疫情问答检索系统v1.0

系统设计文档

日期：2020年6月24日

**目录**

[1. 数据准备 2](#_Toc43918135)

[2. 架构设计 4](#_Toc43918136)

[3. 结果分析 5](#_Toc43918137)

[3.1讨论方向 5](#_Toc43918138)

[3.2问答分析 5](#_Toc43918139)

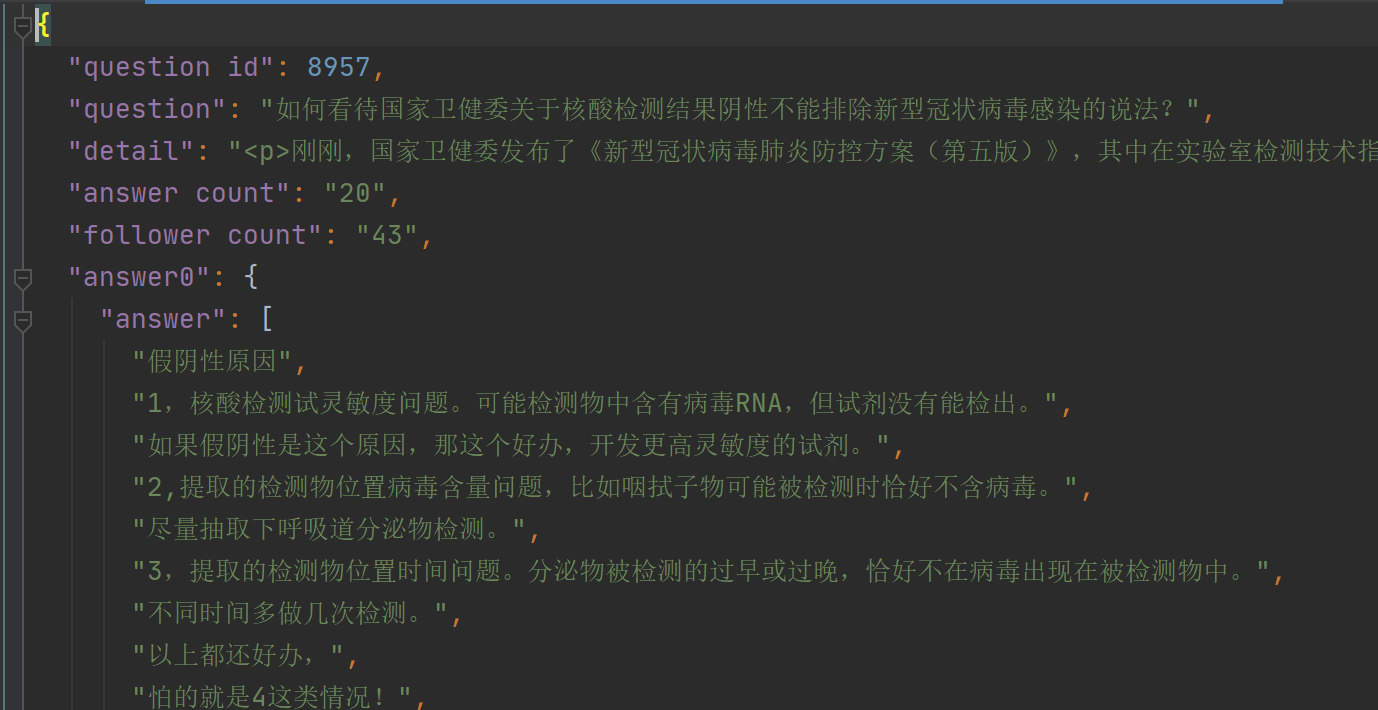
[4. 问题与解决方案 6](#_Toc43918140)

# 数据准备

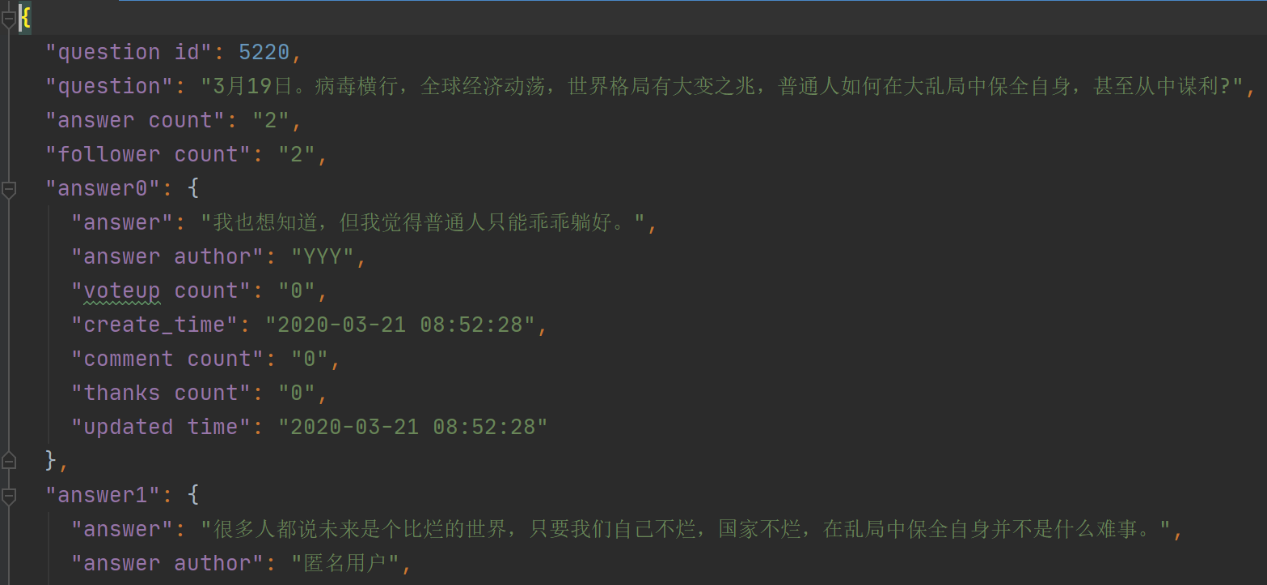
我们的任务是通过爬虫自动采集知乎网站上新冠肺炎疫情相关问题及回答，通过课程上所学知识构建新冠肺炎问答检索系统，检索返回结果可以按多种方式排序，如时间、问题相关度、被赞数量等。完成这个疫情问答系统的构建需要大量的数据作为支撑。经过调研，我们采用爬虫对所需数据进行获取。在调研过程中发现了Zhihu-OAuth库，Zhihu-OAuth库使用方法简单，支持 Python 2 和 Python 3，其中能爬取的数据非常全面丰富，包括：用户动态、答案、文章、收藏夹、专栏、评论、首页信息流、知乎live、特定用户的信息、分享、问题、话题、私信会话和私信信息等等全部囊括在内。可以直接用pip命令装载。

本次爬取的知乎话题为“新型冠状病毒肺炎”，话题ID为21239580。最终爬取的数据放在“zhihu”下的“新型冠状病毒肺炎”文件夹内，以JSON形式存储。我们最终爬取了“新型冠状病毒肺炎”话题下的12477条问题数据，作为后续可视化展示的数据支撑。

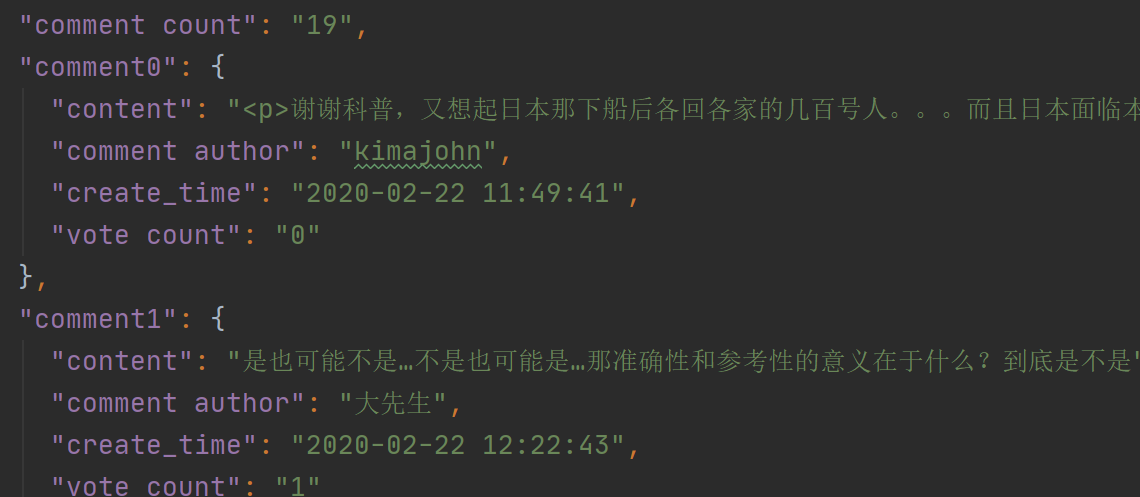
每个问题数据包含问题名称、问题详情、时间、回答数量、关注数以及若干条答案数据。以下是问题部分的JSON数据格式：



每个问题中包含若干回答，每个答案包括回答内容、回答者、点赞数、感谢数、回答时间、更新时间以及若干评论。以下是回答部分的JSON数据格式：



每个回答中包含若干评论，每条评论包括评论内容、评论者、评论时间、点赞数。以下是评论部分的JSON数据格式：



在使用爬虫对数据进行爬取时，文本数据中会出现空白字符、特殊字符、超链接等，这些符号对后续的可视化展示会有影响，所以需要进行数据清洗。在数据清洗过程中，我们清洗了JSON数据中的文本数据，分别去除问题，答案，以及评论内容中的空白字符，特殊字符，超链接，以及标签等。

此外我们的系统展示中设计了每条问题的关键词展示，所以在数据存入MySQL数据库之前，调用了关键词提取算法，取出回答详情中的回答数据中的关键词。在检索系统的排序功能中，我们设置了相关度排序，所以在数据入库之前还调用了相关度计算算法，计算用户回答与问题详情之间的距离，计算的数值同关键词以及问题基本数据一起存入MYSQL数据库。

经数据清洗，共有7413条问题数据，122151条回答数据，589388条评论数据作为有效数据存入数据库中，供可视化展示使用。

# 架构设计

我们的系统选择使用Django开发框架，Django是一款python的web开发框架，与MVC有所不同，属于MVT框架。其中，m表示model，负责与数据库交互；v 表示view，是核心，负责接收请求、获取数据、返回结果；t 表示template，负责呈现内容到浏览器。MVT框架结构如图3-1所示：

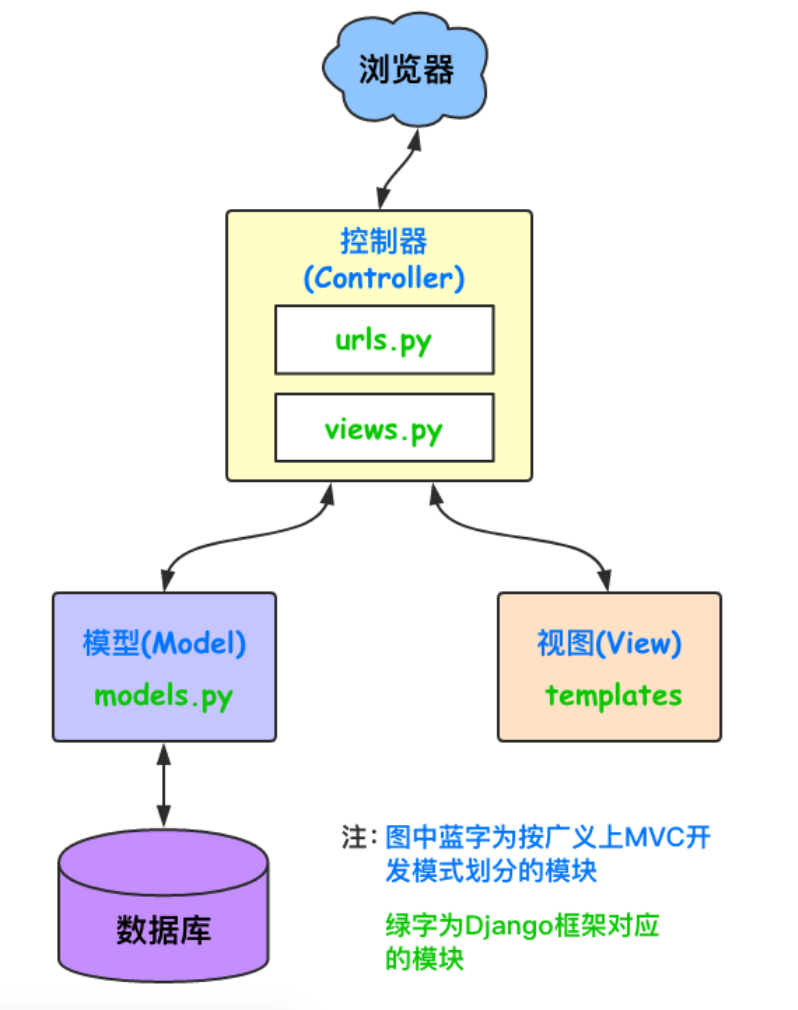


图 3-1 MVT框架结构图

在MVT框架中，客户端请求信息交给视图函数处理，当视图需要数据的时候调用模型Model，然后去数据库拿数据，接着返回给模型Model，再最后返回给视图。模型Model中不编写任何SQL语句。在关系数据库中的一行，相当于模型Model中的一个对象。在关系数据库中的一个表，相当于一个对象的集合。在python中用列表定义关系数据库中的一个表，我们直接在对象中写列表的操作就可以了，ORM会转换成SQL语句。之后我们得到一个列表对象返回给视图View, 视图View会进行一些处理，处理的结果需要用模板Templates呈现（渲染HTML和CSS执行JS），然后返回客户端。

# 结果分析

## 3.1讨论方向

根据爬虫结果，发现大家提出的问题主要集中在如下几个方面：疫情的源头、疫情相关症状与防疫的情况、新冠病毒的原理、疫情情况下催生的新生社会模式、疫情期间居家情况，例如粉丝数最高的问题是“面对疫情足不出户，有哪些可以迅速学会的菜谱吗？”，关于国家应对疫情的措施的讨论，例如“如何看待武汉在三个区连夜开辟方舱医院？”，对于疫情防治情况的讨论，对于国外疫情的讨论等等。

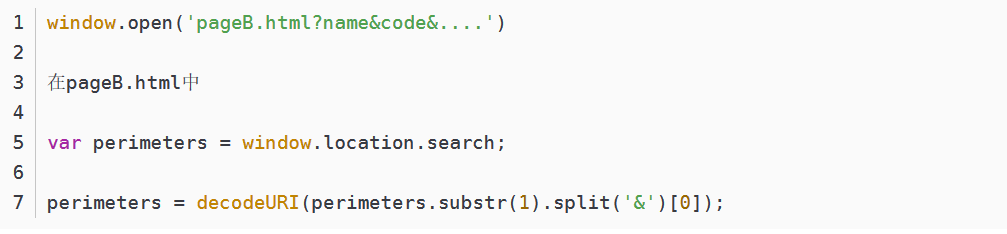
## 3.2问答分析

根据大家的讨论，可以看出大家在疫情期间并没有显示出恐慌、暴动的情绪，积极响应国家号召，进行居家隔离，戴口罩等防控措施，并同时对疫情重灾地区情况表达高度关注与重视。这同时在侧面强有力的反应了中国人民在党的领导下团结一心，众志成城，一起抗疫，度过难关，高度表现了中华民族的强大凝聚力。

# 问题与解决方案

#### 两个不同的html界面之间传参

1. 方法一：



1. 方法二：

方法一的缺点是当有一个以上的值需要传递的时候，需要准确记录每个值的位置。而本方法是以名称值对的方式传递，比较友好。从地址栏获取参数，name(类型string)为要获取的参数名。



1. 方法三：localStorage或者sessionStorage本地缓存

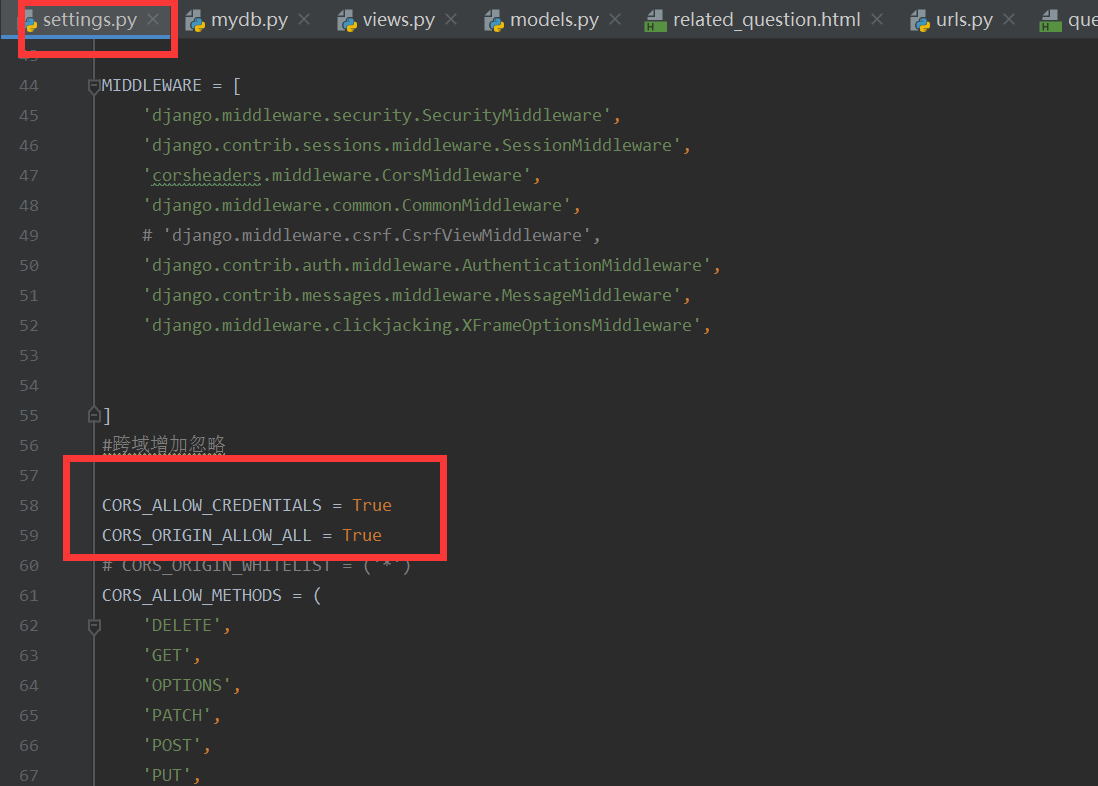


#### 路径存在非法参数

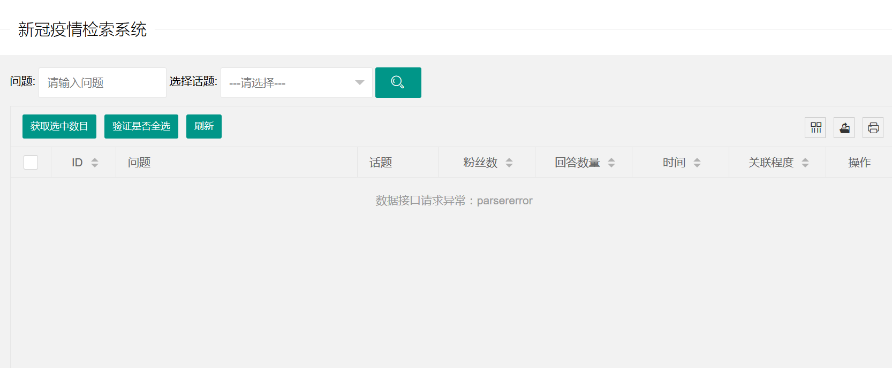
当存储路径中包含英文字符的时候，会出现路径非法。解决方法是利用replace()方法将英文字符转换成中文字符，问题解决。

#### Django解决跨域请求的问题

解决方案是在项目的setting.py文件中，设置“跨域请求忽略”。



#### 内部数据接口异常



解决方案是contentType为application/json时，Django不支持request.POST.get()，但可以通过request.body来获取string类型的参数。

#### layui框架中表格(table)中实现按照星级评分(rate)排序(sort)

解决方案是设置全局变量，进行sort排序。

1、在表单中设置rate的关键点有两个：

(1) 在表单中，设置templet函数转义，返回一个div；

(2)在回调函数done:function()里面进行rate.render操作，对rate赋值。

2、利用sort对星级评分rate排序，且不重载表单，关键点也有三个：

(1)设置一个全局变量，记录表格渲染的值（在回调函数中可以取得这个表单内容）；

(2)在sort中的table.reload()的第一个参数是table的elem值，这样reload的就是DOM了；

(3)在sort的监听事件中，设置for循环，根据全局变量来给rate赋值。