**《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：

# B站、知乎弹幕评论情感分析

## 1.背景介绍

面对新型冠状病毒感染肺炎的全球大流行,我国反应迅速,积极采取“扩大[检测](http://www.dqxxkx.cn/article/2021/1560-8999/1560-8999-23-2-341.shtml)、应收尽收”的防治[策略](http://www.dqxxkx.cn/article/2021/1560-8999/1560-8999-23-2-341.shtml),在短时间内实现了疫情的有效控制。国内的疫情防控形势持续向好,多地已在“内防反弹、外防输入”的[策略](http://www.dqxxkx.cn/article/2021/1560-8999/1560-8999-23-2-341.shtml)指导下,积极推动“复工复产”。近年来,基于实时性、互动式、[多元化](http://www.dqxxkx.cn/article/2021/1560-8999/1560-8999-23-2-341.shtml)等特征的[社交媒体](http://www.dqxxkx.cn/article/2021/1560-8999/1560-8999-23-2-341.shtml),如B站、知乎等,已成为传播热点事件、反映社情民意和[度量](http://www.dqxxkx.cn/article/2021/1560-8999/1560-8999-23-2-341.shtml)网络舆情的重要媒介。疫情背景下,[社交媒体](http://www.dqxxkx.cn/article/2021/1560-8999/1560-8999-23-2-341.shtml)的兴起为公众对热点事件的表达和分享提供了便捷化的主题选择和内容自定义途径,以知乎为代表的[社交媒体](http://www.dqxxkx.cn/article/2021/1560-8999/1560-8999-23-2-341.shtml)中的高现势性和高时序性评论数据作为舆情信息基底,有效支撑了传播机理探索、驱动因素挖掘、演化态势预测、舆论监控引导及事件预警防控等主题研究。因此,基于[社交媒体](http://www.dqxxkx.cn/article/2021/1560-8999/1560-8999-23-2-341.shtml)评论数据的舆情分析,有助于[辅助决策](http://www.dqxxkx.cn/article/2021/1560-8999/1560-8999-23-2-341.shtml)机构尽早掌握群众切实关注的问题,更好地构建“群策群力、群议群定”的立体化、[多层次](http://www.dqxxkx.cn/article/2021/1560-8999/1560-8999-23-2-341.shtml)全民交流网络,促进相关舆情应对和舆论引导工作由事后管理、部件管理和被动管理向事前管理、事件管理和主动管理的全面转变。本文旨在依托疫情背景探究情感变化,利用[社交媒体](http://www.dqxxkx.cn/article/2021/1560-8999/1560-8999-23-2-341.shtml)评论数据,通过[自然语言](http://www.dqxxkx.cn/article/2021/1560-8999/1560-8999-23-2-341.shtml)处理及社团网络算法挖掘和分析疫情期间民众对疫情防控的态度变化。

## 2.内容

本案例的主要内容为对B站和知乎弹幕评论进行情感分析，主要分为五个小节，分别为环境大搭建，爬虫获取数据集，数据统计可视化，数据分析以及数据分析可视化。

### 2.1搭建环境

#### 2.1.1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

#### 2.1.2 Hadoop架构搭建

* 创建用户



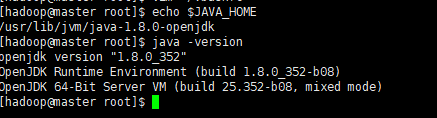
* 切换到Hadoo



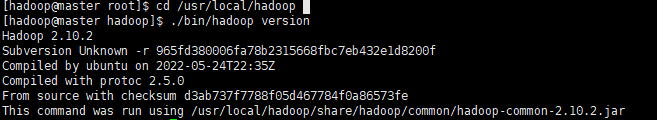
* 配置java环境
* 修改branch



* 验证



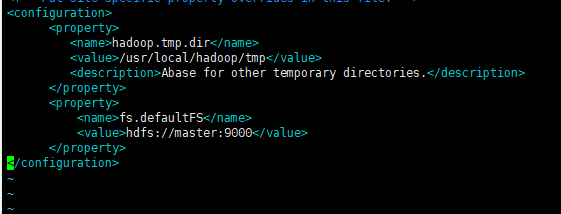
* 下载安装好hadoop



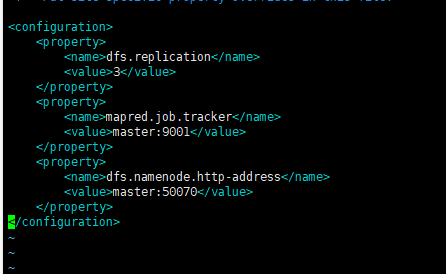
* 添加环境变量



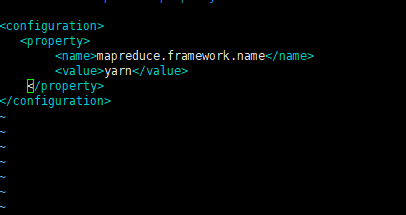
* 配置Hadoop环境下的slave、core-site.xml文件



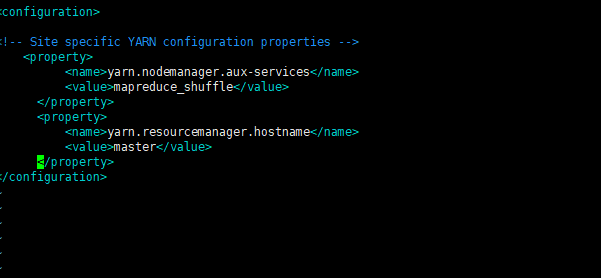
hdfs-site-xml



mapred-site.xml



yarn-site.xml

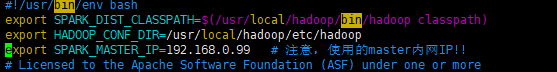


#### 2.1.3 SPARK环境搭建

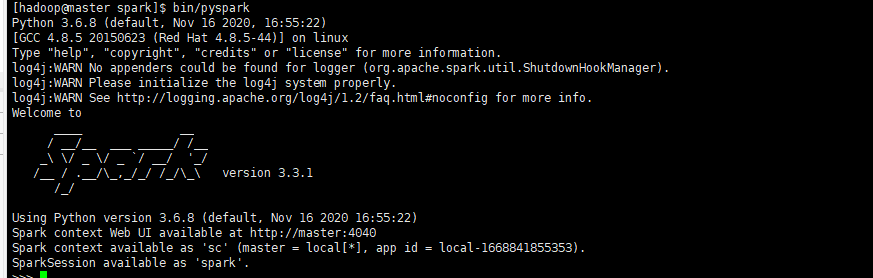
* 下载安装后配置环境变量



* 配置spark-env.sh文件



* 测试spark



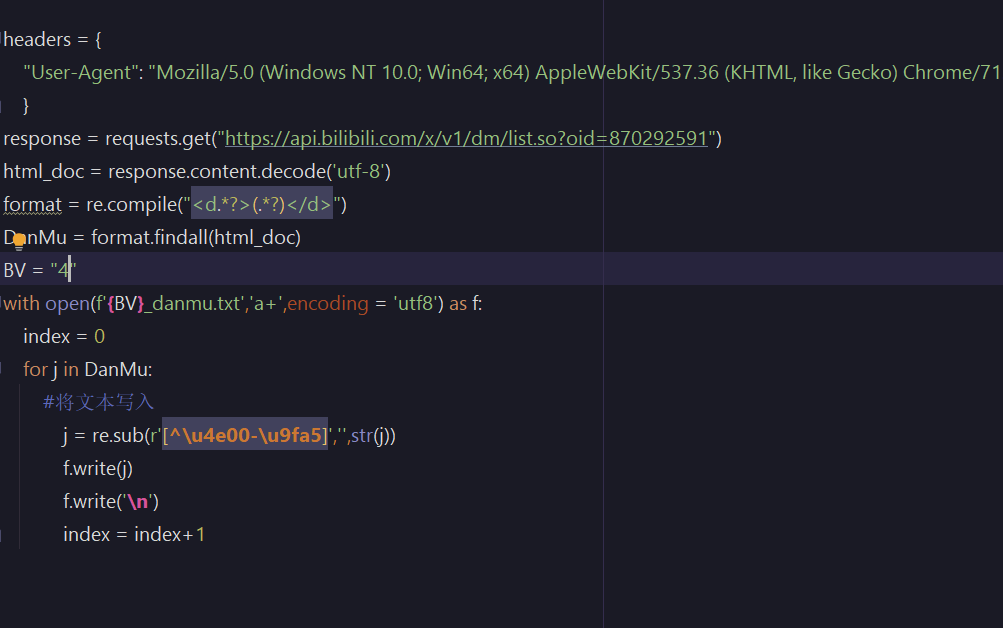
### 2.2 爬虫获取数据集

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

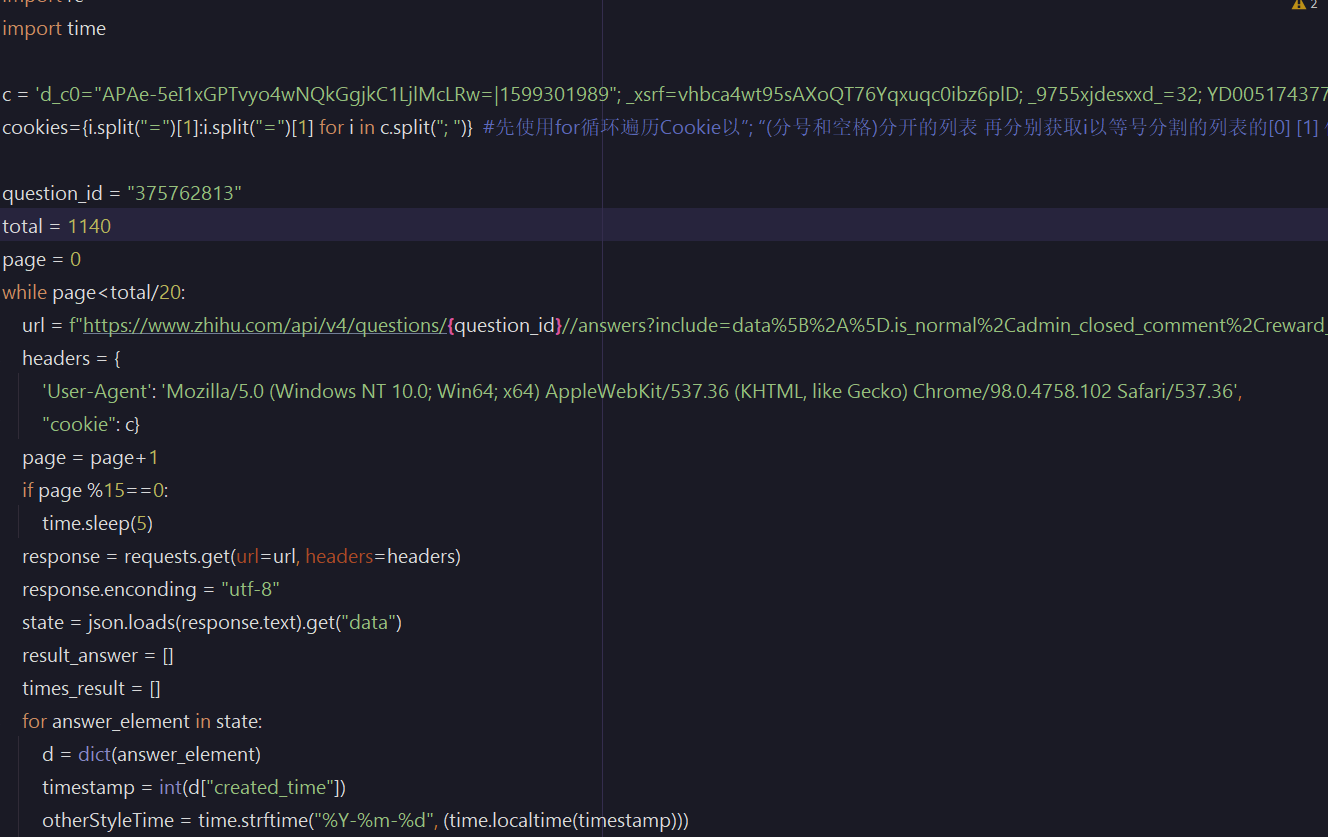
* 爬取B站评论代码部分



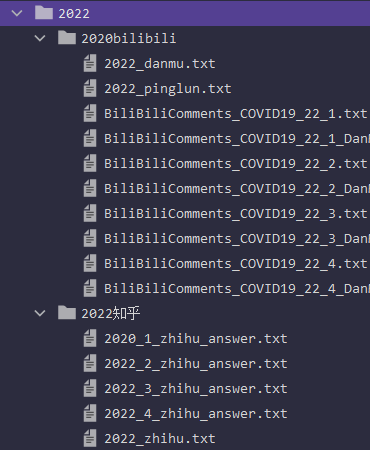
* 爬取B站弹幕部分



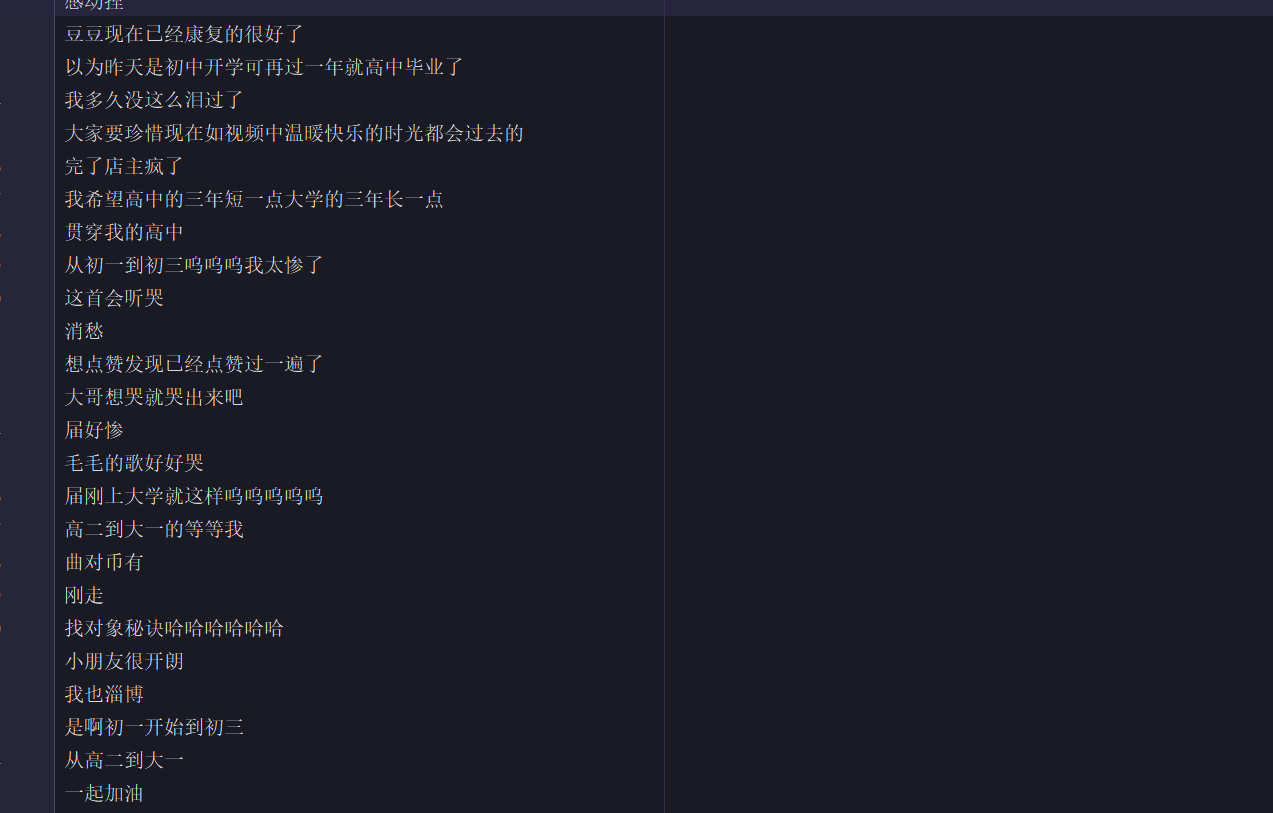
* 爬取知乎回答代码部分

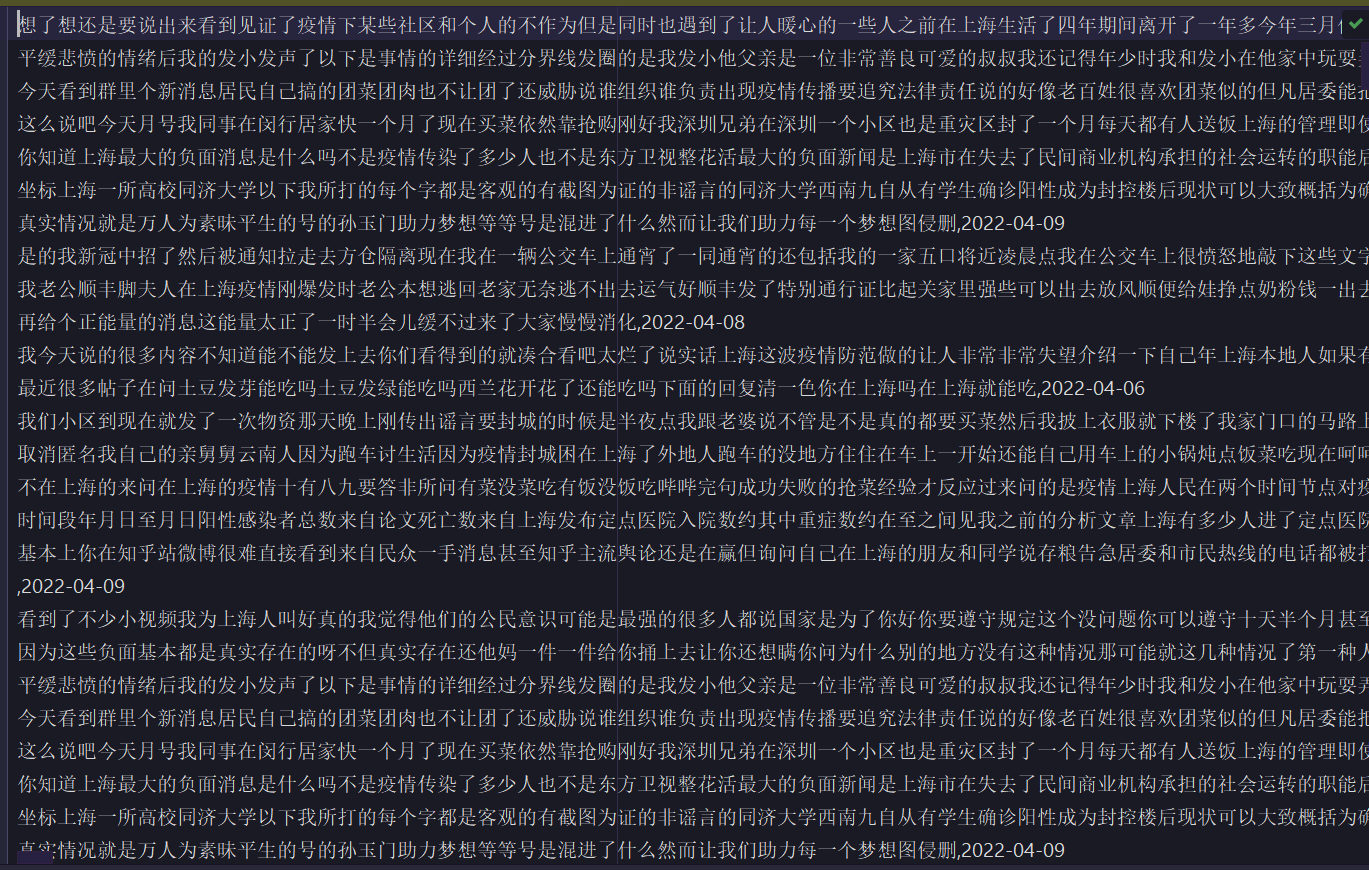


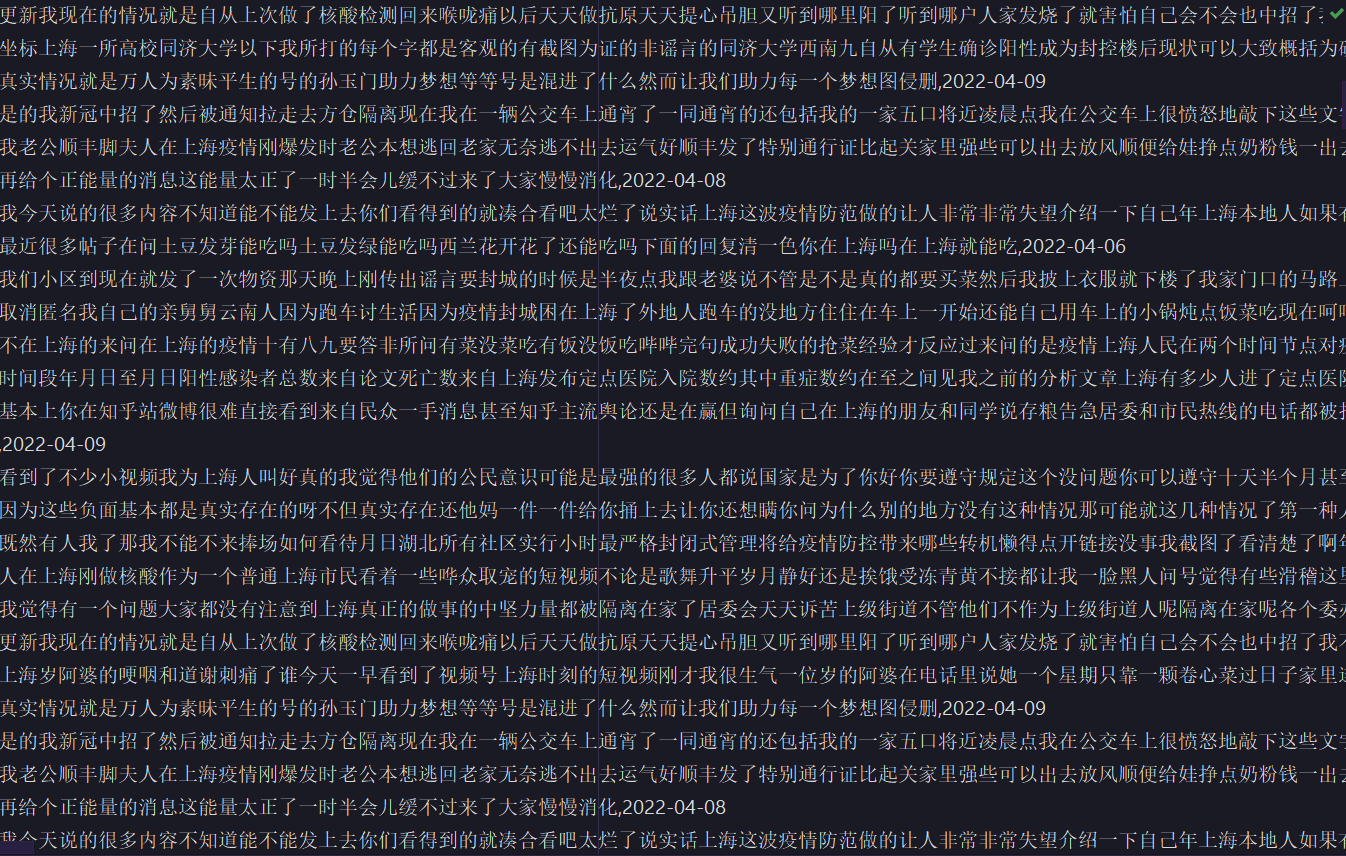
我们们从自己案例需要，在2020-2022三年12个季度分每个季度找取对应视频，存放到对应的文档来处理



* 部分爬取的数据

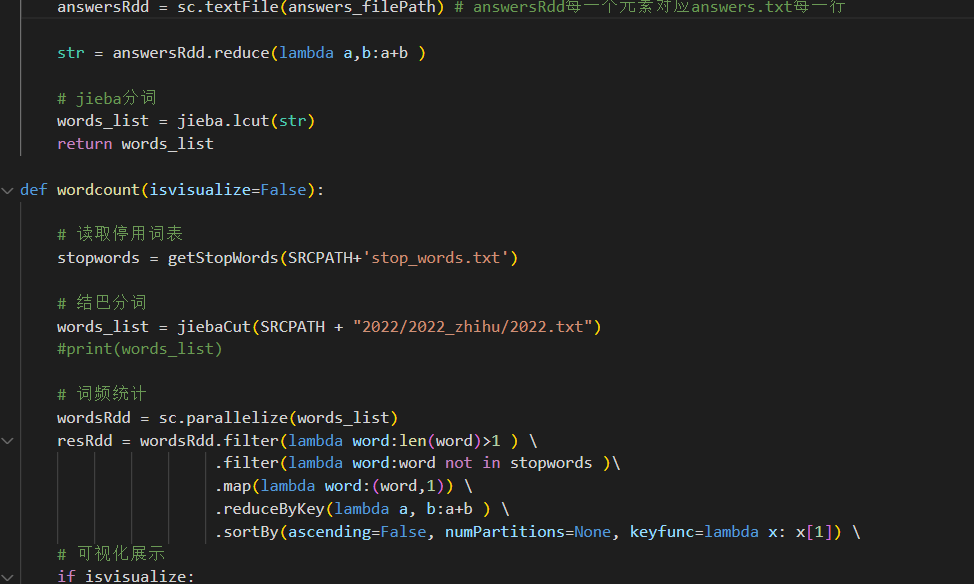






### 2.3数据统计可视化

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

* 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。

**2020 B站弹幕关键词**



**2020 B站评论关键词**

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。

**2021 B站弹幕关键词**



**2021B站评论关键词**

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。



**2022B站弹幕关键词**



**2022 B站评论关键词**

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

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



**2020知乎回答关键词**

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



**2020知乎回答关键词**

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。



**2022知乎回答关键词**

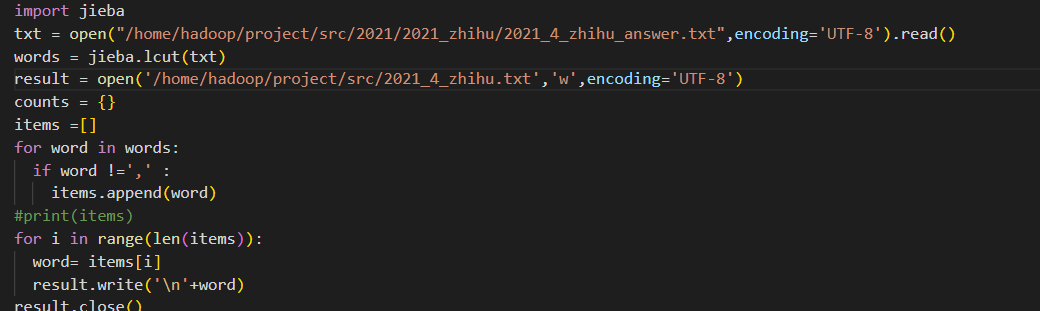
### 2.4 数据分析

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。

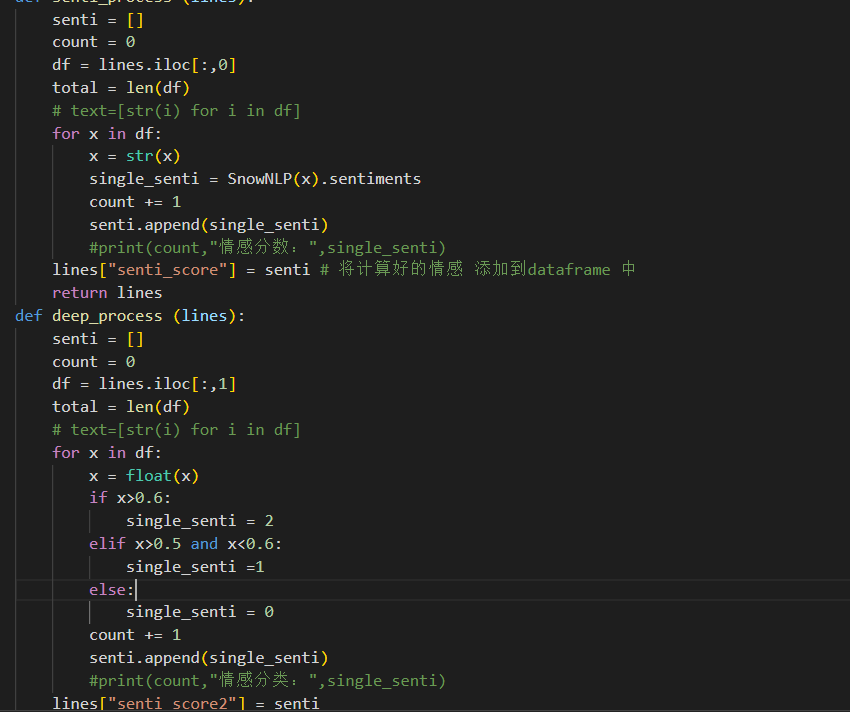
使用SnowNLP进行情感分析，只需要引入该模块，之后直接使用函数SnowNLP(j)，这里的j代表句子，返回的对象调用sentiments属性即可获得情感分数，这里的情感分数是0到1的一个小数，越小代表情感是消极的概率越大。

#### 2.4.1数据预处理

* 我们使用结巴分词先对数据进行清洗处理

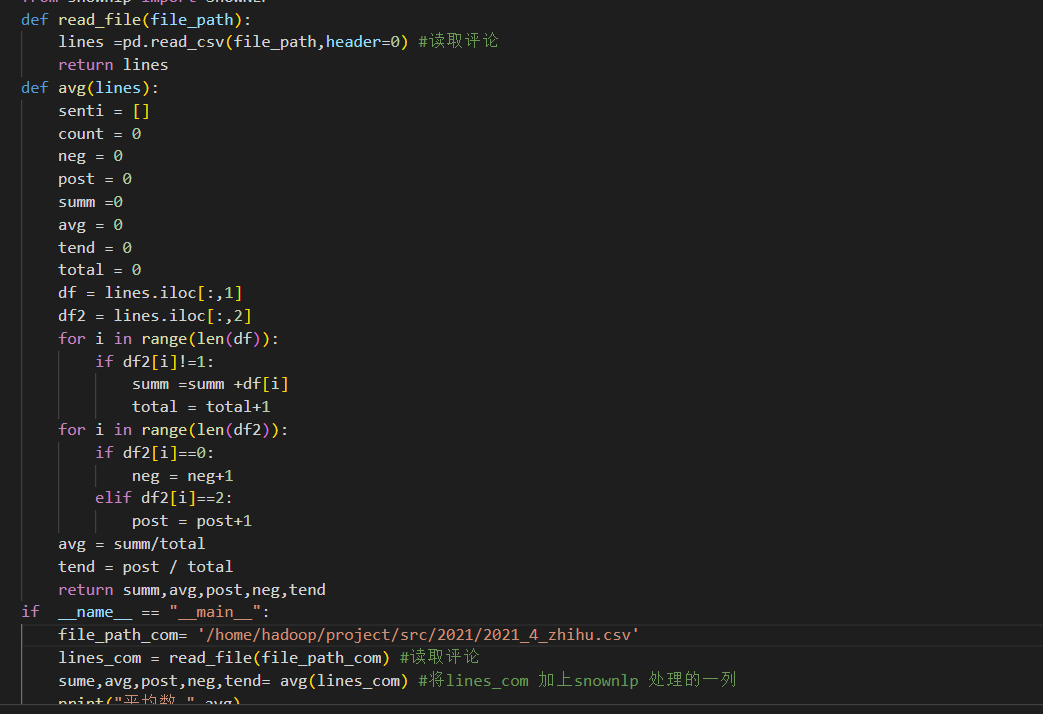


* 代入模型



* 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



### 2.5 数据可视化

可视化对比的主要有三个方面

* B站和知乎平台弹幕、评论情感指数的对比，主要用到折线图
* B站和知乎平台评论中积极弹幕和消息弹幕数量对比，主要用到河流图
* B站和知乎平台评论感情趋向对比，主要用到南丁格尔玫瑰图

#### 2.5.1可视化工具介绍

[Echarts](https://github.com/ecomfe/echarts" \t "https://pyecharts.org/?spm=a2c6h.12873639.article-detail.6.1c665027a9wXT5" \l "/zh-cn/_blank) 是一个由百度开源的数据可视化，凭借着良好的交互性，精巧的图表设计，得到了众多开发者的认可。而 Python 是一门富有表达力的语言，很适合用于数据处理。当数据分析遇上数据可视化时，[pyecharts](https://github.com/pyecharts/pyecharts" \t "https://pyecharts.org/?spm=a2c6h.12873639.article-detail.6.1c665027a9wXT5" \l "/zh-cn/_blank)诞生了。主要特点有，简洁的 API 设计，使用如丝滑般流畅，支持链式调用；囊括了 30+ 种常见图表；支持主流 Notebook 环境；可轻松集成至 Flask，Django 等主流 Web 框架；高度灵活的配置项。

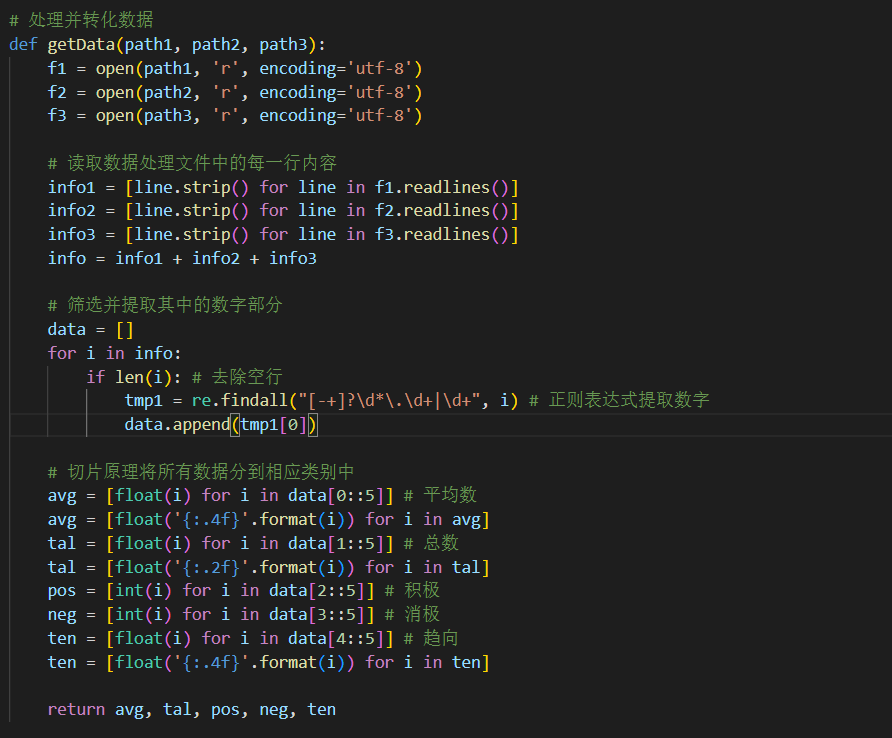
此次项目中，主要用到了pyecharts中的Pie，Line和ThemeRiver包

#### 2.5.2流程介绍

首先我们需要对之前步骤中数据分析得到的数据进行读取。根据传入的数据格式，我们需要先提取出所有数字信息，并每五个归为一类。



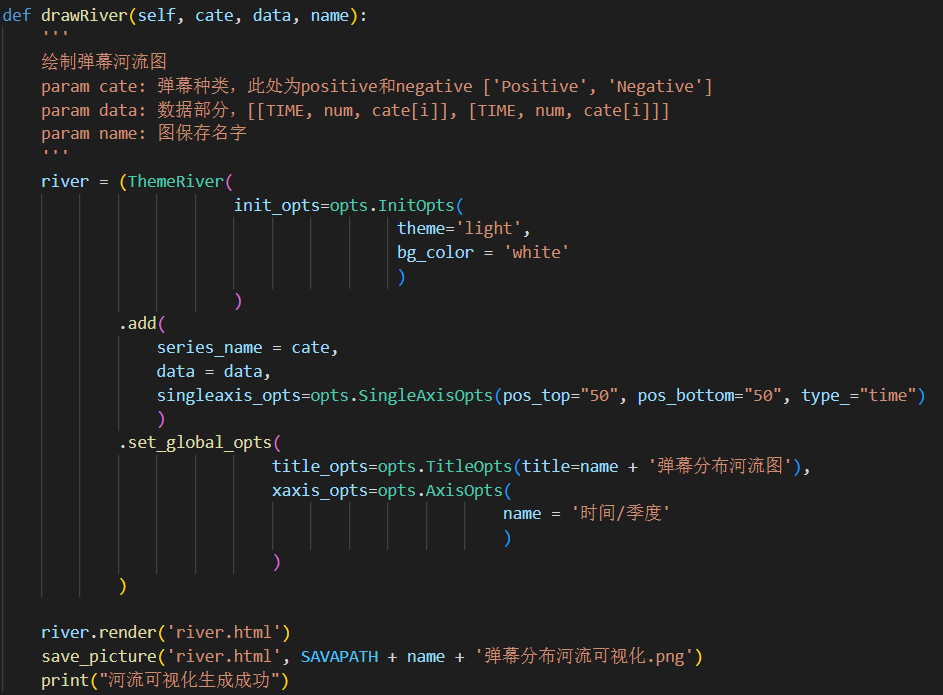
**数据分析数据集格式**



**数据接收与处理**

然后我们需要对我们需要绘制的图进行配置，主要配置项为InitOpt，配置了图的主题和背景颜色；xaxis，yaxis，根据每个图的格式传入了所要展示的x和y数据；TiTleOpts，配置了图的名字；AxisOpt，配置了坐标轴的范围，名字；LegendOpts，配置了图例的展示和位置。







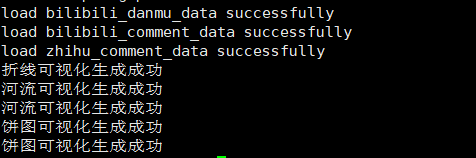
**模型定义**

之后再根据模型需要的参数格式，将数据传入即可生成成功。



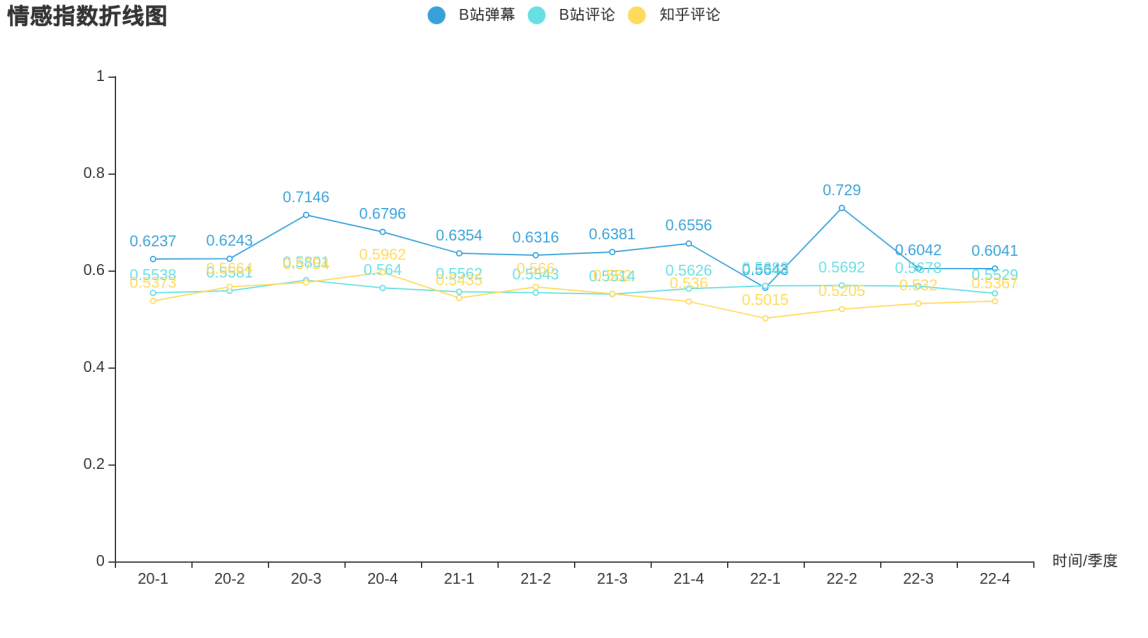
**导入数据生成模型**

成功运行示例如下：



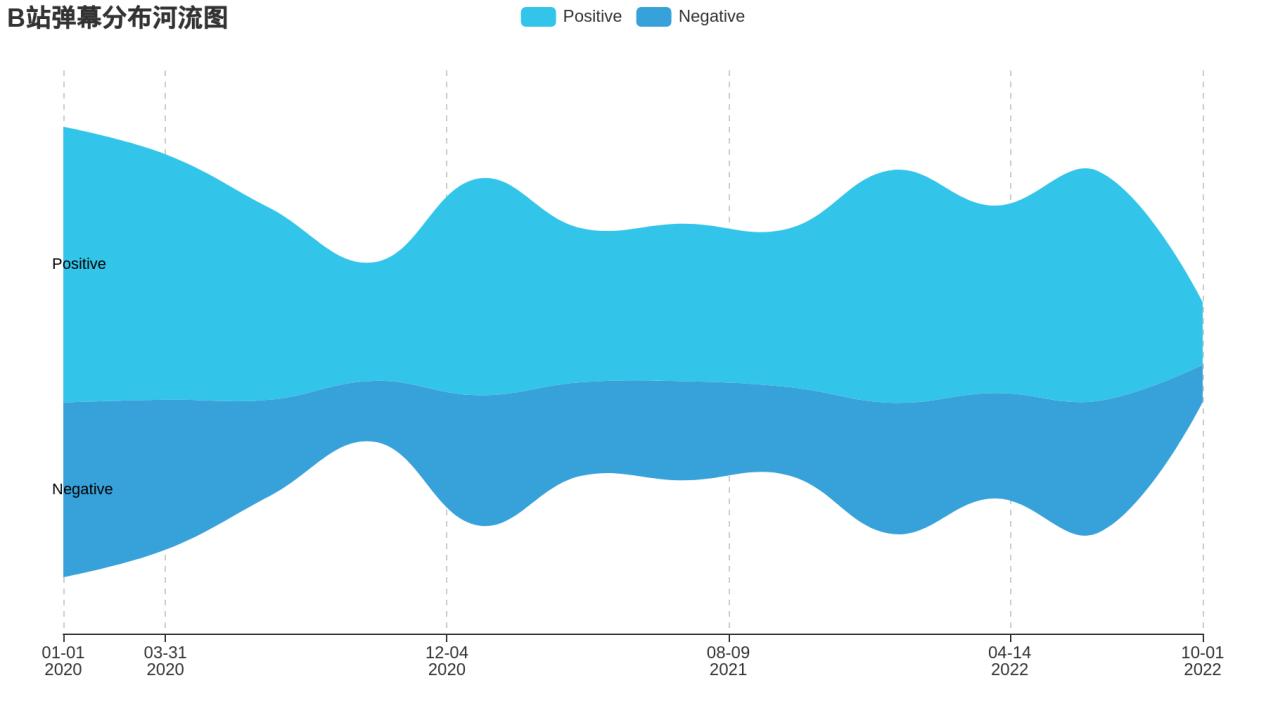
#### 2.5.3可视化展示

B站和知乎平台弹幕、评论情感指数的对比

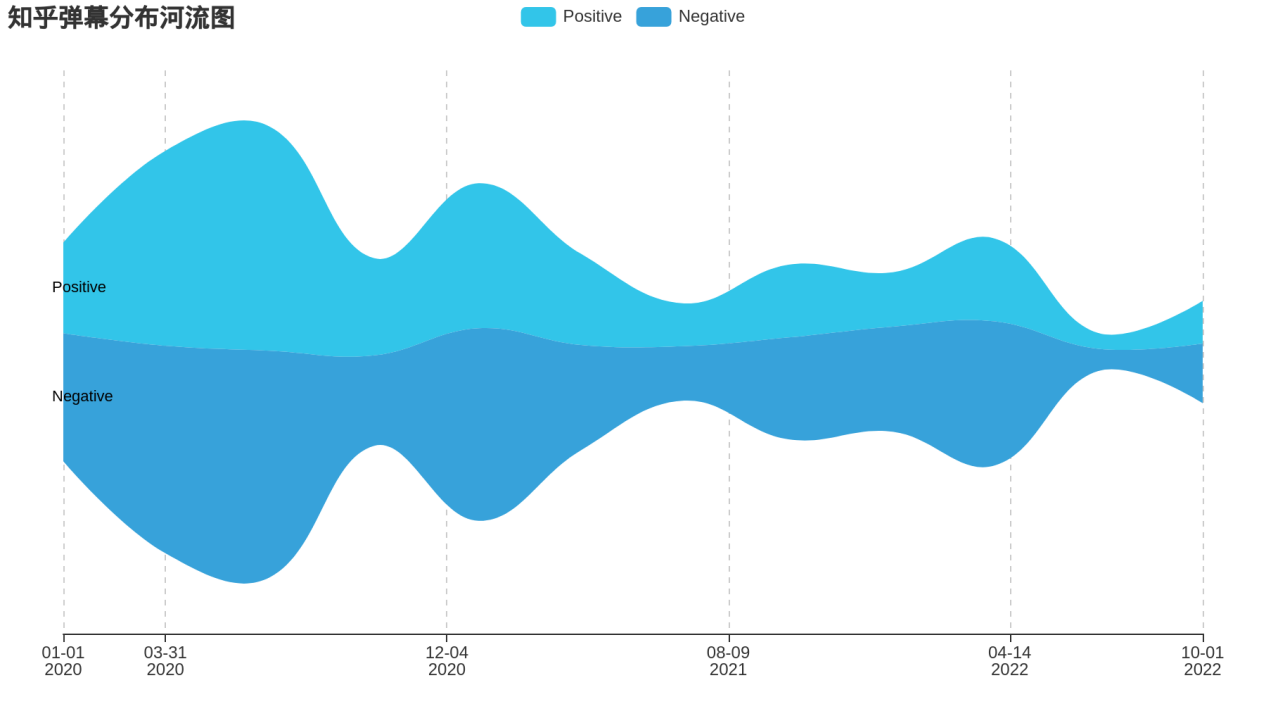


**情感指数折线图**

B站和知乎平台评论中积极弹幕和消息弹幕数量对比

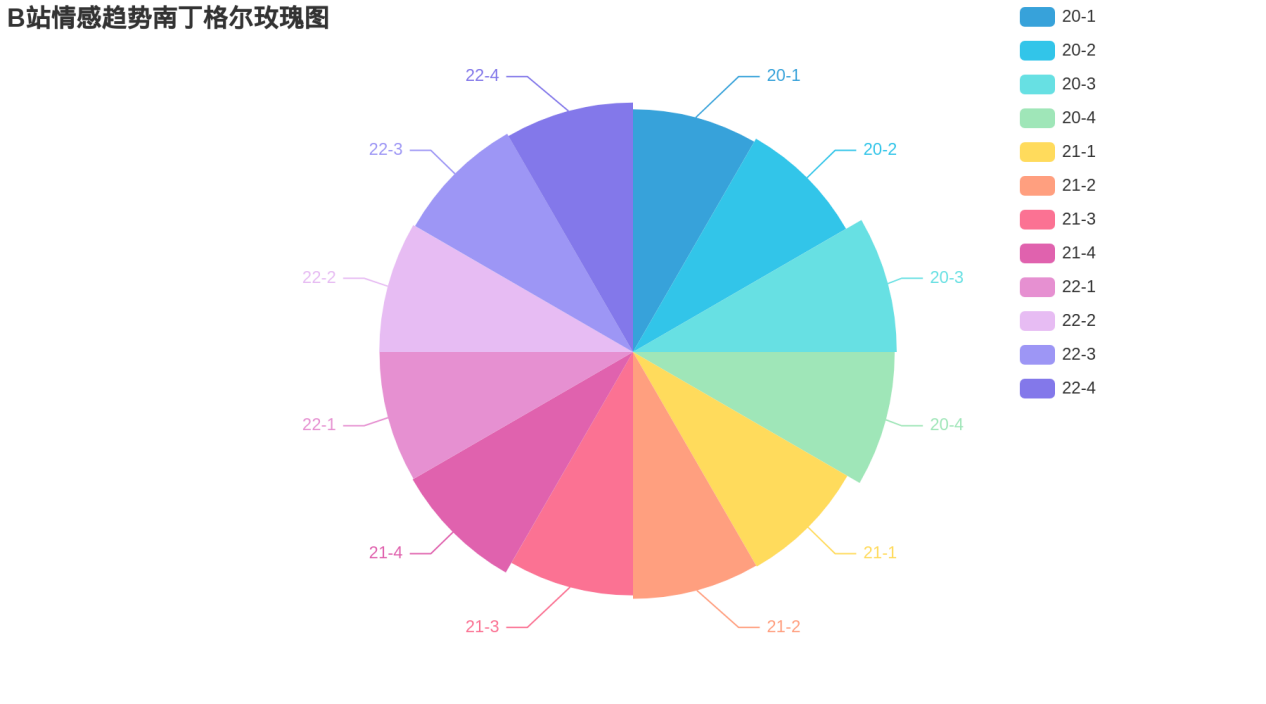


**B站弹幕分布河流图**

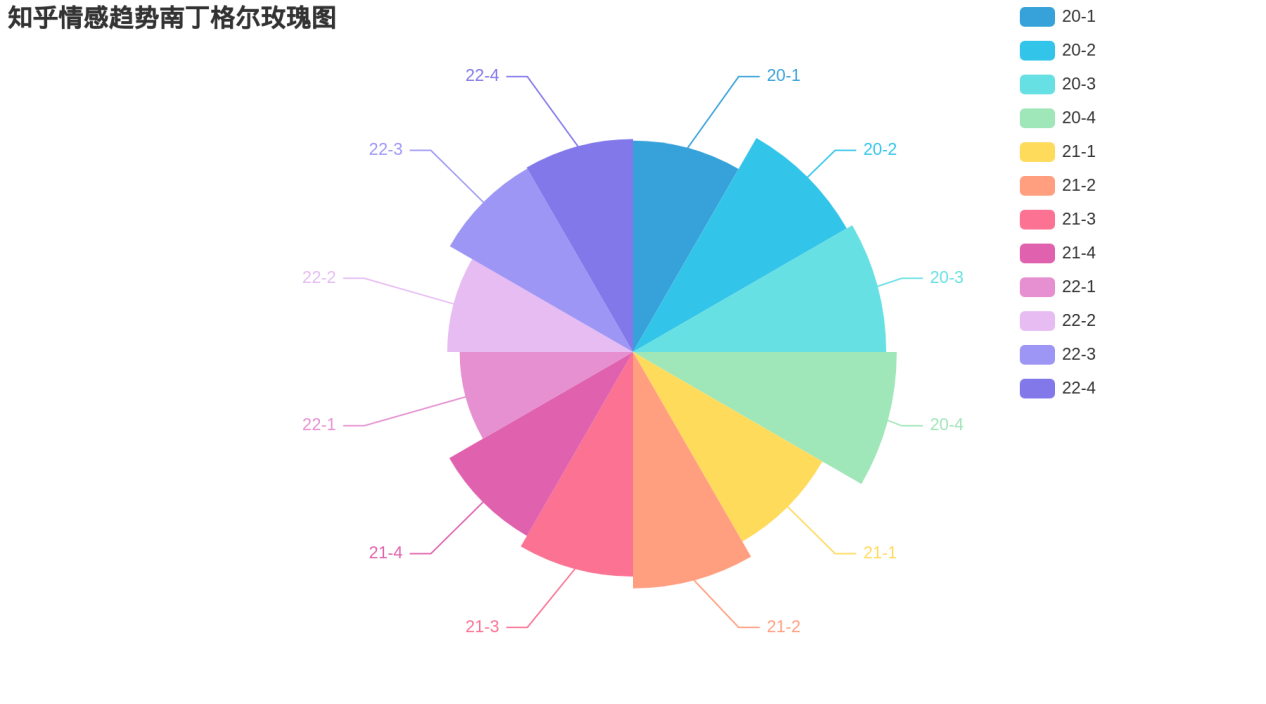


**知乎弹幕分布河流图**

B站和知乎平台评论感情趋向对比



**B站情感趋势南丁格尔玫瑰图**



**知乎情感趋势南丁格尔玫瑰图**

## 3.小结

本案例以B站和知乎评论情感分析作为主题，基于Hadoop&Spark集群环境，使用分布式爬虫爬取到的数据作为实验数据，结合snowNLP模型来对数据进行情感分析和比较主要内容包括原始数据集的爬取与预处理、数据情感分析、数据可视化。基于以上实验结果和可视化展示，我们可以知道如下结论：

* 根据折线图，B站弹幕整体情感指数高于B站评论和知乎回答，且呈现中高周低的趋势，说明疫情防控形式下，民众心态成前期后期较消极，中期较积极的态势；B站评论情感指数较为稳定；但知乎回答整体有下降趋势。
* 根据河流图，B站评论中积极数的比例明显高于知乎回答中的比例，说明B站用户的心态整体优于知乎用户。
* 根据玫瑰图，B站评论的感情趋向较为稳定，而知乎回答的感情趋向是先升高后降低，比较符合现实中民众的心态变化

疫情当前，人民至上，生命至上。我们在打赢疫情防控阻击战、攻坚战、持久战的同时，为降低新冠肺炎疫情给我国经济社会和居民生活带来的不利影响，需稳定居民收入、营造良好环、加强心理建设，疏解人民群众紧张焦虑心理。