数据采集子系统——测试文档

（以甘肃省博物馆为例）

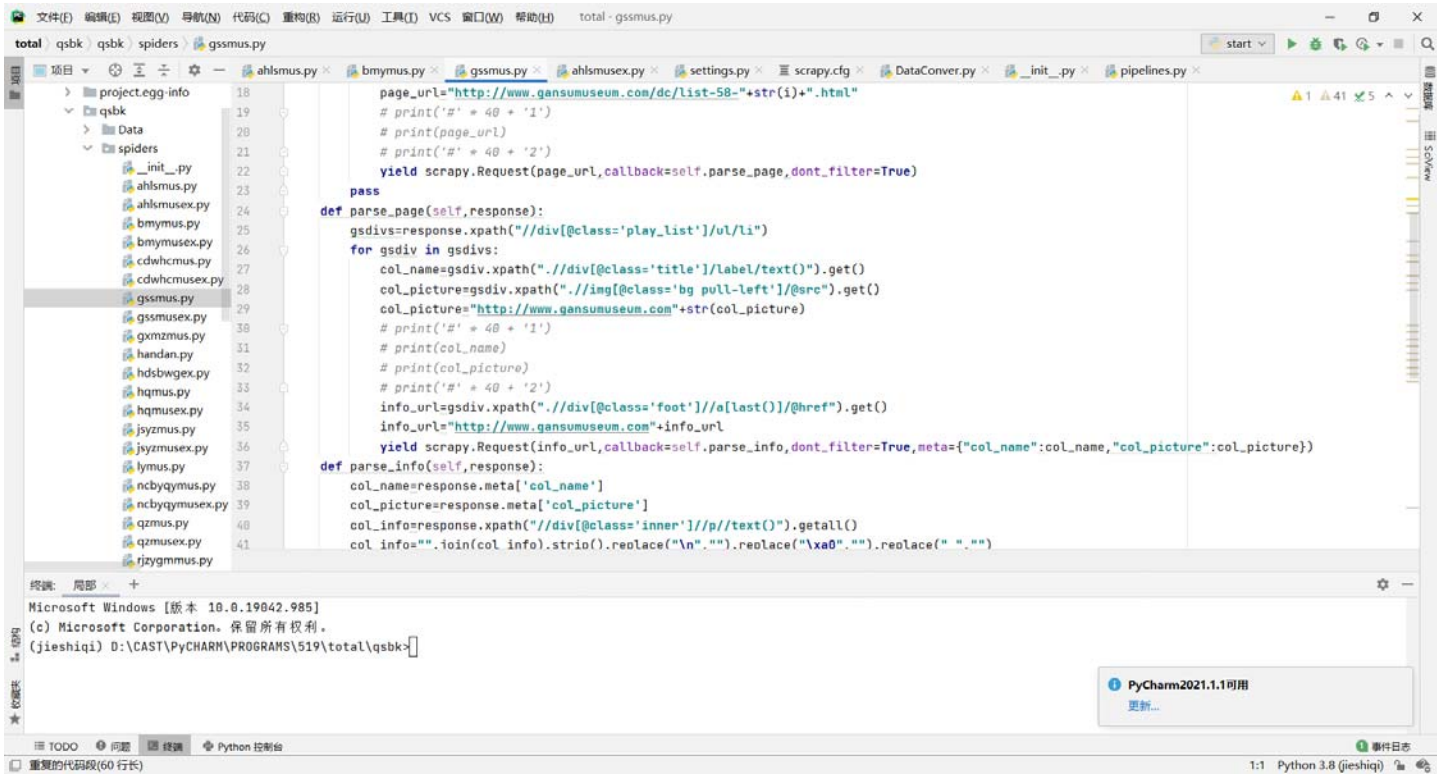
成功爬取数据

1、查看网站结构

分析网站类型，找到要爬取内容的路径

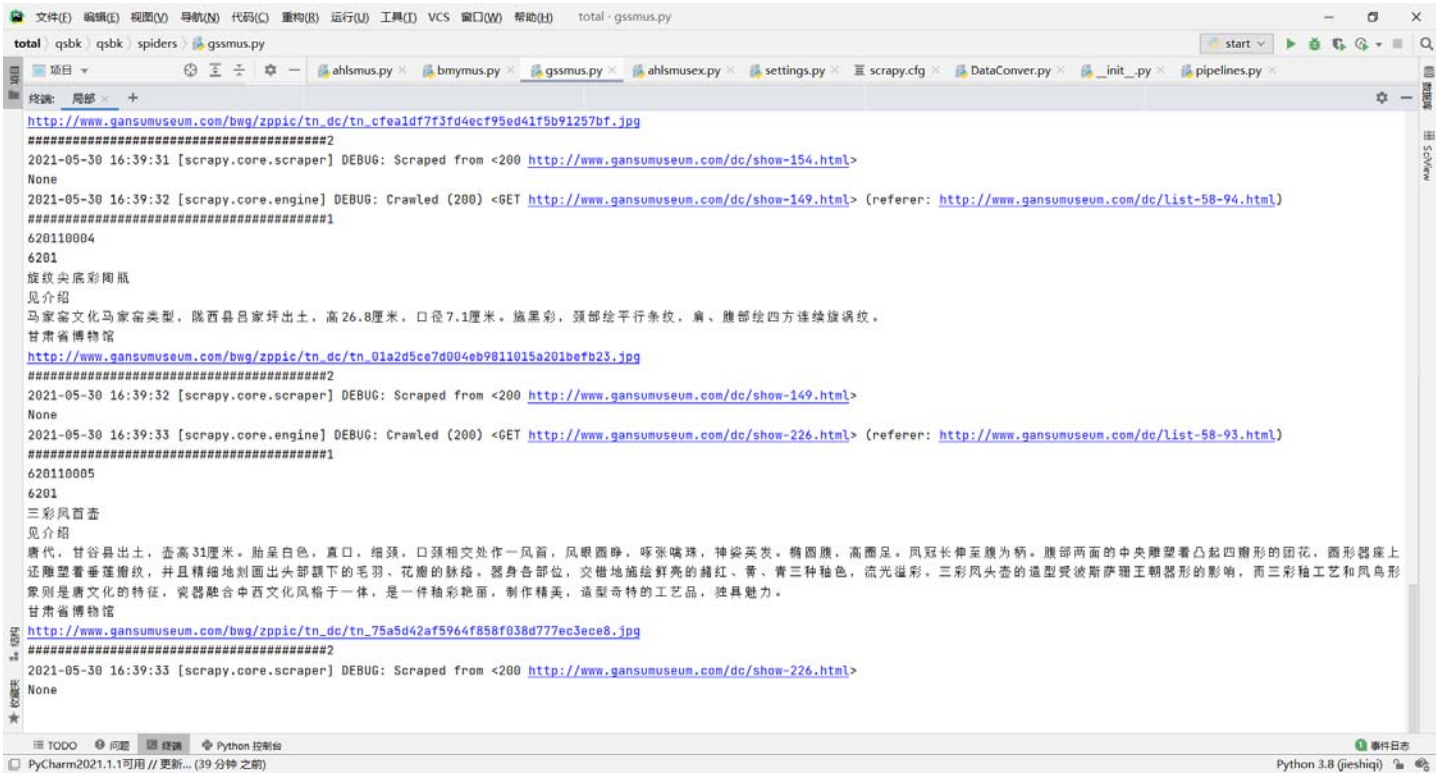
2、编写爬虫代码

记录网站地址，新建爬虫文件，根据爬取内容路径编写代码



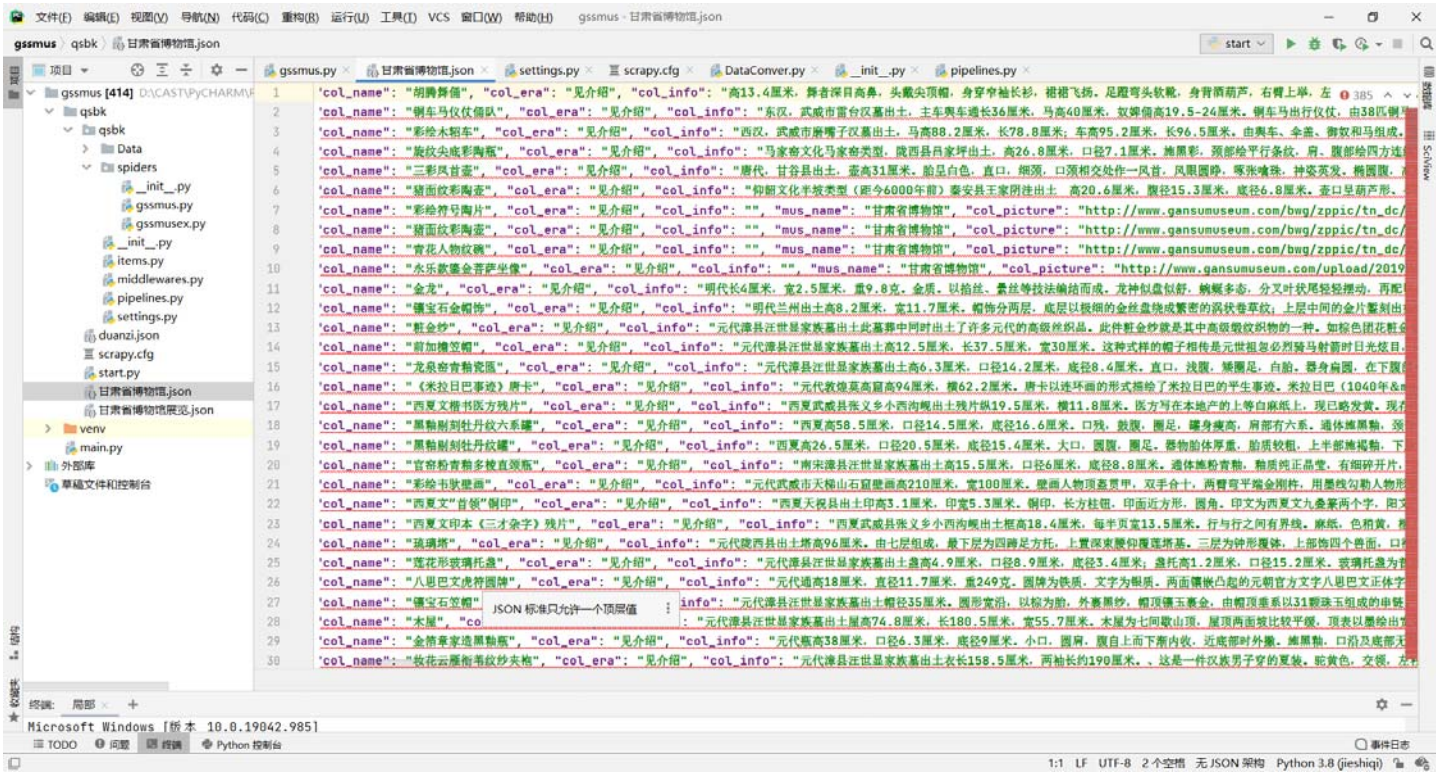
3、运行爬虫

终端编写代码： `scrapy crawl gssmus` ， 回车运行



数据处理成功

对爬取到的数据，提取重要内容、分类、整理、存储

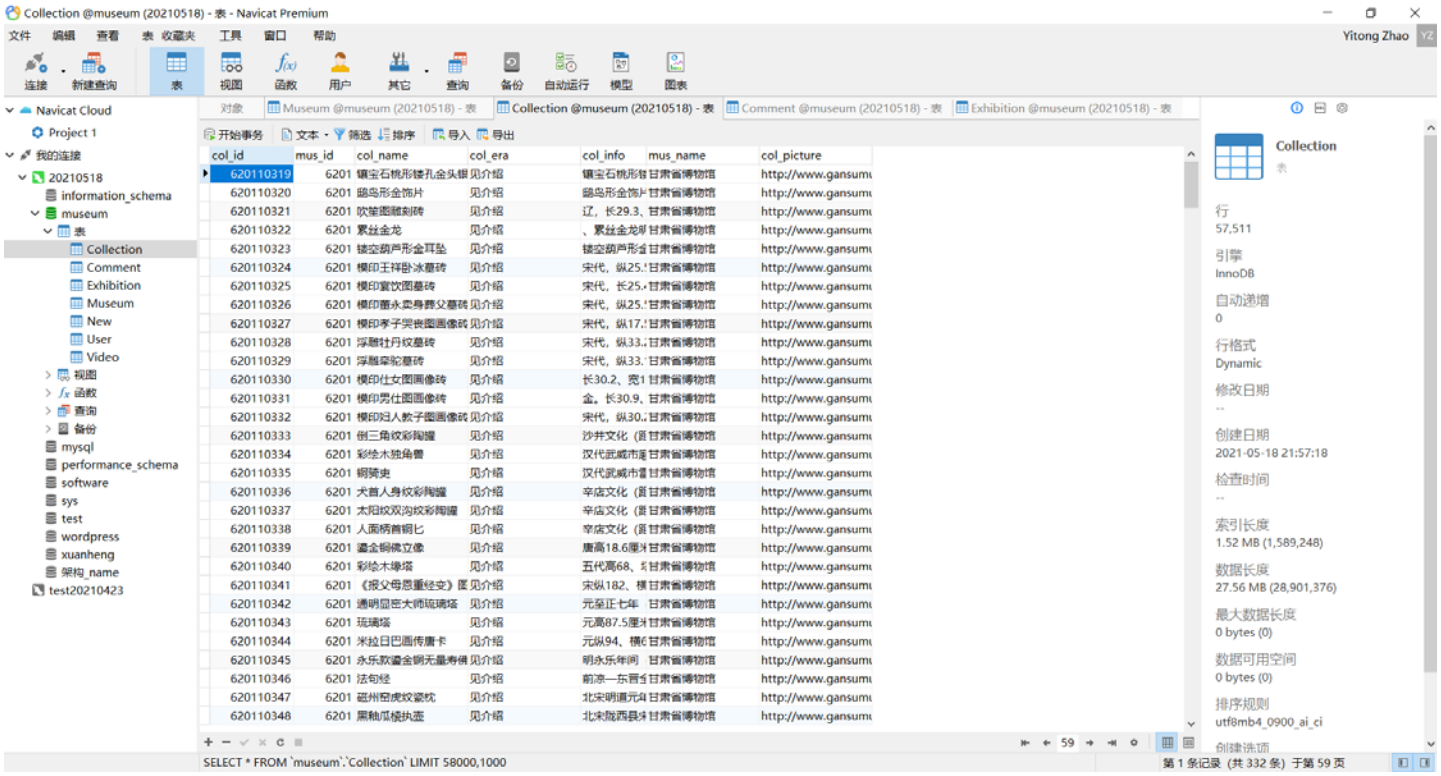


数据上传到数据库

修改管道： `pipelines.py` 文件

自定义管道：开启与MySQL数据连接，处理item数据，实现存储，编写sql语句实现存储，确认提交事

务，释放资源连接



将爬虫部署到服务器上并上传成功

安装对应组件、修改 scrapy.cfg

scrapy.cfg 文件

```
[settings]
default = qsbk.settings

[deploy:test]
url = http://123.56.13.242:6800/
project = Data_成员名
```

在scrapydWeb上运行爬虫