# 实验2：数据集集成实验

## **一、 实验目的**

1. 使用网络爬虫技术，爬取态势感知—舆情热点大数据项目所需数据；
2. 上传原始数据，建立数据仓库，使用kettle，将原始CSV格式数据导入至指定数据库；
3. 使用Kettle对同步的数据进行简单的数据清洗工作，并输出清洗后的数据至数据库。

## **二、实验内容**

1. 模拟服务器发送请求
2. 爬取数据及代码实现；
3. 文本信息的抽取、数据清洗、将数据导入数据库
4. 数据集成。

## **三、实验过程**

**任务一:模拟服务器发送请求**

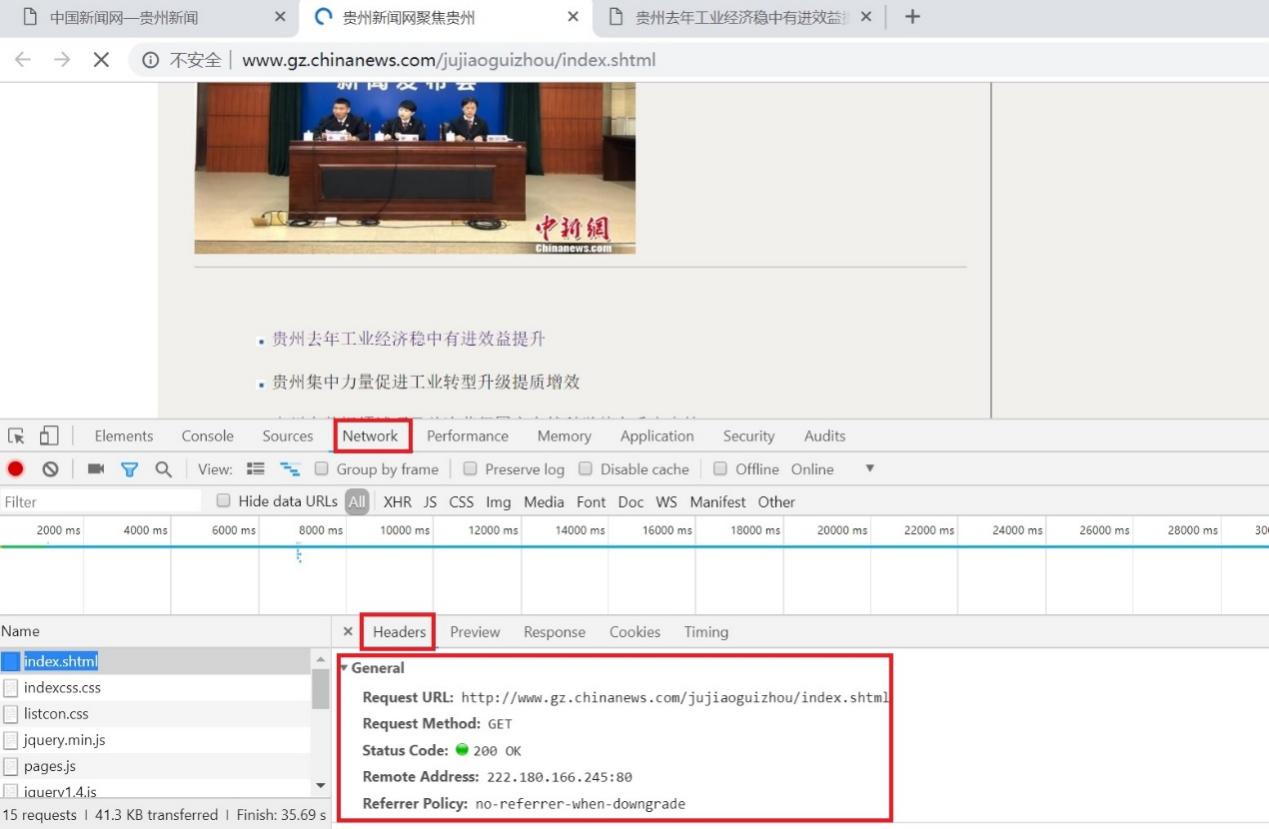
1.打开浏览器，在浏览器地址栏中输入地址“www.gz.chinanews.com”，可以看到网站界面。按【F12】打开浏览器开发者工具,点击『Network』，选择 “index\_shtm”,点击『Header』,查看请求头的内容。如图所示:

图5-1 请求头页面信息

服务器在收到请求后，会给用户端响应，点击『response』查看响应，如图所示：

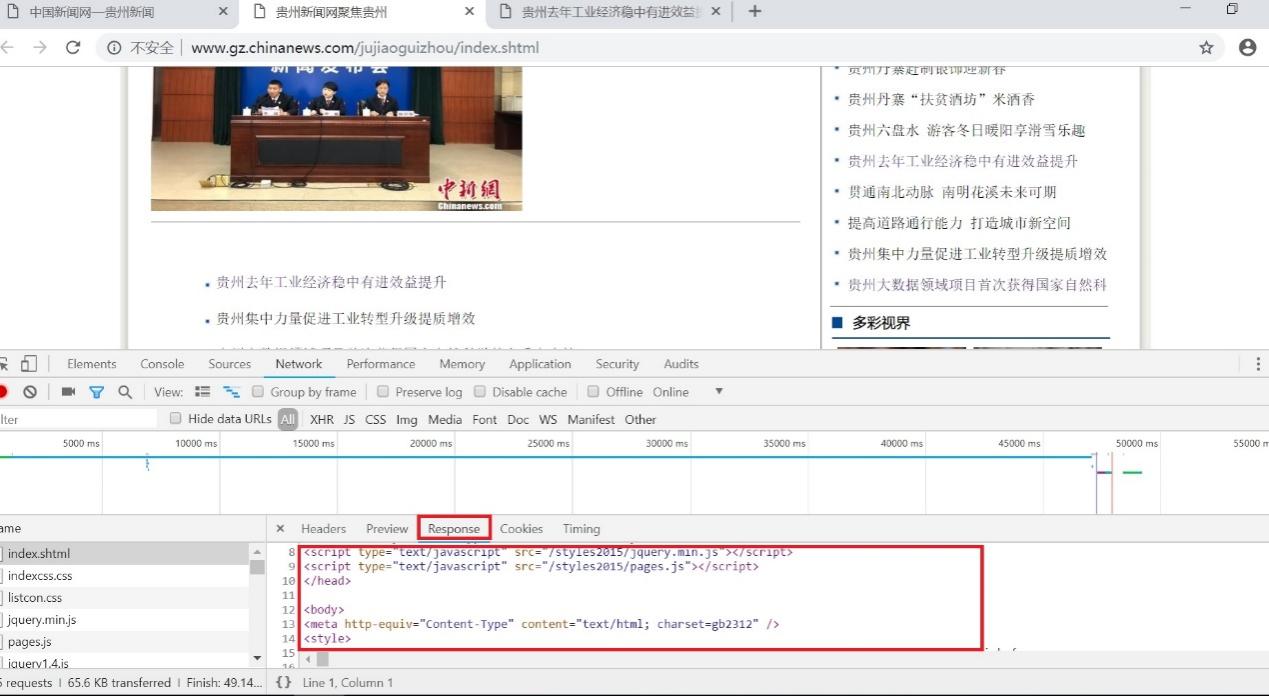


图5-2 响应页面信息

2.按【F12】打开开发者工具,点击左上角的选取工具，选取标题，看到标题对应的是<div class=“pagetitle”>的内容。如图所示：**h4c**

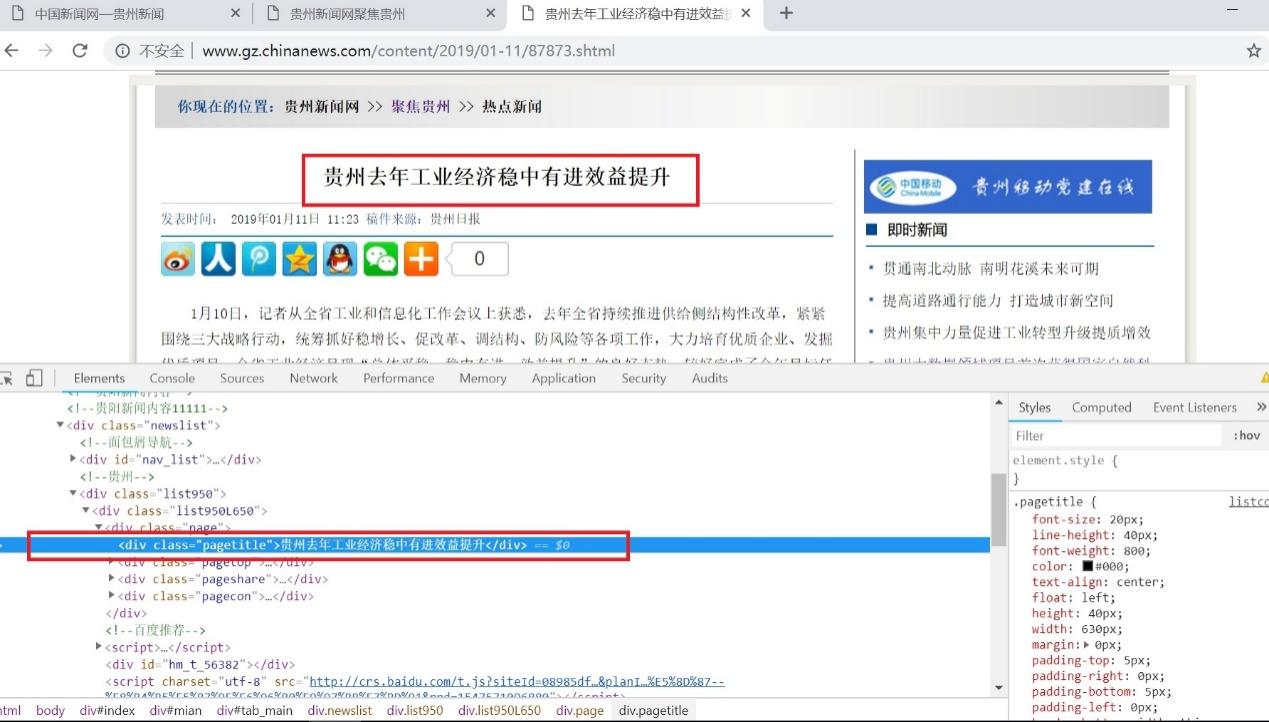
****

图5-3 选取标题

3.选择发表时间和稿件来源，可以看到对应的<div class=“pagetop”>的内容。如下图所示：**二者都在‘pubtime’**

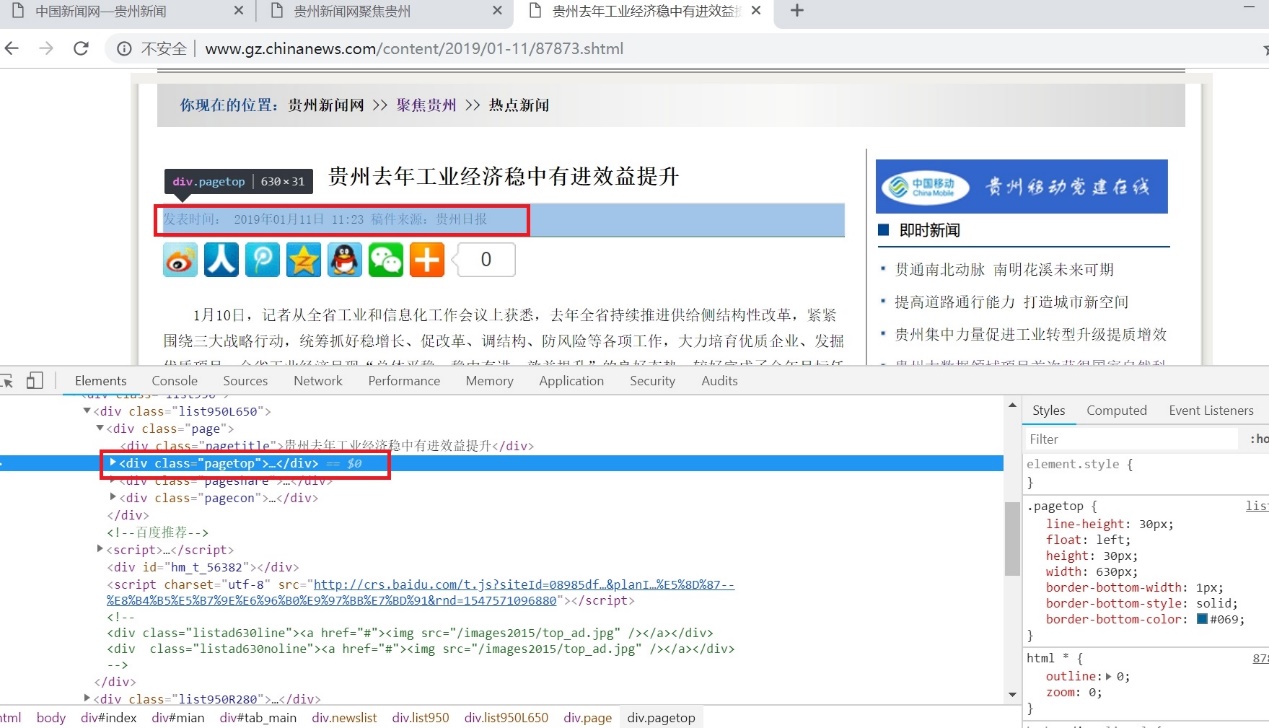
****

图5-4 选取时间和稿件来源

选取稿件内容，可以看到对应的元素是<div class=“pagecon”>，如图所示：

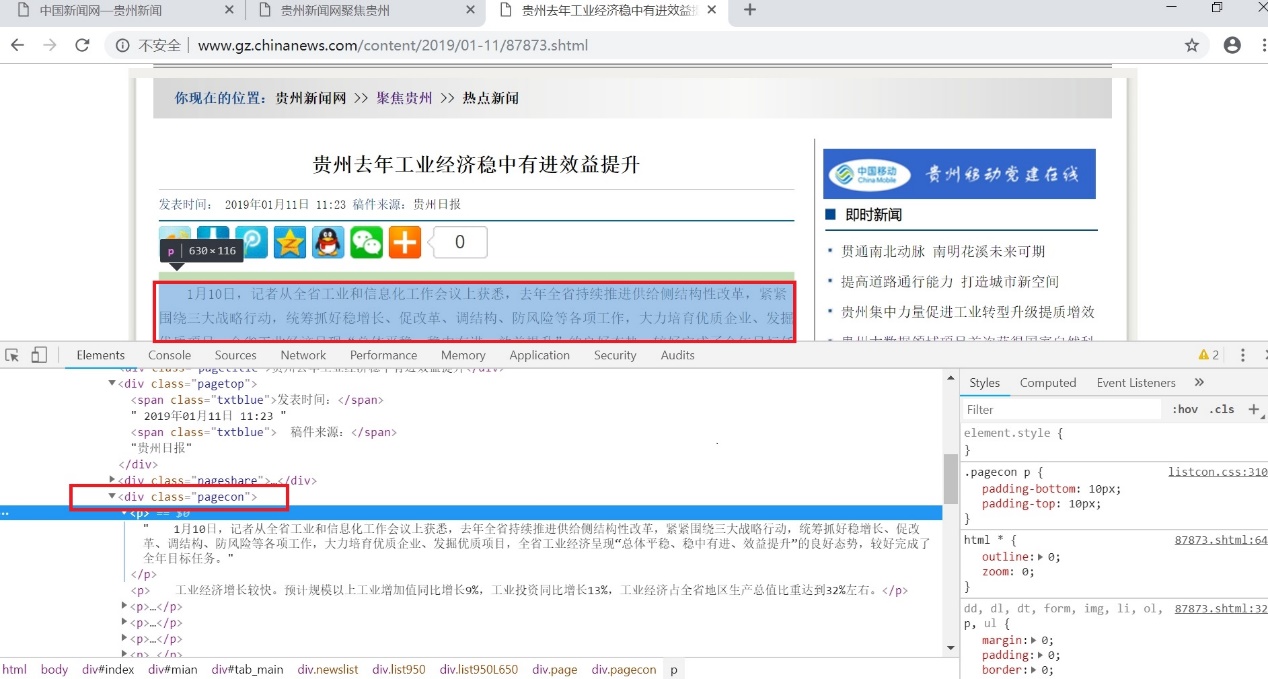


图5-5 选取稿件内容

**任务二:代码实现**

1.打开PyCharm，打开事先写好的爬虫，点击『report』，看到我们编写此所需要用到的库：csv、requests、re、etree、detetime。如图所示：

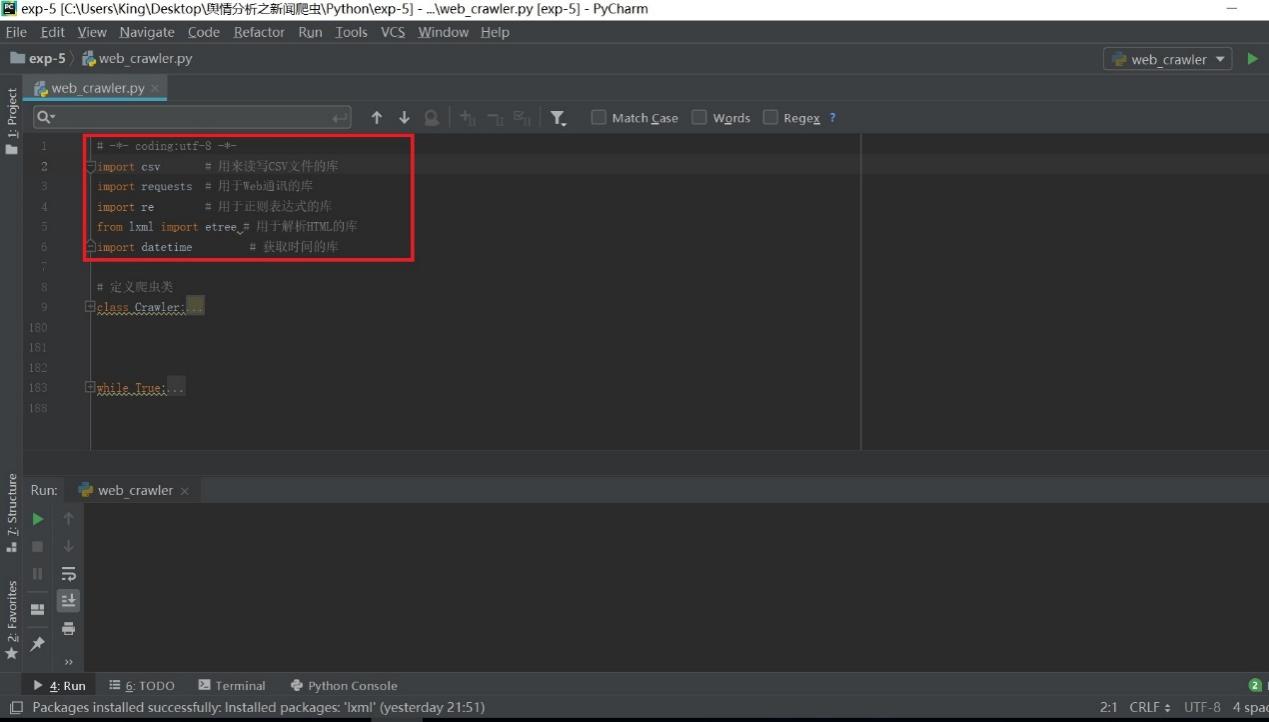
****

图5-6 爬虫所用到的库

2.进入程序头部，首先创建爬虫对象，点击『start\_request』，可以看到该方法的内容如下图所示：

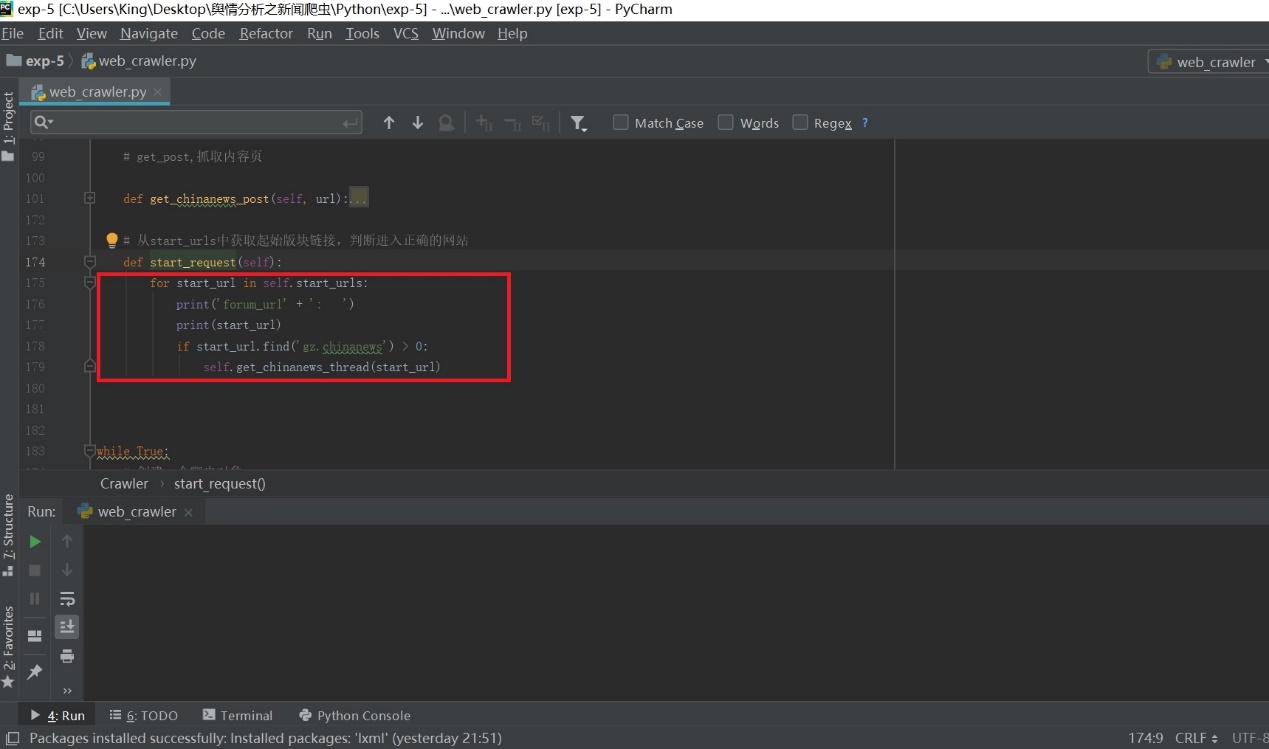


图5-7 start\_request方法内容

判断进入了正确的网站，则进入『chinese\_thread』方法，点击『chinese\_thread』，该方法的内容如图所示：

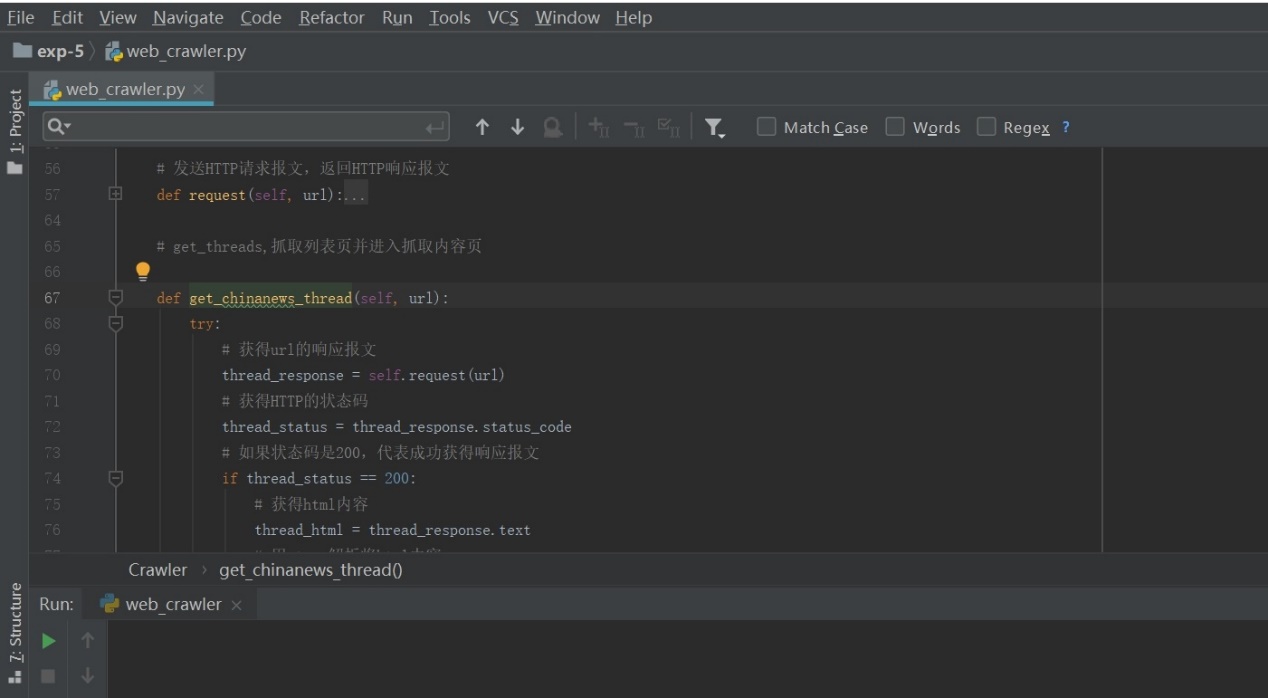


图5-8 chinese\_thread方法内容

打开浏览器，按【F12】选择li元素列表，其中“a”标签对应的“href”属性就是对应新闻的相对url地址，如图所示

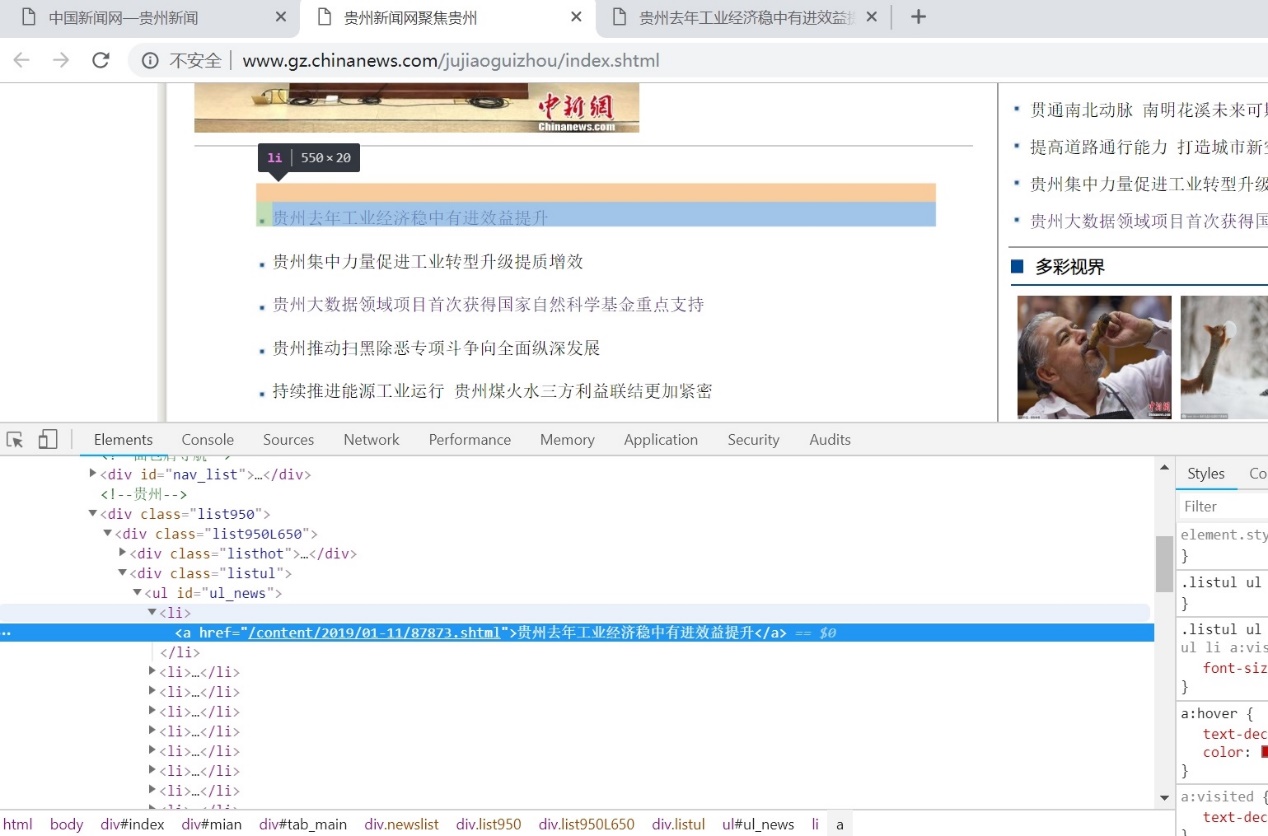


图5-9 相对url地址

由于不能使用相对url地址直接访问，需要跟域名拼接成绝对url地址，并输出绝对url地址。如果状态码不是二百就记录错误，将异常保存为csv格式。代码如下

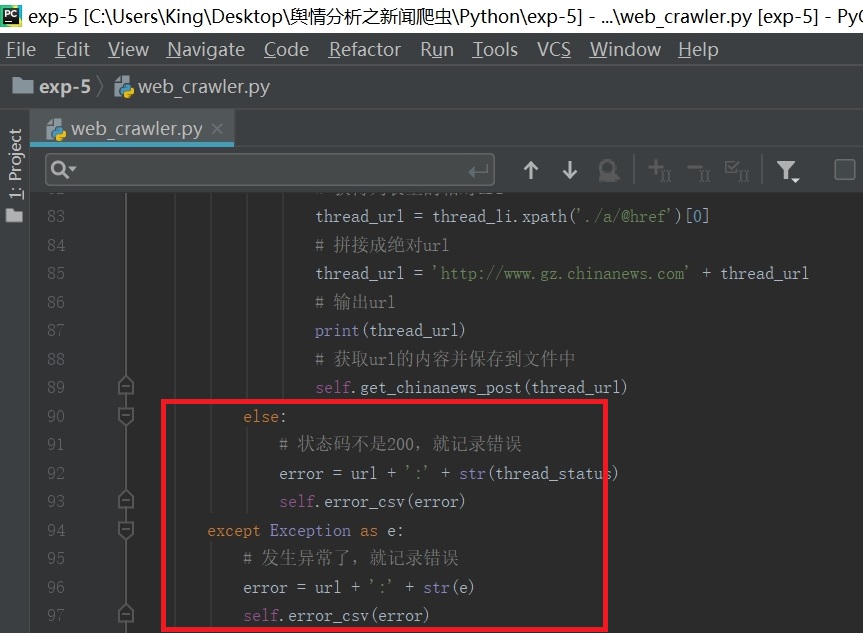


图5-10 chinese\_thread方法内容

如果成功获取url地址并保存，则调用『get\_chinanews』方法，代码如图所示

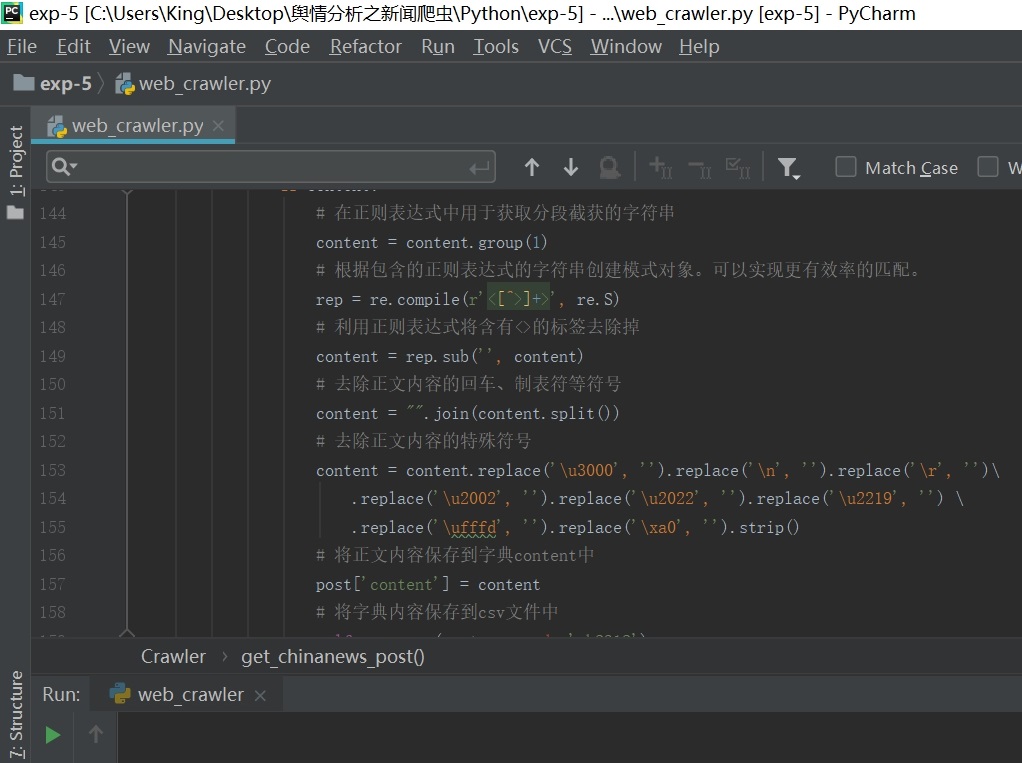


图5-11 get\_chinanews方法内容

经过以上处理后，点击进入『save\_csv』方法，代码如图所示

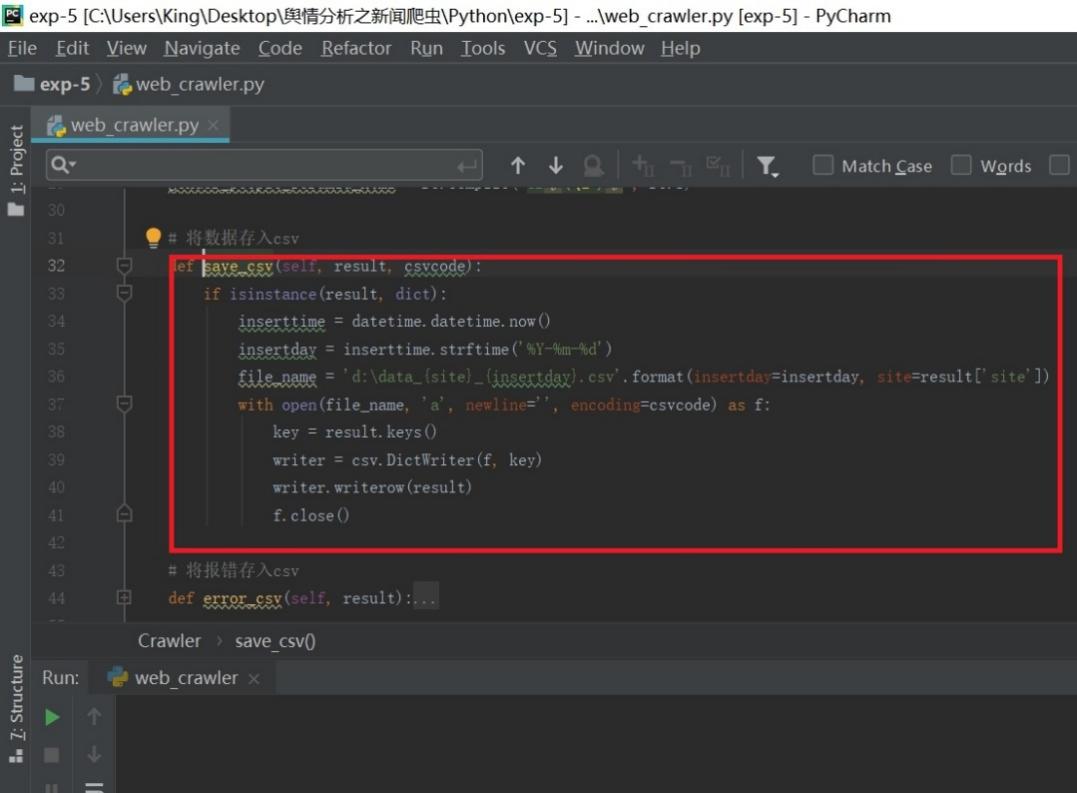


图5-12 save\_csv方法内容

点击『运行』，程序开始爬取。如图所示

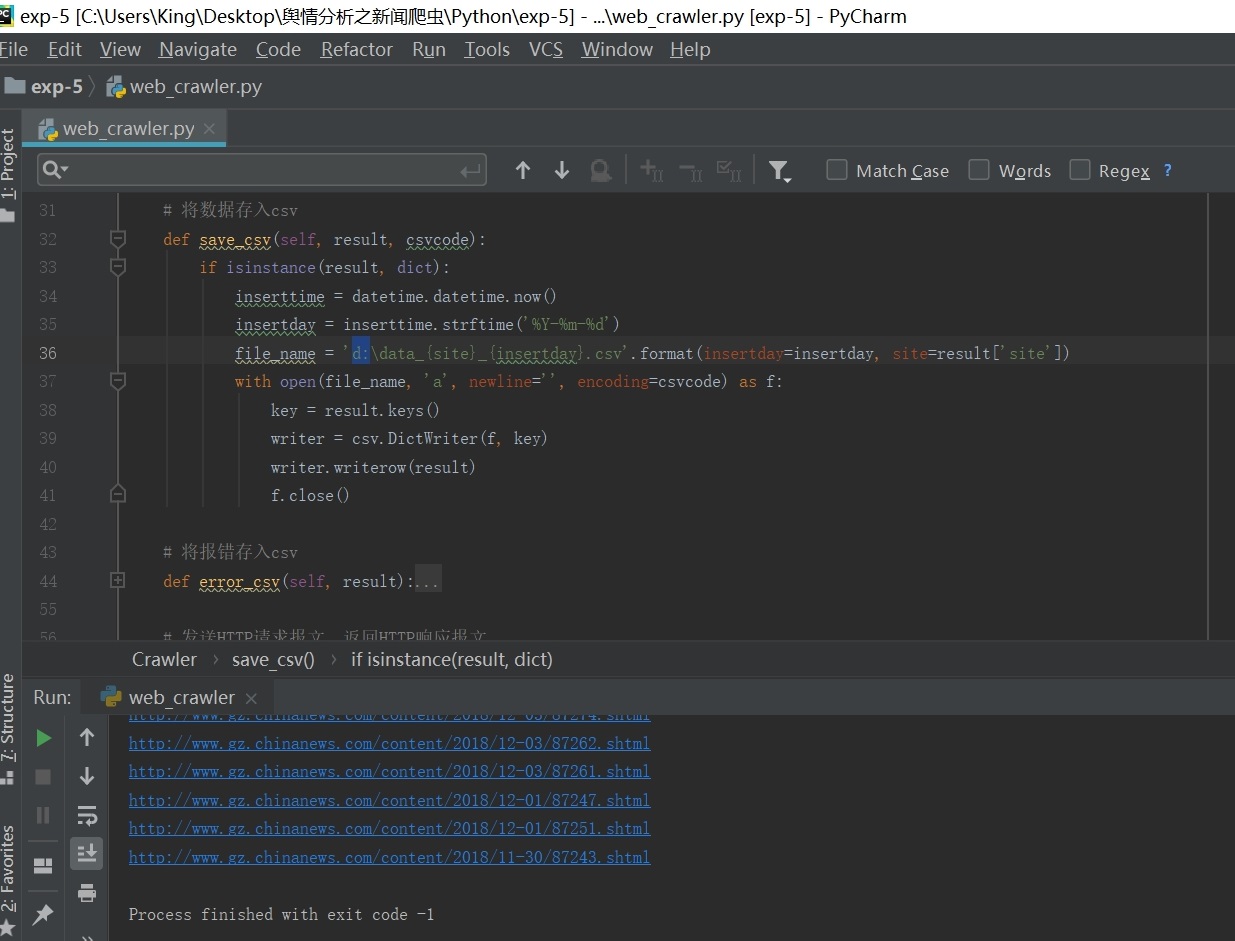


图5-13 Python爬取数据

打开我们存储数据文件的根目录，使用“notepad”打开。如图所示

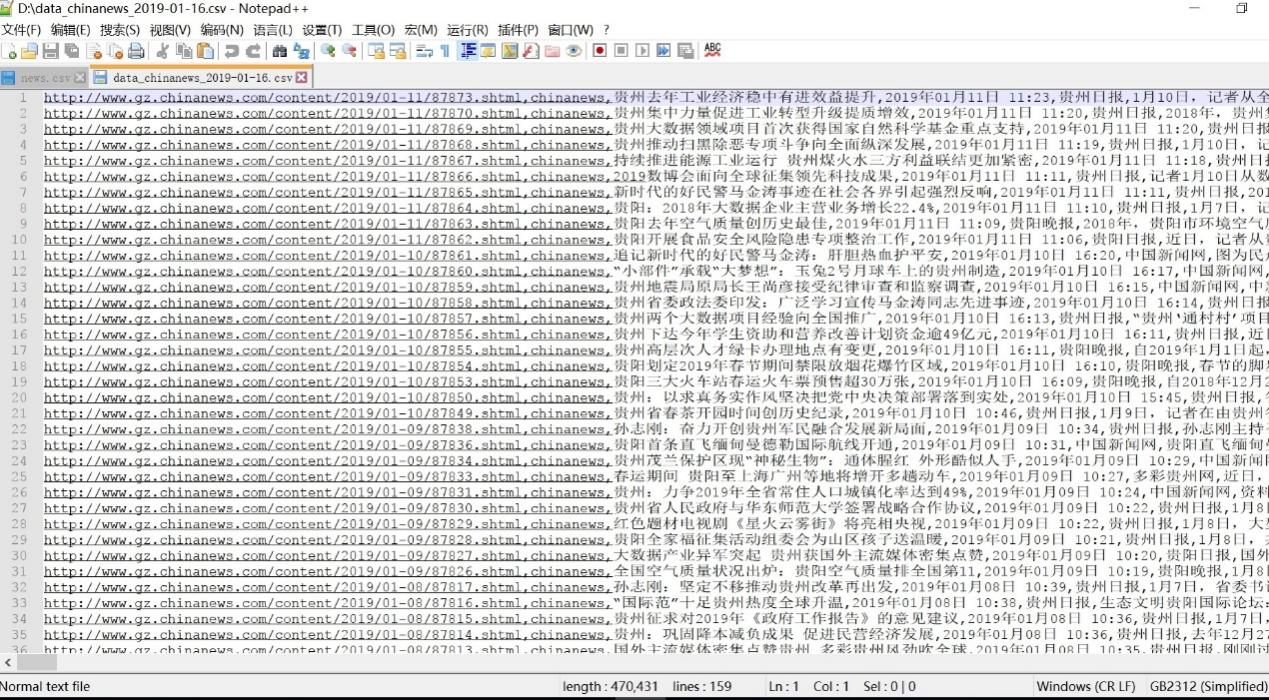


图5-14 Python爬虫结果

打开原网页，将爬虫结果与网页信息相对比，结果表明信息吻合，爬虫成功。如图所示

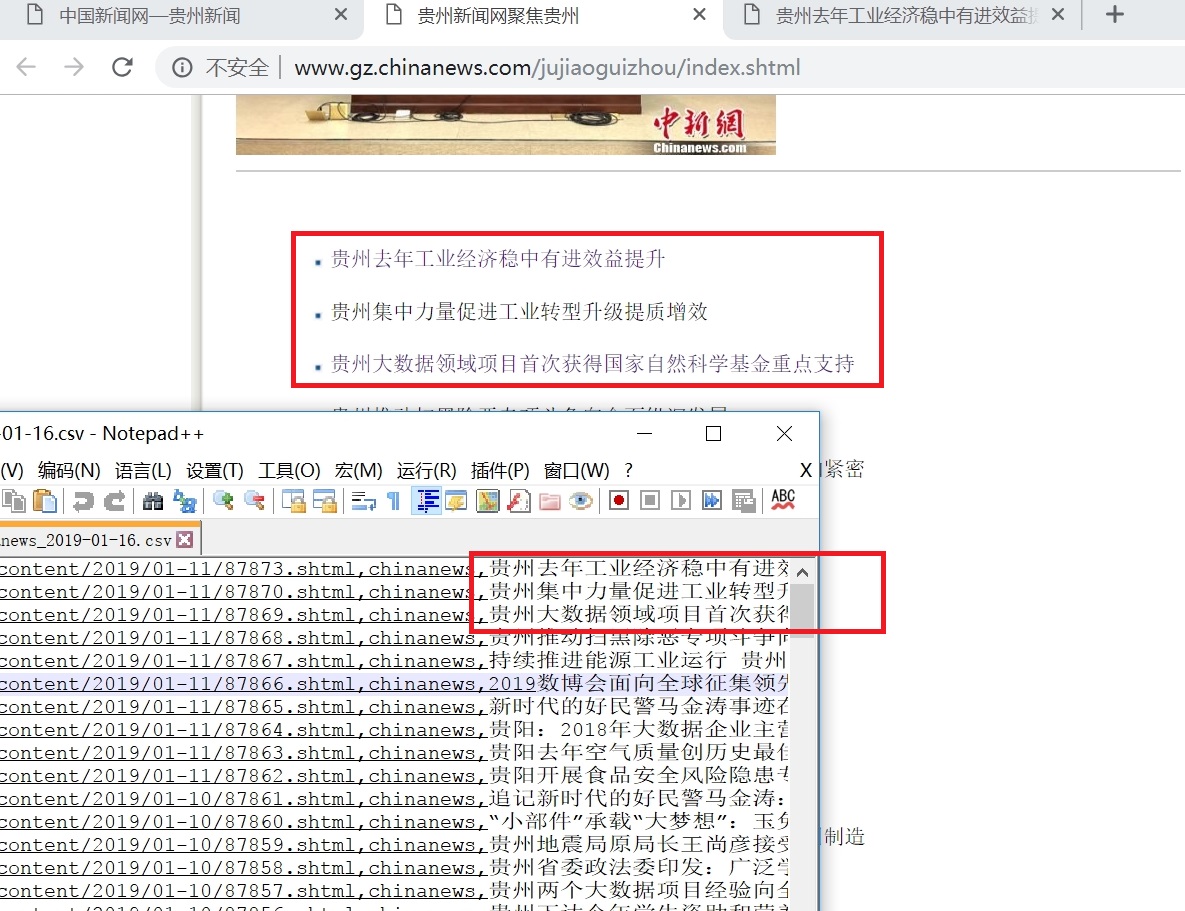


图5-15 Python爬虫结果与原数据对比

**任务三:将数据导入数据库**

打开“Nabvcat”，点击打开“demo2yuqing”数据库，使用新建脚本新建数据库中的数据表。打开已经写好的SQL脚本“creat\_ods\_news.sql”，复制脚本内容。在打开“Nabvcat”，点击『新建查询』，将脚本内容粘贴上，点击『运行』，如图所示

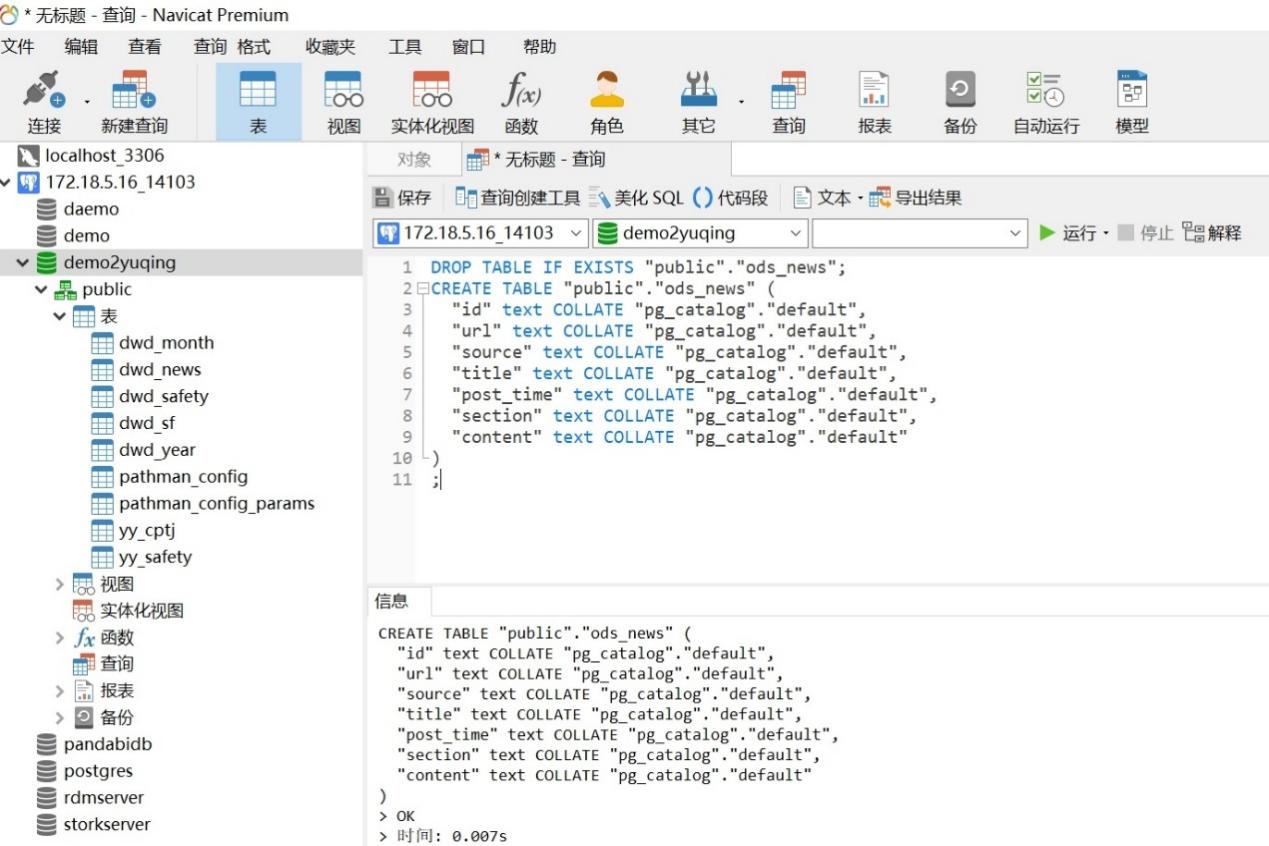


图5-16 新建数据表“ods\_news”

打开脚本“creat\_base.news.sql”,复制脚本内容。打开“Navicat”，点击『新建查询』，粘贴脚本内容，点击『运行』。如图所示

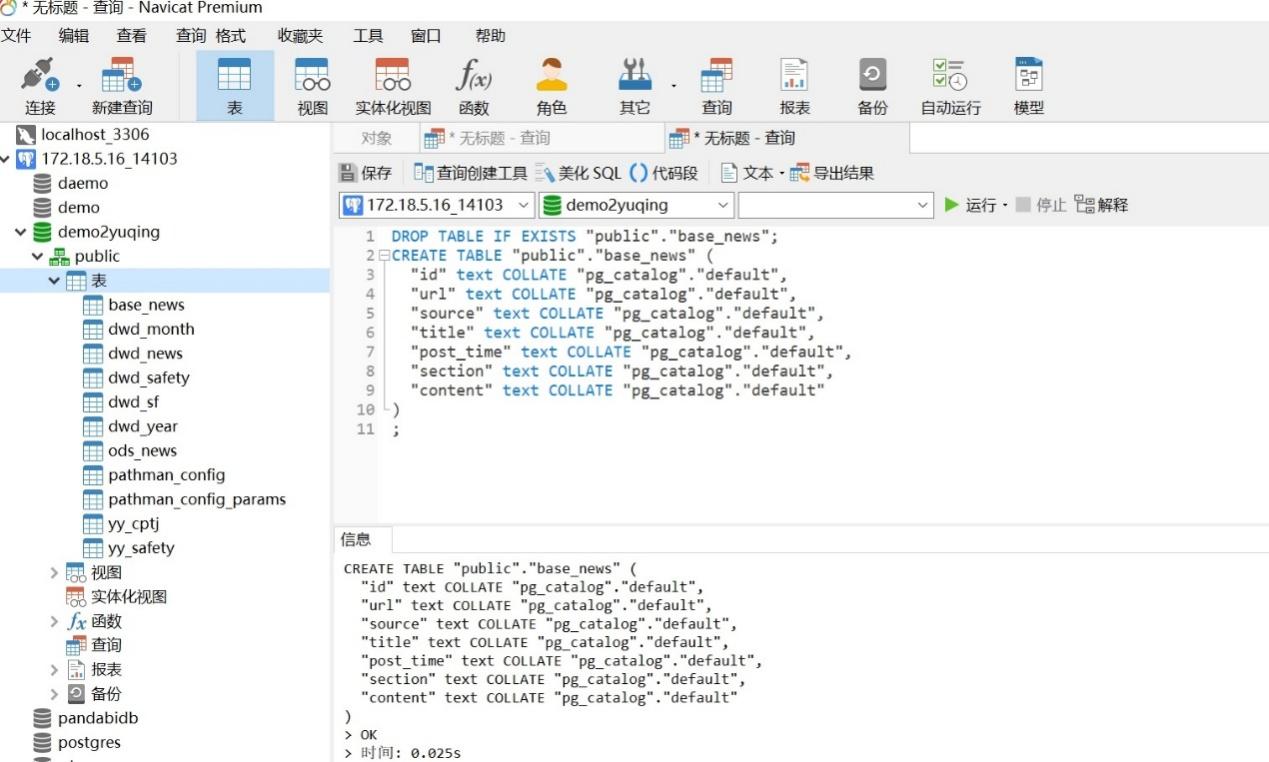


图5-17 新建数据表“base\_news”

打开新建的数据表，发现数据表为空，只有字段名称。如图所示

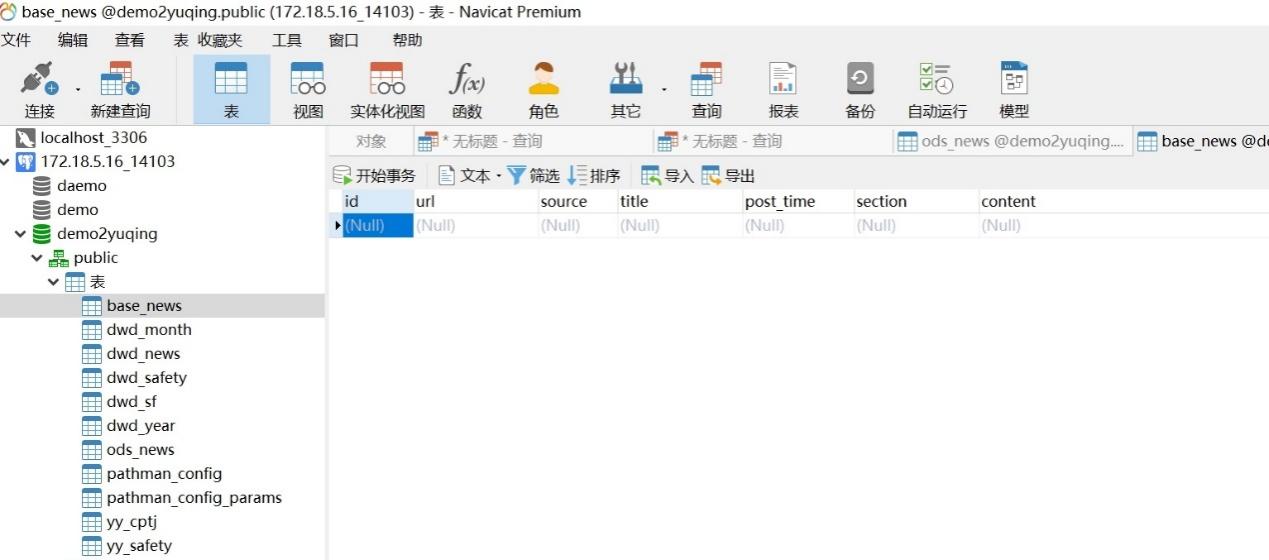


图5-18 新建数据表内容

打开“Kettle”，点击『核心对象』，选择『csv文件输入』、『字段选择』、『表输出』拖拽至右边的工作区域。如图所示

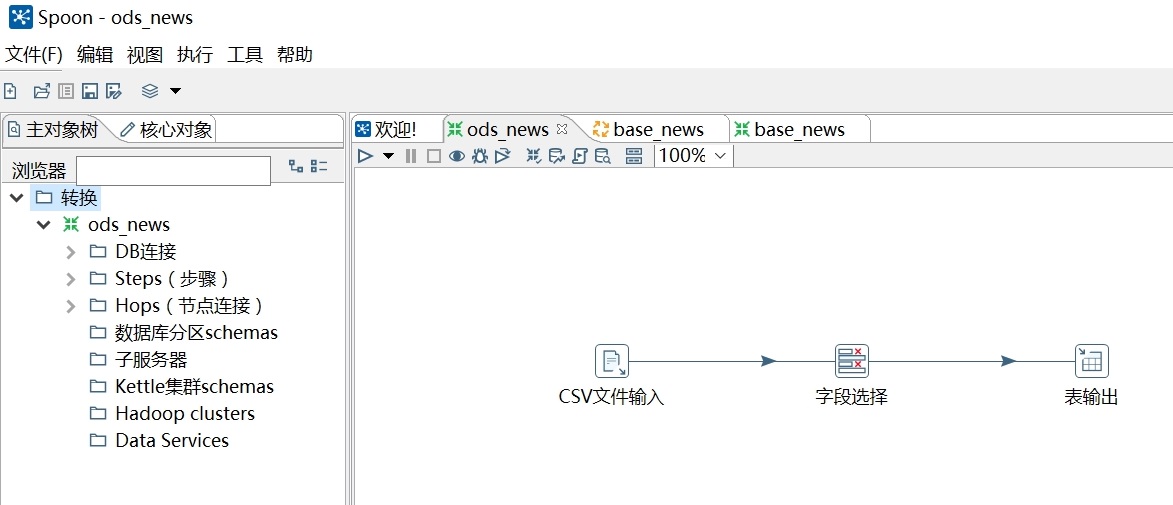


图5-19 新建数据表内容

选择『csv文件输入』，右键点击选择『编辑步骤』，选择上文爬取的数据信息，点击确定。如图所示

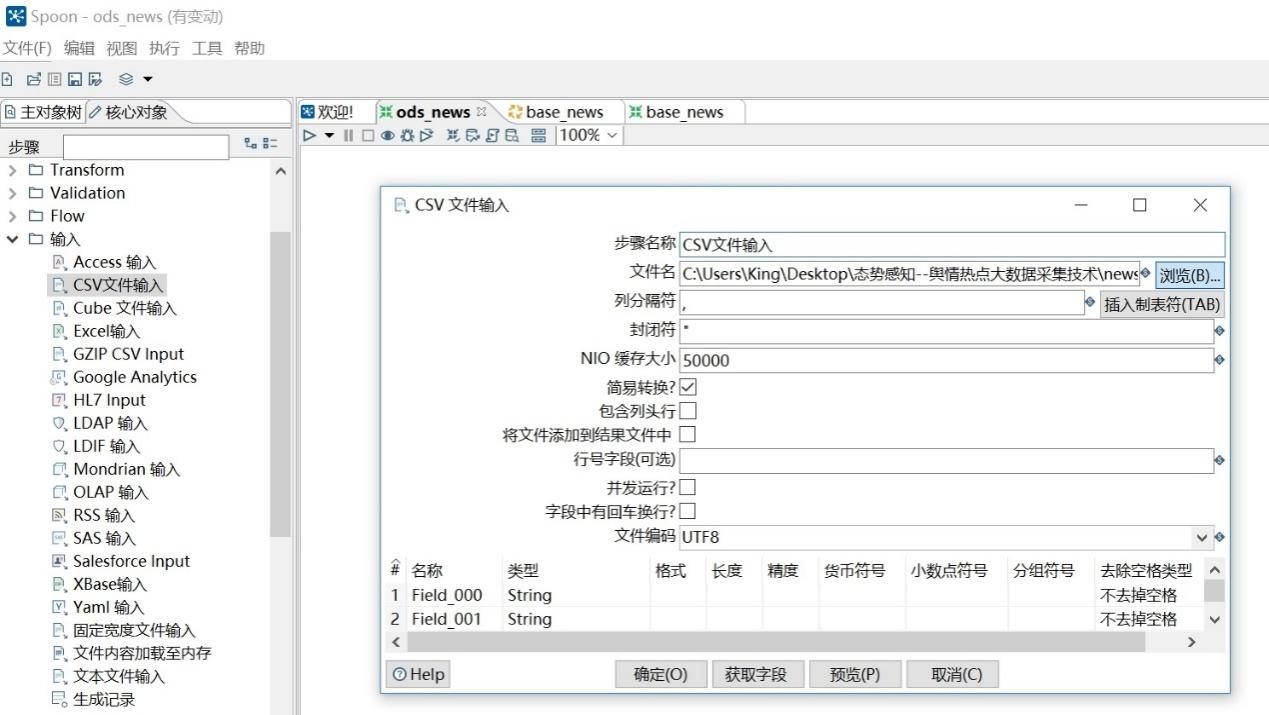


图5-20 定义csv文件输入

点击『执行』，得到执行结果。如图所示



图5-21 执行结果

打开“Navicat”，刷新并打开“ods\_news”数据表，可以看到实验结果正确。如图所示

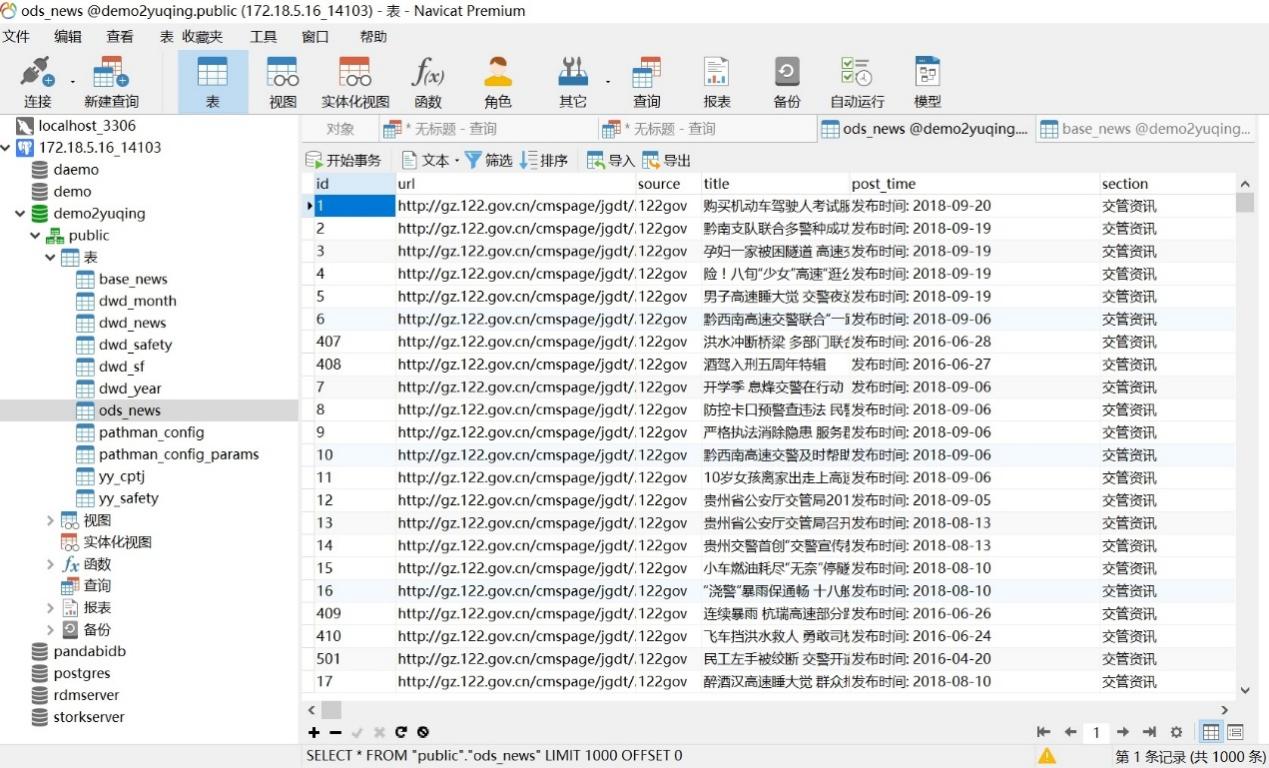


图5-22 “Navicat实验结果”

**任务四:清洗数据**

打开进入“Kettle”，打开已经写好的kettle脚本，进行简单的数据清洗操作。右键点击『去除脏数据』，选择“Open Referenced Object”，进入脚本内容。如图所示

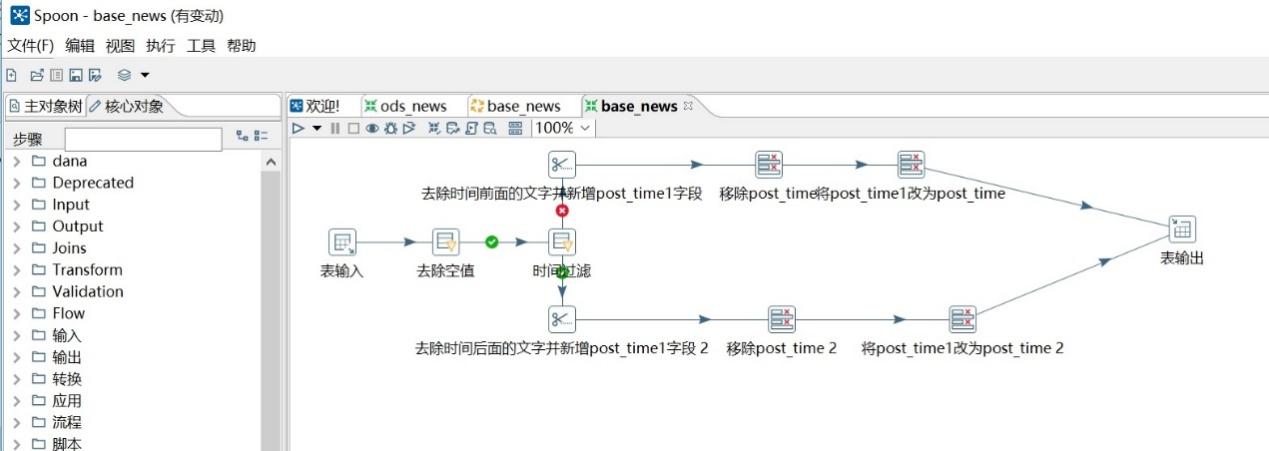


图5-23 去除脏数据

点击『执行』，得到执行结果。如图所示

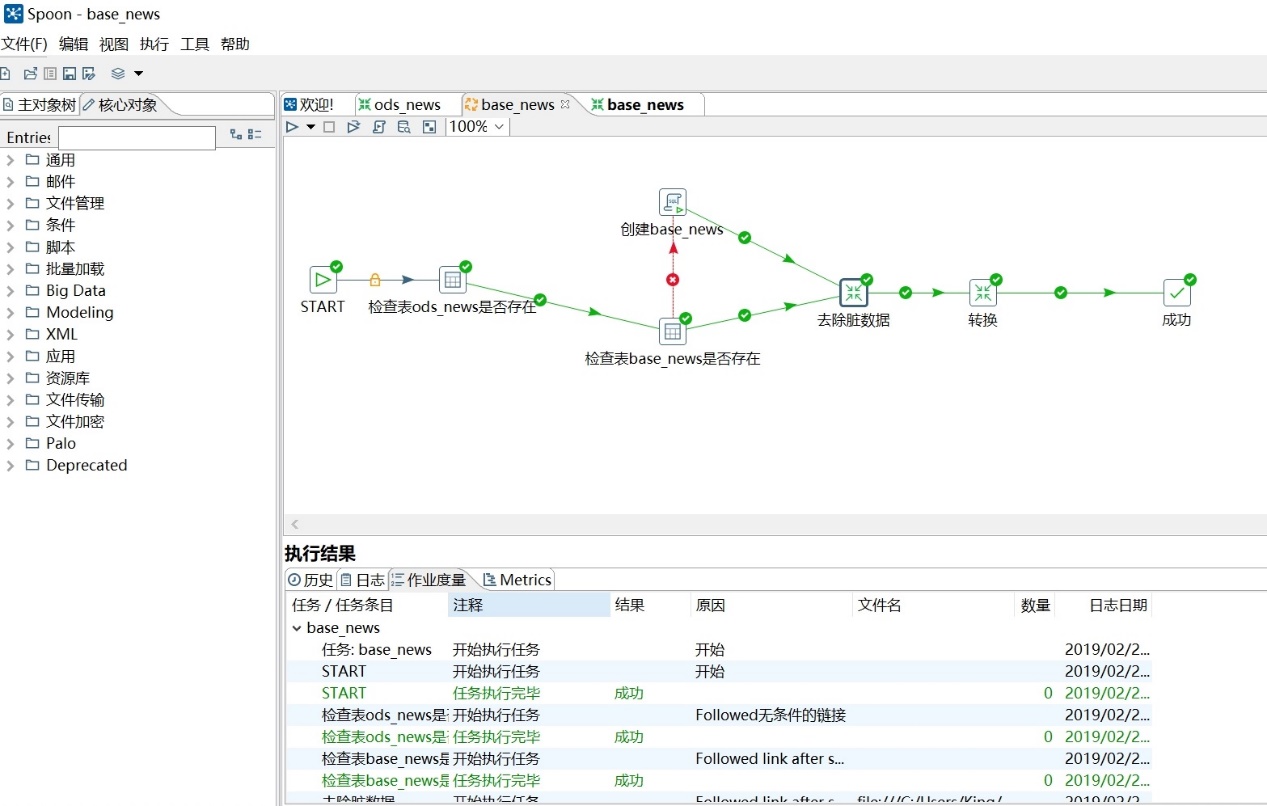


图5-24 执行结果

打开“Navicat”，刷新并打开“base\_news”数据表，对比两张数据表的内容，实验数据清洗完成，实验成功。如图所示

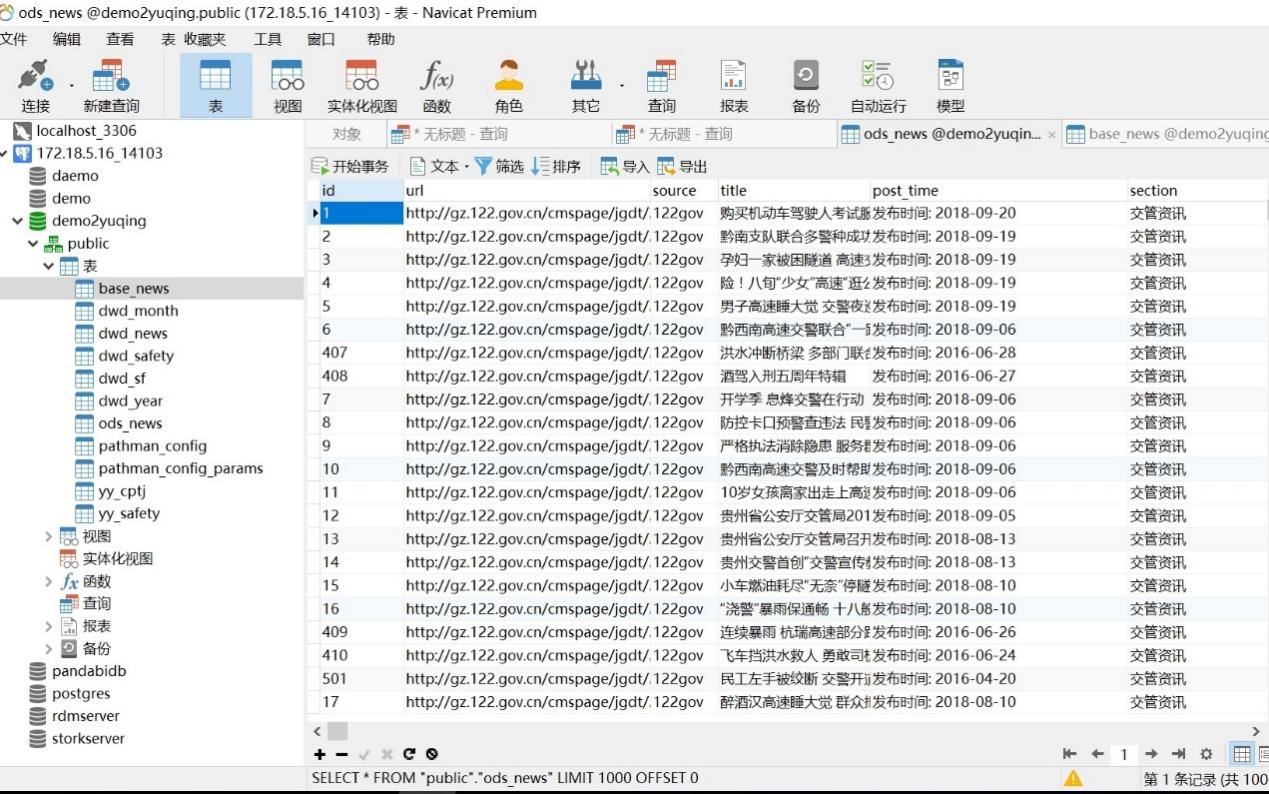


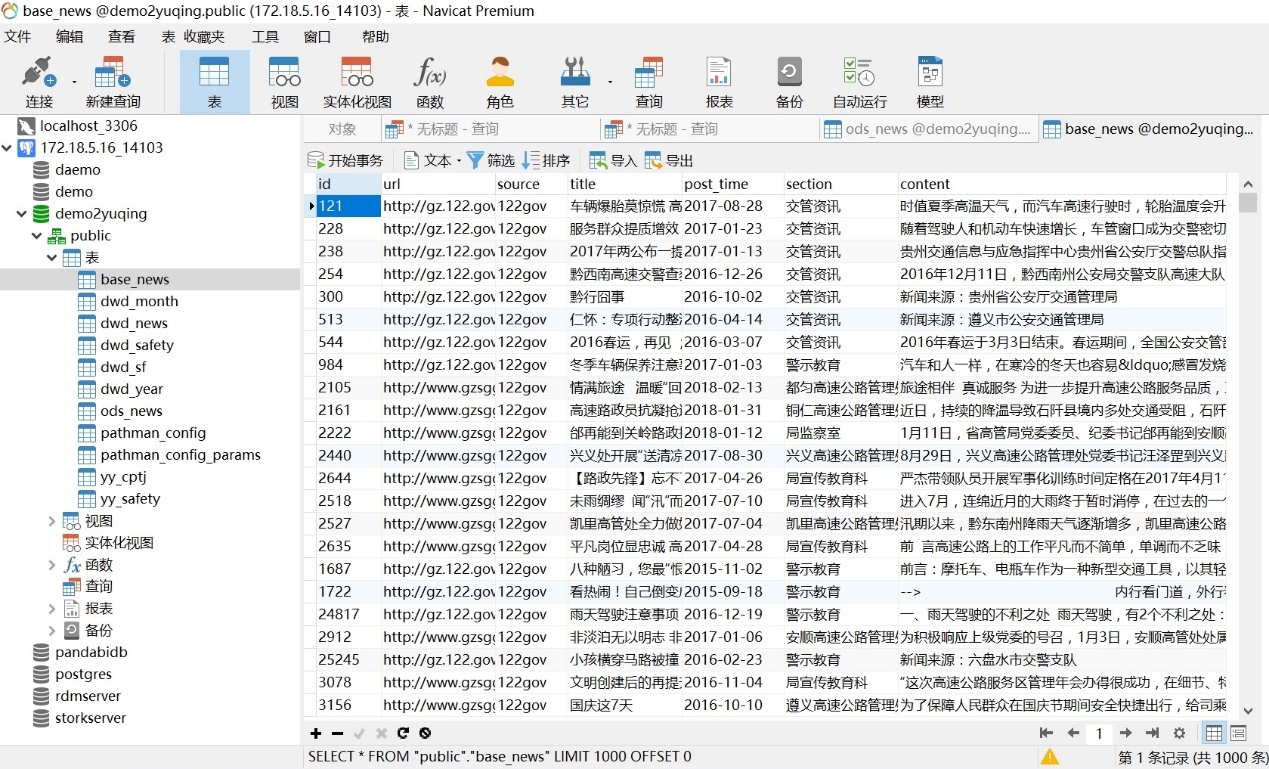
图5-25 清洗前数据

图5-26 清洗后数据

## **四、实验小结**

通过本次实验对态势感知--舆情热点大数据案例数据采集的演练，掌握数据从爬取，导入，清洗，输出各个工作流程。加深大数据相关采集技术的理解与掌握。完整演练数据采集的整个过程。