**贵阳理工学院**

**论文**

**题 目:**  全球疫情分析

|  |  |
| --- | --- |
| **学 院:** | 大数据学院 |
| **学生姓名:** | 舒俊杰，刘星，廖彦，沈晓珺，施雄，梅家晨 |
| **学 号:** | 2017250126，2017250116，2017250114，2017250124，2017250125，2017250121 |
| **班 级:** | 大数据173 |
| **专 业:** | 数据科学与大数据技术 |
| **任课教师:** | 罗剑 |

**中国·贵州·贵阳**

目录

[全球疫情分析 II](#_Toc28111)

[摘要 III](#_Toc27292)

[关键词：全国疫情 III](#_Toc5870)

[1绪论 1](#_Toc1208)

[1.1 研究的目的和意义 1](#_Toc4677)

[1.1.1 研究的目的 1](#_Toc5968)

[1.1.2 研究的意义 1](#_Toc10183)

[1.2 国内外研究现状 1](#_Toc23944)

[2相关技术原理和基础知识 2](#_Toc29957)

[2.1 开发平台 2](#_Toc31170)

[2.1.1 硬件平台 2](#_Toc27038)

[2.1.2 软件平台 2](#_Toc16575)

[2.2 Python 2](#_Toc9324)

[2.3 Django 2](#_Toc26280)

[3系统设计 3](#_Toc7902)

[3.1 需求分析 3](#_Toc25964)

[3.1.1对世界疫情感染进行确诊人数和治愈人数分析 3](#_Toc16480)

[3.1.2本次项目功能需求分析 3](#_Toc28447)

[3.2 可行性分析 4](#_Toc445)

[3.2.1技术可行性 4](#_Toc5181)

[3.2.2 操作可行性 4](#_Toc18659)

[3.2.3社会可行性 4](#_Toc17651)

[3.3 总体设计 4](#_Toc10095)

[3.3.1 系统流程图 4](#_Toc19943)

[系统流程图 5](#_Toc15079)

[4 系统的实现和测试 5](#_Toc13338)

[4.1历史数据分析 5](#_Toc9113)

[4.2实时更新的数据爬取 7](#_Toc12488)

[4.3 hive进行数据分析 9](#_Toc28516)

[4.4hadoop环境搭建 10](#_Toc21889)

[4.5 Django编写后端数据 11](#_Toc8)

[4.6 sqoop的数据导出和导入 11](#_Toc25035)

[4.7 使用elasticsearch 12](#_Toc28841)

[4.8 前端开发 13](#_Toc24332)

[4.9 github的使用 14](#_Toc23715)

[4.10 定时运行爬虫更新数据 15](#_Toc31885)

[5总结和展望 15](#_Toc14832)

**全球疫情分析**

# 摘要

新型冠状病毒感染所致肺炎起初发热、乏力、干咳，逐渐出现呼吸困难等具备人传染人的能力，正确佩戴口罩加强防护多数预后良好，少数病情危重，甚至可致死亡疾病定义新型冠状病毒肺炎是一种急性感染性肺炎，其病原体是一种先前未在人类中发现的新型冠状病毒，即2019新型冠状病毒。2020年2月7日，国家卫健委决定将“新型冠状病毒感染的肺炎”暂命名为“新型冠状病毒肺炎”，简称“新冠肺炎”。2月11日，世界卫生组织（WHO）将其英文名称为Corona Virus Disease 2019（COVID-19）。2月22日，国家卫健委决定将“新型冠状病毒肺炎”英文名称修订为“COVID-19”，与世界卫生组织命名保持一致，中文名称保持不变。2020年1月30日，WHO宣布将新型冠状病毒肺炎疫情列为国际关注的突发公共卫生事件。患者初始症状多为发热、乏力和干咳，并逐渐出现呼吸困难等严重表现。多数患者预后良好，部分严重病例可出现急性呼吸窘迫综合征或脓毒症休克，甚至死亡。目前，缺乏针对病原体的有效抗病毒药物，以隔离治疗、对症支持治疗为主。如今全球感染人数逾1506万例，造成了很大的经济损失，而感染人数还在不断增加。可想而知，疫情如不能控制，那么无法想象。

关键词：全国疫情

# 

# 1绪论

## 1.1 研究的目的和意义

### 1.1.1 研究的目的

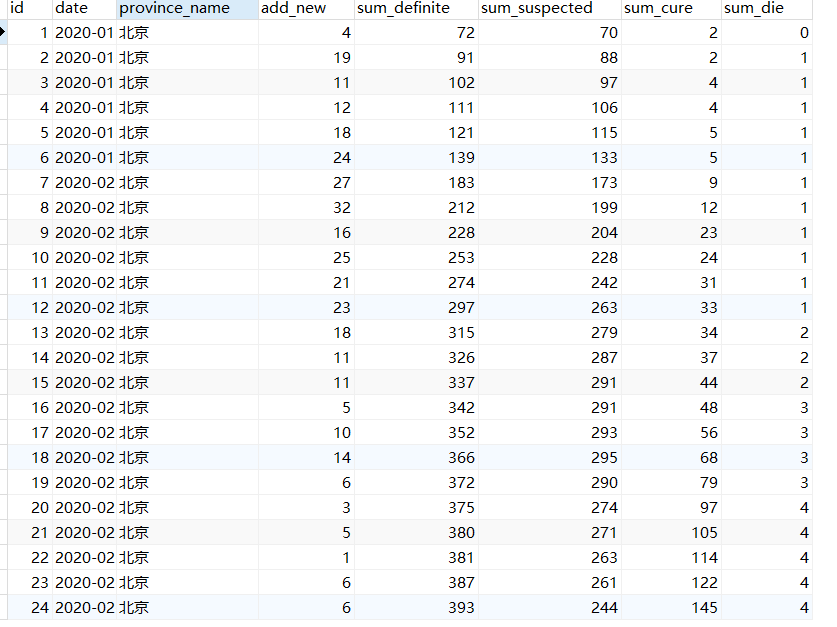
告诫人们新型冠状病毒对我们的危害，放眼全球只有国家间相互团结，才能和疫情一战，现在经济损失非常严重，唯有将传染病毒彻底消灭，世界才能重新按下开启键。

### 1.1.2 研究的意义

虽然现在疫情得到了较好的控制，但是仅限在中国而言，外国感然人数会造成中国输入病例有所增加，所以现在我们也不能掉以轻心，疫苗还在研发中，也有不少地区对疫情反复的情况。

## 1.2 国内外研究现状

国内疫情现趋于稳定；国外疫情大部分国家疫情趋势都逐渐稳定，美洲地区和印度每日确诊人数还未减少。



# 2相关技术原理和基础知识

## 2.1 开发平台

### 2.1.1 硬件平台

硬件平台是为了保证系统开发环境的基础。要根据开发的需求来选择更高性价比的硬件配置。本设计开发的硬件环境如下：

（1）开发用计算机Lenovo S40；

（2）处理器：Intel i5-4210U @2.4GHz；

（3）内存：4GByte；

（4）硬盘：机械硬盘500GB。

### 2.1.2 软件平台

软件平台是爬虫运行以及数据展示所需的集成软件环境，设计选择一个高效率、多功能、便捷性强、易操作的软件平台，对软件的开发具有十分重要的意义。结合自我适用平台能力和系统所需具备的开发环境软件要求，本系统的软件平台选择如下：

（1）操作系统：Windows10；

（2）开发环境：Python ；

（3）Python集成开发环境IDE：PyCharm；

（4）数据库：MySQL；

（5）数据库工具hive；

（6）web开发框架:Django;

（7）搜索工具：elasticsearch

（8）爬虫：分布式爬虫，定时爬虫

## 2.2 Python

Python语言是一种面向对象的解释型计算机程序设计语言，由荷兰人Guido van Rossum于1989年发明。Python的第一个公开发行版问世于1991年，是一种先编译后解释的编程语言，其基本原理是通过解释器对程序逐行做出解释，然后直接运行生成的字节码文件。Python既可以像 C语言那样调用计算机的功能接口，又可以像shell一样，通过人机之间的命令交互达到轻松编程的目的。

## 2.3 Django

Django的主要目的是简便、快速的开发数据库驱动的网站。它强调代码复用，多个组件可以很方便的以“插件”形式服务于整个框架，Django有许多功能强大的第三方插件，你甚至可以很方便的开发出自己的工具包。这使得Django具有很强的可扩展性。它还强调快速开发和DRY(Do Not Repeat Yourself)原则。

**2.4 hive**

hive是基于[Hadoop](https://baike.baidu.com/item/Hadoop/3526507" \t "https://baike.baidu.com/item/hive/_blank)的一个[数据仓库](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E4%BB%93%E5%BA%93/381916" \t "https://baike.baidu.com/item/hive/_blank)工具，用来进行数据提取、转化、加载，这是一种可以存储、查询和分析存储在Hadoop中的大规模数据的机制。hive数据仓库工具能将结构化的数据文件映射为一张数据库表，并提供[SQL](https://baike.baidu.com/item/SQL/86007" \t "https://baike.baidu.com/item/hive/_blank)查询功能，能将[SQL语句](https://baike.baidu.com/item/SQL%E8%AF%AD%E5%8F%A5/5714895" \t "https://baike.baidu.com/item/hive/_blank)转变成[MapReduce](https://baike.baidu.com/item/MapReduce/133425)任务来执行。Hive的优点是学习成本低，可以通过类似SQL语句实现快速MapReduce统计，使MapReduce变得更加简单，而不必开发专门的MapReduce应用程序。hive是十分适合数据仓库的统计分析和[Windows](https://baike.baidu.com/item/Windows/165458" \t "https://baike.baidu.com/item/hive/_blank)注册表文件

**2.5 elasticsearch**

Elasticsearch是一个基于[Lucene](https://baike.baidu.com/item/Lucene/6753302" \t "https://baike.baidu.com/item/elasticsearch/_blank)的搜索服务器。它提供了一个分布式多用户能力的全文搜索引擎，基于RESTful web接口。Elasticsearch是用Java语言开发的，并作为Apache许可条款下的开放源码发布，是一种流行的企业级搜索引擎。Elasticsearch用于[云计算](https://baike.baidu.com/item/%E4%BA%91%E8%AE%A1%E7%AE%97/9969353" \t "https://baike.baidu.com/item/elasticsearch/_blank)中，能够达到实时搜索，稳定，可靠，快速，安装使用方便。官方客户端在Java、.NET（C#）、PHP、Python、Apache Groovy、Ruby和许多其他语言中都是可用的。根据DB-Engines的排名显示，Elasticsearch是最受欢迎的企业搜索引擎，其次是Apache Solr，也是基于Lucene。

# 3系统设计

## 3.1 需求分析

3.1.1对世界疫情感染进行确诊人数和治愈人数分析

（1）对世界疫情进行分析，可以实时更新数据使人们可以看到世界的疫情情况:其中美国人数感染最多，达到400多万。巴西感染人数200多万，印度感染人数100多万，俄罗斯感染人数80多万，南非感染人数40多万。（柱状图：可以更加直观的反应各个国家感染人数对比）

（2）疫情治愈对比，可以看到各国对疫情的感染人员进行的医疗情况让人们看到各国的医疗情况和对疫情的重视程度，其中德国和中国的医疗条件，和国家重视程度都居高，而英国感染人数却只能达到德国的十分之一。（柱状图：使人们更容易看出对比程度）

（3）中国疫情地图，通过hive进行实时分析爬取网页上中国各个省份最新数据进行汇总分析并通过中国地图来显示感染人数（中国地图：标明各个地区感染人数提醒人们出行注意）

（4）美国疫情发展，可以实时看到疫情的变化情况，也可以看到疫情的发展趋势，7月27日之前都可了解到美国的治愈人数，感染人数，死亡人数。图中可看出治愈人数和感染人数都是不断增加的。（折线图：更好反应各个时间段的治愈人数，确诊人数和死亡人数）

（5）每日数据展览，使用hive清洗爬虫爬取网页的数据后可以通过展示进行各个国家的每日感染人数，治愈人数和死亡人数展示，并且可工具搜索框来搜索各个国家的疫情情况。（表格形式：按全球每日确诊人数，死亡人数，治愈人数进行排表）

3.1.2本次项目功能需求分析

（1）模拟浏览器发起一个HTTP请求，获取返回的网页页面的具体信息，并对返回的网页页面进行解析。使用正则表达式提取返回网页中需要的信息。完成对全球疫情信息数据的爬取；

（2）对提取好的关键信息数据进行存入数据库，使用MySQL数据库存储数据以及从爬虫状态数据信息，在使用MySQL数据对职位信息进行结构化存储；

（3）将MySQL中数据用sqoop传输至hive中进行数据处理。

（4）从hive中将处理好的数据再传输至MySQL中。

（5）再使用django为Web搭建框架。

（6）使用Web页面的形式来展示爬取下来的全球疫情信息、从爬虫的运行状态和爬取的实时进度进行可视化展示。

## 3.2 可行性分析

### 3.2.1技术可行性

本设计选用的是Python语言，根据国内外Python爬虫的研究现状，爬虫技术已经逐步成熟稳健。网络爬虫目前已经比较普遍，国内外有众多对网络爬虫的研究成果，大部分的技术难题已经有相应的解决方法。Ajax根据系统的功能、性能及实现系统的各项约束条，现有的技术及硬件软件都能够满足。根据hive、sqoop、django的发展和技术我们可以从其他地方和依靠老师的讲解我们可以较轻松的达到项目的要求技术。

### 3.2.2 操作可行性

本设计使用Python语言，在具备Python环境的操作系统就能运行，对于Python进行爬虫而言，很多领域已经使用到，操作是可行的。

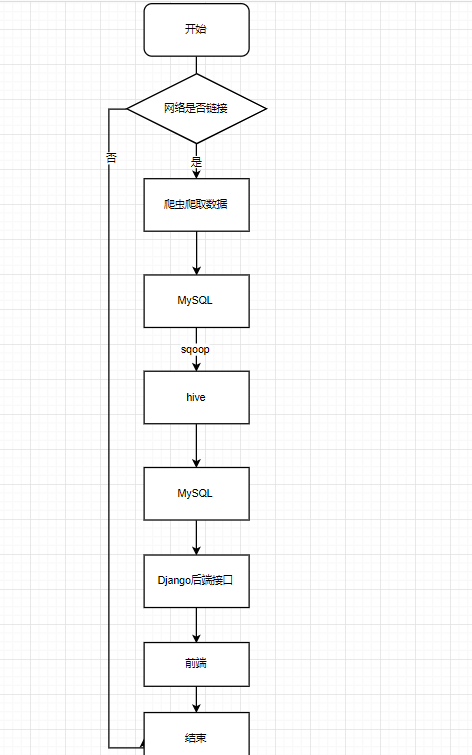
### 3.2.3社会可行性

本设计所使用的开发软件均属于开源项目，不存在违反知识产权保护的问题。网络爬虫所爬取的数据是互联网上公开的数据，不存在侵犯个人隐私、传播违法信息等违背国家政策的违法行为，因此具有社会可行性。

## 3.3 总体设计

### 3.3.1 系统流程图

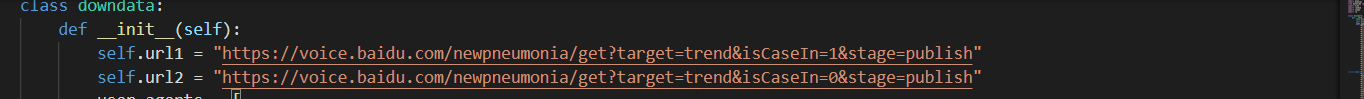
系统流程图是用统一规定的标准符号描述程序运行具体步骤的图形表示。程序框图的设计是在处理流程图的基础上，通过对输入输出数据和处理过程的详细分析，将计算机的主要运行步骤和内容标识出来。程序框图是进行程序设计的最基本依据，因此它的质量直接关系到程序设计的质量。本设计的系统流程图如图所示。



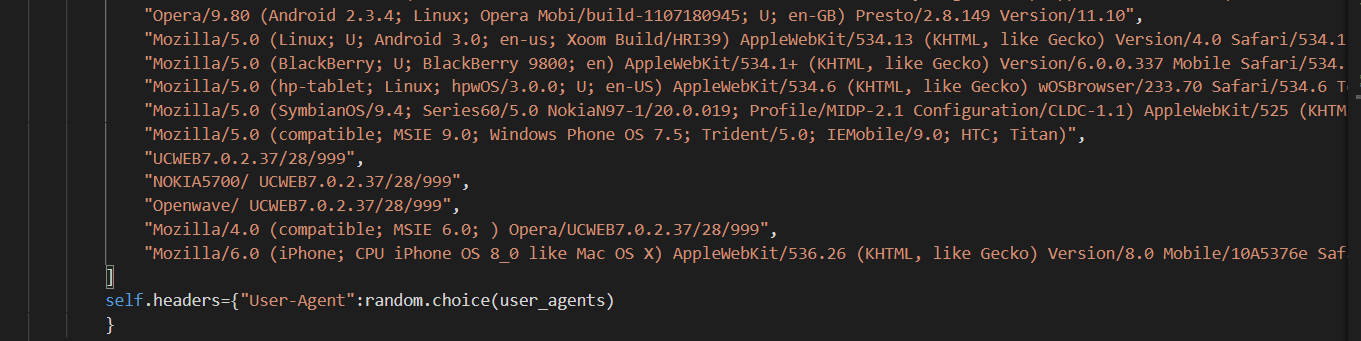
**系统流程图**

### 4 系统的实现和测试

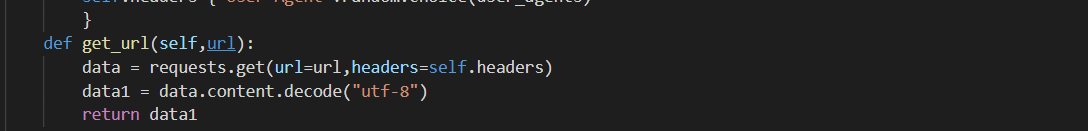
**4.1历史数据分析**



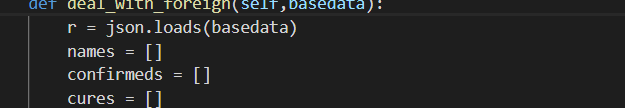
首先空过分析网页页面获取到数据接口的地址



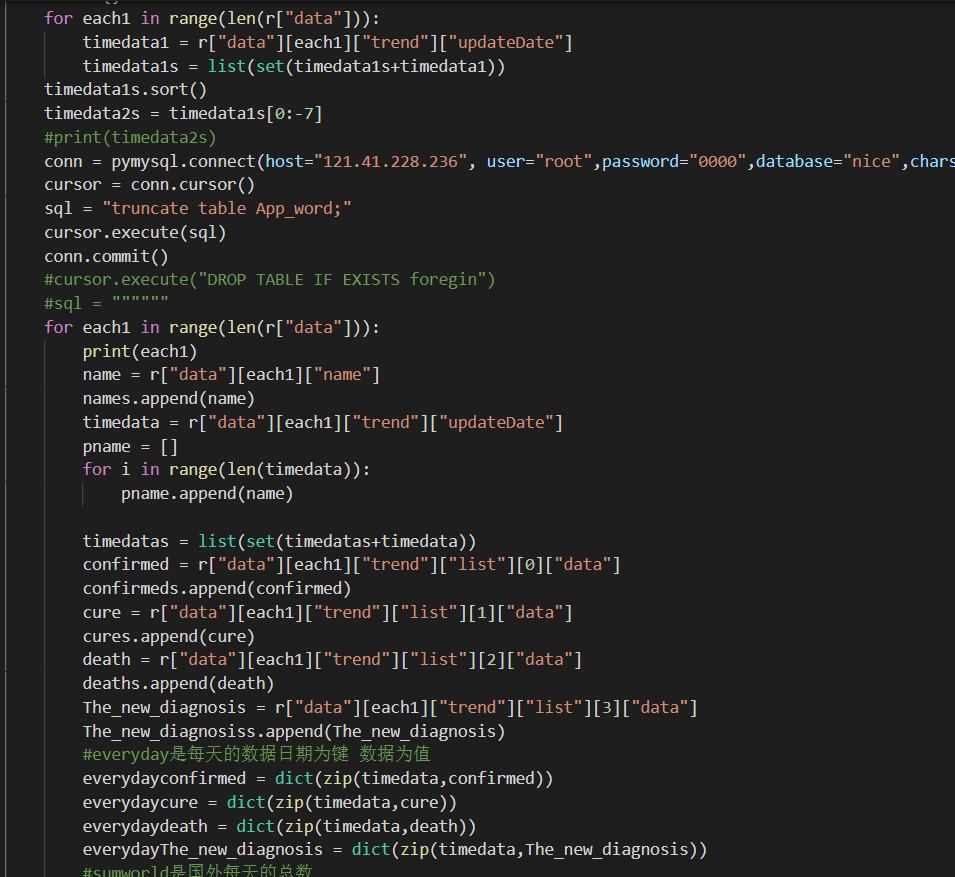
通过写一个user\_agent库用random向里面每一次随机调用一个useragent来防止ip被封



将页面上的数据返回



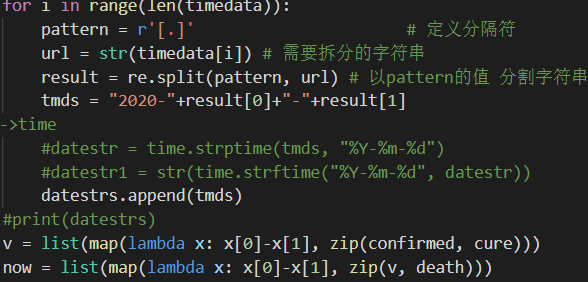
数据转化为json字符串



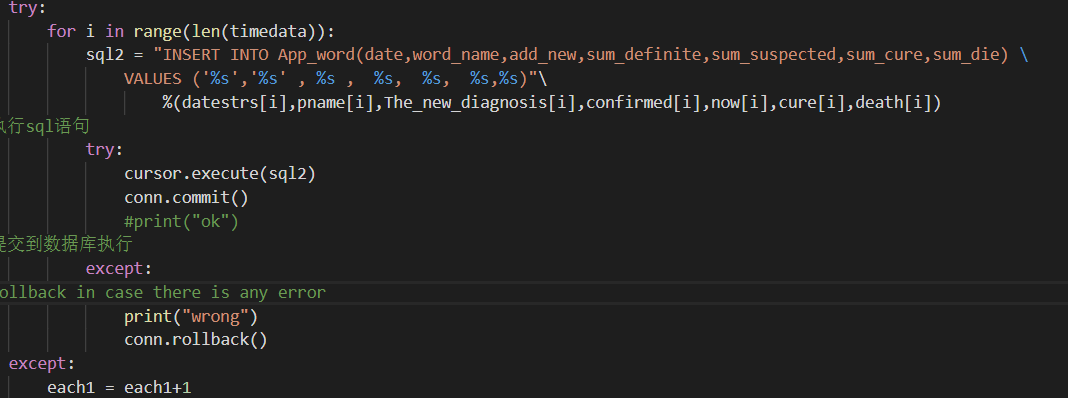
遍历json字符串 筛选出我们需要的数据

ELI0C3L2D$6EX2(UAG[U1)3

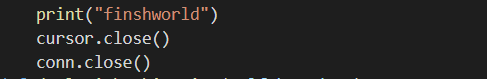
连接mysql数据库



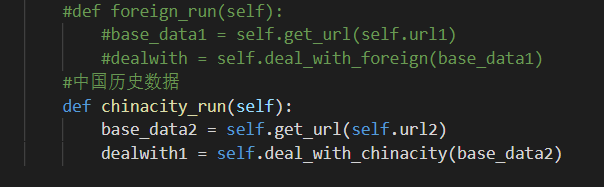
进一步处理数据成为数据库中所对应的字符类型

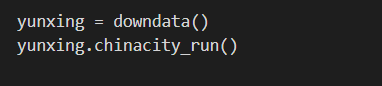
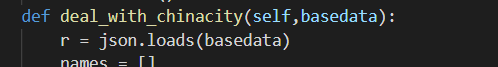


向数据库插入数据



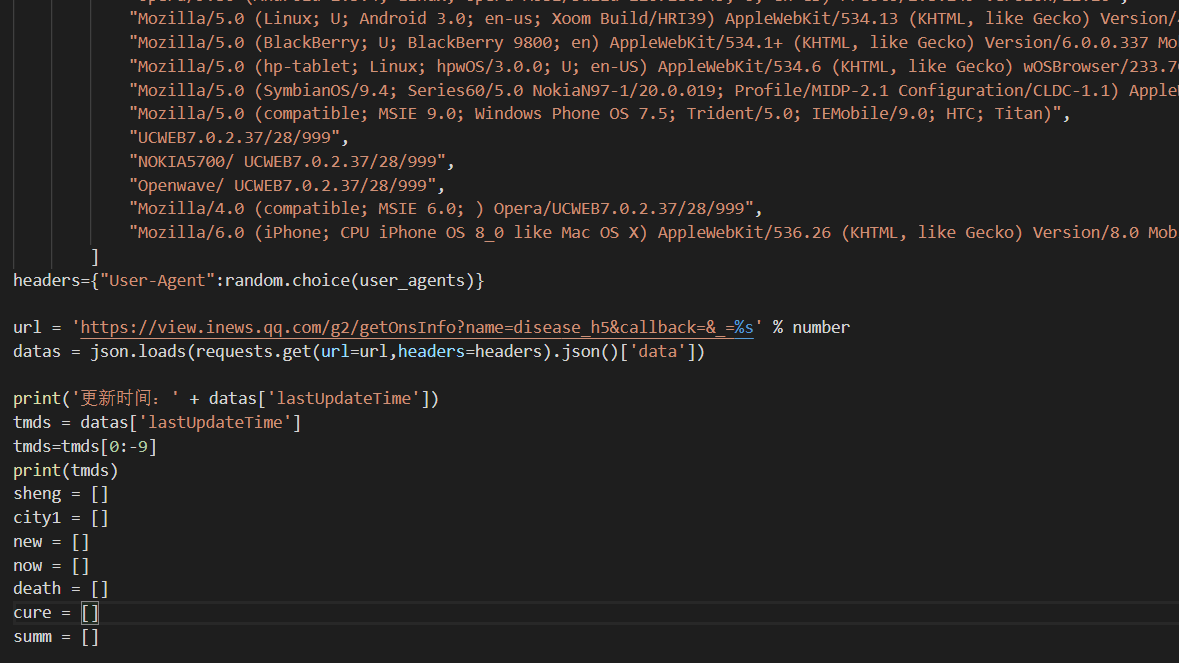
完成输出finish关闭数据库连接

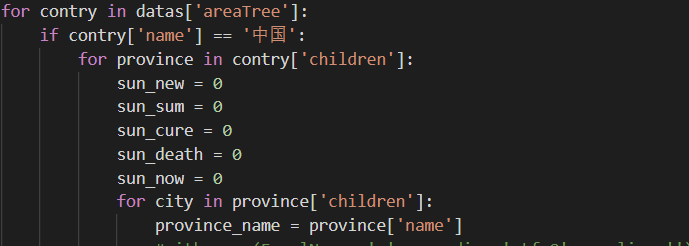
[YMJTYZ61G}%@H~[Y]0(2AY



通过调用函数来完成对国外和中国历史疫情数据的爬取

**4.2实时更新的数据爬取**

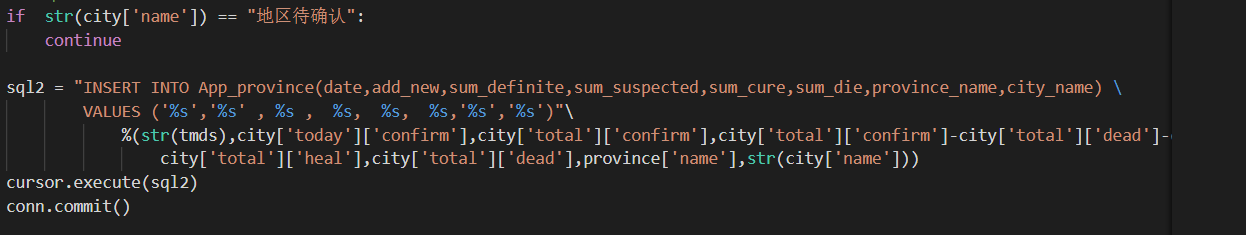
****

随机从库中抽取headers以及接口url进行get请求并且将返回的文本转化为json 

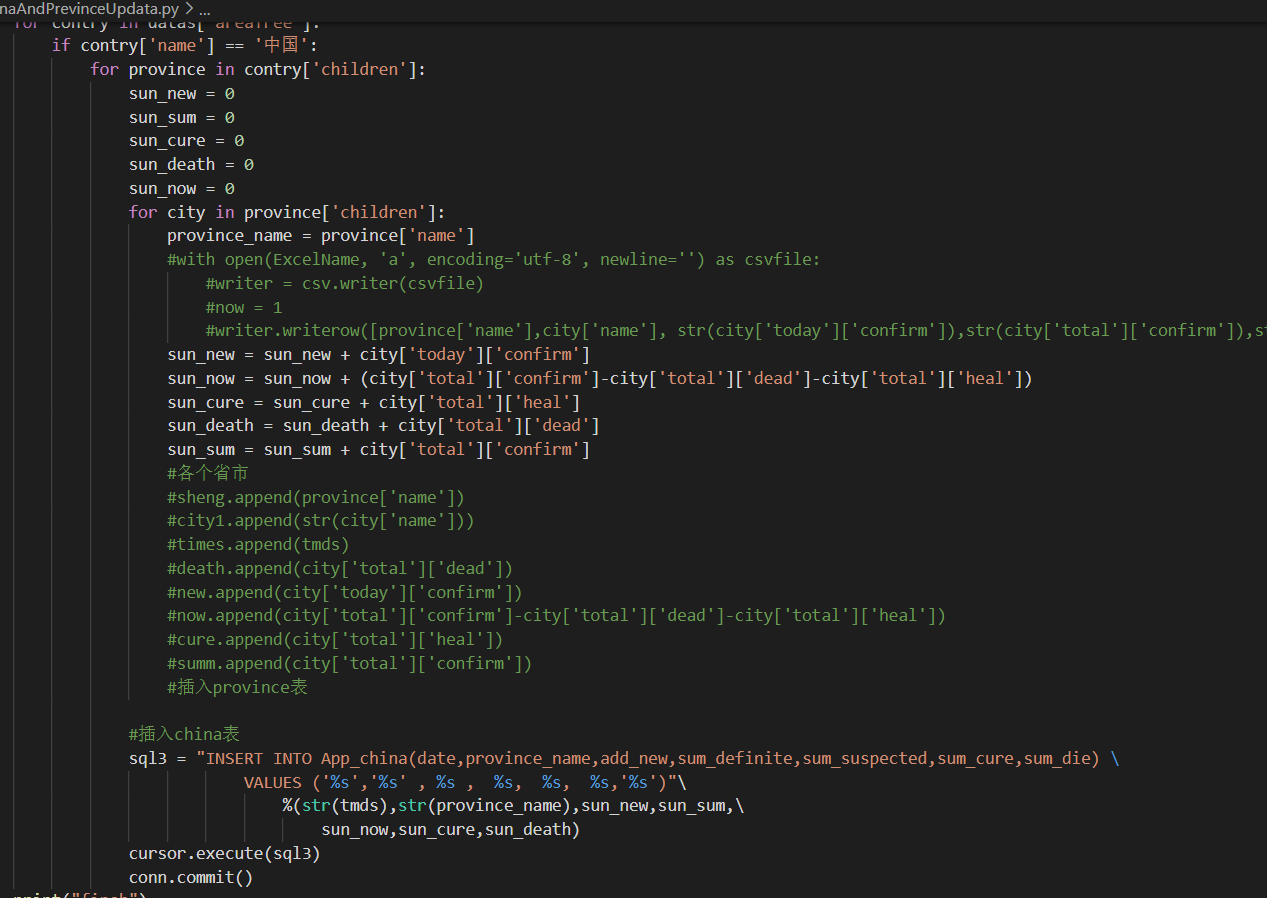
遍历json

_B89H7F45V@KU7QDZSZS2}0

链接接数据库

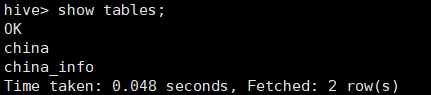


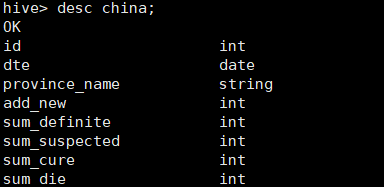
提取到每个地区的数据处理之后加入数据库循环中要是出现cityname==地区待确认不是需求数据跳过

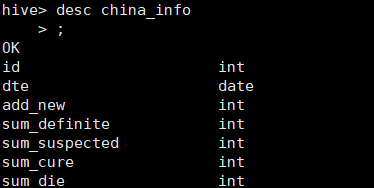


将地区数据进行处理得到省份数据处理过后插入数据库完成之后输出fisish关闭数据库连接

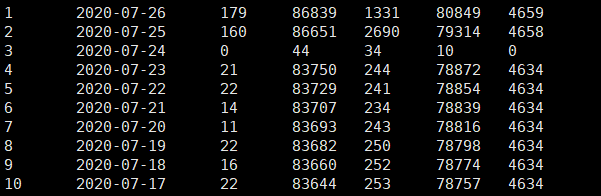
**4.3 hive进行数据分析**

****

****

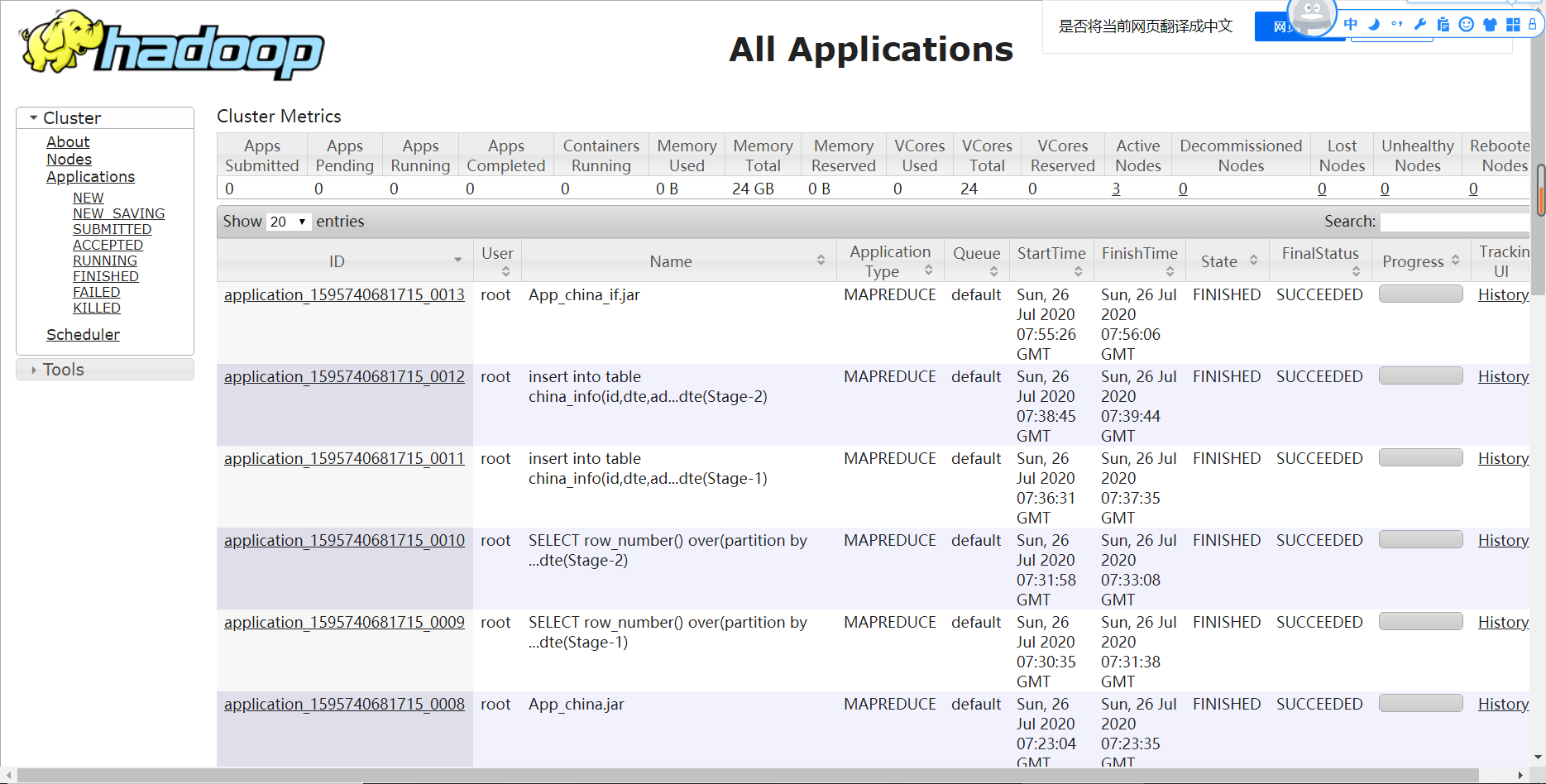
****

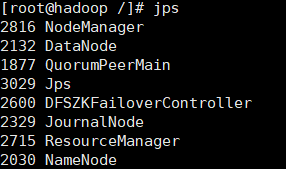
****

****

Hive将MySQL中数据导出进行数据分析清洗

**4.4hadoop环境搭建**

****

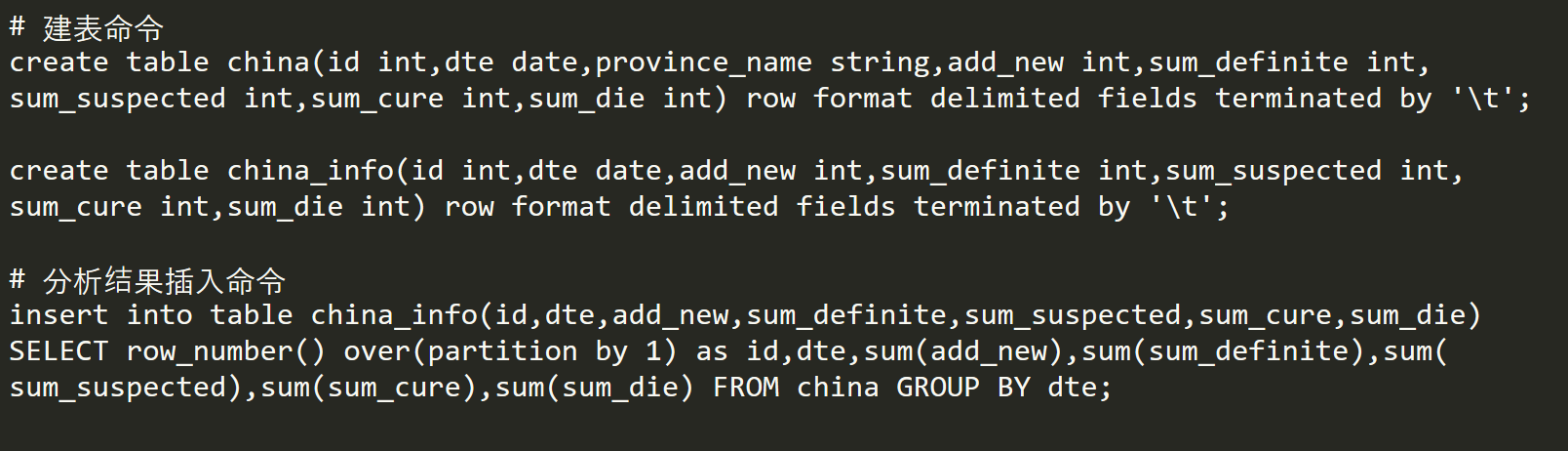
****

**4.5 Django编写后端数据**

****

Django进行前后端分离方式后台主要给前端提供接口，前端JS调用后台的接口，根据接口定义的传参进行传参，得到返回值，然后展现在页面上，或者对数据进行了操作，把操作后的数据传给后端，后端进行数据的更新等

**4.6 sqoop的数据导出和导入**

****

使用sqoop对MySQL进行数据的导入和导出

**4.7 使用elasticsearch**

****

使用elasticsearch进行搜索



elasticsearch同步mysql数据库代码

****

****

数据展览可根据搜索框改变国家疫情数据

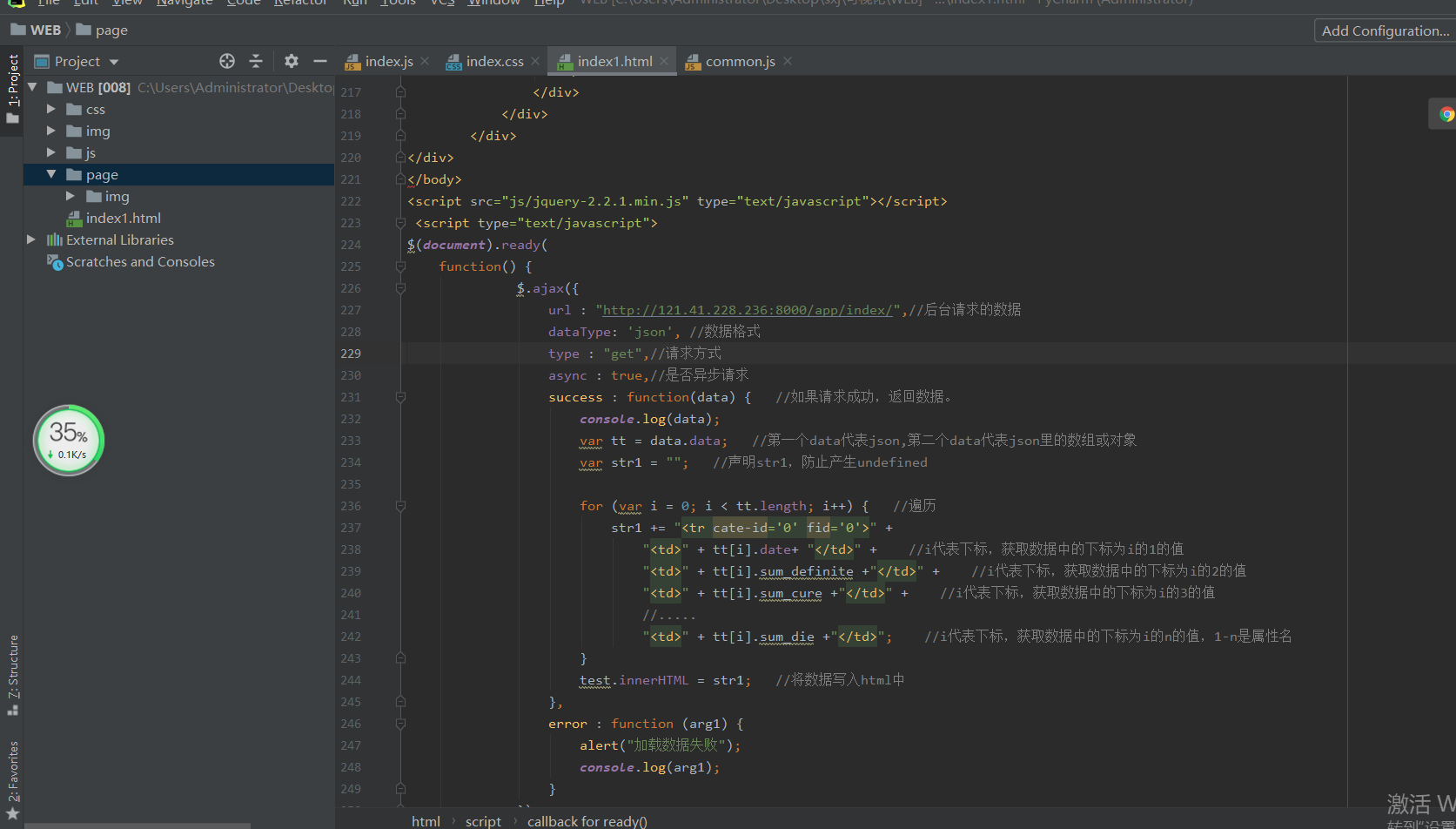
**4.8 前端开发**

网址**：**http://121.41.228.236/

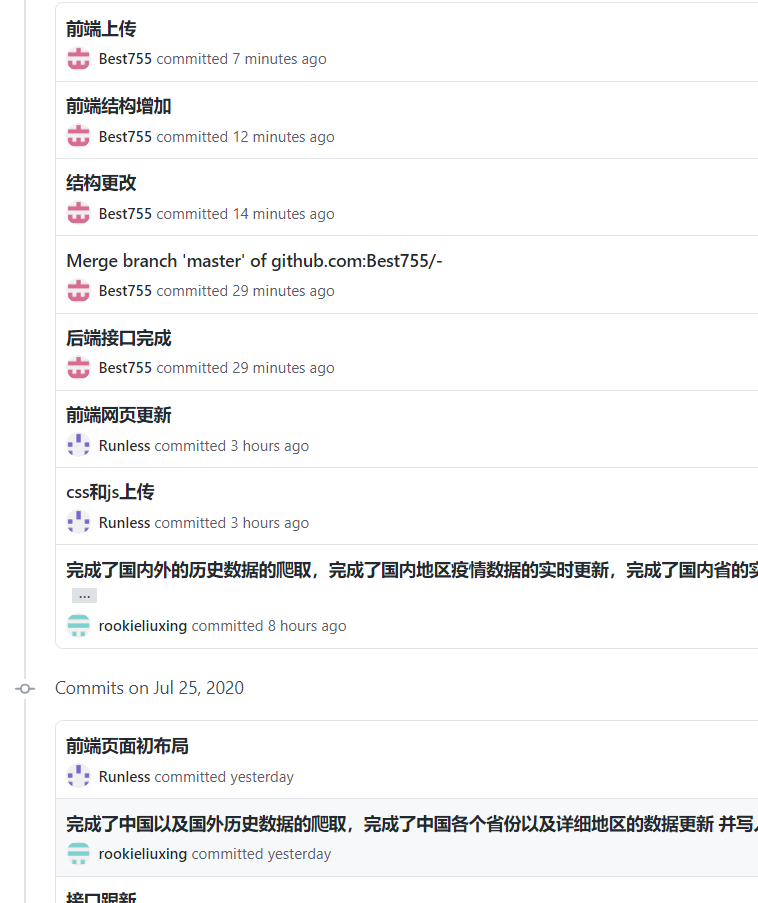
该页通过使用ajax请求后端的接口数据来引入到前端页面的图表，能够在页面上实时的看到疫情的变化

得到中国疫情地图和世界疫情详细数据

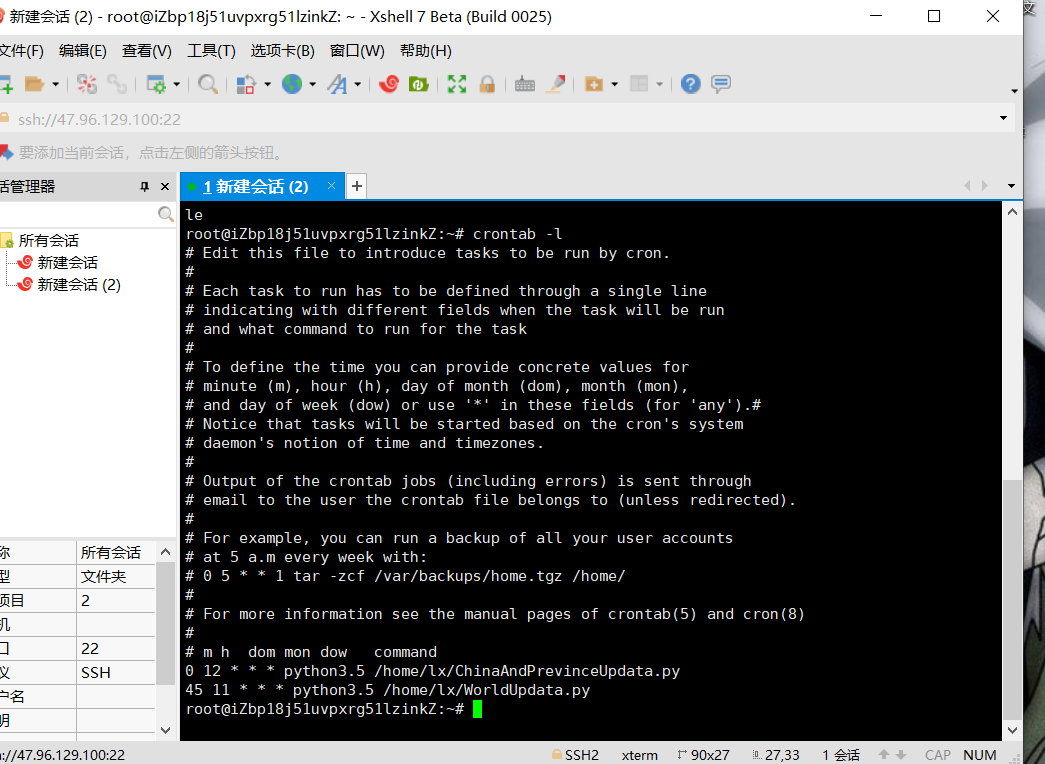




**4.9 github的使用**

****

# 4.10 定时运行爬虫更新数据



服务器上定时运行爬虫更新数据

# 5总结和展望

本次项目的目的是用爬虫爬取数据，hive环境搭建，实时更新数据爬取，Django编写后端接口,基于elasticsearch技术编写的一个搜索框所提供的技术做出一个关于全球疫情的PPT。团队里面分工明确，各施其职。此次做这个全球疫情也是现如今对我们生活的一个小提示，疫情还没有完全褪去，我们也不能放松警惕。直到传染病毒完全消散，才是世界重启的时候。