豆瓣图书top250爬虫说明

本爬虫用到的模块有：

1. requests模块用来发起请求
2. bs4模块用来解析, bs将复杂HTML文档转换成一个复杂的树形结构，每个节点都是Python对象
3. pandas模块用来处理数据，放入excel中

爬取的网站主页为 <https://book.douban.com/top250>

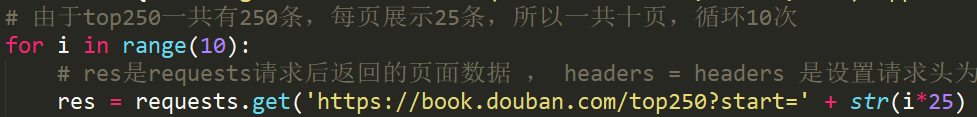
页面如下



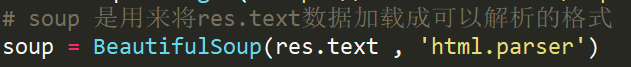
第二页的url为 <https://book.douban.com/top250?start=25>

第三页的url为 <https://book.douban.com/top250?start=50>

所以可以得到规律，用一个循环来遍历



之后将请求得到的结果转换为bs对象

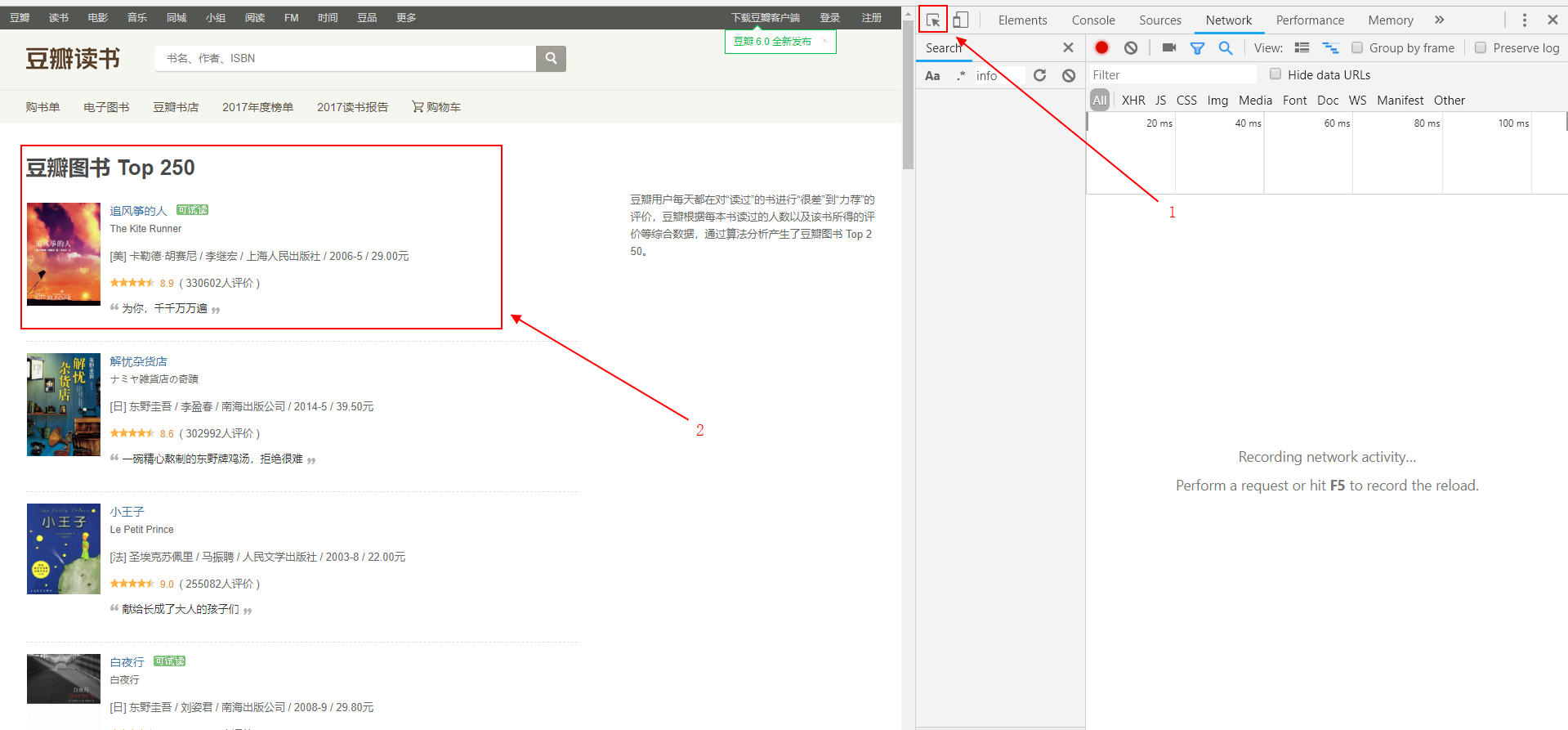


得到bs对象后，就可以对其进行解析，得到自己想要的数据

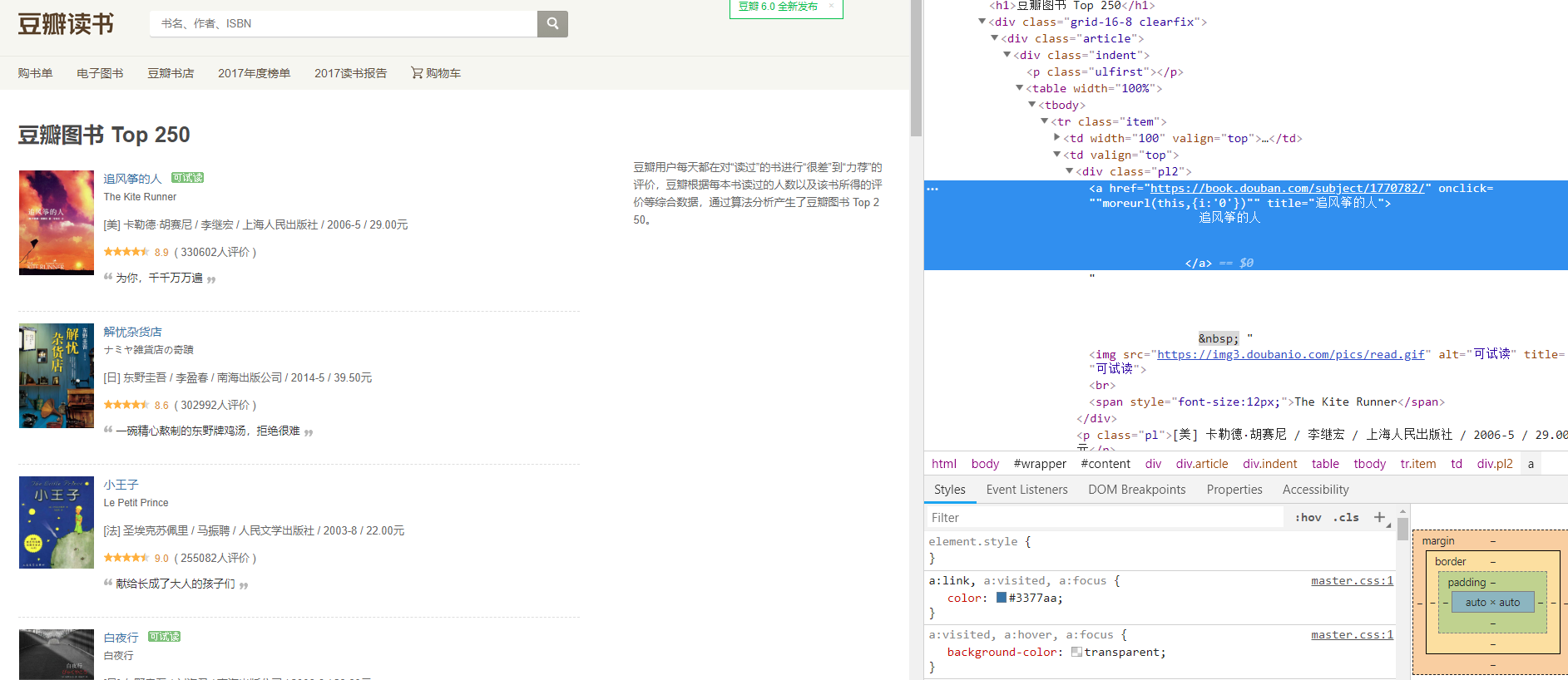


这些数据。

在Chrome浏览器中按F12进入审查元素



鼠标点击1处后再点击2处标题， 追风筝的人

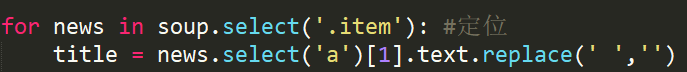


右侧就显示了页面对应的结构

首先我们取出标题



看到这个 table 下的 tr的class=“item”，所以代码如下



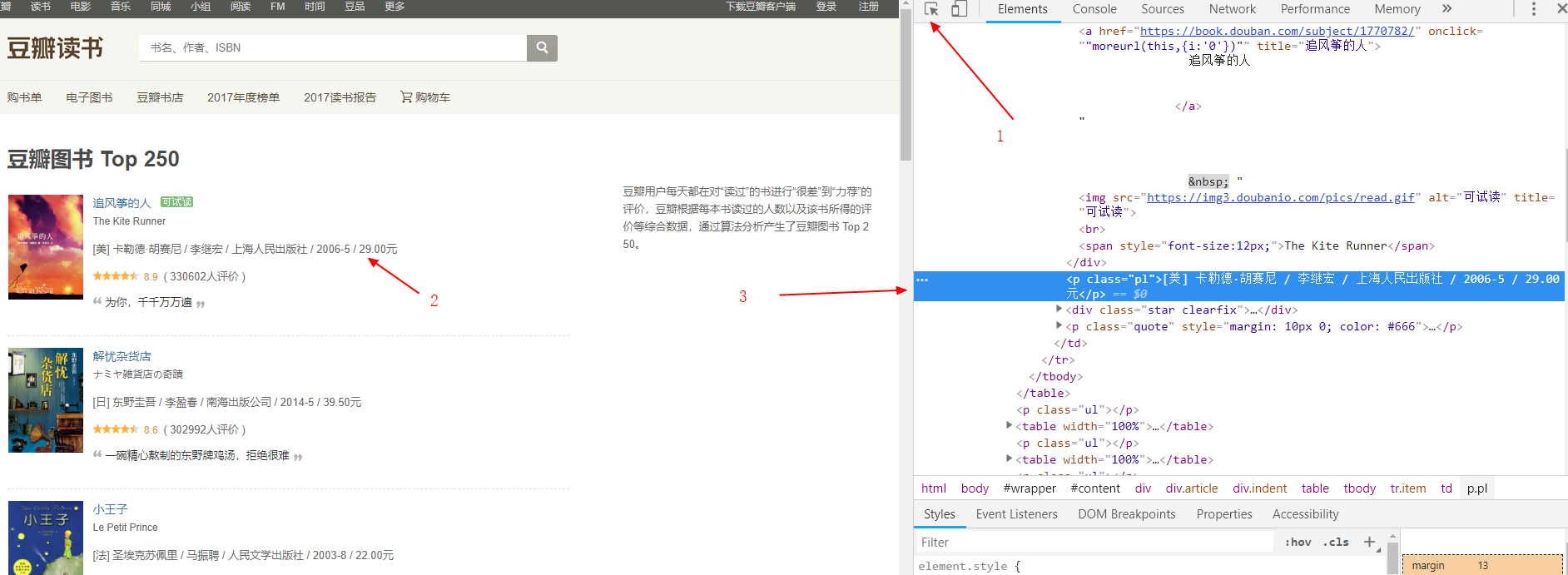
意思是遍历soup中所有class=”item”的对象，

title = news.select(‘a’)[1]

意思是找出news对象中第二个 a 标签，但是下表从0开始，所以是[1]， .text 是取出它的文本，犹豫文本里面存在空格，

所以用replace(‘ ’,’’)方法将空格替换为空,这样就取出了title

接下来取作者等信息



依次点击1，2 ，就会在右侧显示3的数据，这就是作者等信息，需要取出来

由于作者，出版社，时间，价格等很多内容在一个p标签中，我们要想办法将其分割开，

我们先用a 取到整个p标签中的文字。



即这些文字，我们需要分割开，使用split方法，根据 / 进行分割，最后一个是价格，倒数第二个是出版时间，倒数第三个是出版社

a = news.select('p')[0].text

price = a.split('/')[-1]

time = a.split('/')[-2]

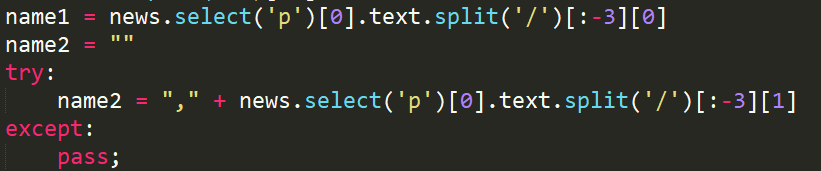
store = a.split('/')[-3]

然后我们要取作者的名字了 ，有的作者有一个名字，有的作者是一个外国作者和一个中国译者，为了解决这个问题，我们先根据(‘/’)分割，取出[0:-3]的内容，即出版社之前的内容。

用name1 接收所取出的第一个元素，也就是第一个名字。

第二个名字不一定存在，怎么办呢，我们先默认它为空 ：””

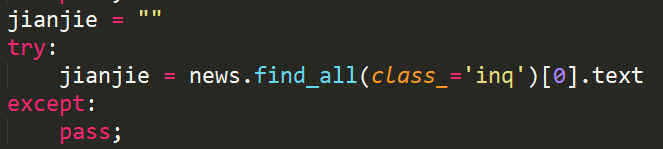
然后用try，except语句将接下来的内容进行包括，将有第二个名字的取出，没有的则pass掉。如果不用try语句，则会出现下标越界，程序出错。



然后要取出书籍的简介如图：

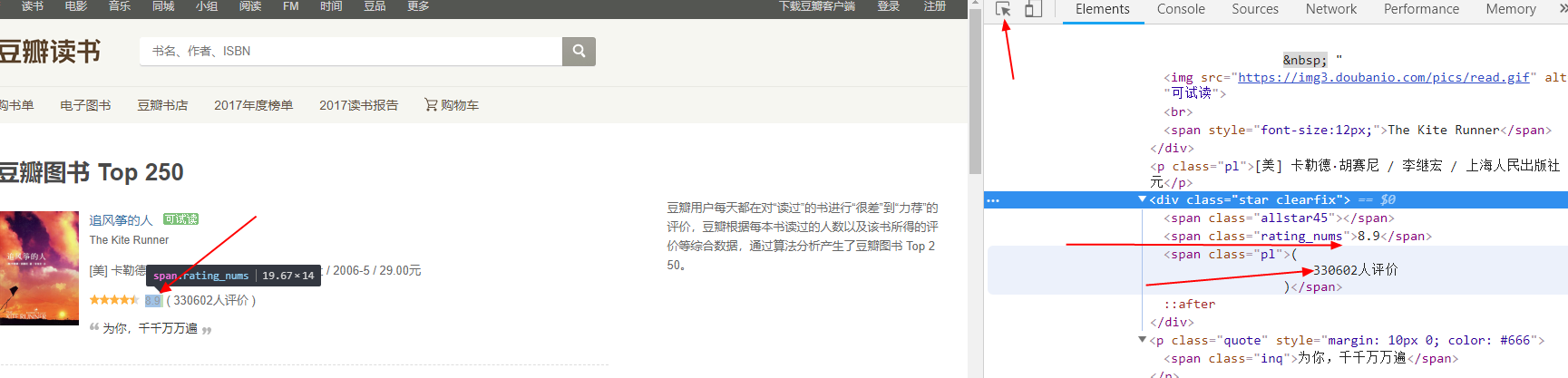


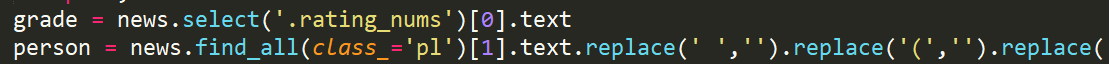
照着之前的方法，找到他的class



由于有的书没有简介，所以用try except包括，捕获异常

接下来找到评价人数和评价分数





分别取出这两条数据

然后将上面得到的所有数据组成一个dict，填充入之前声明的list 中

newsary.append({'title': title , 'name': name1 + name2 , 'person':person , 'jianjie': jianjie , 'price' : price , 'time' : time , 'store' : store ,'grade':grade})

最后将newsary 转换为pandas的DataFrame，以便存入excel中，

最后调用newsdf.to\_excel方法，将数据写入excel

