**爬取有路网里的图书信息**

1. **爬虫思路分析**

**1、网络数据采集所用到的技术**

所用到的技术和第一个项目一样，有Scrapy爬虫框架,静态网页爬取技术以及反反爬技术

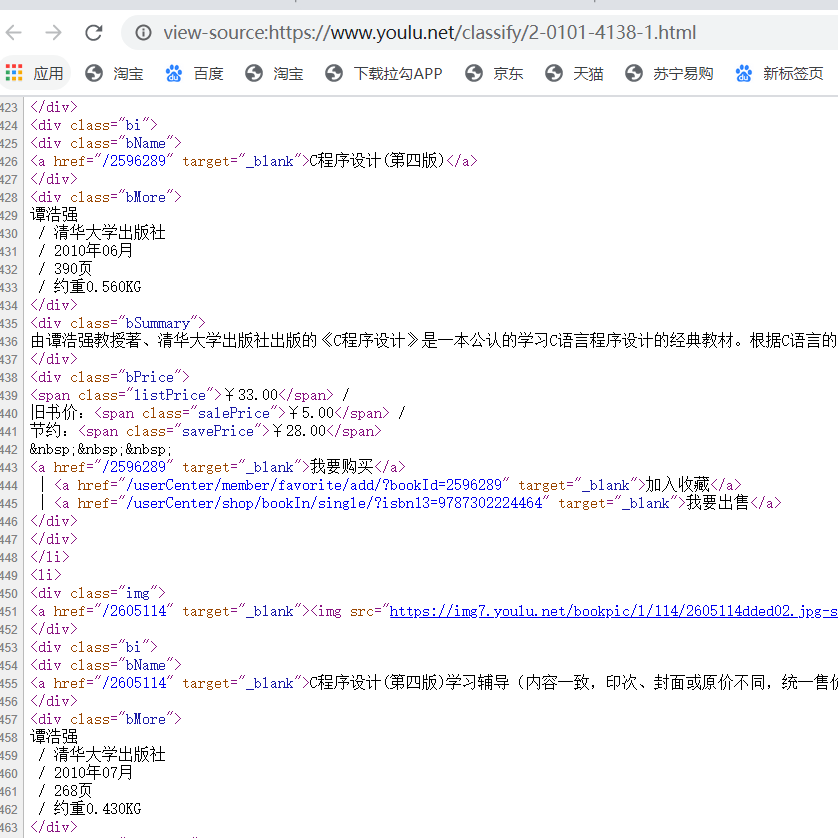
**2、爬虫实现思路**

1. 判断网页是静态网页还是动态网页

打开网页进入图书分类页面



判断网页是静态网页还是动态网页查看网页源代码

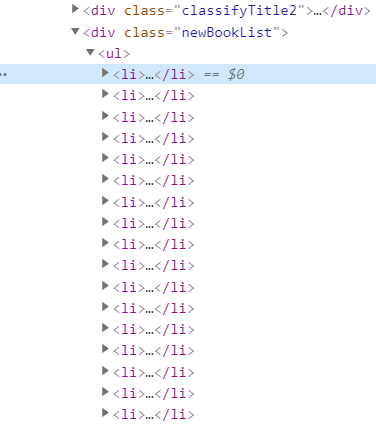


发现网页源代码有每个图书的相关信息，故为静态页面。

1. 选取需要爬取的数据

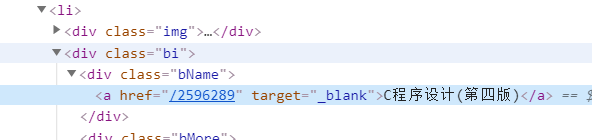
确定爬取的数据有：书籍名字name，书籍链接book\_link，作者author，出版社press，出版时间press\_data，书籍页数number，书籍重量weight，新书价格new\_price，旧书价格price，节约的钱save，书籍简述resume

查看网页结构:



发现每项工作的信息都在ul标签下的li标签里，所以可以先爬取li标签，然后从li标签爬取每项信息。先用xpath定位到li标签。

书籍名字name：



书籍名字是在a标签的文本内容，所以xpath语法后面需要加上text()。

书籍链接book\_link：



书籍链接在a标签的href里，但发现该链接不是全部链接，单独拿出来打开不了。先手动打开几个详细链接看看

<https://www.youlu.net/2596289>

<https://www.youlu.net/2605114>

<https://www.youlu.net/3383983>

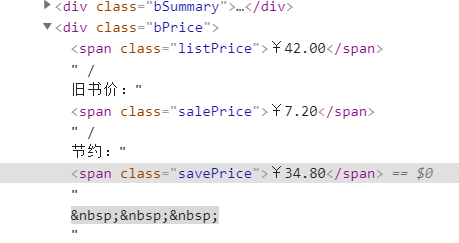
发现书籍详细链接和我们爬取的a标签里的href多了主页的链接https://www.youlu.net 所以需要将主页链接和爬取a标签里的href拼接起来。因此先爬取a标签里的href，再将其拼接起来。

作者author，出版社press，出版时间press\_data，书籍页数number，书籍重量weight：



这些都在一个标签内，所以需要split（）函数对爬取的信息进行切割。

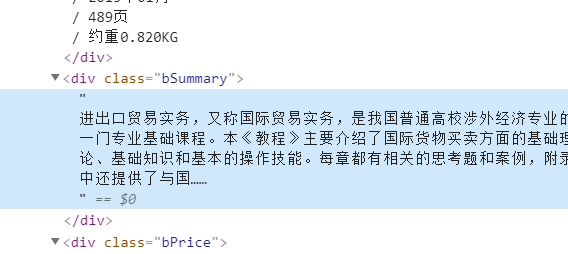
新书价格new\_price，旧书价格price，节约的钱save：



这些分别是span里的文本内容。故比较简单。

书籍简述resume：

在div里的文本内容，但有换行符，所以也要用split（）函数进行切边分割



1. 查看多页爬取需要的技术

查看前几页的网址，手动翻页进行查看：以下是前3页的网址

<https://www.youlu.net/classify/2-0315-1810-1.html>

<https://www.youlu.net/classify/2-0315-1811-2.html>

<https://www.youlu.net/classify/2-0315-1811-3.html>

发下每页的网址只有后面的数字发生改变，故爬取多页只需要改变数字即可。

爬取的数据过多可能会被网址禁止访问，所以需要反反爬技术即随机ip,随机userAgent。

随机ip在西刺代理网找到<https://www.xicidaili.com/>，还需要对ip进行验证可不可用。

打开其他类型的图书的网页信息，对其每页的网址，发现也是这样的变化。所以其他类型的翻页操作一样。

查看其他类型的图书，发现其网页地址不一样，但其爬取的内容所在的标签一样。故此，爬取其他图书的内容只需要换网址即可。

1. **爬虫实现步骤**

（1）打开控制台创建名称为niuke的yuolu项目

输入scrapy startpoject yuolu创建名称为yuolu的scrapy项目,

进入yuolu后输入scrapy genspider book youlu.net 创建一个book的脚本文件。

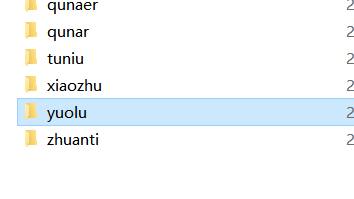
1. 打开PyCharm,打开创建好的yuolu项目。
2. 打开itmes.py，创建爬取内容字段
3. 打开book.py,创建爬取内容
4. 打开pipelines.py，写入代码对爬取的内容进行保存在csv中
5. 打开settings.py，创建项目所需要的的参数
6. 创建main.py,以代表控制台运行程序
7. 爬取几页试试代码可不可行并进行修改。
8. 进行反反爬，创建iptest.py,对在西刺网找的ip进行验证。多找几个。
9. 再次打开settings.py，用（8）里找到的ip构建一个ip池。
10. 打开middlewares.py。构造使用代理ip进行爬取网页。
11. 运行程序爬取所有信息。
12. **爬虫代码实现**

**1、第一种爬虫技术Scrapy爬虫框架、代码及注释**

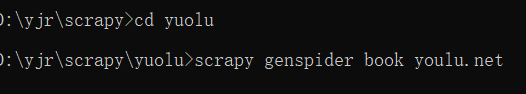
创建scrapy项目：

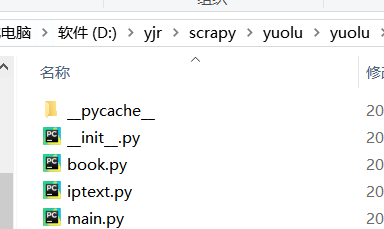
打开控制台输入scrapy startpoject youlu创建名称为youl的scrapy项目：



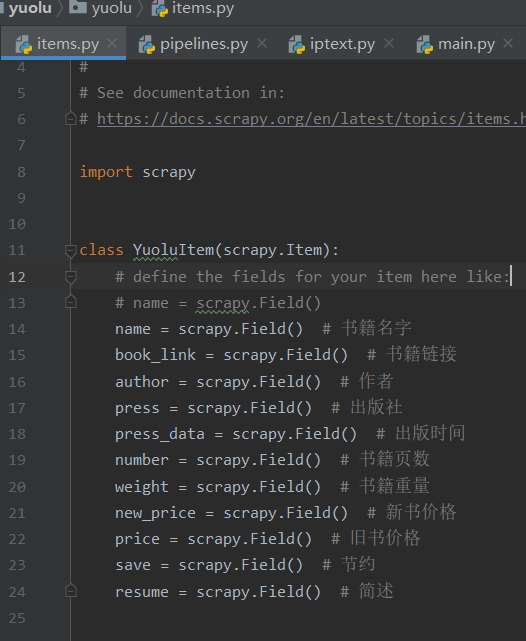


进入yuolu后输入scrapy genspider book youlu.net 创建一个book的脚本文件。



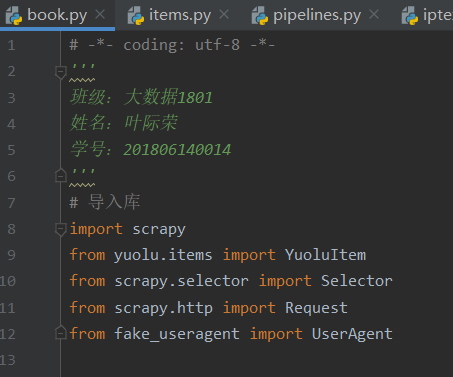


打开PyCharm,打开创建好的niuke项目，打开itmes.py，创建爬取内容字段：

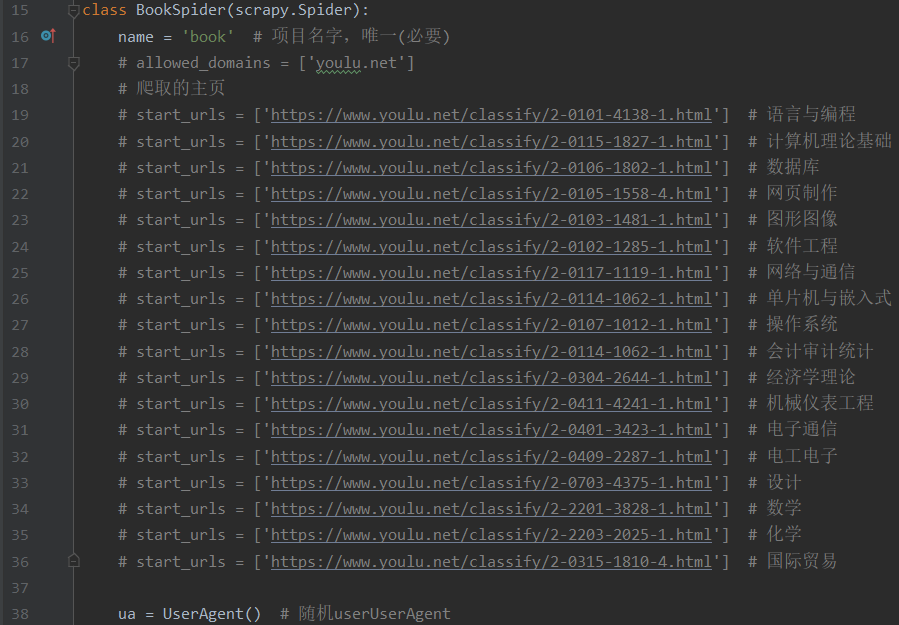


**2、第二种爬虫技术静态网页爬取技术、代码及注释**

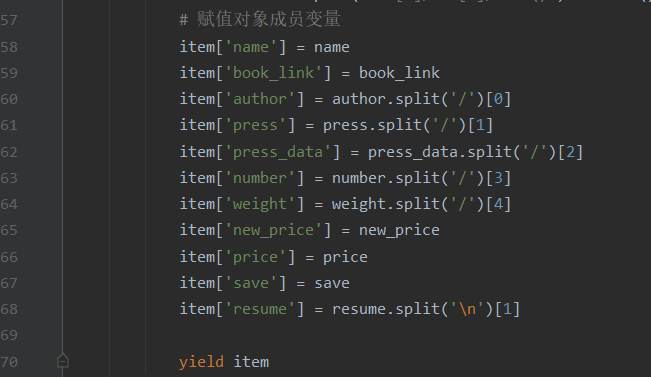
打开book.py，写入爬取的主页及用xpath爬取所有相关信息

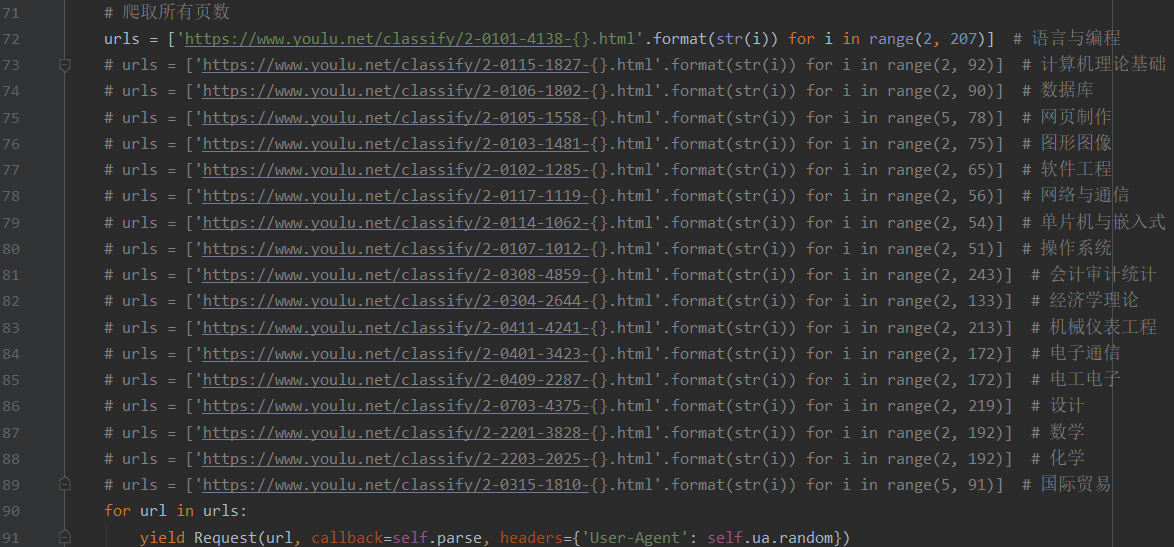


爬取的相关链接

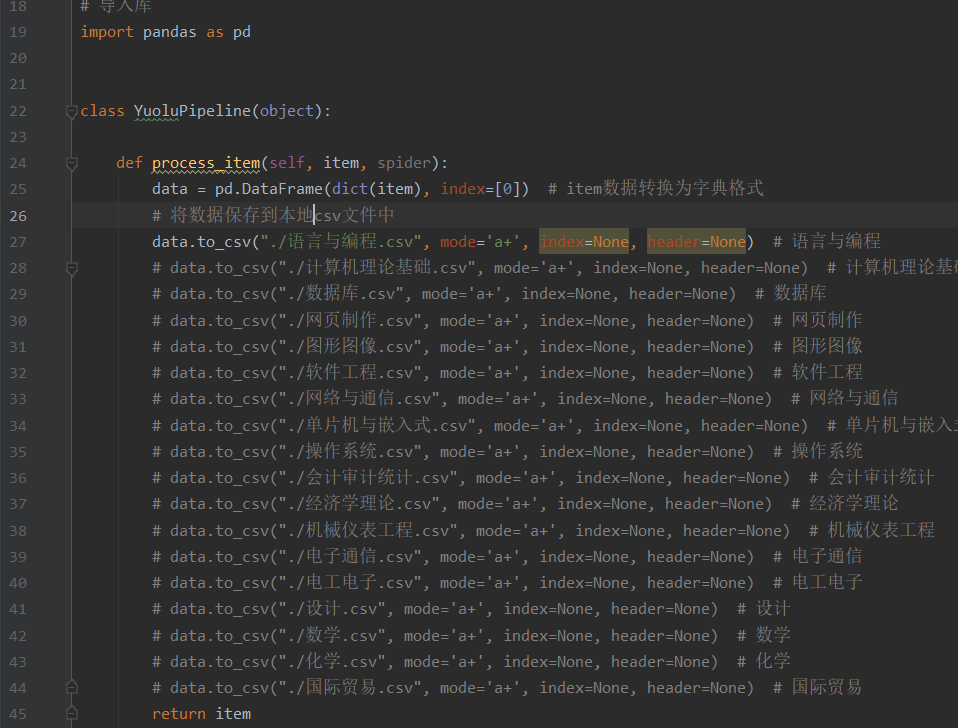




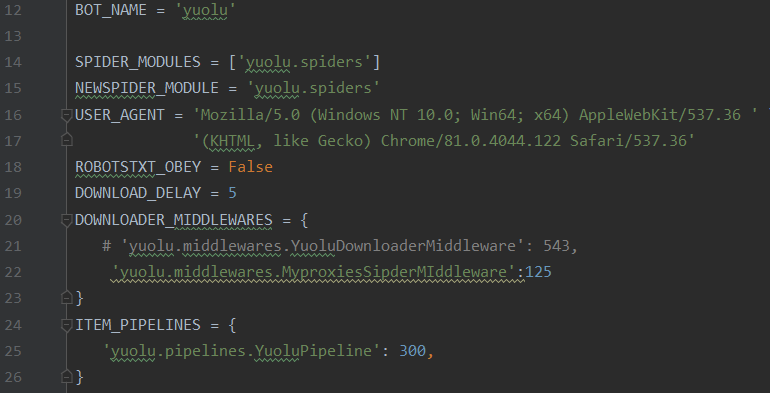




打开pipelines.py，写入代码对爬取的内容进行保存在相关的csv中



打开settings.py，创建项目所需要的的参数



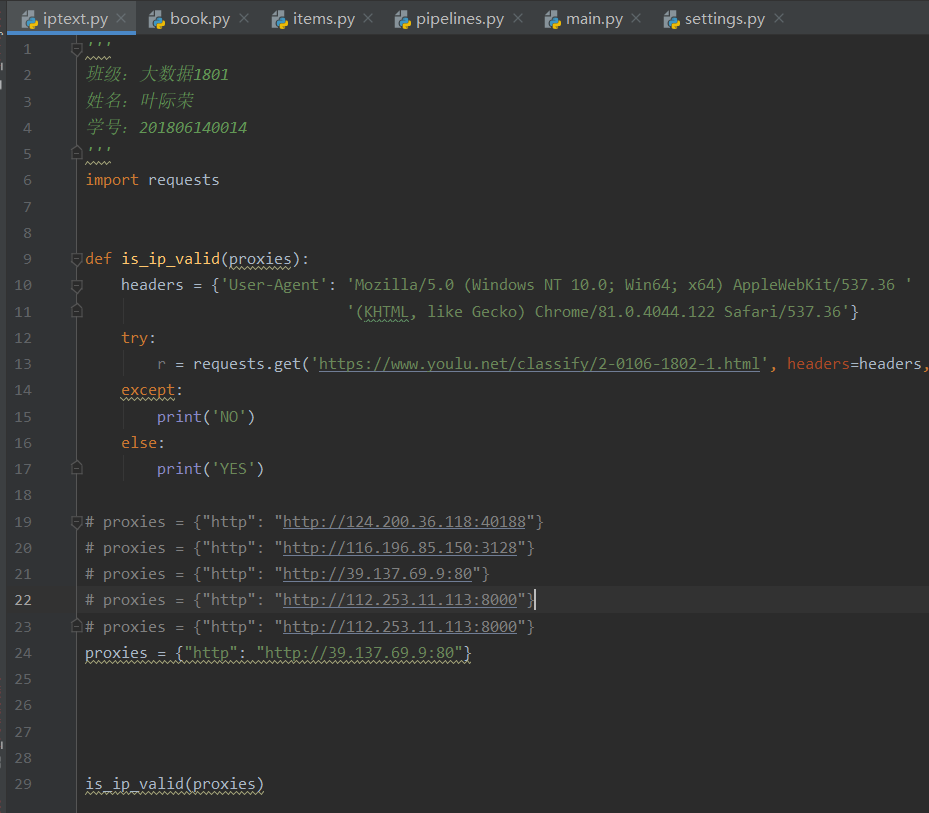
创建main.py,以代表控制台运行程序



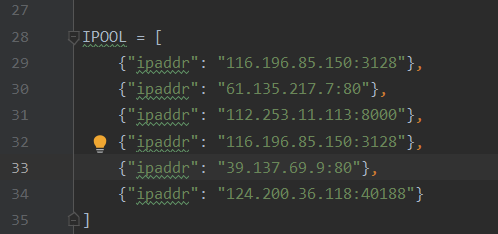
**3、第三种爬虫技术反反爬技术、代码及注释**

创建iptest.py,对在西刺网找的ip进行验证

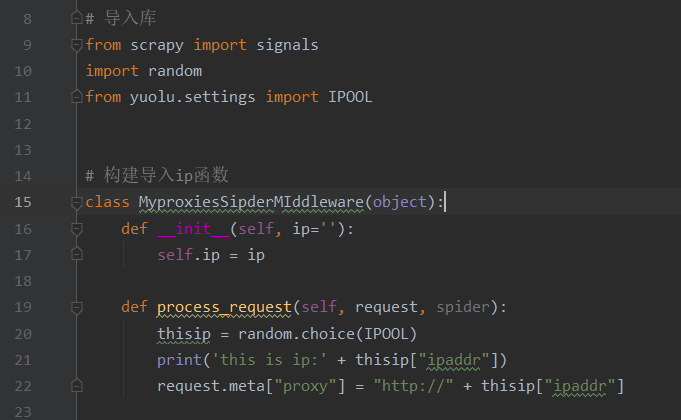
ip可用输出yes,不可用则输出no。



再次打开settings.py，用找到的ip构建一个ip池



打开middlewares.py。构造程序运行时使用代理ip进行爬取网页



共爬取大约3万条信息，暂未对此数据进行处理