****

**《可视计算基础》**

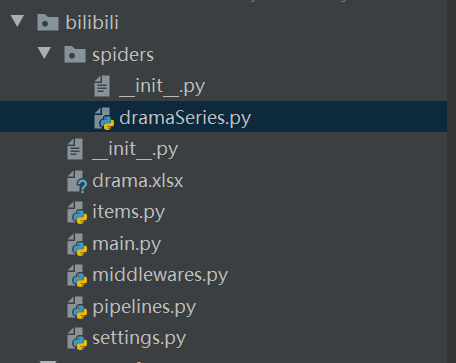
**代码说明文档**

# Bilibili番剧爬取

**实验人员：**

**18051624 沈建鑫**

首先用scarpy projectstart name创建一个scarpy工程，工程目录如下：



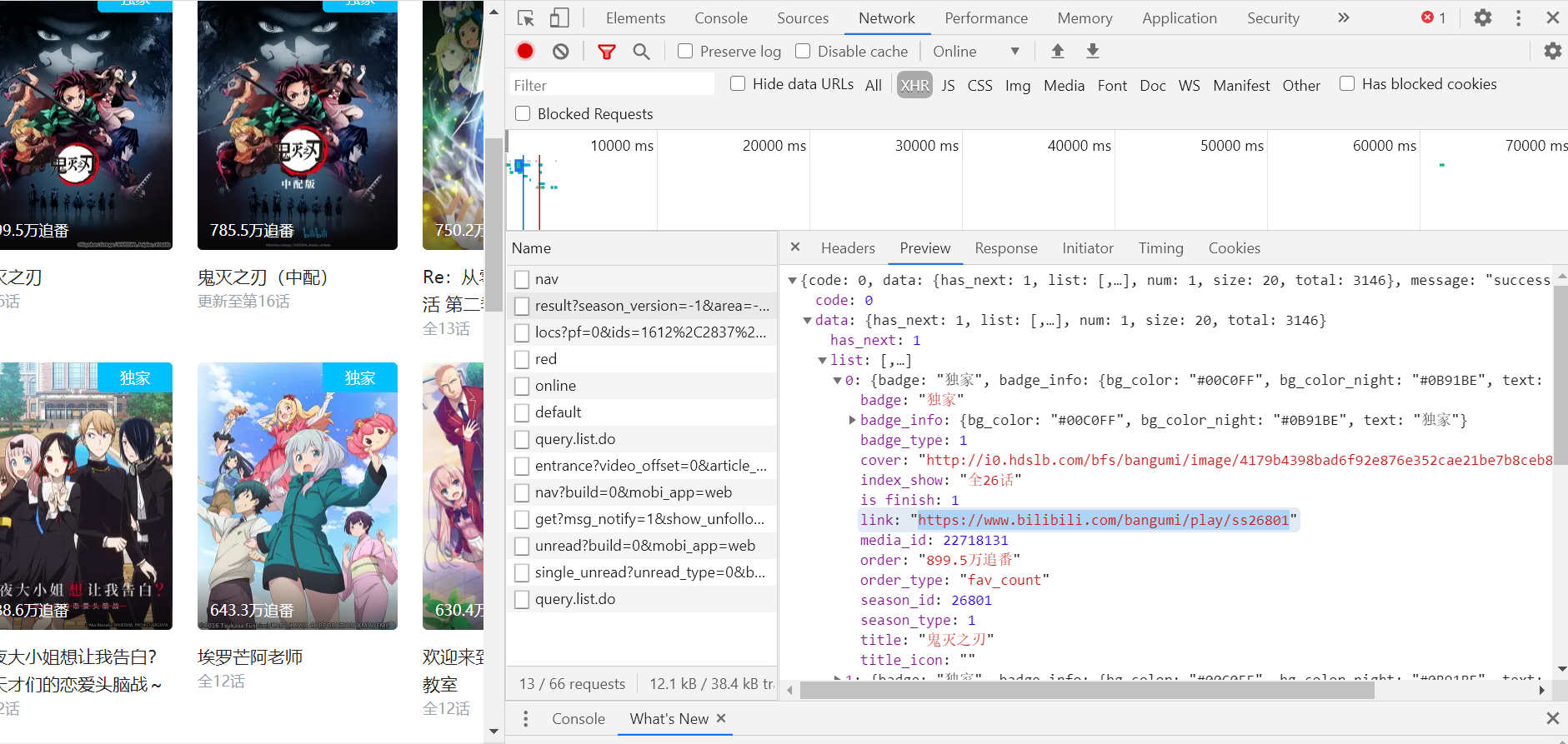
其中，spiders里面放爬虫代码，items中定义了爬取的字段，main是用来代替命令行输入crawl，middlewares是中间层处理函数，pipelines是数据处理相关文件，settings存放基本配置。

第一步，我们确定要爬取的网址：

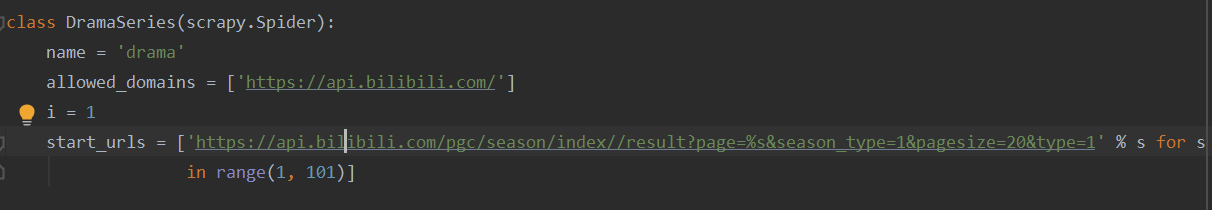
<https://www.bilibili.com/anime/index/#season_version=-1&area=-1&is_finish=-1&copyright=-1&season_status=-1&season_month=-1&year=-1&style_id=-1&order=3&st=1&sort=0&page=1>

剔除一些没必要的信息后：

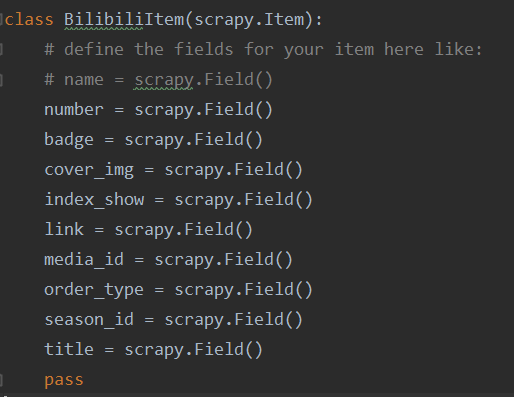
https://api.bilibili.com/pgc/season/index//result?page=1&season\_type=1&pagesize=20&type=1



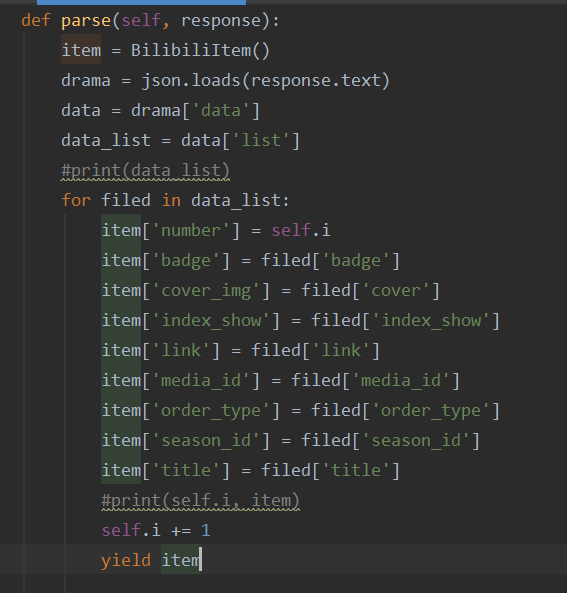
通过观察可以得到，要想爬取多页的数据只要改变page的信息即可，因此可以确定我们的url，这里我们爬取100页。



接下来在items中定义我们需要爬取的字段，通过观察该文档的第一幅图即可得到哪些字段是需要爬取的。



定义完后，回到爬虫代码，可以完成parse函数，实现爬虫解析相应的主要逻辑了。

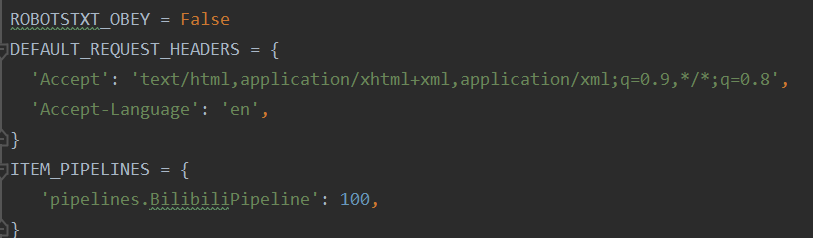


到这一步结束已经基本实现爬虫的功能了，接下去完善一下用scrapy projectstart 创建出来的工程文件中的其他代码文件。

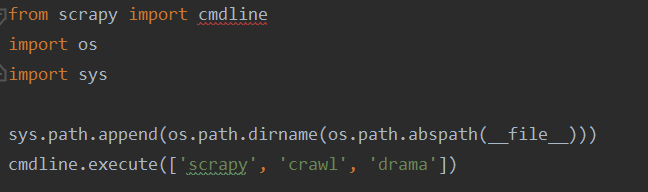
对于文件写入，根据刚刚解析出来的字段构建列表，然后用openpyxl的 Workbook库来实现文件的直接写入。



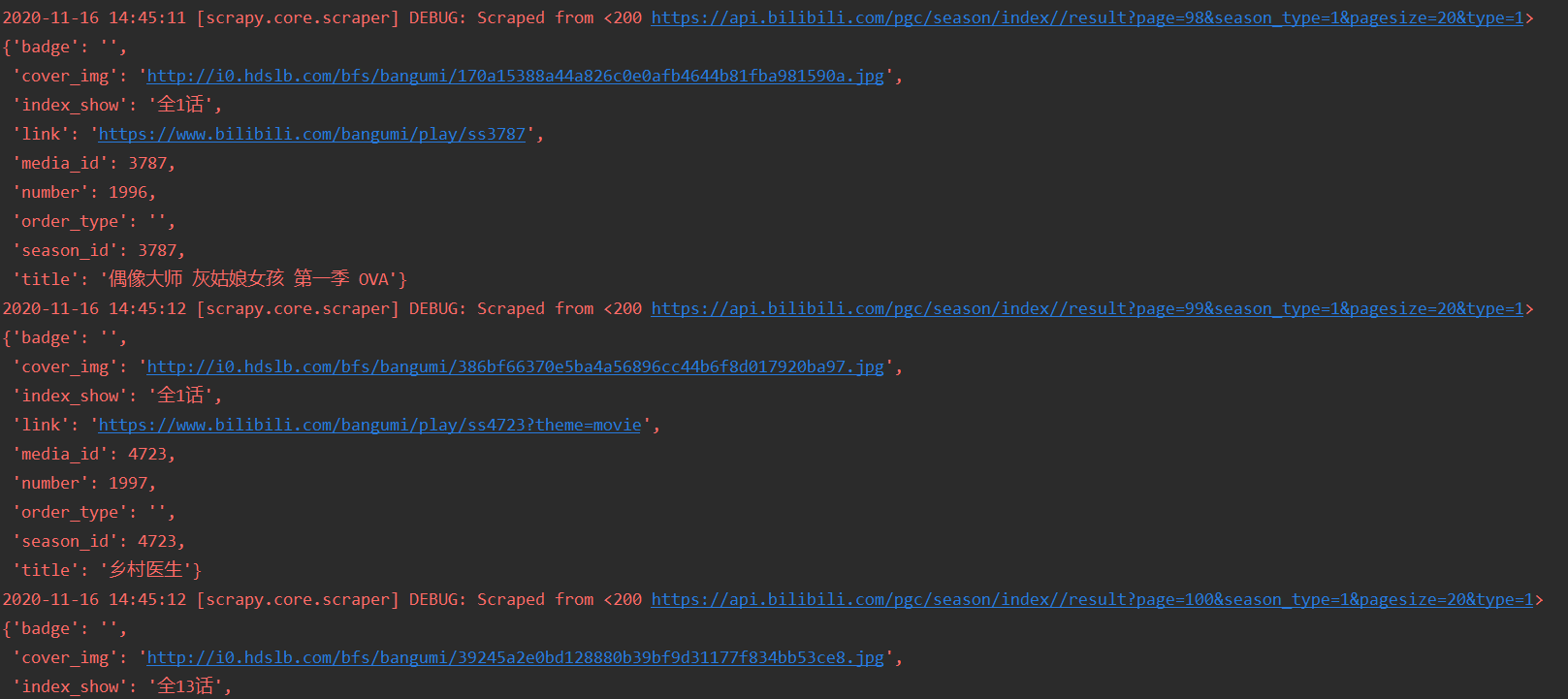
在settings中，我们添加scrapy发送http请求默认使用的请求头以及pipelines的设置。



最后写一个简单的main函数来代替每次命令行运行输入scrapy crawl name。



运行main就可以看到爬取出的结果了。



Xlsx：

