

**课程报告**

**课程名称：大数据导论**

**专业班级：**

**学 号：**

**姓 名：**

**指导教师： 王蔚**

**报告日期： 2021/12/5**

**计算机科学与技术学院**

**目 录**

[1 报告题目 B站视频之大数据处理（数据级） 3](#_Toc7536117)

[1.1 团队介绍与本人角色分工 3](#_Toc7536118)

[1.2 心得体会与总结 3](#_Toc7536119)

[1.3 汇报PPT 3](#_Toc7536120)

1 报告题目：B站视频之大数据处理（数据级）

* 1. 团队介绍与本人角色分工

我们六个人因缘相识相知，共同完成这个难度高但是蛮有乐趣的项目。

团队里面有两人负责系统级，两人负责测试分析级，其中我和L同学负责数据级的工作，L同学负责编写爬虫在B站上爬取视频数据，我负责编写程序进行数据清洗，具体工作内容如下：

① 数据爬取：L同学对于bilibili专门编写了一个爬虫类，使用爬虫借助scrapy框架来进行数据爬取：首先通过B站各排行榜的页面获取数量足够大的视频的链接，利用lxml包帮助解析网页HTML，其次，对每个视频，利用爬虫程序获取所需数据，综合利用requests包访问api接口，lxml解析HTML，以及F12开发者工具来获取数据，最后处理获得的数据，将其输出为一个JSON格式的文件data.json。下面是部分程序和运行控制台输出的截图：



② 数据清洗：我使用Java搭配maven框架，编写了一些工具类：数据导入类DataImport、数据处理类DataSolve、数据导出类DataOutput。开始没有确定数据存储和传输形式的时候，是写了对于.csv、.xlsx(.xls)、.json三种不同形式数据的导入和输出方法，之后经过商议使用JSON文件格式，于是就具体根据原始数据的格式与特点对总计403595个视频的数据集进行以下清洗：

1. 改写数据标题；

2. 类型转换：“万”字数据串变成确切数字；

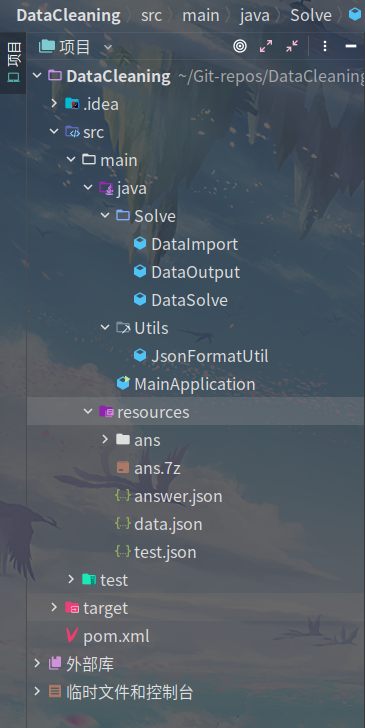
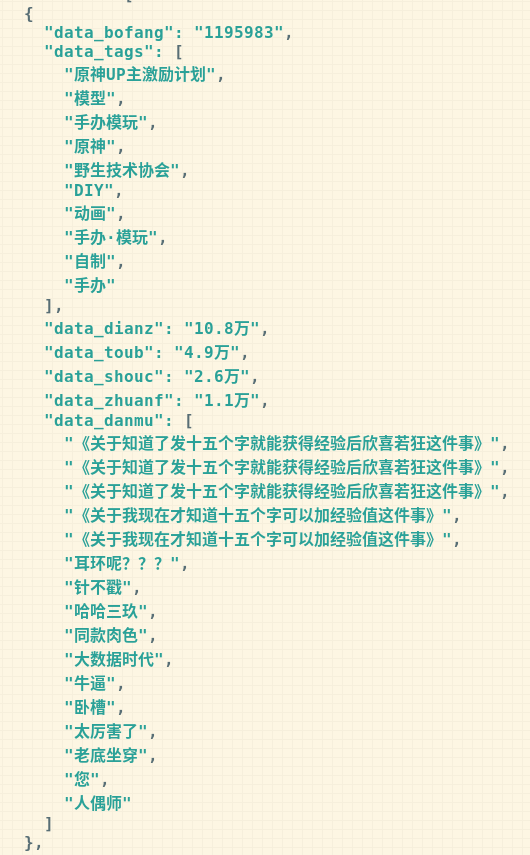
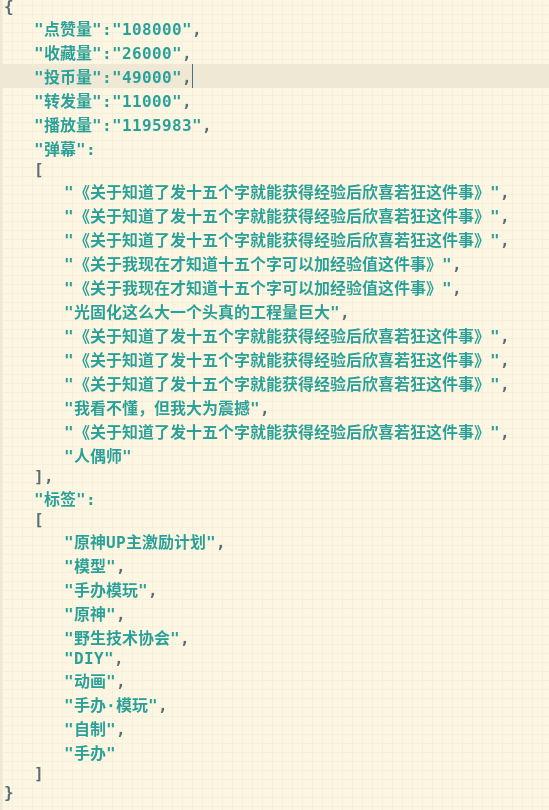
3. 处理缺省数据；

4. 语法错误解决：处理不正确的标点符号，删除字符串开头结尾的空格，模糊匹配算法修正拼写错误。

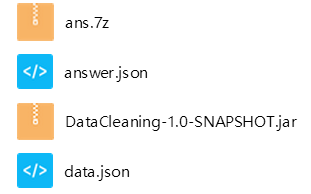
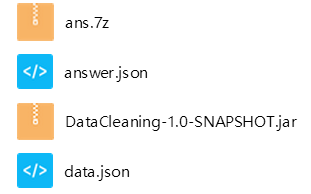
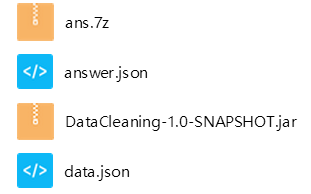
主要过程是将导入的数据格式分割为不同JsonObject对象，先对此对象处理，其次转换成String型处理并合并，最后按照预定格式分割字符串，获得不同的处理好的JsonObject对象并输出。

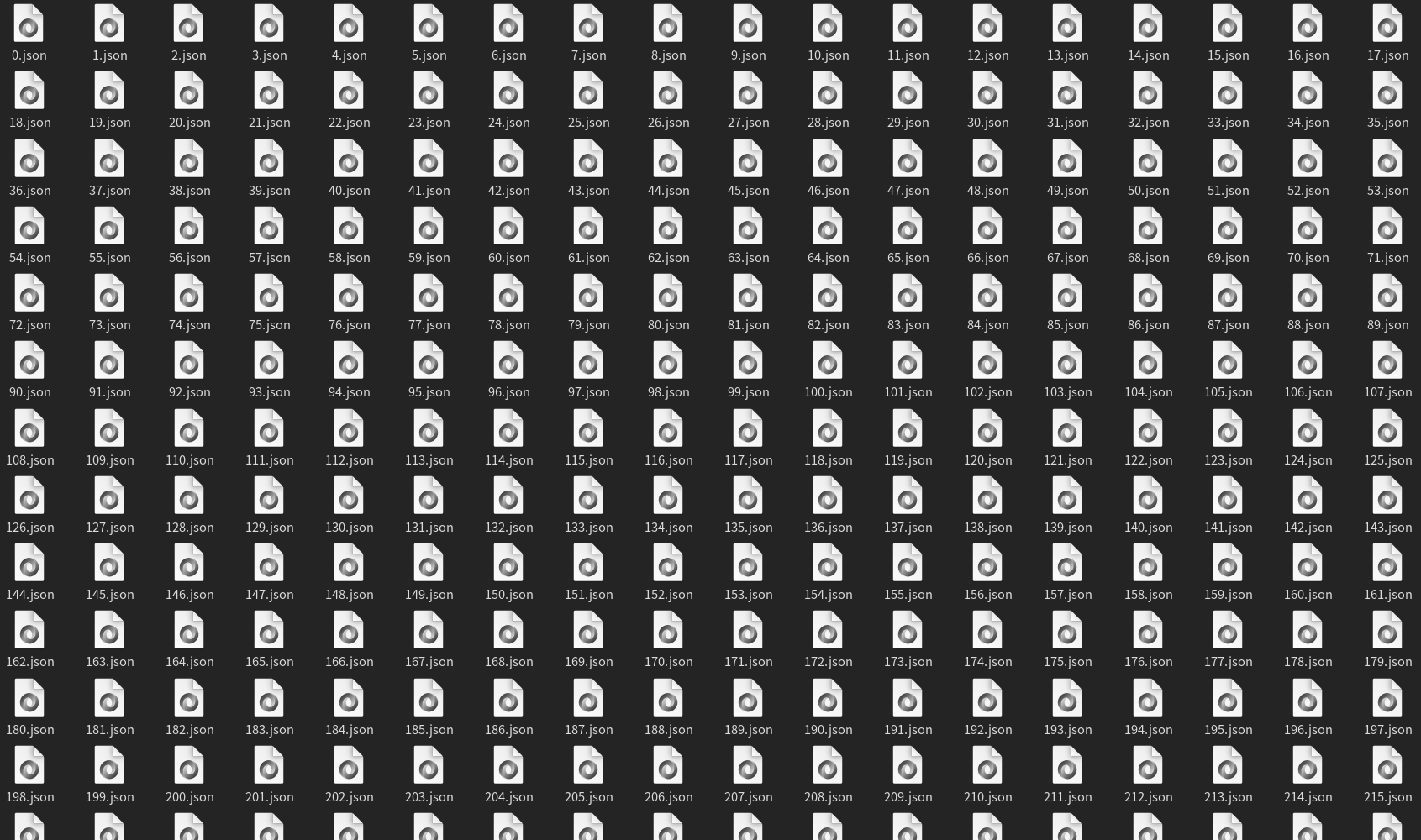
数据清洗前为约11.2GB，清洗之后有两种输出形式，分别为一整个文件输出（answer.json），文件大小为约9.6GB，以及文件夹输出（ans），其中文件夹内包含403421个文件（其余视频由于大篇幅数据缺失或者字符串格式错误或者乱码过多而丢弃），每个文件为一个视频的相关信息，文件夹大小为约8.0GB。

下图自左到右分别为代码框架、处理前数据、处理后数据截图：

下面是清洗前后的数据集，其中data.json是爬取的数据，answer.json和ans.7z是数据清洗后的数据集，以及ans包下部分数据文件截图。





* 1. 心得体会与总结

**路昊东心得体会与总结：**

总体来讲，本次大数据导论大作业的确使我收获了很多。

首先是在解决问题中实践学习。过程中有一些问题和挑战：首先难点是如何读入数据，通过阅读一些官方文档，写出了读入.json、.csv、.xlsx、.xls四种格式的数据的工具类，由于四种文件格式不同，读取方式也不尽相同，通过本地测试逐步调通代码，最后协商选取JSON文件格式作为数据集存储格式。其次的难点在于如何分析清理数据，通过观察数据，结合网上查找到的一些清理数据的方法，写了数据处理的工具类。最后由于要求是读出.json文件，而处理后的是数据为String格式，所以自定义撰写了将字符串分割并转化为适合格式无乱码的的输出工具类。

细节上，在处理不同格式的数据的时候，通过看一些官方的文档以及stackoverflow上的一些帖子，我对于不同的数据存储格式的本质都有了比较深刻的了解，对各自的优缺点和处理方式也有了一些研究。对于这次的数据处理，由于格式太大，所以处理用时非常漫长，即使是开了多线程。下一步可以换成导入mysql或者mogoDB等数据库进行更快速而方便的处理，同时借助一些可视化工具能够更直观的检查数据处理的正误。同时在数据处理的过程中，我认真学习了不同的数据清洗方法，虽然只是使用了其中的一小部分方法，但无论是学习还是实践的过程都让我受益良多。

其次，本次任务切实锻炼了我独立思考的能力，对于一些清洗的方法，网上只是有一些思路，真正实现的源码还是由自己思考后编写，同时对于不同策略的选取也是由自己决定，最后由此课外拓展甚至写了有关数据格式转换和加速的脚本，也算是收获颇丰。

除此之外，本次任务锻炼了我团队合作的能力，无论是与系统级的同学沟通，还是与同小组队友的沟通，都极大的锻炼了我的沟通能力、协调能力，数据集的选取、传输与存储数据格式的商讨，也都让我逐渐学会如何高效的在开发过程中与人沟通，最后做出优等的解决方案。

当然，本次作业也让我大开眼界，仿佛更加靠近了看似虚幻的大数据，对于课堂上讲解的知识有了具体而真切的应用。B站视频网站发展迅速，每个视频一经发出，就面临着各种各样数据的冲击。标签体现了up主以及观众对视频类型的偏好；弹幕则可以在一定程度上反映出网络用语的流行，以及不同观众对该视频的看法；点赞数等则反映了视频在观众中的受欢迎程度。b站最初是二次元爱好者聚集地，随着b站的发展，b站的视频逐渐变的多元化并吸引着越来越多元化的用户，通过对热门视频的标签数量的统计来比较直观的得到用户更喜欢带有哪些标签的视频和带有哪些标签的数据更能受大家喜欢。除此之外，我们还可以对这些数据分析出B站对外运营的模式，向我们展示了B站那些视频具有很好的潜力以及哪些视频深受用户喜欢，不仅对UP主的视频制作提供方向，更好的把握用户的喜好，也让我们用户更好的了解自己的喜好类型，对那些致力于想成为UP主的用户提供了数据支持。

**L心得体会与总结：**

在本次大数据作业中，我熟悉了python的语法，掌握了一个新的爬虫框架，在数据存储格式、网页解析、开发者工具调试等方面更加拓展了自己的能力。

当然在团队合作中，我逐渐更加学会如何在技术层面和大家交流沟通，和同组组员搭配合作的非常愉快。

当然最终也是了解到了有趣的大数据知识，丰富了个人的能力，立志在大数据层面出自己的一份力！

* 1. 汇报PPT （每页打印4张PPT）