# 爬虫抓取

## 爬虫框架 webcontroller

## 项目运行环境

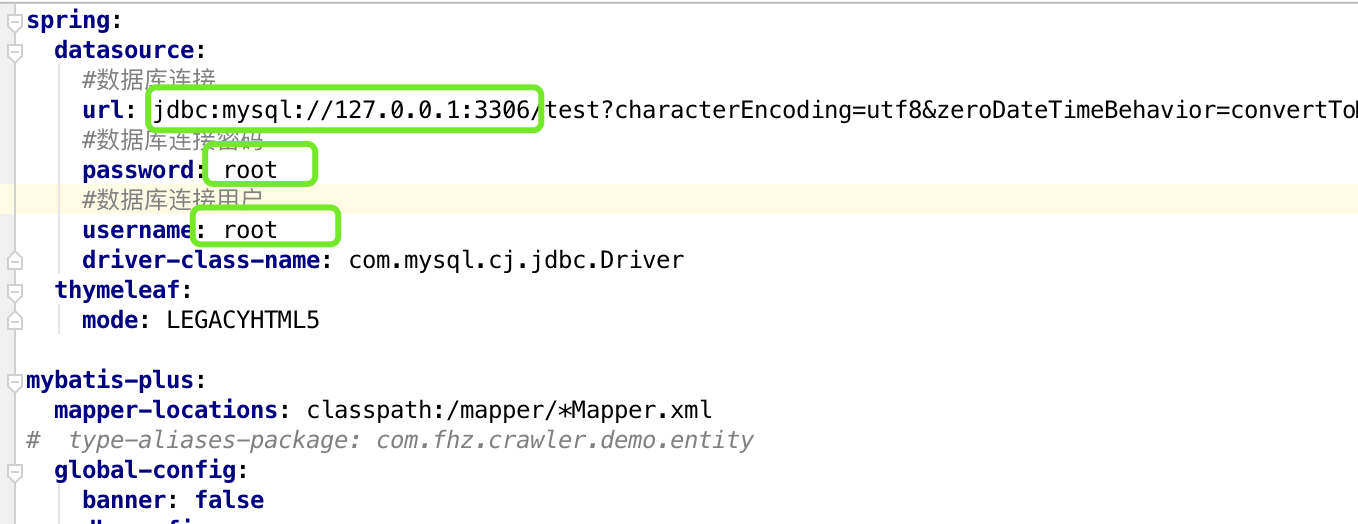
Jdk8，mysql，eclipse、idea

## 项目运行

1. 运行项目下的init.sql文件，初始化数据库，
2. 修改配置文件，

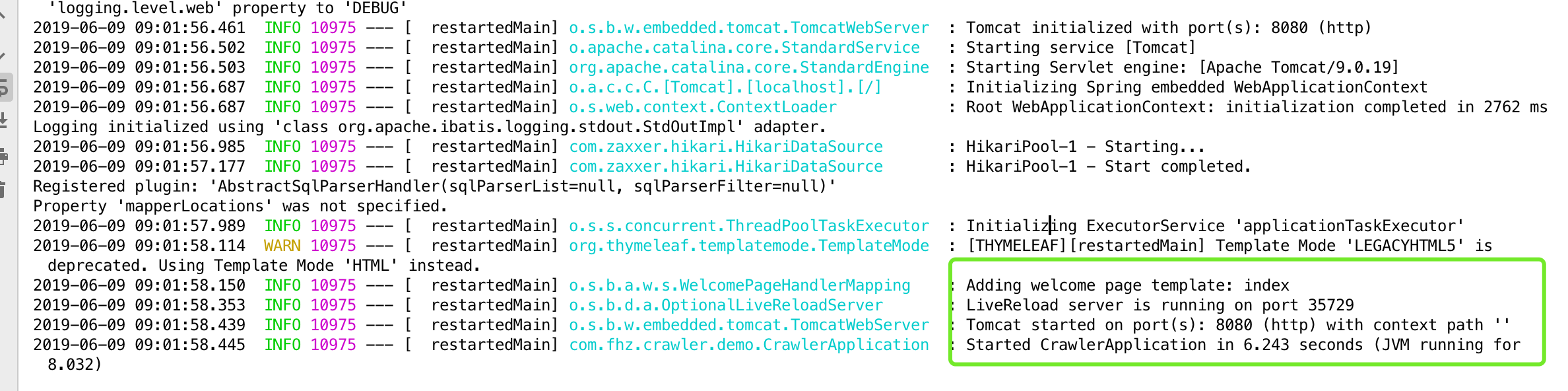
在文件application.yml中，修改到您本地配置如下配置

（数据库连接、登录密码、用户名）



1. 运行项目

在CrawlerApplication主类下，运行main函数，出现如下界面说明项目启动成功



1. 访问页面

<http://lcoalhost:8080/index>



按钮说明：

确定：保存页面修改的数据

清除：删除页面表单保存的数据，不持久化到数据库

开启爬虫：运行爬虫，爬取数据

文件保存：进入文件保存查看界面

音频抓取

在index界面配置网站集合：*http://sc.chinaz.com/tag\_yinxiao/XinWen.html*

点击确定，然后点击开启爬虫，等运行完毕，点击文件保存，查看抓取的数据保存地址

视频抓取

在index界面配置网站集合：<http://www.v1.cn，>

点击确定，然后点击开启爬虫，等运行完毕，点击文件保存，查看抓取的数据保存地址

文本、图片抓取

在index界面配置网站集合：**https://news.baidu.com**

点击确定，然后点击开启爬虫，等运行完毕，点击文件保存，查看抓取的数据保存地址

（注：每个网站都有不同的爬虫抓取规则，请按照以上程序操作文本）