

# 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

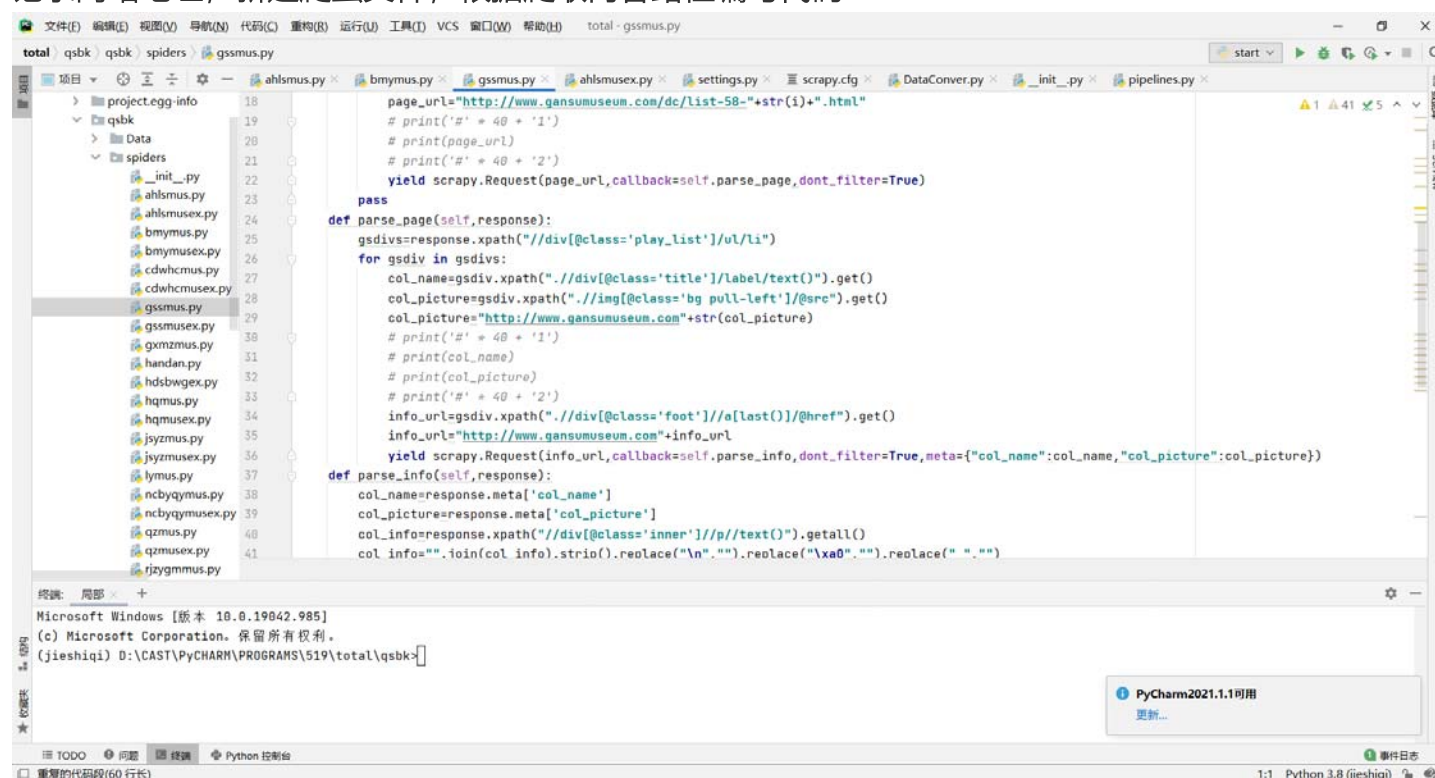
### 1、查看网站结构

分析网站类型，找到要爬取内容的路径



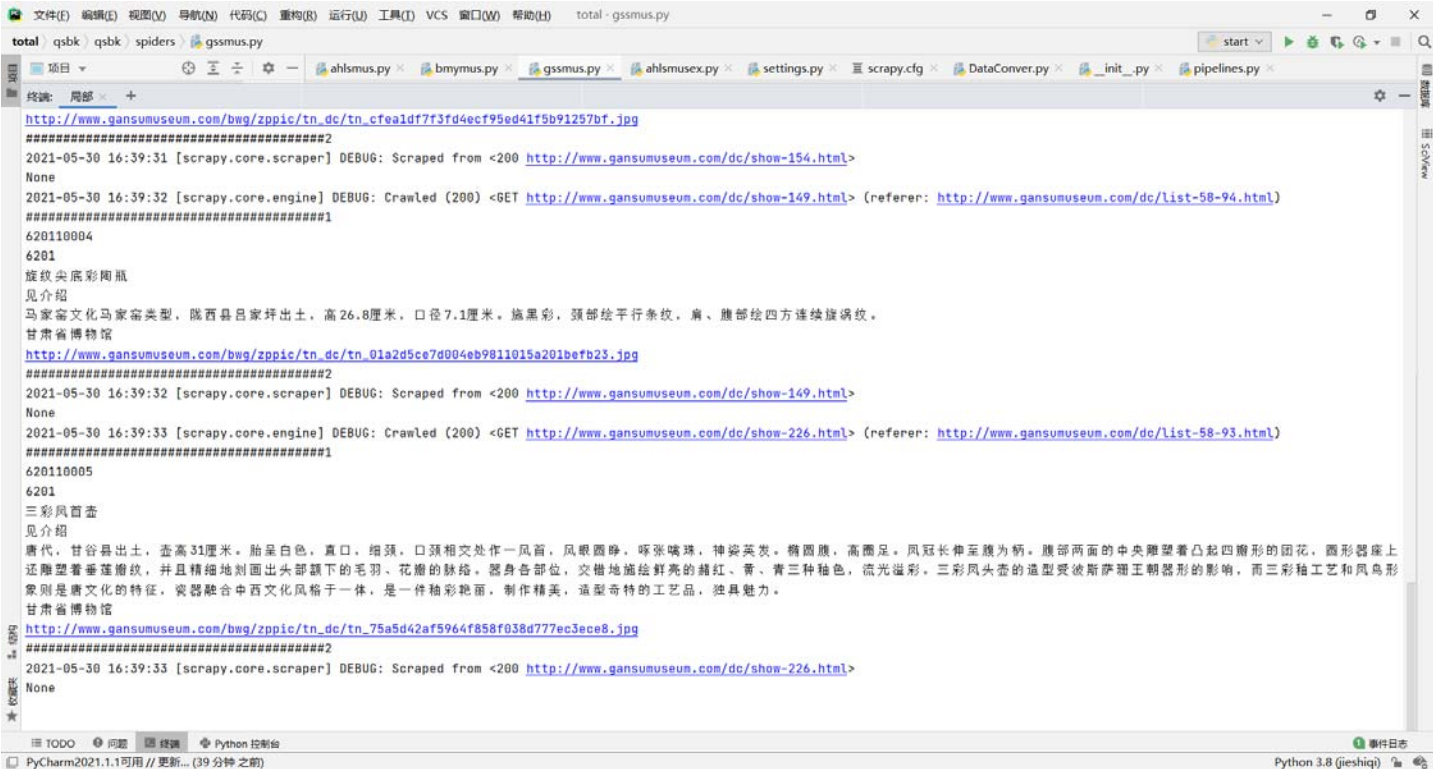
### 2、编写爬虫代码

记录网站地址，新建爬虫文件，根据爬取内容路径编写代码



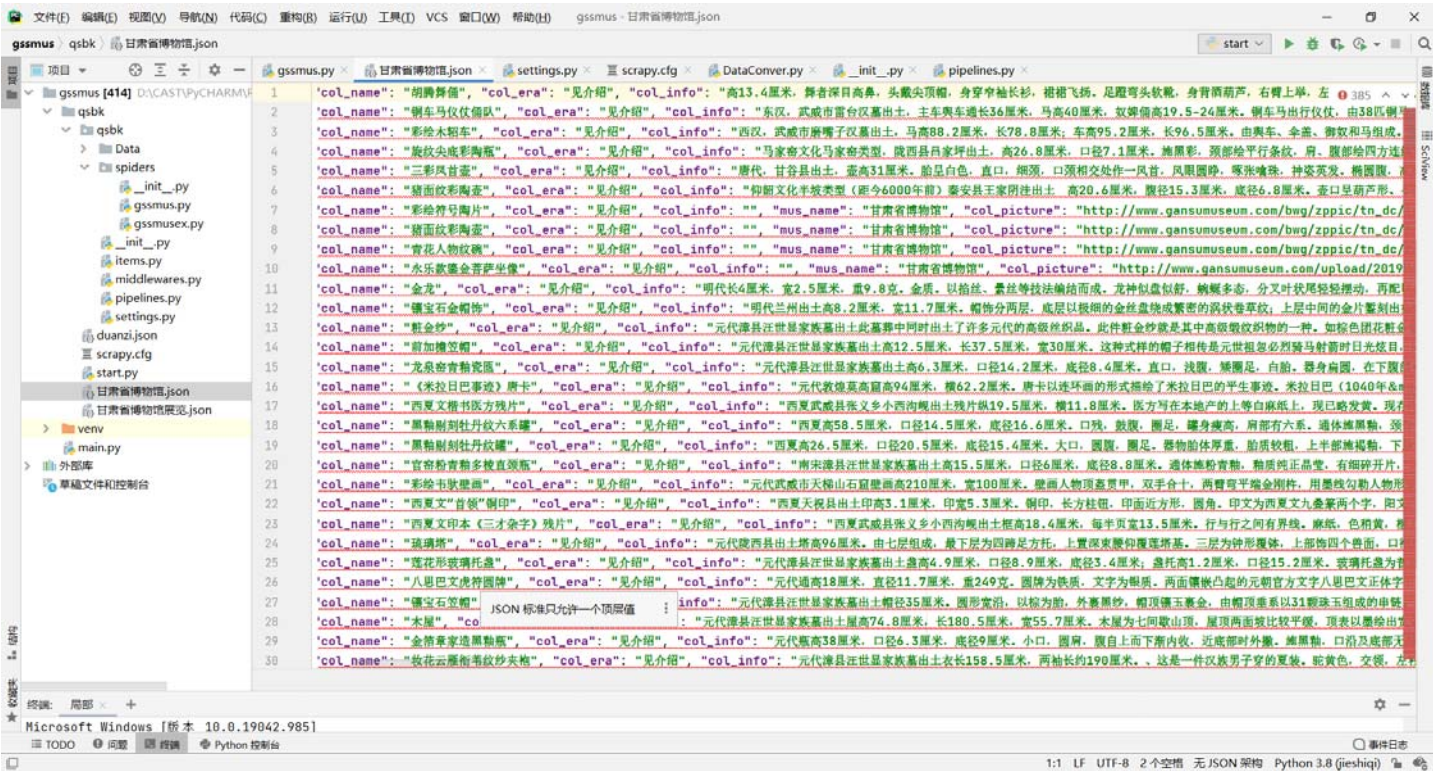
3、运行爬虫

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

Collection @museum (20210518) - 表 - Navicat Premium

文件 编辑 查看 表 收藏夹 工具 窗口 帮助

连接 新建查询 表 视图 函数 用户 其它 查询 备份 自动运行 模型 图表

Navicat Cloud

Project 1

我的连接

20210518

information\_schema

museum

表

Collection

Comment

Exhibition

Museum

New

User

Video

视图

函数

查询

备份

mysql

performance\_schema

software

sys

test

wordpress

xuanheng

架构\_name

test20210423

对象

Museum @museum (20210518) - 表

Collection @museum (20210518) - 表

Comment @museum (20210518) - 表

Exhibition @museum (20210518) - 表

开始事务 文本 筛选 排序 导入 导出

col_id	mus_id	col_name	col_era	col_info	mus_name	col_picture
620110319	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110320	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110321	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110322	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110323	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110324	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110325	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110326	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110327	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110328	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110329	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110330	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110331	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110332	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110333	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110334	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110335	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110336	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110337	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110338	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110339	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110340	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110341	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110342	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110343	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110344	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110345	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110346	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110347	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.
620110348	6201	镇宝石桃形镂空金头簪	见介绍	镇宝石桃形镂空金头簪	甘肃省博物馆	http://www.gansum.

SELECT \* FROM `museum`.`Collection` LIMIT 58000,1000

第 1 条记录 (共 332 条) 于第 59 页

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

ScrapydWeb

127.0.0.1:6800

1/2

0 0 14

ip install logparser on host "127.0.0.1:6800" and run command "logparser". Or wait until LogParser parses the log.

PROJECT (Data\_ZhaoYitong), SPIDER (gssmus)

Log analysis Log categorization Progress visualization View log Crawler stats

project	Data_ZhaoYitong
spider	gssmus
job	2021-05-30T16_51_17
first_log_time	2021-05-30 16:51:32
latest_log_time	2021-05-30 17:26:14
runtime	0:34:42
crawled_pages	476
scraped_items	0
shutdown_reason	N/A
finish_reason	finished
log_critical_count	0
log_error_count	378
log_warning_count	1
log_redirect_count	5
log_retry_count	2