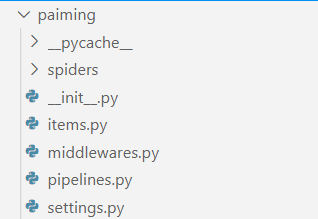
## 1.程序的框架与结构

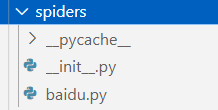
项目采用scrapy爬虫爬虫框架对数据进行采集。

在命令行输入scrapy startproject paiming 后就可自动生成这个项目文件，如下：

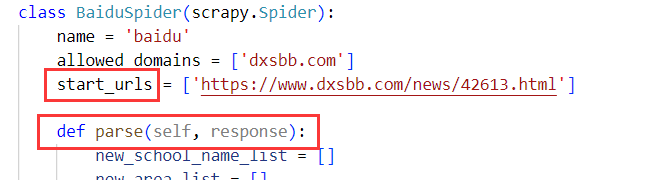


|  |  |
| --- | --- |
| Spiders文件夹 | 存储爬虫文件，编写核心爬虫代码 |
| Items.py | 定义数据结构的地方 |
| Middlewares.py | 中间件、代理文件 |
| Pipelines.py | 管道文件 |
| Setting.py | 配置文件 |

在命令行输入 scrapy genspider baidu baidu.com 就会在spiders文件夹下创建一个叫baidu.py的爬虫文件，如下：



此后只用定义需要爬取网站的url以及写parse函数对返回的html进行解析即可：

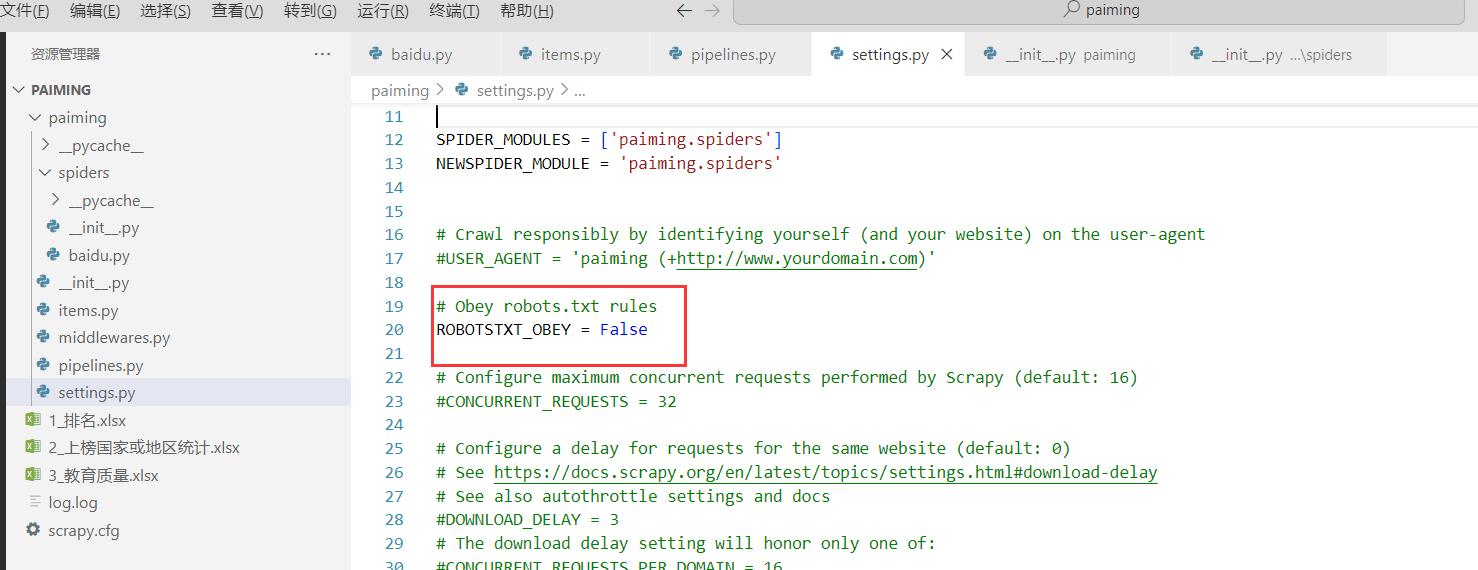


## 2.项目的代码说明

由于大多数文件都是项目自动生成的，因此只需要解释baidu.py里的parse函数与setting.py里的配置内容即可。

### 2.1 Settings.py的说明

首先需要在setting里把是否遵守robot协议改为否。



并且设置一下爬取时debug 信息的输出，改为输出到日志文件中：



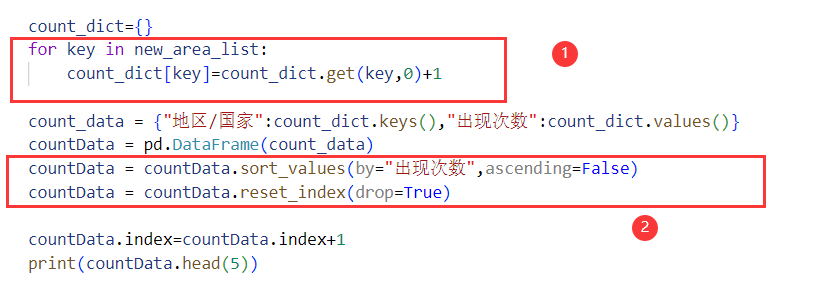
### 2.2 对parse函数的说明



第一部分，对爬取到的html使用xpath进行解析，分别提取学习名称与地区。

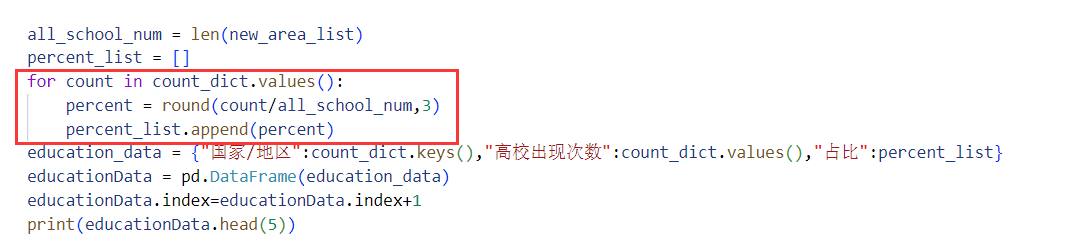
第二部分，由于提取到的数据带有<td>xxx</td>的标签，因此对他们进行了去除。

第三部分，将获取到的学校名称的列表与其国家列表进行处理，组合成了一个DataFrame，即rankData，方便后续生成excel文件。

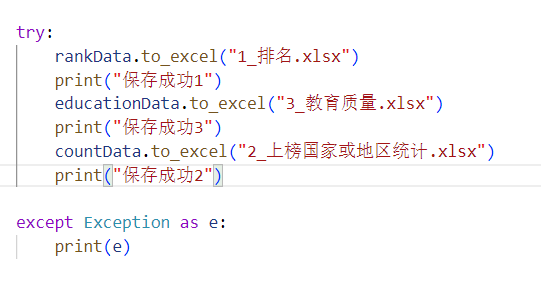


第一部分是对区域数据进行统计，以字典的数据结构进行存储。

第二部分，首先把字典生成了DataFrame,即countData。再对这个DataFrame进行降序排序，并重新调整了其序号。



这个部分主要是统计每个国家高校出现的占比，以评估教育质量。这里我对占比结果保留了3位小数。



最后这个就是对上述的三个dataframe生成了excel进行保存。