4
* 虚拟化：docker 1 namenode + 2 datanode
* 框架：Hadoop
  1. 数据集特征说明

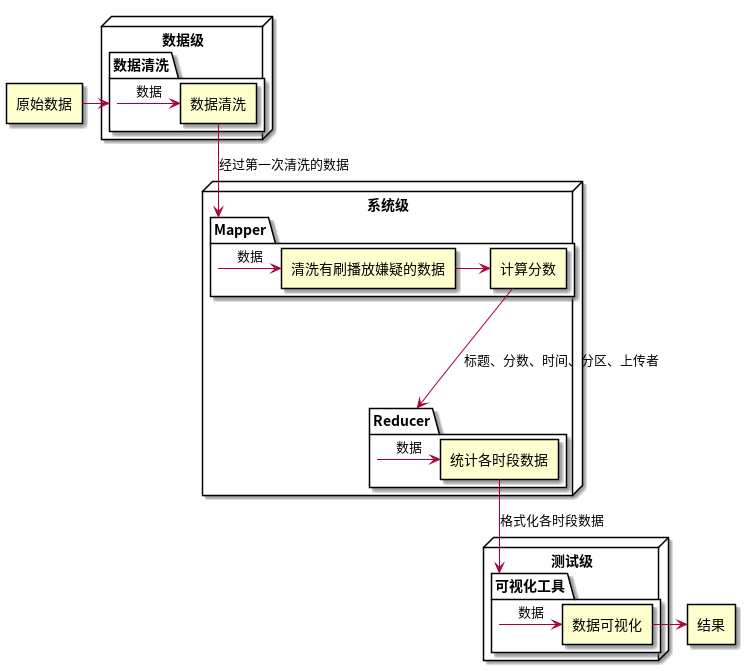
我们的实验主要是对bilibili 的视频的受欢迎程度进行一定的分析，所以想要获得bilibili的视频的相关信息，我们采用python网络爬虫的方法获得了bilibili的视频的数据。使用python网络爬虫工具网址(<https://pypi.org/project/bilispider>)

获得数据截图如下图所示：



**图1-1 bilibili视频信息数据截图**

* 1. 测试应用说明



**图1-2 测试应用说明**

* 1. 研究目的及意义

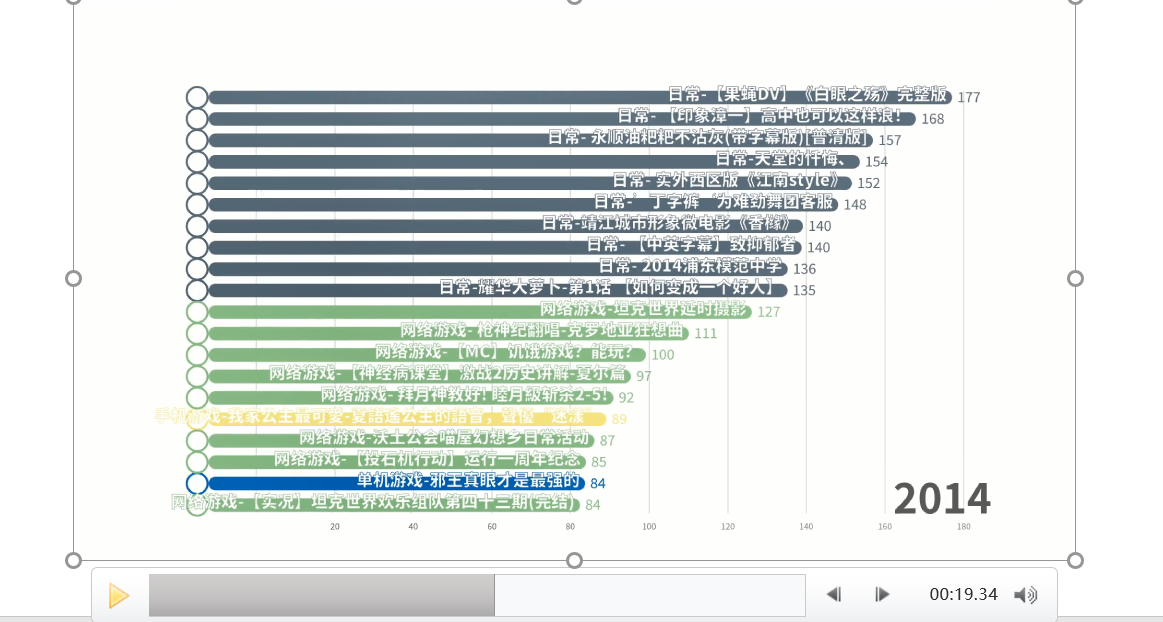
随着bilibili在年轻一代中的走红，如何有效的研究和分析用户访问行为，对bilibili的维护和用户推荐设计有着重要的意义。有了经数据集和系统级同学清洗后的数据，我们数据级的工作就是将这些数据可视化，并通过可视化的结果作相应的分析。

* 1. 问题挑战

如何将数据可视化，将数据做什么样的可视化？这是一大问题。经bilibili上大量数据可视化的视频启发，我们决定做一个随时间增长而变化的视频，这个用到了github上的一个工程Jannchie/Historical-ranking-data-visualization.js，通过修改里面的config.json文件，将我们爬取的数据以视频形式呈现处理。同时为了进一步分析，也利用excel绘制了变化曲线图。

* 1. 测试结果

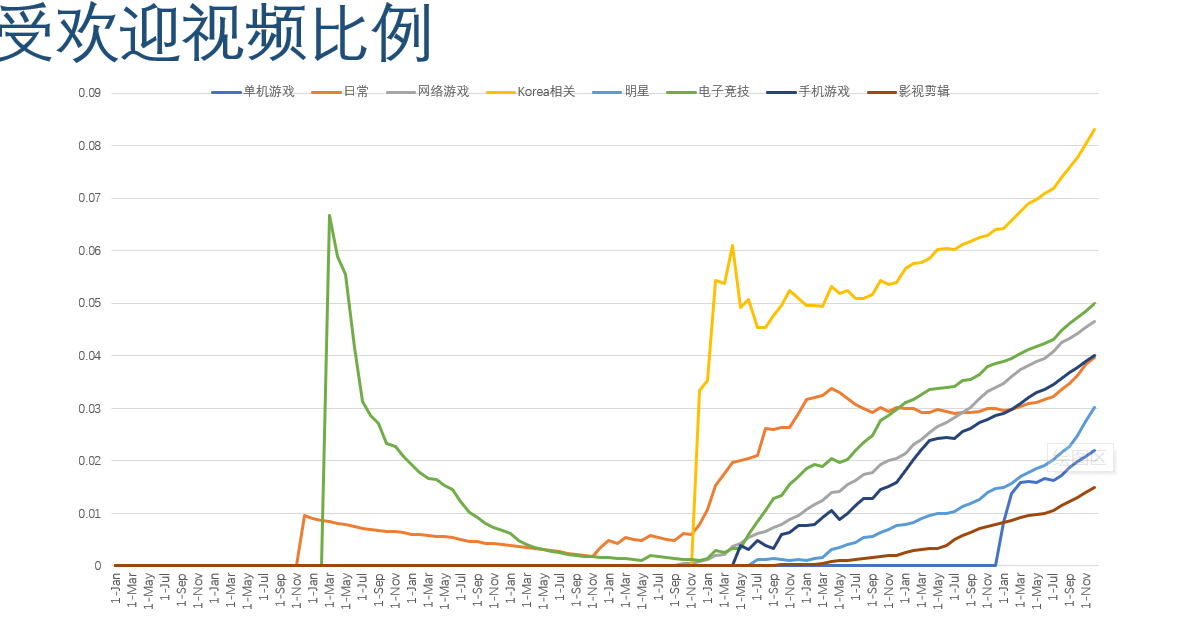
生成的视频如图所示：



**图1-3 生成视频**

通过对此分析，我们发现，在12年之前，明星及单机游戏板块位列前茅，12年之后，生活区板块占据了前几名，17年网游板块位列前茅，18年手游板块位列前茅，19年生活区板块重回前列。基本符合近几年潮流趋势，如17年时的电竞热，18年的手游热，同时12年之后，开始有许多人在网络上上传一些有意思的生活视频，up主职业也随之诞生。

生成的表格如图所示：



**图1-4 生成表格**

从表中数据可以看出，15年4月左右，手机游戏板块热度开始持续增长，同时电子经济和网络游戏板块也随之兴起，而Korea相关板块的持续增长也与近年来饭圈这一集体的扩大息息相关，图中有些板块在某一时段突然爆发主要是因为在这一时间bilibili才开设了相关分区。电子竞技板块在11年呈现一个持续性的下滑可能与当时国内对电子游戏视若毒品 的看法有关，而在近年出现增长也与电竞职业选手广为大众所接受有关。

* 1. 角色分工

数据集：西月栋，张佰鑫

系统级：张敬天，郭适，尹洪波

测试级：梁一飞，贺子杰

* 1. 心得体会与总结

本次测试级的工作让我们学到了很多，小组组员大致了解并实践了大数据信息的几种展示方式，并对所得出的可视化数据进行了简单分析，通过数据再结合国内的发展趋势可以看出很多东西，同时因为我们都是bilibili的用户，在分析数据的过程中也能想到自己关注的up主的变化，从一开始关注的游戏区up主，到现在关注的越来越多的生活区up主，也基本符合数据所展现出来的趋势。

由于时间有限，做的分析比较粗糙，但是收获也很大。在团队合作过程中，我们互相帮助，互相进步，知识在交流中增长了许多。