数据虫巢说明

【实现技术】

**一、Maven**

[Maven](http://baike.baidu.com/view/336103.htm)项目对象模型(POM)，可以通过一小段描述信息来管理项目的构建，报告和文档的软件[项目管理工具](http://baike.baidu.com/view/3169783.htm)。当时有一些项目（有各自Ant build文件），仅有细微的差别，而JAR文件都由[CVS](http://baike.baidu.com/view/20982.htm)来维护。于是希望有一种标准化的方式构建项目，一个清晰的方式定义项目的组成，一个容易的方式发布项目的信息，以及一种简单的方式在多个项目中共享JARs。

二**、Spring Boot**

Spring Boot是由Pivotal团队提供的全新框架，其设计目的是用来简化新Spring应用的初始搭建以及开发过程。该框架使用了特定的方式来进行配置，从而使开发人员不再需要定义样板化的配置。通过这种方式，Spring Boot致力于在蓬勃发展的快速应用开发领域（rapid application development）成为领导者。它还有一些其他的优点，例如**对MVC的支持、外部Properties的注入等。**

**三、Echarts**

ECharts，一个纯 JavaScript 的图表库，可以流畅的运行在 PC 和移动设备上，兼容当前绝大部分浏览器（IE8/9/10/11，Chrome，Firefox，Safari等），底层依赖轻量级的 Canvas 类库[ZRender](https://github.com/ecomfe/zrender)，提供直观，生动，可交互，可高度个性化定制的数据可视化图表。

ECharts 3 中更是加入了更多丰富的交互功能以及更多的可视化效果，并且对移动端做了深度的优化。

**四、WebCollector**

WebCollector是一个无须配置、便于二次开发的JAVA**爬虫框架（内核）**，它提供精简的的API，只需少量代码即可实现一个功能强大的爬虫。WebCollector-Hadoop是WebCollector的Hadoop版本，支持分布式爬取。

**五、ANSJ分词**

这是一个[ictclas](http://www.oschina.net/p/freeictclas)的java实现.基本上重写了所有的数据结构和算法.词典是用的开源版的ictclas所提供的.并且进行了部分的人工优化，内存中中文分词每秒钟大约100万字(速度上已经超越ictclas)，文件读取分词每秒钟大约30万字，准确率能达到96%以上，目前实现了

. 中文分词

. 中文姓名识别

. 用户自定义词典

可以应用到自然语言处理等方面,适用于对分词效果要求高的各种项目.

**六、MySQL**

MySQL是一个[关系型数据库管理系统](http://baike.baidu.com/view/1450387.htm)**，**MySQL 是最流行的[关系型数据库管理系统](http://baike.baidu.com/view/1450387.htm)之一，关系数据库将数据保存在不同的表中，而不是将所有数据放在一个大仓库内，这样就增加了速度并提高了灵活性。

【环境配置】

操作系统：win10

Maven：3.3.9

Jdk：1.7

Mysql：5.7.17

Eclipse：4.3

QQ浏览器：9.5.4

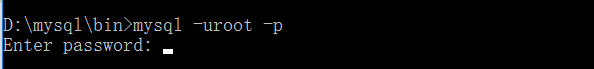
【操作步骤】

**一、 导入项目**

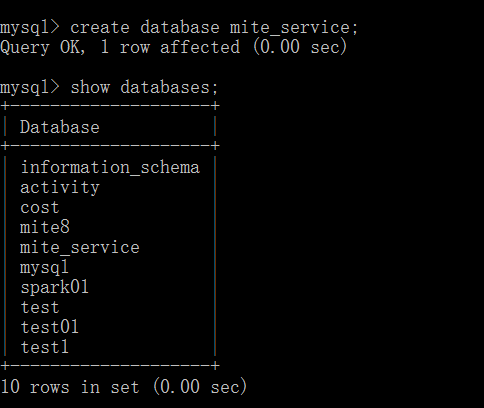
1.1从https://github.com/blogchong/mite8-com下载项目

1.2将解压好的文件夹mite8-com-master\SQL\_DATA中的数据导入到mysql数据库

使用命令行进入到你安装mysql的bin目录下进入数据库。



进入后创建数据库并且查看是否成功。

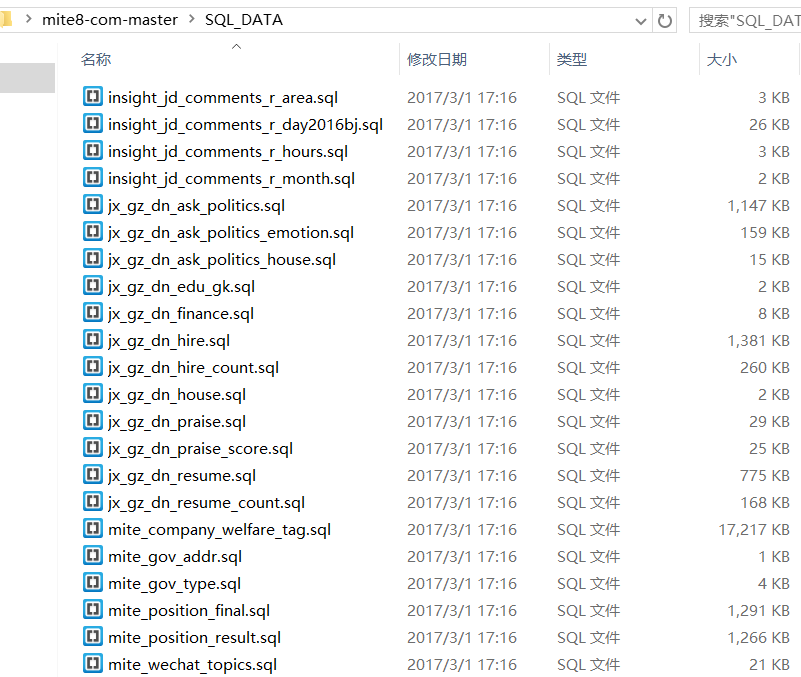


切换到创建的数据库并且使用source加上SQL的文件路径就可以导入数据库

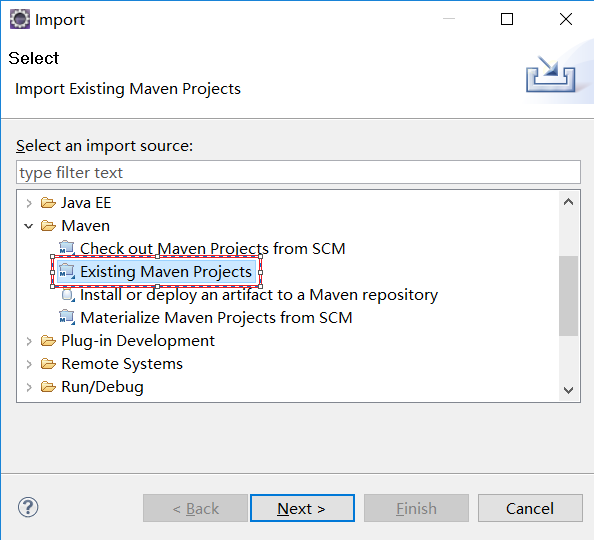
注意：创建的数据库名称和配置文件中相同。



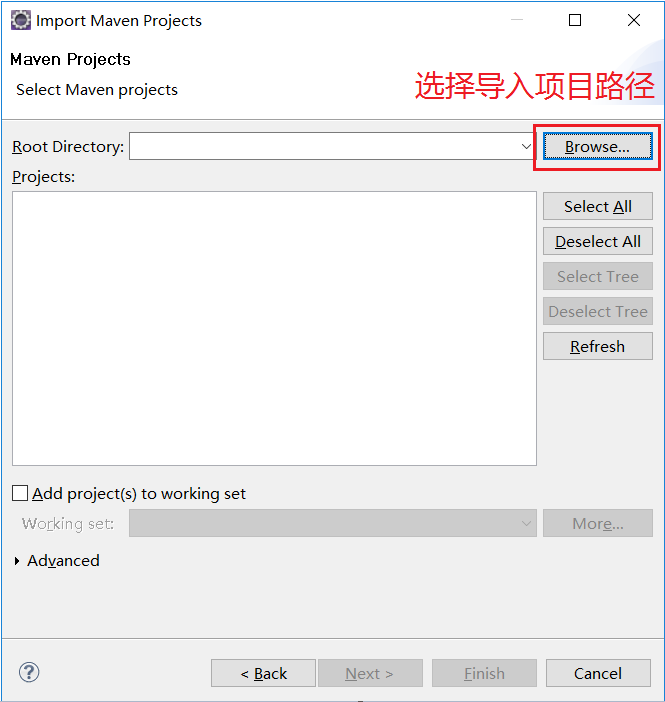
SQL\_DATA文件夹



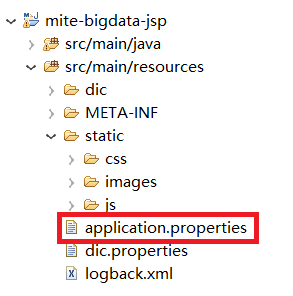
1.3进入eclipse 找到File菜单》》》Import



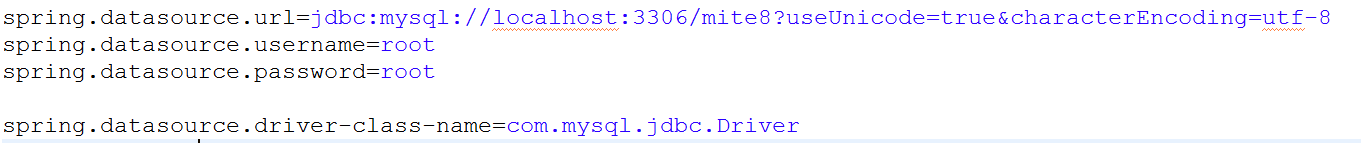
1.4选择路径导入解压后的mite-bigdata-jsp文件夹



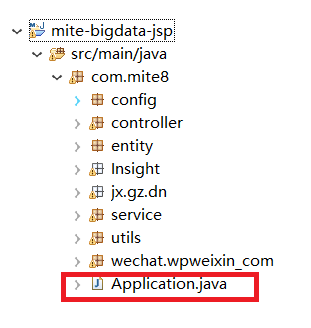
1.5找到文件application.properties



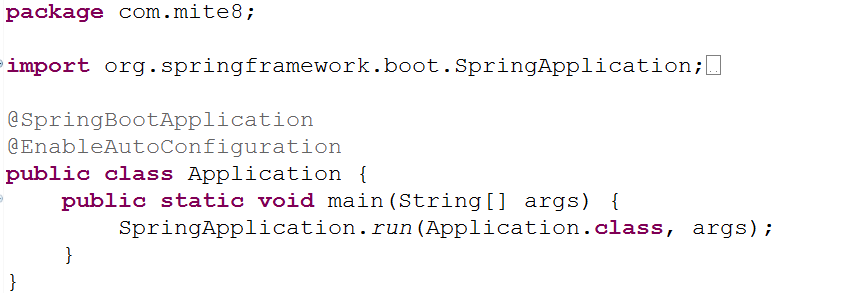
1.6修改数据库连接（将其修改成1.2中你导入的数据库）



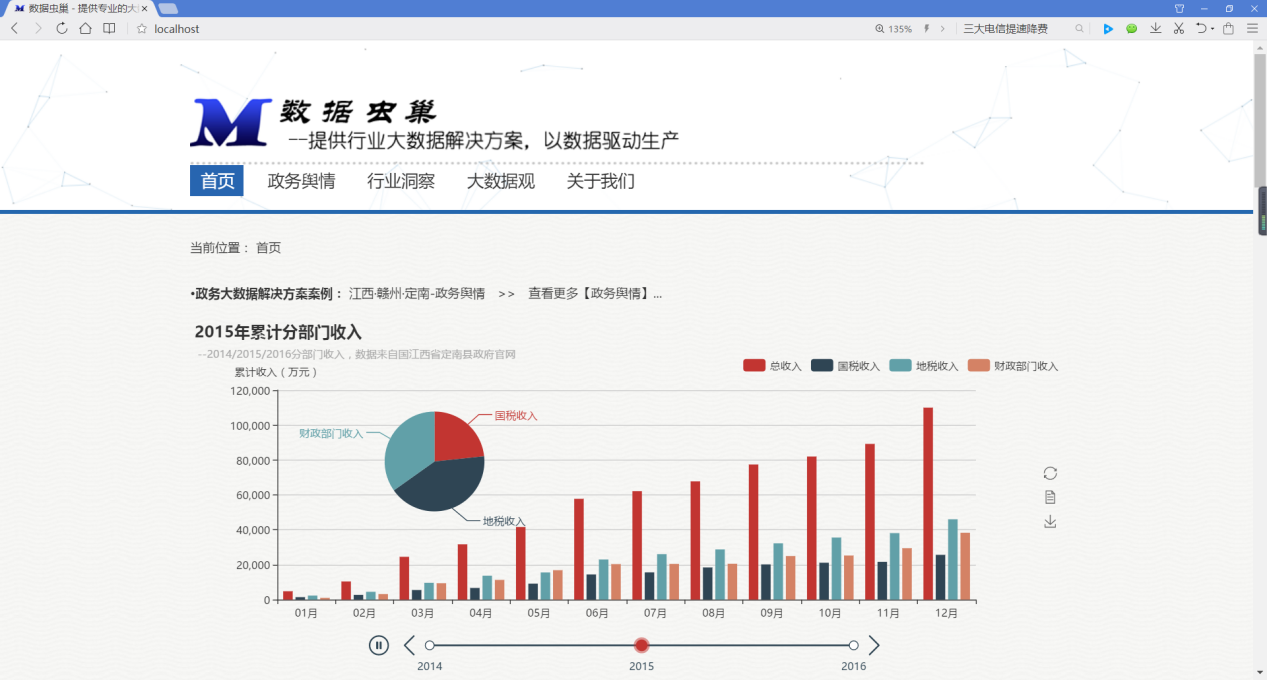
1.7找到其中的Application类



1.8右键run As 运行



1.9在浏览器输入http://localhost即可进图页面



1. **案例分析**

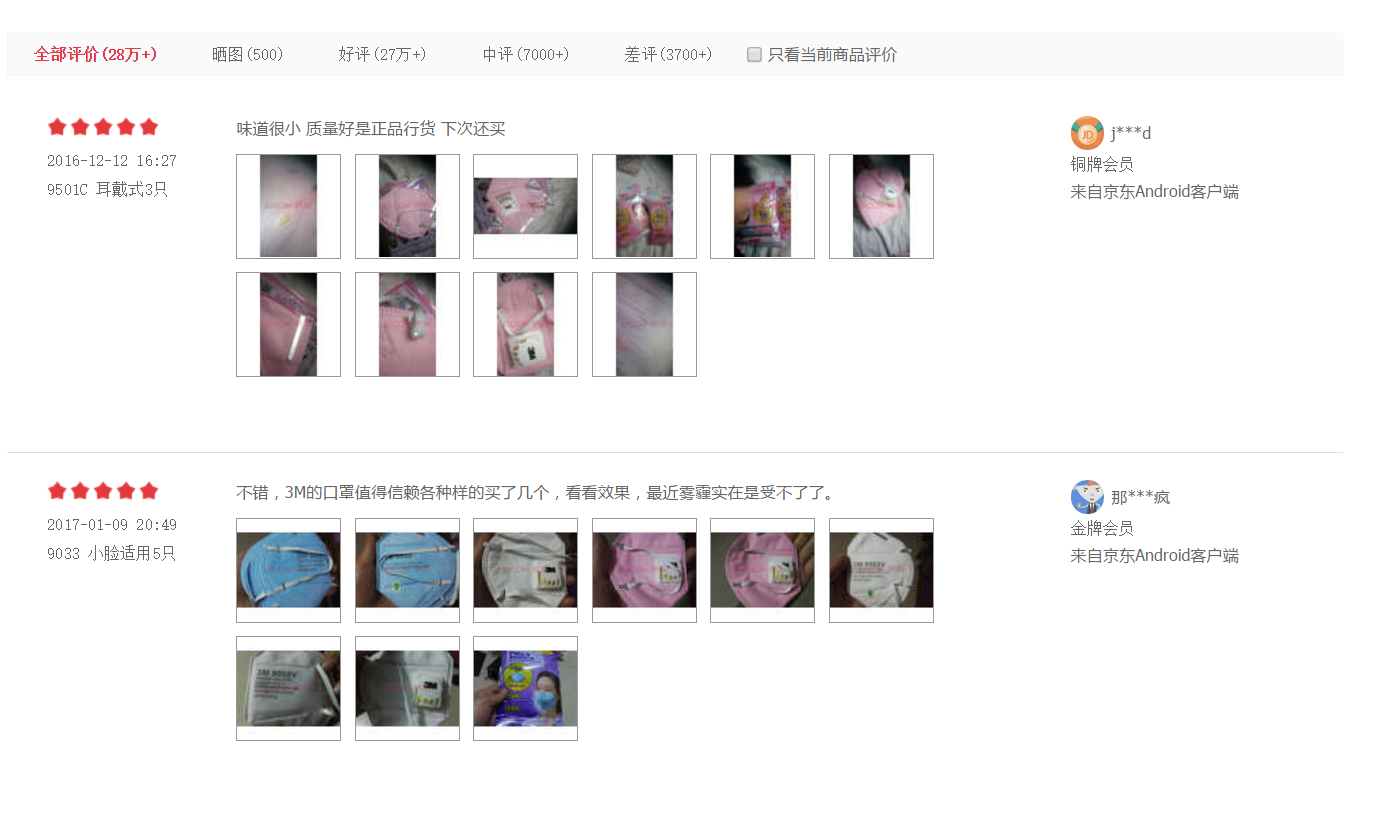
1.1响应雾霾影响地域分布

该案例爬取的是京东商城雾霾口罩的相关信息：

https://search.jd.com/Search?keyword=%E9%9B%BE%E9%9C%BE%E5%8F%A3%E7%BD%A9&enc=utf-8&qrst=1&rt=1&stop=1&vt=2&wq=%E9%9B%BE%E9%9C%BE%E5%8F%A3%E7%BD%A9&psort=4&page=1&click=0

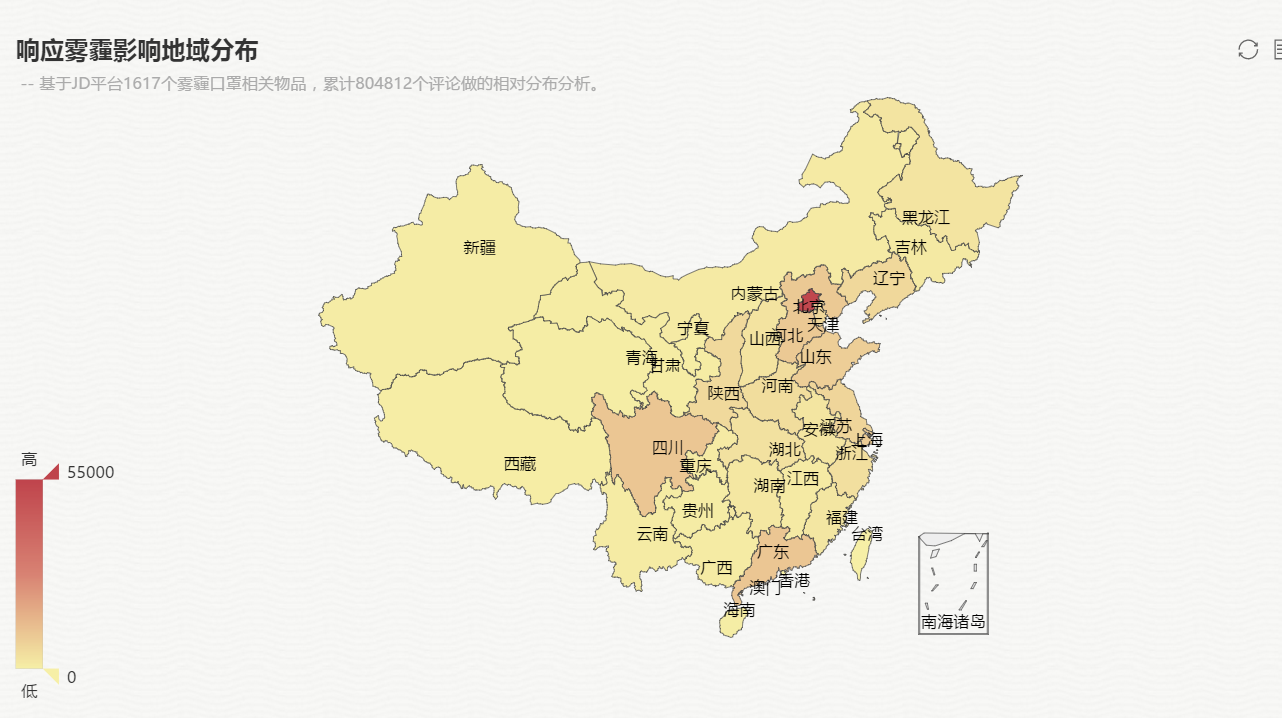


爬取每个相关店铺的评论人的地址分析出雾霾严重区域的分布情况：



将爬取的数据存入到mysql数据库中，使用spring-boot控制查询数据流程，把数据返回到页面用echarts显示出来。

显示结果：



1.2问政舆情数据分析

该案例是从定南县人民政府爬取的数据：

<http://www.dingnan.gov.cn/wlwz/lxxd/>

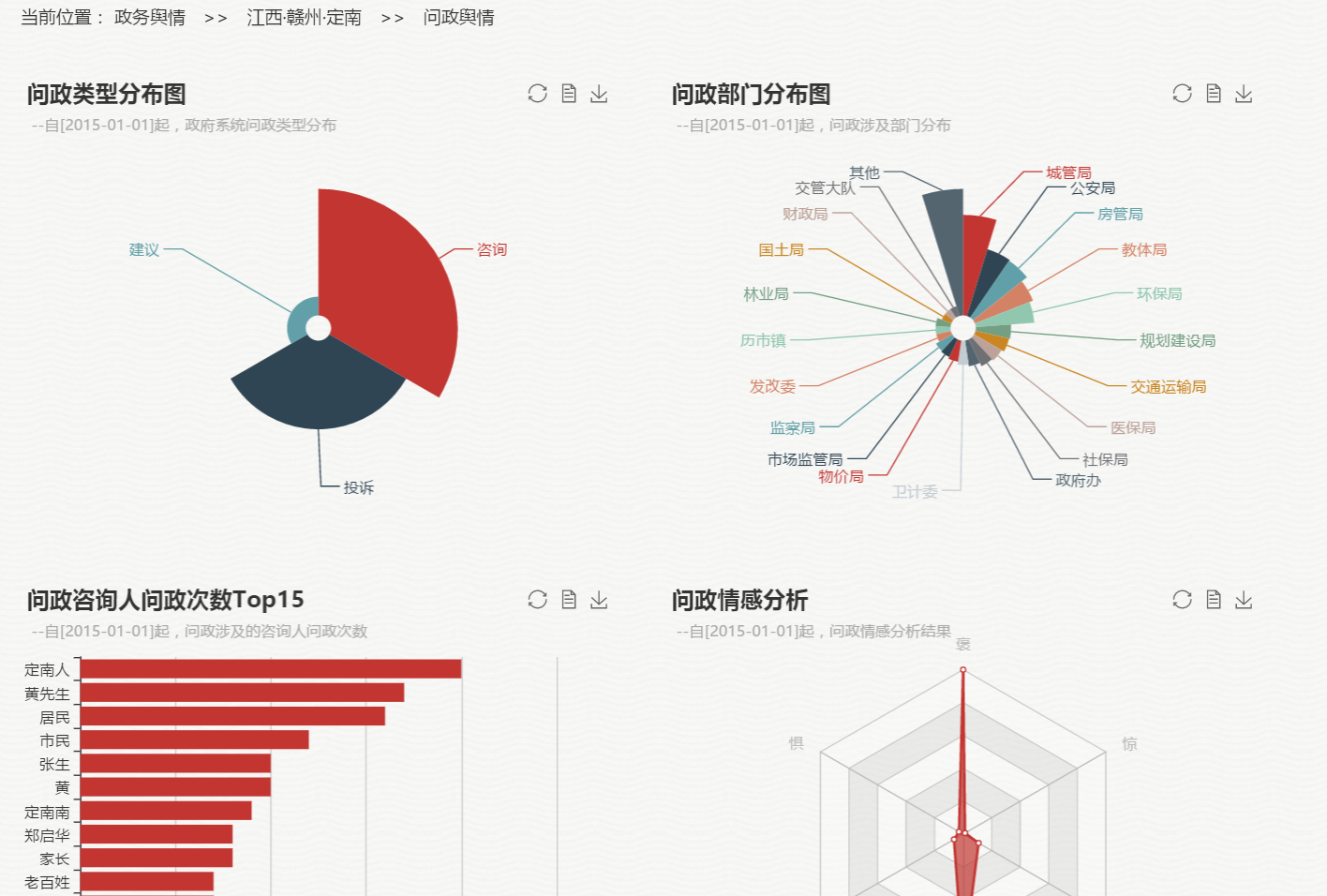


爬取的内容是来信选登板块：



将爬取的数据经过ANSJ存入到mysql数据库中，使用spring-boot控制查询数据流程，把数据返回到页面用echarts显示出来。

显示结果：



1.3人才&招聘舆情

该案例是爬取的是定南在线的招聘求职板块的相关数据：



主要爬取的内容是：



将爬取的数据存入到mysql数据库中，使用spring-boot控制查询数据流程，把数据返回到页面用echarts显示出来。。

显示结果：

