**《大数据导论》**

**课程报告**

学 院： 计算机学院

专业班级： 计算机科学与技术

年 级： 2020级

姓 名（所有人）： 艾孜海尔·米日萨力，喻鹏飞，陈鹏宇

学 号（所有人）：

完成时间： 年 月 日

成 绩：

指导教师：

## 1.1搭建环境及配置介绍

操作系统：CentOS 7.6

语言环境：Python3.7，Java8

相关软件：Hadoop 2.8.5、Spark2.4.7

1.2 Hadoop架构搭建

创建用户

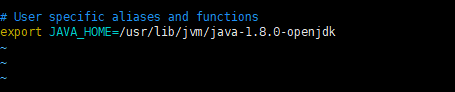


切换到Hadoo

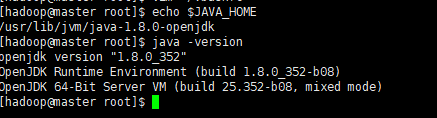


配置java环境

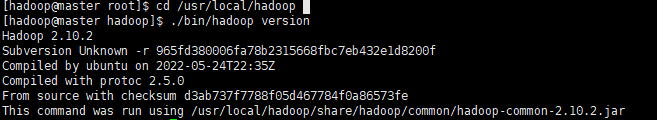
修改branch



验证



下载安装好hadoop

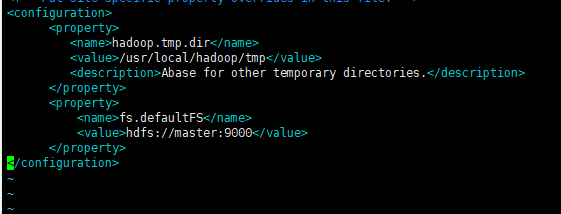


添加环境变量

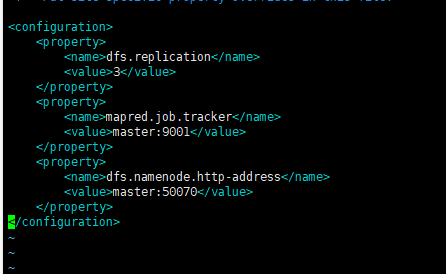


配置Hadoop环境下的slave文件

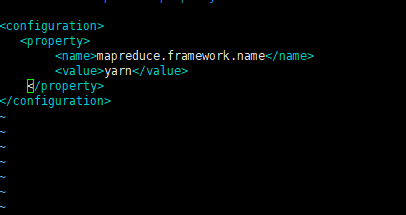
core-site.xml文件



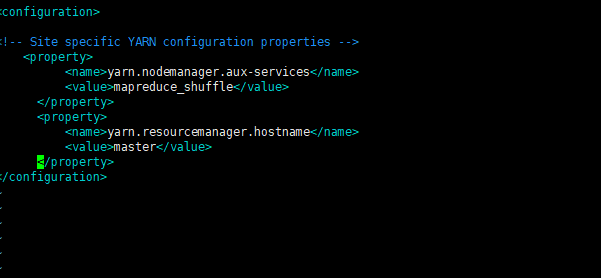
hdfs-site-xml



mapred-site.xml



yarn-site.xml



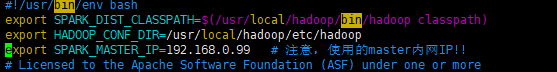
1.3

SPARK环境搭建

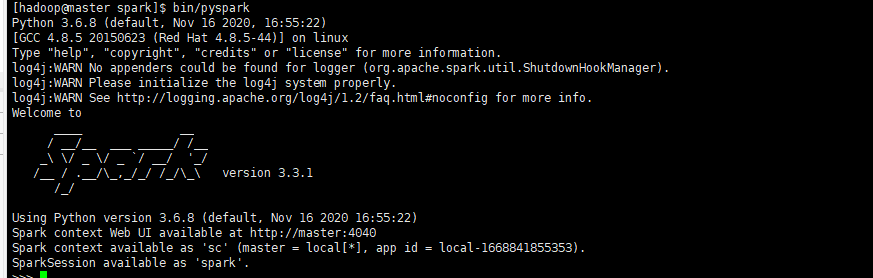
下载安装后配置环境变量



配置spark-env.sh文件



测试spark



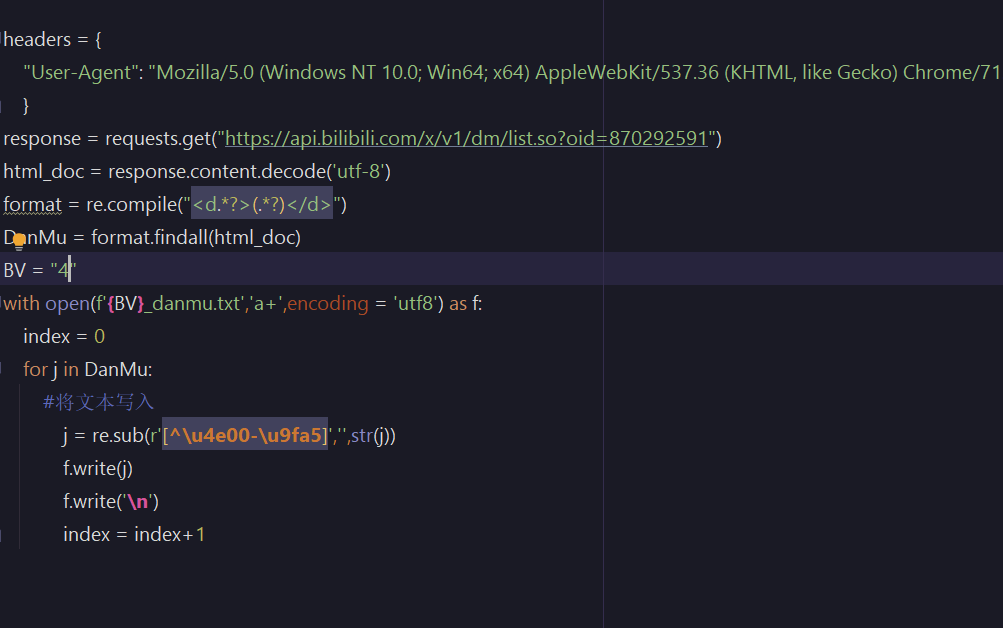
2.1 爬虫获取数据集

我们这次选择获取数据来源为B站弹幕评论和知乎的回答，设计三个爬虫程序分别获取B站弹幕,B站评论,知乎回答

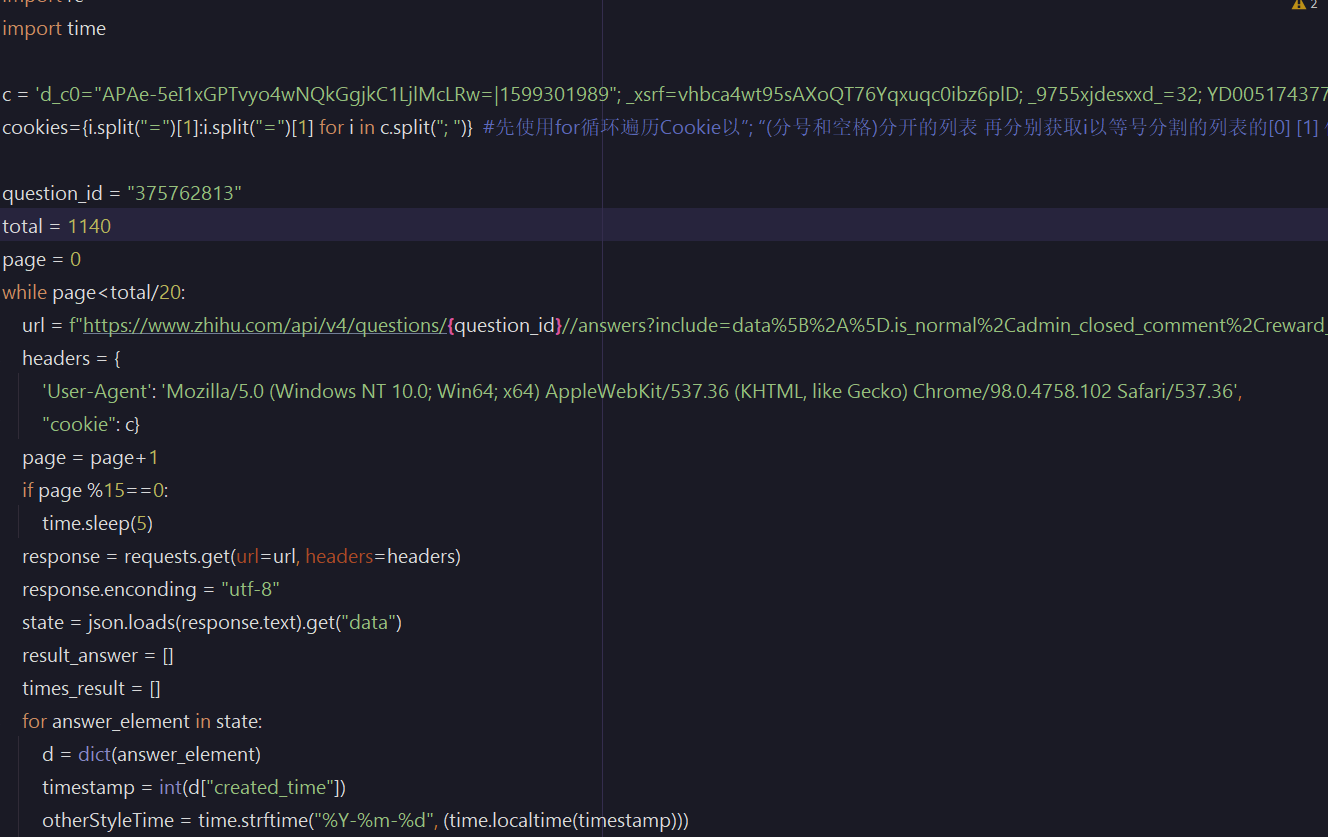
爬取B站评论代码部分



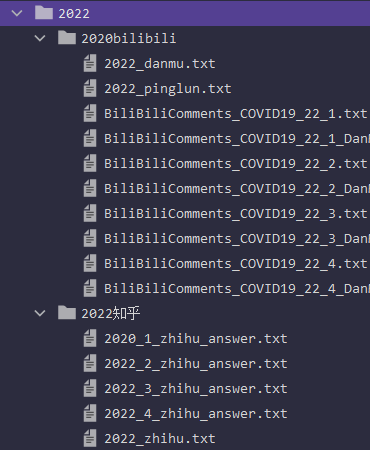
爬取B站弹幕部分



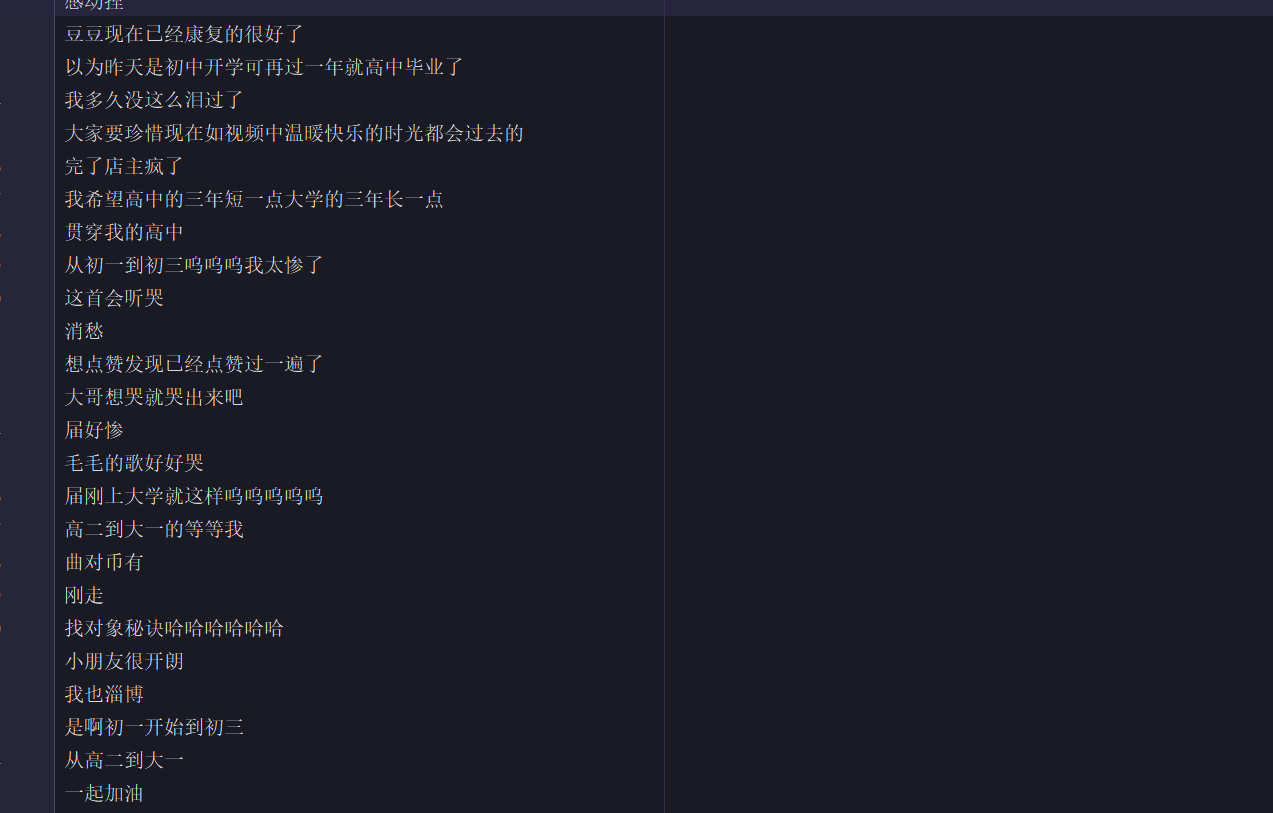
爬取知乎回答代码部分

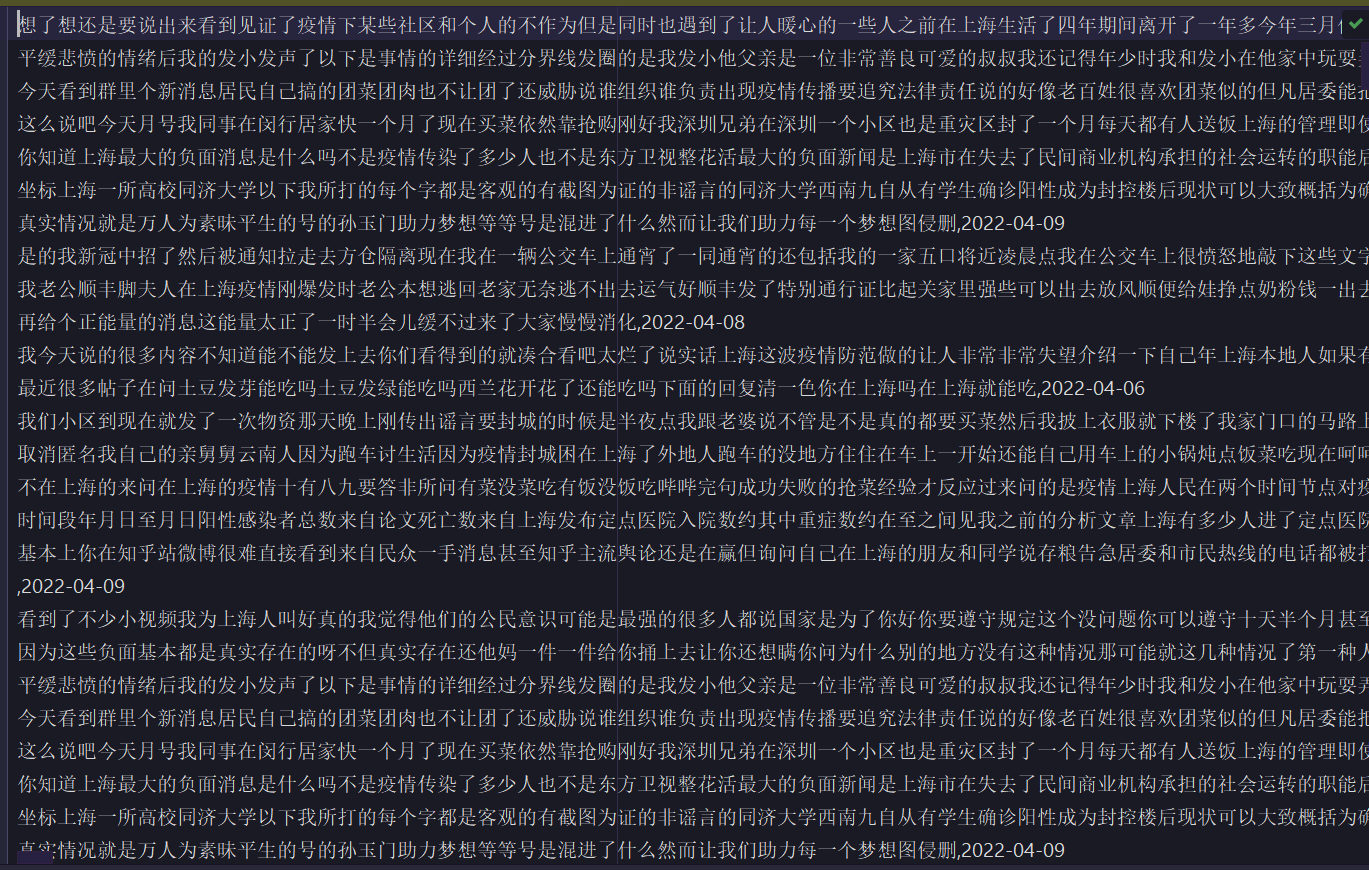


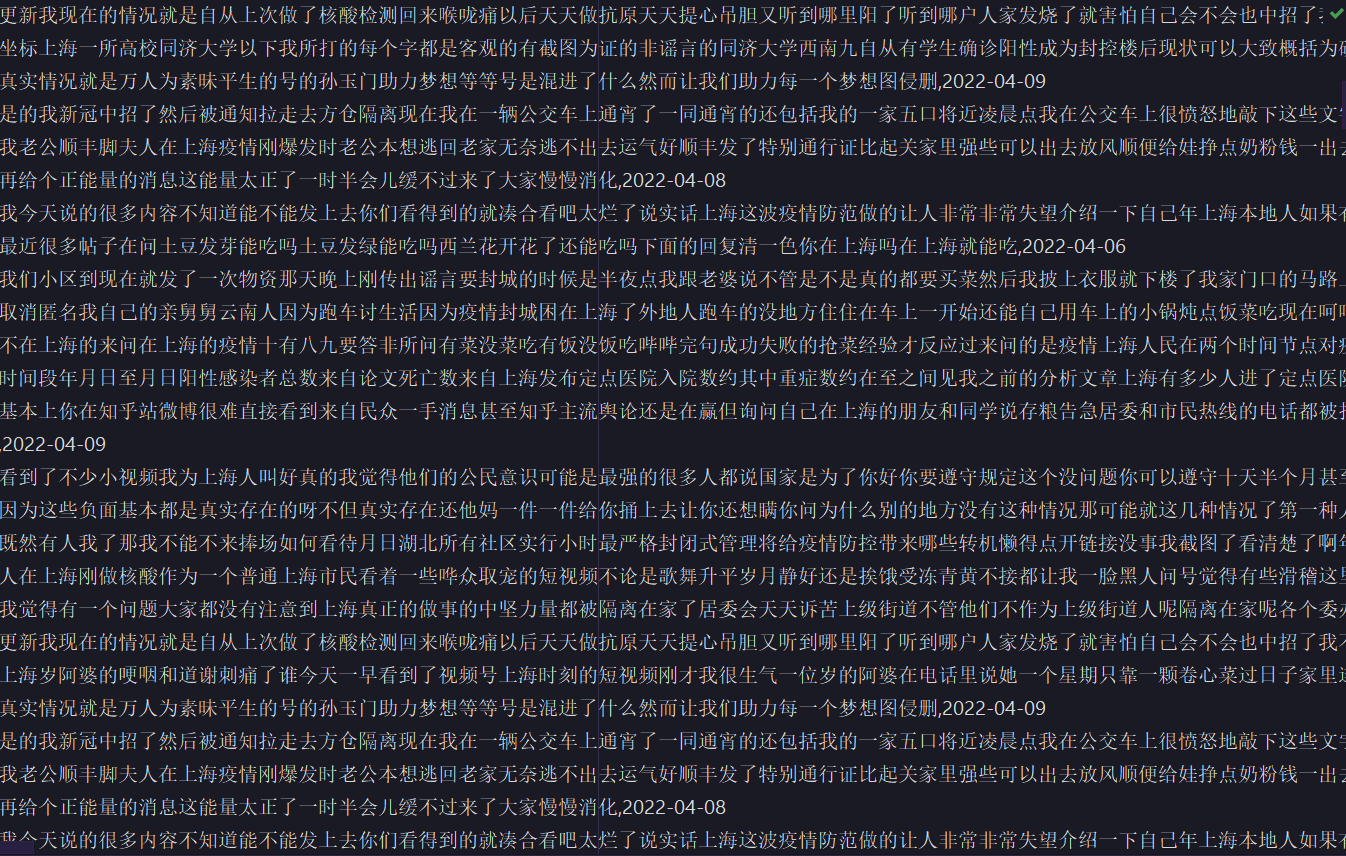
我们们从自己案例需要，在2020-2022三年12个季度分每个季度找取对应视频，存放到对应的文档来处理



部分爬取的数据

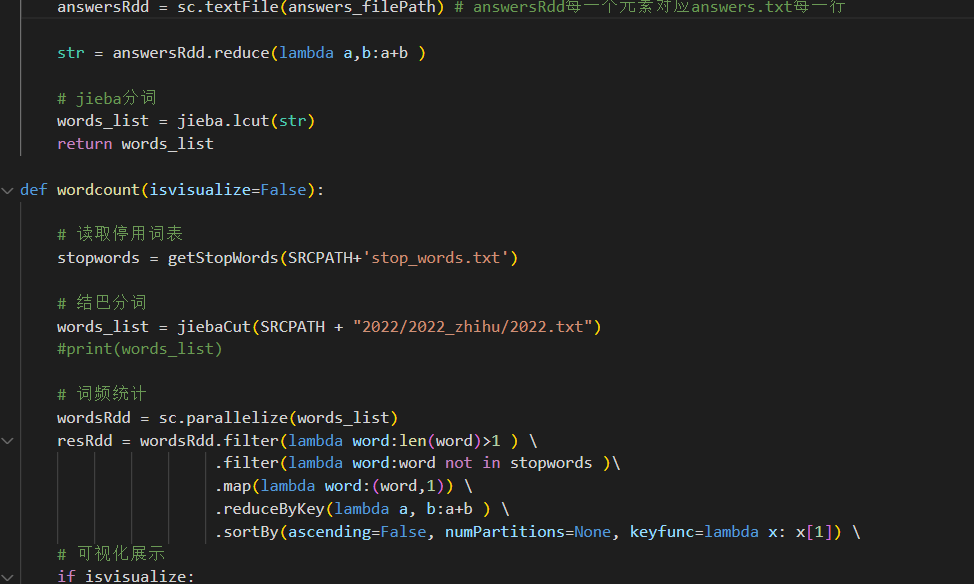






3.1数据统计可视化

我们使用了WordCount来进行词频统计



分平台将每一年所有数据汇总起来，分别做出年度关键词词云，从词云可分析出群众三年来心路历程变化，以及可以看出两个平台用户区别

2020 B站弹幕和评论关键词

2020年是病毒刚爆发一年，爆发地武汉自然成为了大家最关心的话题，全民众志成城，抗击疫情，





2021年B站最多的是讨论新冠病毒的传染是不是美国的阴谋，大量视频都与此有关，以及疫苗的开发进度也是大家最为关心的



2022年，已经是疫情第三年，大家对于新冠疫情都有点麻木，而B站用户大部分是青少年，其中已大学生为主力，疫情三年让大学生活变的不完整，也留下了许多遗憾。群众对病毒已经不耐烦，讨论共存的声音越来越多。病毒变异体奥密克戎也进入了大家的视野



知乎受众多为已经进入社会工作的成年人，对于疫情看法与青少年群体会有明显不同，

三年知乎词云

2020

疫情爆发后，知乎用户群体对于疫情对于未来经济市场和百姓生活方面以及国家发展方向有着更多的思考



2021

2021国内疫情趋于稳定，对于全球疫情的讨论成为主体，最重要的还是讨论美国是否为病毒背后制造者一事



2022

2022年从上海爆发大面积疫情后，全国各地陆续爆发疫情，疫情防控进入常态化，每日做核酸成为了司空见惯的事。



4.1 数据分析

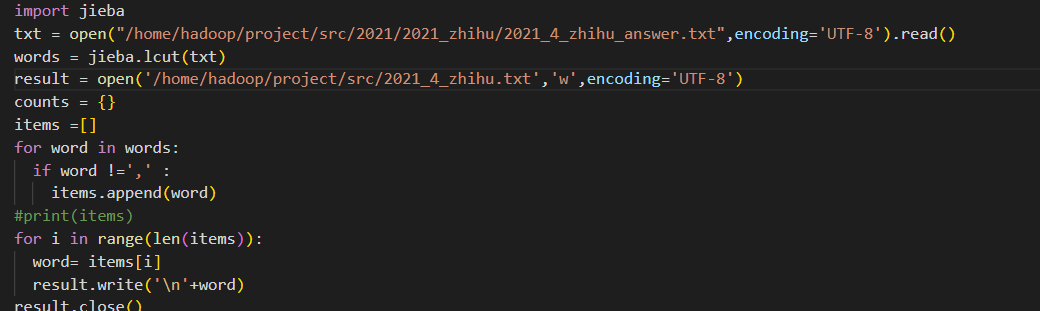
数据分析部分我们选择了SnowNLP模型来对三年12个季度数据进行情感分析，探测群众对于疫情的态度看法转变过程

使用SnowNLP进行情感分析，只需要引入该模块，之后直接使用函数SnowNLP(j)，这里的j代表句子，返回的对象调用sentiments属性即可获得情感分数，这里的情感分数是0到1的一个小数，越小代表情感是消极的概率越大。

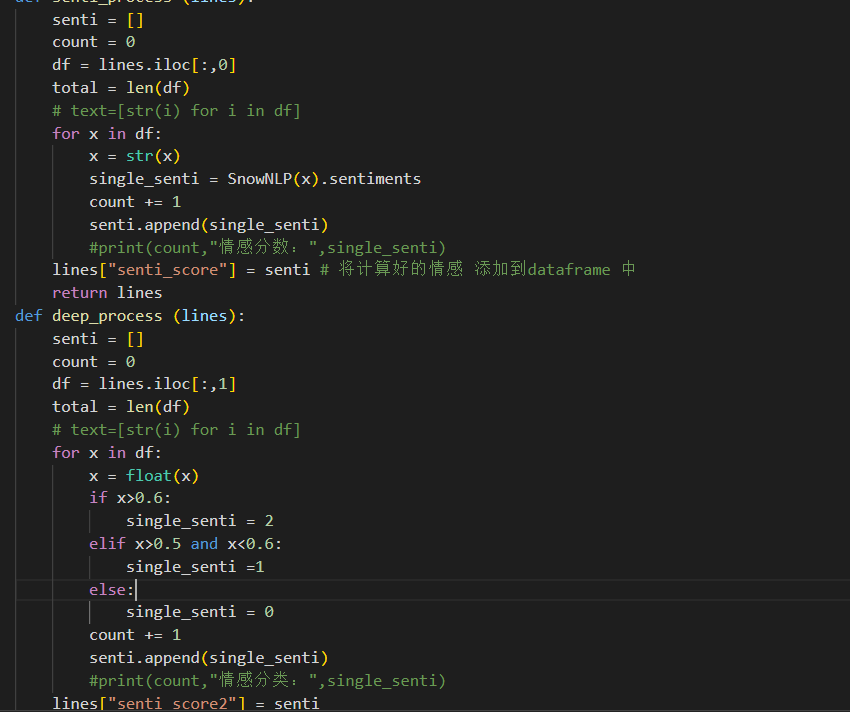
4.1.1

数据预处理

我们使用结巴分词先对数据进行清洗处理



代入模型



对计算后的结果汇总

汇总每一年情感分数的平均数和积极评论和消极评论的比例

