
# “新冠肺炎”微博话题数据可视化分析

课程名称：大数据处理基础

# 目录

# 1.需求分析

## 1.1 引言

### 1.1.1 背景

2020 对每个人来说都是特殊的一年，因为新冠疫情的爆发，让这个新的一年的开端，蒙上了一层阴影，大家因为疫情都停止外出，赋闲在家，完全依靠手机或电视新闻来了解目前的防疫情况，而官方发布信息的手段也在与时俱进，除了每天固定时间段的新闻会汇总发布，不同地区的人们也会格外关心自己当地的疫情防控情况和社会外界情况以及政府应对措施。而微博是现在大众主流信息获取和交流的途径，拥有着巨额的用户量和阅读量，时效性高，覆盖面广，而微博热搜也是当今社会了解时事和热点话题的有效途径，于是我们小组选择了微博作为平台来进行我们此次项目的数据来源，并对采集到的相关数据进行处理分析，得出人们最关注的有关疫情方面的讨论情况。

### 1.1.2 用户特点

疫情期间，在全国人民都十分关心有关疫情的各类信息的大环境下，还有对疫情相关所包含的真假信息的讨论和分辨言论众多，我们此次项目面向的是全体关心疫情发展动态的用户，我们的数据采取的是实时数据，并且对实时数据进行分析处理，最后采用可视化技术更为直观的展现出来。

## 1.2 需求分析

### 1.2.1 总体需求分析

对爬取到的微博数据进行分析，以图形化的方式直观展现各类信息中人们最关心的疫情问题类别情况，以及展现各类疫情热点问题的被关注讨论的热度。通过数据分析，对采集到的信息做可视化设计，以更加直观易懂的方式展现。

### 1.2.2 需求描述

#### 1.2.2.1 微博数据统计

分别统计采集微博正文的转发量、评论量、点赞量和评论内容。

评论量：累计的微博正文评论量。

转发量：微博正文的转发量。

点赞量：微博正文点赞量。

评论内容：微博的评论内容（包括转发内容）

#### 1.2.2.2 结果展示效能

通过数据处理得到“新冠肺炎”这个关键词发博文的地理位置的可视化图；对于某一种

内容的博文转发量、点赞量和评论量的排行榜图以及各种关键词出现频度的热词分布图。

展示效能：通过百度地图支持得到发博文地理位置图，博文转发量点赞量以及评论量的排行榜柱状图和热词分布图。

# 2.概要设计

## 2.1 引言

### 2.1.1 编写目的

利用比较抽象的语言对整个设计进行概括，确定爬取手段、数据处理流程、外接端口接口、图像显示界面、人机交互界面等的初步设计。

### 2.1.2 背景

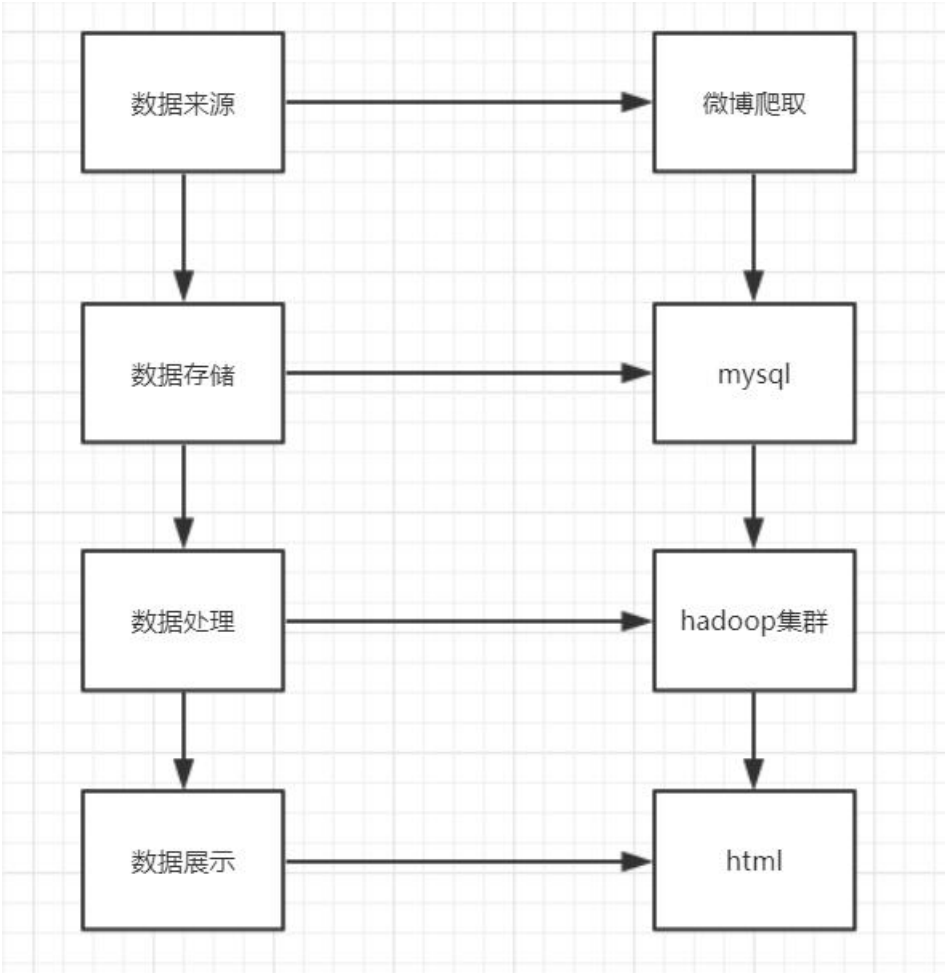
因为新冠疫情的爆发，让这个 2020 的一开始，就显得风雨欲来，大家因为疫情都停止外出，赋闲在家，完全依靠手机或电视新闻来了解目前的防疫情况，而官方发布信息的手段也在与时俱进，除了每天固定时间段的新闻会汇总发布，不同地区的人们也会格外关心自己当地的疫情防控情况和社会外界情况以及政府应对措施。我们选择了微博这个大平台进行数据爬取，然后利用可视化手段将数据进行汇总处理，以更加直观生动的方式展示疫情期间大家最为关心的疫情方面的焦点问题。

### 2.1.3 术语解释

| 术语 | 解释 |
| --- | --- |
| Hadoop | 分布式系统的基础架构 |
| Springboot | 一个通过集成大量的框架使得依赖包的版本冲突，及引用的不稳定性等问题得到了很好的解决的轻量级框架 |
| Mybats | 支持定制化 SQL、存储过程以及高级映射的持久层框架 |
| HTML | 超文本标记语言，用于数据显示 |
| 数据可视化 | 用 SpringMVC+Ajax+Echarts+MySQL 技术，把统计出的数据展示出来 |

## 2.2 系统设计描述

根据前面的功能需求分析和系统整体的需求，将整个系统划分为以下几层：数据来源层、数据存储层、数据处理层和数据展示层。图 3-1 为从数据来源层到数据展示层的结构图：




本系统的总体设计是从微博那里爬取有用的信息，并设计该系统将爬取到的的信息进行加工处理，最后将分析得到的数据显示在前端页面上。在总体设计时都需要考虑到这些方面，还要在进行详细设计时给出相关架构。为了最后实现该系统，在总体设计中把握以下重点，然后在详细设计中可以分别对每个细节难点和重点一一解决。总体设计中包括以下核心的问题，在进行程序详细设计时需要针对这些个难点一一对应的进行分析：

(1) 微博在疫情期间每天产生的数据量是非常庞大的，要从中获取所有的疫情数据并且进行过滤、清洗是一项很复杂的工作，因此我们选择了一个其中包含面最广泛的一个话题进行数据爬取和后续处理，既保证了数据的覆盖广度，也保证了数据的爬取量。

(2) 页面与数据库 MySQL 接口的对接问题。在对数据进行操作时，Web 要与 MySQL 进行交互，它们两者的接口就需要互相统一，并且要有较高的效率。所以在编写 MySQL 接口的时候，需要考虑到它们的性能、效率、优化等问题，让接口符合系统的要求。

## 2.3 各个模块设计描述

### 2.3.1 数据来源模块描述

不管一个系统是多么复杂、多么简单，它们的共同点就是都需要符合要求的数据，这样才可以接下来的操作。也可以这么说，如果一个系统代码是实现系统功能的主要部分，那么数据就是核心。因此该系统的数据也是一个值得关注的部分。该项目的数据来源于微博数据爬取：

主要是微博文章和评论，以及点赞和转发数。

#### 2.3.2 数据存储模块描述

本模块主要有两部分，第一部分是数据源放入到 Hadoop 的 HDFS 分布式文件系统中，第二部分就是把处理好的数据存储到 MySQL 数据库中。

#### 2.3.3 数据处理模块描述

该模块用到的技术是 Hadoop 集群。运用 Hadoop 对 HDFS 上的清洗并进行过滤处理，最后把处理好的数据存储到 MySQL 数据库中。

#### 2.3.4 数据展示模块描述

通过数据可视化技术将分析结果绘制图表和关键词词云展示到 Web 页面，为 html 页面提供支持。

# 3.详细设计描述

## 3.1 引言

本详细设计说明书是针对疫情期间“新冠肺炎”博文内容研究系统而编写，目的是对该系统进行详细设计，在概要设计的基础上进一步明确系统结构，详细介绍各个系统模块，为后面的实现和测试做准备。

## 3.2 集群设计

硬件：64 位计算机，1 台主节点，1 台从节点虚拟机
软件：HDFS、MapReduce、Hive、Sqoop、Oozie、HBase

## 3.3 数据来源模块的详细设计

#### 3.3.1 数据描述

表 1：person

| Id | Int | 序号 |
|---|---|---|
| Name | varchar | 博主昵称 |
| Gender | varchar | 博主性别 |

| Location | varchar | 博主发文地址 |
|---|---|---|
| Birth | varchar | 博主出生日期 |

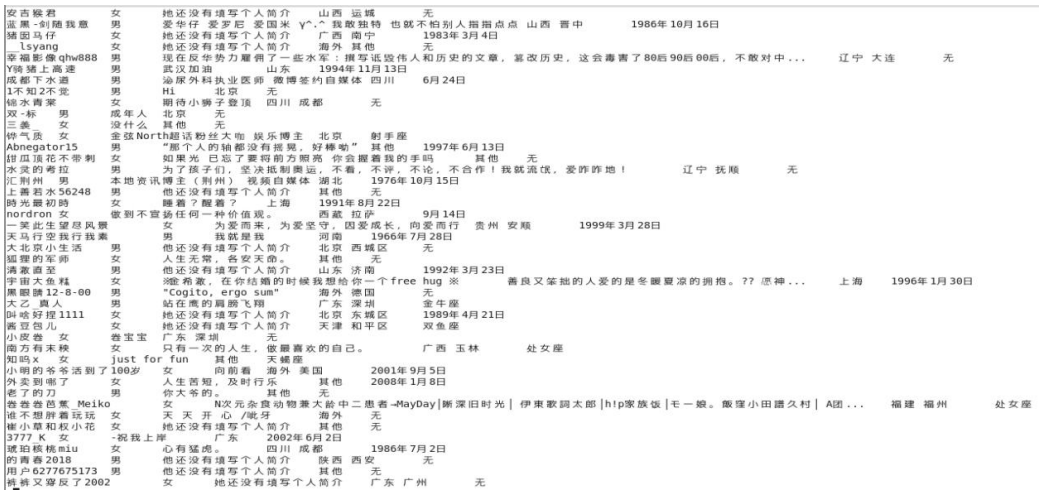
表 2：xinguan

| Id | int | 序号 |
|---|---|---|
| Name | varchar | 博主昵称 |
| User_site | varchar | 博主网址 |
| Content | varchar | 博文内容 |
| Content_site | varchar | 博文网址 |
| Time | varchar | 发文时间 |
| Forward | varchar | 博文转发量 |
| Comment | varchar | 博文评论量 |
| Love | varchar | 博文点赞量 |

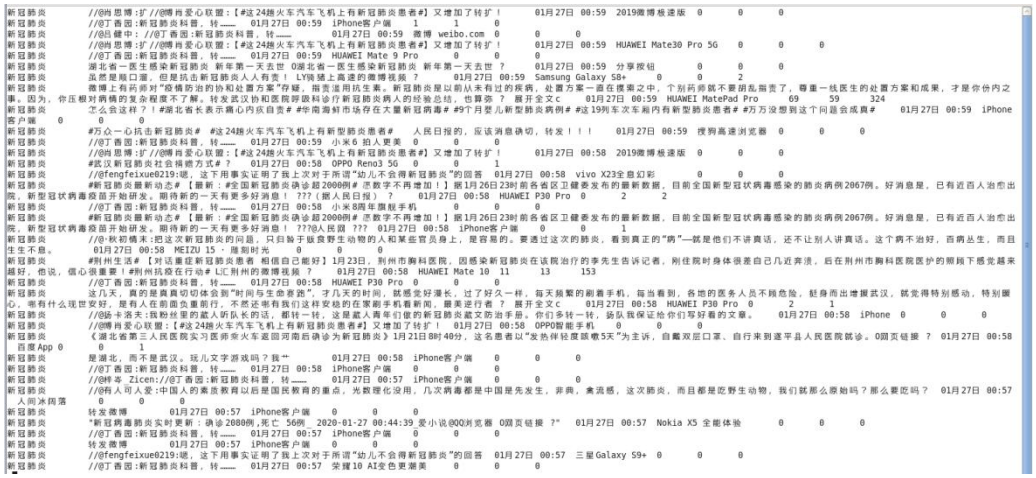
### 3.3.2 目标数据描述

改系统所需数据如下

Person 表数据来源：



Xinguan 表数据来源：



### 3.4 数据存储模块详细设计

该模块的详细设计，首先通过集搜客 app 爬取微博数据，即通过关键词“新冠肺炎”爬取一月二十七日至一月三十一日的微博的博主、博文、网址、发文地址等信息；其次对爬取的 excel 数据进行转码导入到虚拟机 CentOs 中，并上传到 HDFS 进行处理，将处理完的数据下载到本地之后通过 sqlyog 和 mysql 配合构建后台数据库。

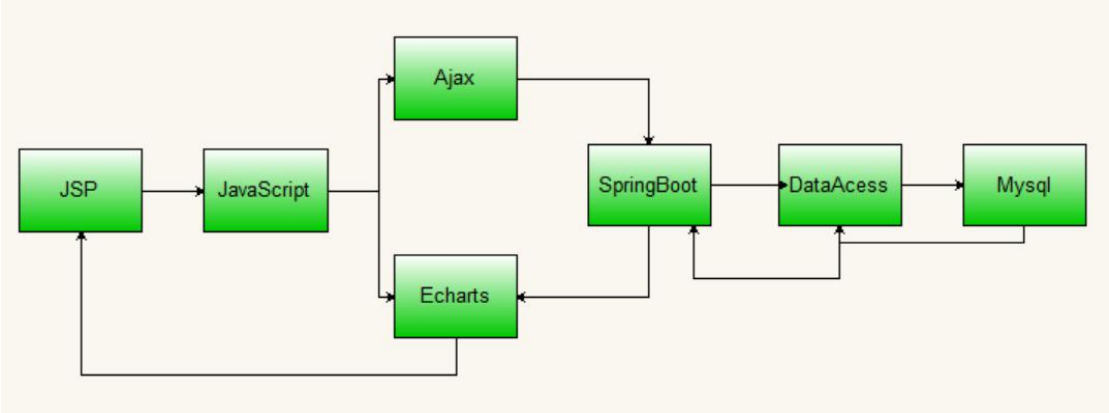

### 3.5 数据处理模块详细设计

数据处理分为三个大阶段：
第一次处理包括三个部分：数据预处理、构建数据仓库、数据分析。
第二大阶段分为：将数据导入到 mysql、Hbase shell 操作、将数据导入到 Hbase 和 Hbase java API 操作。
第三阶段分为：对微博全文内容分词，并进行词频统计找出热点词，最后组合成热词分布图：
1、 python 进行预处理
使用 python 对数据进行清洗和过滤。过滤掉非中文字符及特定字符。
2、 python 方法生成词云
使用 ps 等工具对某一图片进行处理，使其形状轮廓为白底黑色。使用 python 中的 jieba、wordcloud、nump 等库对处理后的文本进行处理，设置色系，按照刚刚处理好的图片轮廓形状生成词云。
3、 词频统计
使用 python 进行分词和统计处理，统计出数据中高频词汇。将数据导入数据库，使用 echarts 模板生成可视化文件。

## 3.6 数据展示模块详细设计

### 3.6.1 数据可视化

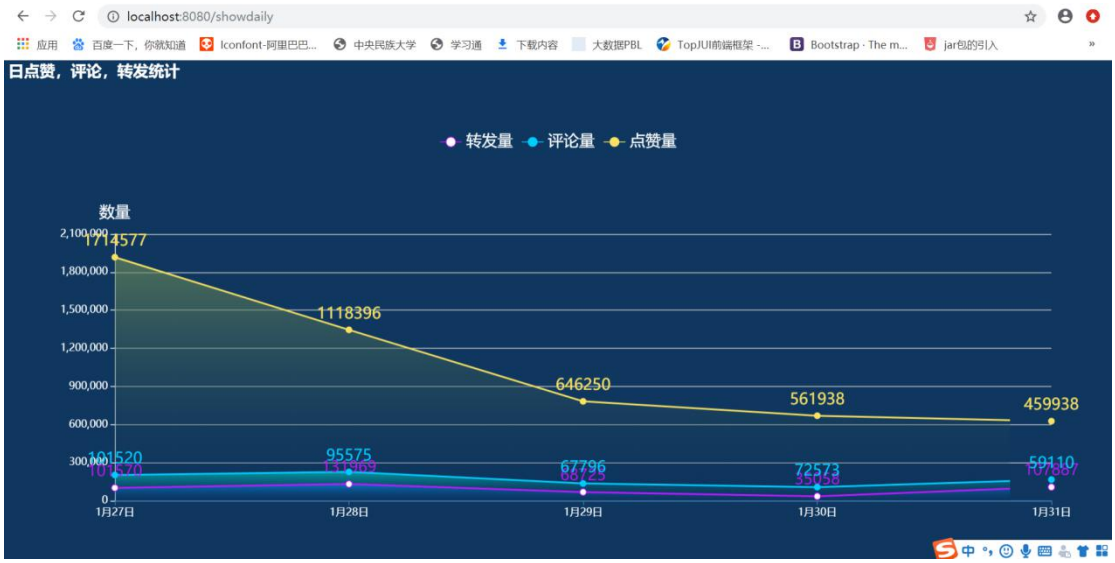
本次系统的设计我们运用的是 springboot 框架：
在 HTML 中内嵌 JavaScript，然后在 JavaScript 中写入 Ajax，Ajax 中的 URL 指定到 Springboot 的 Controller 中的函数接口，执行数据的获取 mapper，mapper 通过 MySQL 的连接从 MySQL 中获取到数据。若以上步骤成功，Ajax 则会获得一个 JSON 数据并传输到 Echarts 上。最终 Echarts 在 JSP 上显示，如图：




Springboot：它基于 Spring4.0 设计，不仅继承了 Spring 框架原有的优秀特性，而且还通过简化配置来进一步简化了 Spring 应用的整个搭建和开发过程。其次它可以创建独立的 Spring 应用程序，并且基于其 Maven 或 Gradle 插件，可以创建可执行的 JARs 和 WARs，同时它内嵌 Tomcat 或 Jetty 等 Servlet 容器。
JavaScript：是一种属于网络的高级脚本语言,已经被广泛用于 Web 应用开发,常用来为网页添加各式各样的动态功能,为用户提供更流畅美观的浏览效果。通常 JavaScript 脚本是通过嵌入在 HTML 中来实现自身的功能的。
Ajax：即“**A***synchronous* **J***avascript* ***A****nd* **X***ML*”（异步 JavaScript 和 XML），是指一种创建交互式、快速动态网页应用的网页开发技术，无需重新加载整个网页的情况下，能够更新部分网页的技术。

### 3.6.2 界面设计


localhost:8080/showhot
三个
流感
冠状病毒
患者
丁香园
接收
症状
感染
新冠肺炎
确诊病例
一群
新型肺炎
武汉加油
求助
万众一心
感冒
朋友
新型冠状病毒
小朋友
肺炎疫情
"新冠疫情"主题图谱


localhost:8080/showdaily
日点赞，评论，转发统计
转发量
评论量
点赞量
数量
2,100,000
1,800,000
1,500,000
1,200,000
900,000
600,000
300,000
0
1714577
1118396
646250
561938
459938
101520
95575
67796
72573
59110
1月27日
1月28日
1月29日
1月30日
1月31日


localhost:8080/showzhuzhuang
评论量排行榜
First
23463
Second
10178
Third
8907
0
3,000
6,000
9,000
12,000
15,000
18,000
21,000
24,000
点赞量排行榜
评论量排行榜
转发量排行榜

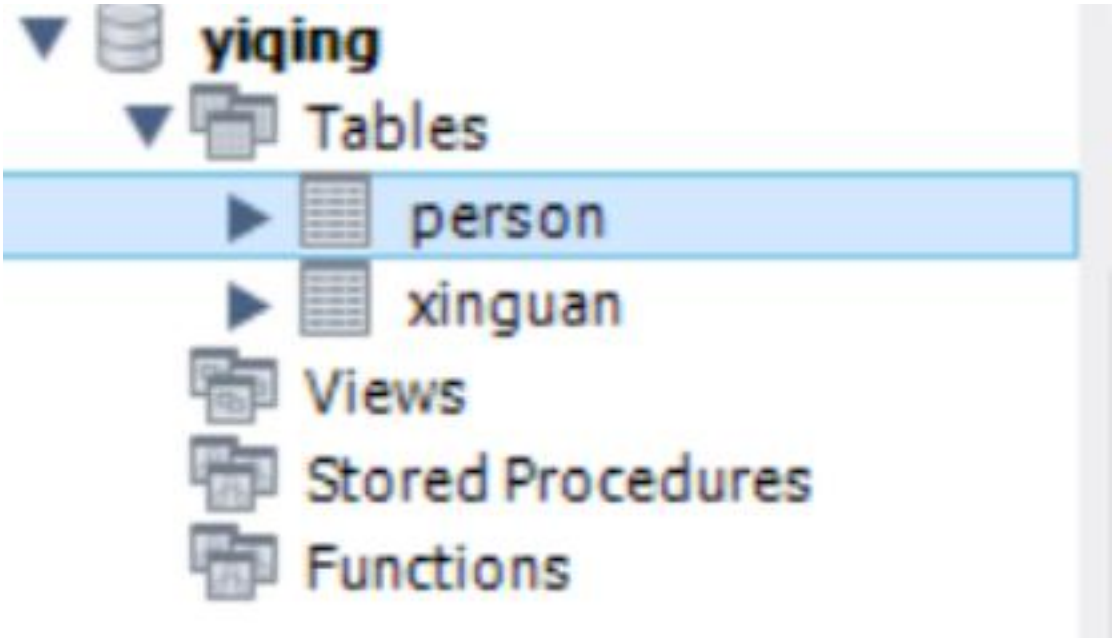


# 4.编码文档

## 4.1 数据分析

### 4.1.1 数据库脚本分析

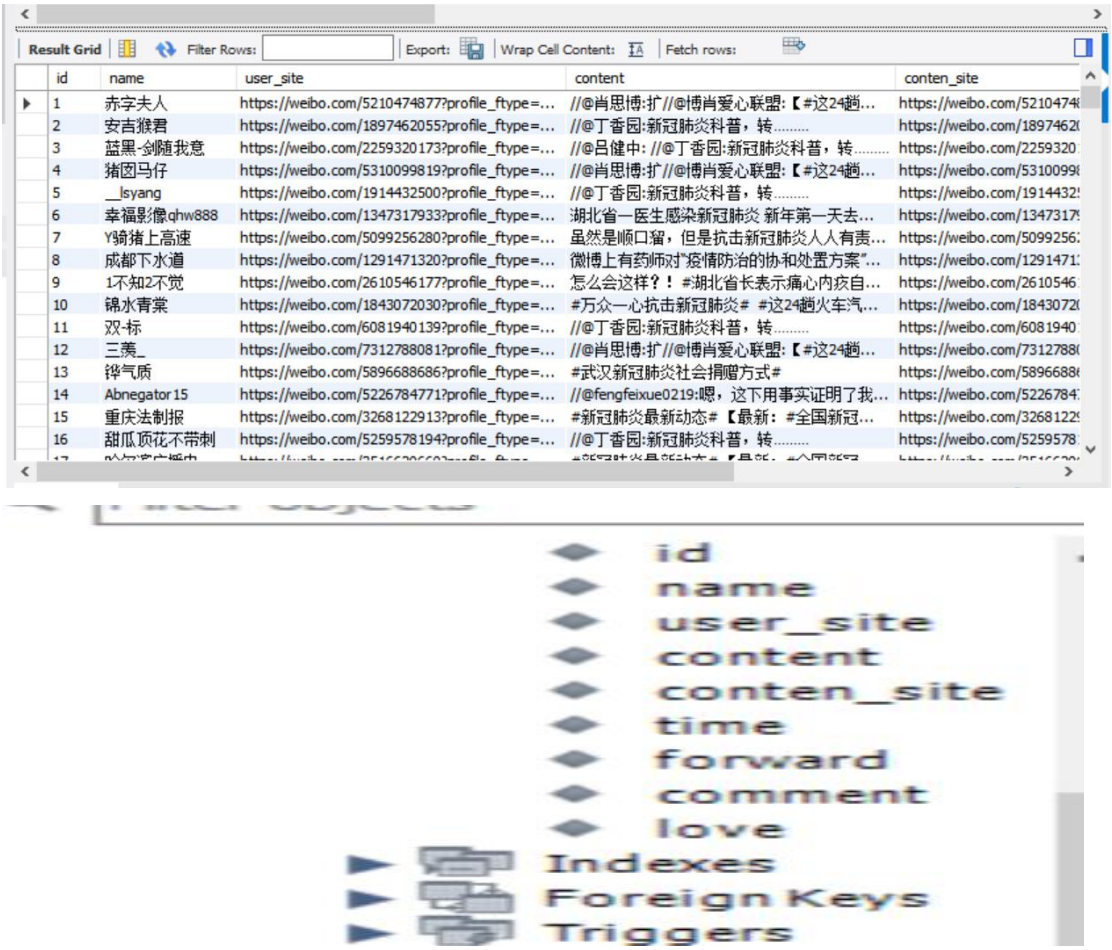
将爬取到的数据转换为表格的形式导入到了 SQL 中
数据库脚本文件：yiqing.sql
建立数据库：yiqing
数据库包含两个表：person,xinguan




person表：存放了微博用户的用户名，性别，址和生日等信息

| id | name | gender | location | birth |
|---|---|---|---|---|
| 1 | 安吉猴君 | 女 | 山西 运城 | 无 |
| 2 | 蓝黑-剑随我意 | 男 | 山西 晋中 | 1986年10月16日 |
| 3 | 猪囡马仔 | 女 | 广西 南宁 | 1983年3月4日 |
| 5 | 幸福影像qhw888 | 男 | 辽宁 大连 | 无 |
| 6 | Y骑猪上高速 | 男 | 山东 | 1994年11月13日 |
| 7 | 成都下水道 | 男 | 四川 | 6月24日 |
| 8 | 1不知2不觉 | 男 | 北京 | 无 |
| 9 | 锦水青棠 | 女 | 四川 成都 | 无 |
| 10 | 双-标 | 男 | 北京 | 无 |
| 12 | 铮气质 | 女 | 北京 | 射手座 |
| 15 | 水灵的考拉 | 男 | 辽宁 抚顺 | 无 |
| 16 | 汇荆州 | 男 | 湖北 | 1976年10月15日 |
| 18 | 時光最初時 | 女 | 上海 | 1991年8月22日 |
| 19 | nordron | 女 | 西藏 拉萨 | 9月14日 |
| 20 | 一笑此生望尽… | 女 | 贵州 安顺 | 1999年3月28日 |
| 21 | 天马行空我行… | 男 | 河南 | 1966年7月28日 |
| 22 | 大北京小生活 | 男 | 北京 西… | 无 |

xinguan 表：存放了微博用户用户名，用户地址，用户发布相关疫情的内容，内容地址，转发量，评论量，点赞量。



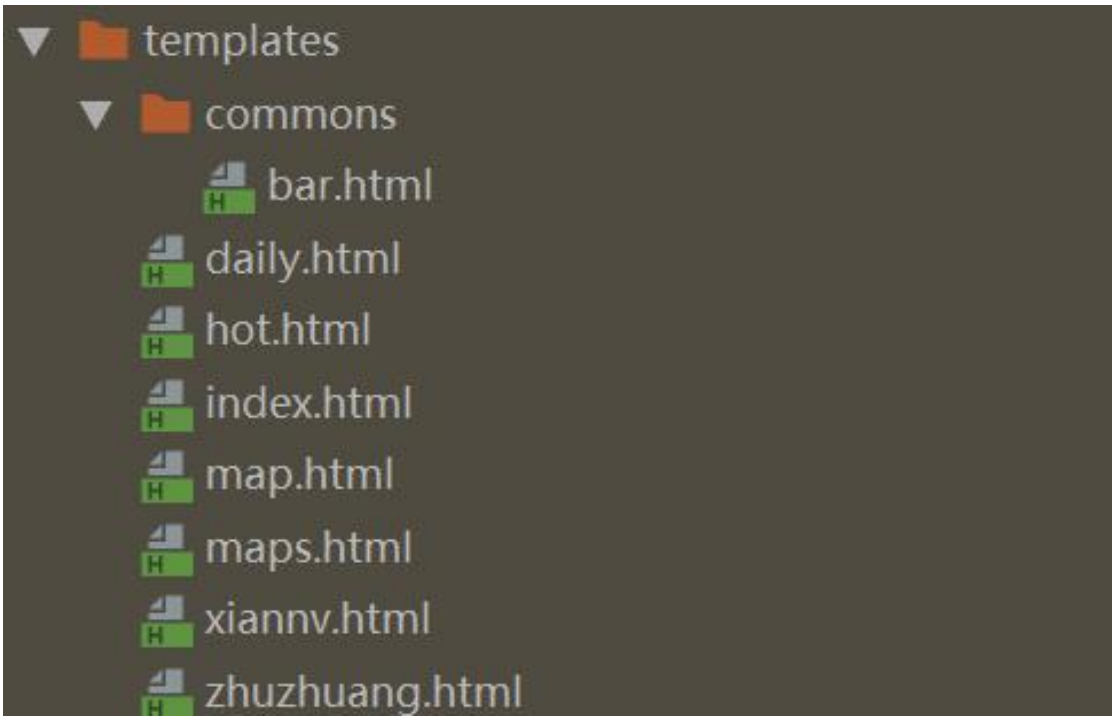
Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

| id | name | user_site | content | conten_site |
|---|---|---|---|---|
| 1 | 赤字夫人 | https://weibo.com/5210474877?profile_ftype=... | //@肖思博:扩//@博肖爱心联盟:【#这24趟... | https://weibo.com/5210474 |
| 2 | 安吉猴君 | https://weibo.com/1897462055?profile_ftype=... | //@丁香园:新冠肺炎科普，转..... | https://weibo.com/1897462 |
| 3 | 蓝黑-剑随我意 | https://weibo.com/2259320173?profile_ftype=... | //@吕健中: //@丁香园:新冠肺炎科普，转..... | https://weibo.com/2259320 |
| 4 | 猪囡马仔 | https://weibo.com/5310099819?profile_ftype=... | //@肖思博:扩//@博肖爱心联盟:【#这24趟... | https://weibo.com/5310099 |
| 5 | _lsyang | https://weibo.com/1914432500?profile_ftype=... | //@丁香园:新冠肺炎科普，转..... | https://weibo.com/1914432 |
| 6 | 幸福影像qhw888 | https://weibo.com/1347317933?profile_ftype=... | 湖北省一医生感染新冠肺炎 新年第一天去... | https://weibo.com/1347317 |
| 7 | Y骑猪上高速 | https://weibo.com/5099256280?profile_ftype=... | 虽然是顺口溜，但是抗击新冠肺炎人人有责... | https://weibo.com/5099256 |
| 8 | 成都下水道 | https://weibo.com/1291471320?profile_ftype=... | 微博上有药师对"疫情防治的协和处置方案"... | https://weibo.com/1291471 |
| 9 | 1不知2不觉 | https://weibo.com/2610546177?profile_ftype=... | 怎么会这样？！ #湖北省长表示痛心内疚自... | https://weibo.com/2610546 |
| 10 | 锦水青棠 | https://weibo.com/1843072030?profile_ftype=... | #万众一心抗击新冠肺炎# #这24趟火车汽... | https://weibo.com/1843072 |
| 11 | 双-标 | https://weibo.com/6081940139?profile_ftype=... | //@丁香园:新冠肺炎科普，转..... | https://weibo.com/6081940 |
| 12 | 三羡_ | https://weibo.com/7312788081?profile_ftype=... | //@肖思博:扩//@博肖爱心联盟:【#这24趟... | https://weibo.com/7312788 |
| 13 | 锌气质 | https://weibo.com/5896688686?profile_ftype=... | #武汉新冠肺炎社会捐赠方式# | https://weibo.com/5896688 |
| 14 | Abnegator15 | https://weibo.com/5226784771?profile_ftype=... | //@fengfeixue0219:嗯，这下用事实证明了我... | https://weibo.com/5226784 |
| 15 | 重庆法制报 | https://weibo.com/3268122913?profile_ftype=... | #新冠肺炎最新动态#【最新：#全国新冠... | https://weibo.com/3268122 |
| 16 | 甜瓜顶花不带刺 | https://weibo.com/5259578194?profile_ftype=... | //@丁香园:新冠肺炎科普，转..... | https://weibo.com/5259578 |

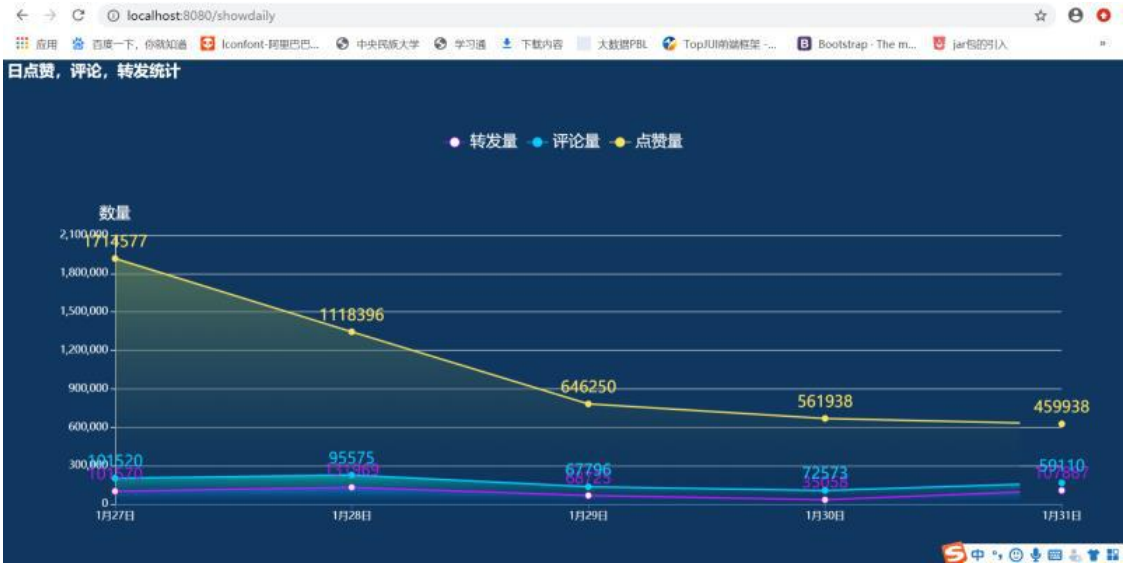



## 4.2 数据可视化

### 4.2.1 html 展示界面



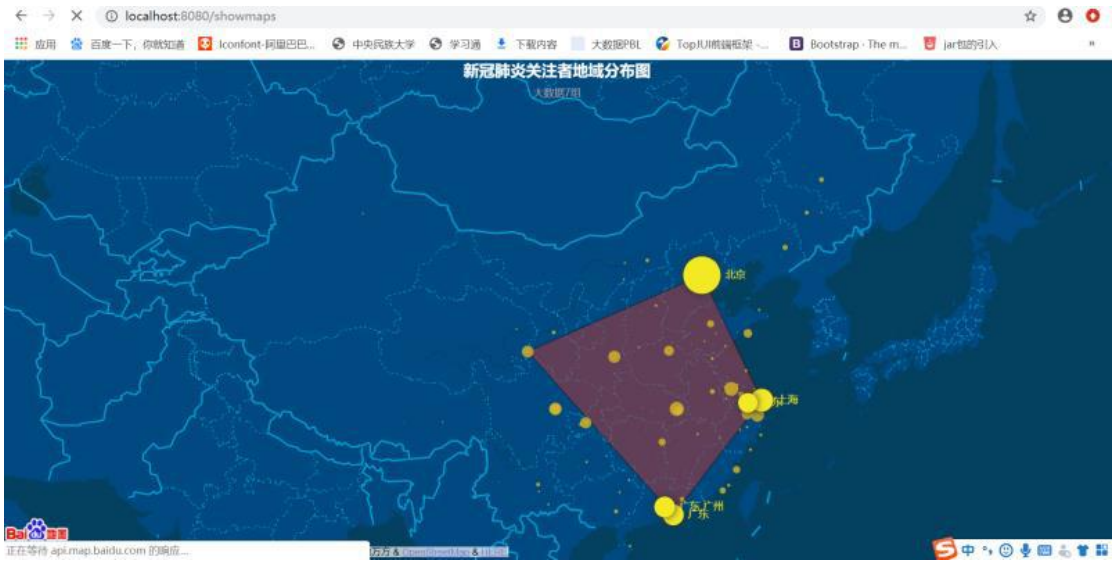

Daily.html 为日点赞，评论，转发量统计界面：统计了 1 月 27 日到 31 日对于新冠肺炎相关话题评论、点赞、转发量。




Hot.html:“新冠疫情”主题图谱界面：统计提取了新冠疫情相关话题的主题


Zhuzhuang.html：评论量排行榜界面：选取并显示前三的评论量。

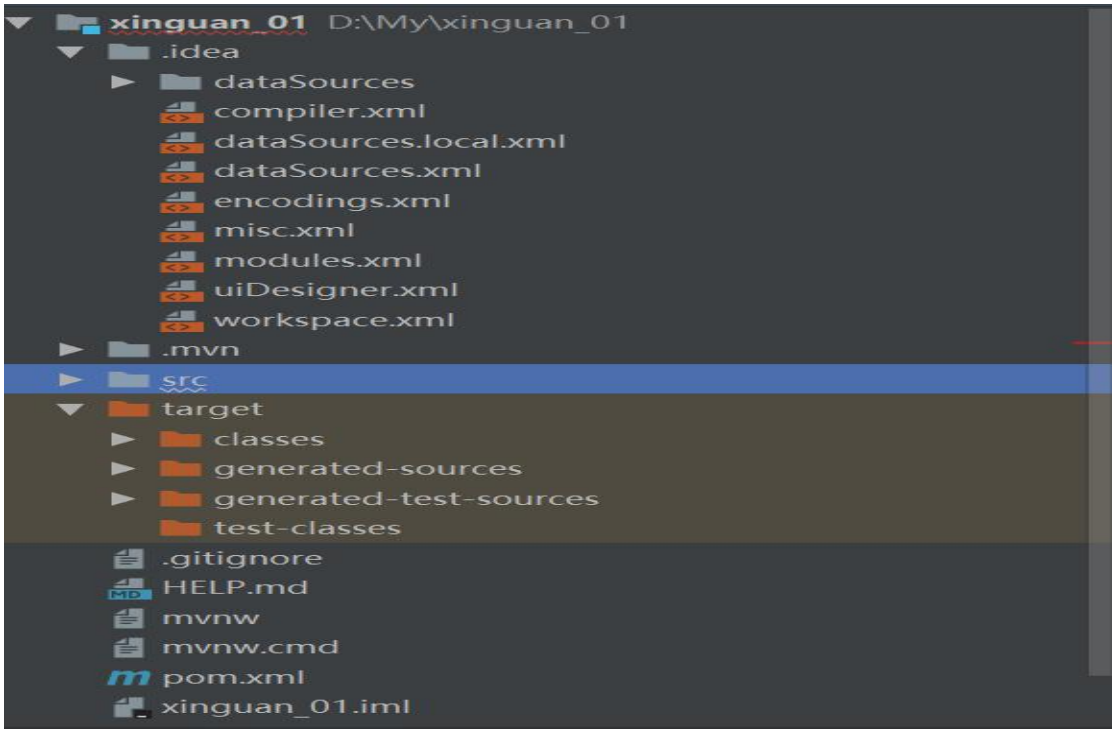

Maps.html:新冠肺炎肺炎关注者地域分布图：即关注新冠肺炎话题的微博用户所在地区分布。



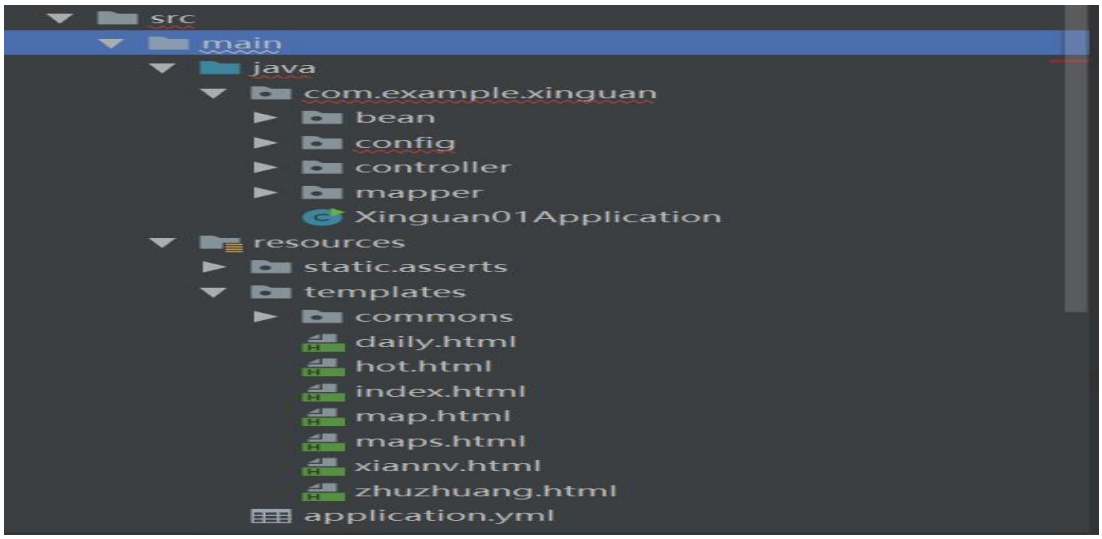

Xiannv.html：新冠肺炎相关话题关键字即词频图


### 4.2.2 包

目录：



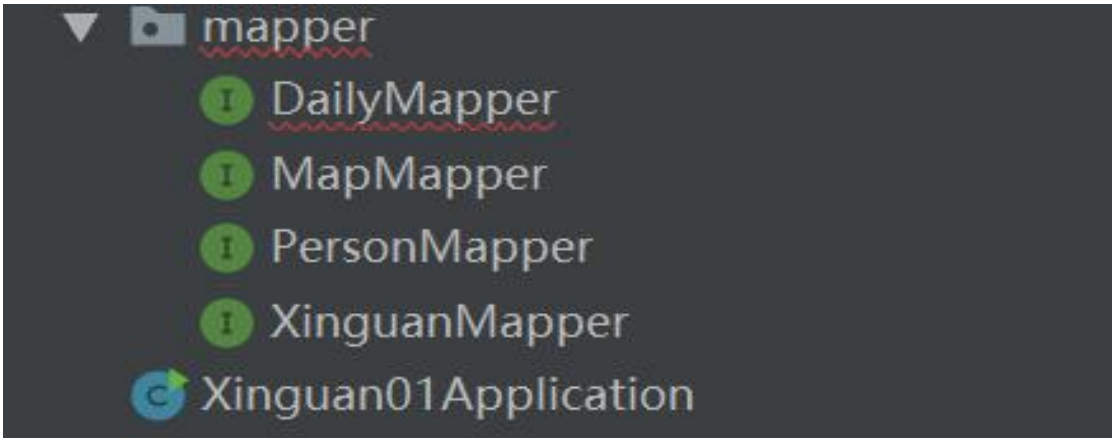

包：




### 4.2.3 类

Controller:实现页面跳转、获取 mapper 返回的数据，html 获取此数据

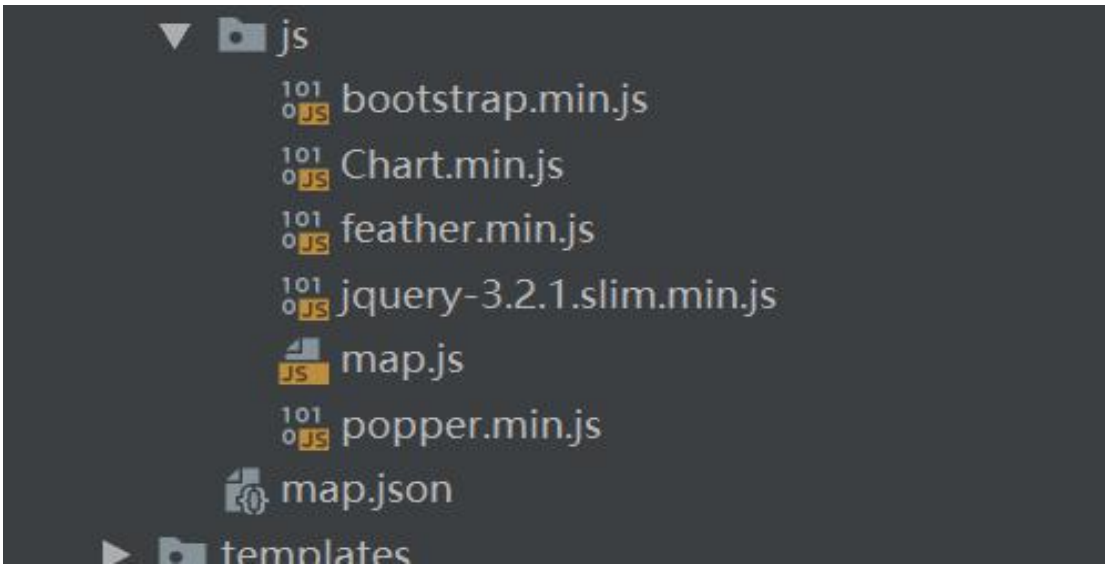

Mapper:从数据库获取我们要的数据并返回给列表




Bean:实体类


bean
Daily
Map
Person
Xinguan

Js lib:


js
bootstrap.min.js
Chart.min.js
feather.min.js
jquery-3.2.1.slim.min.js
map.js
popper.min.js
map.json
templates