## 一．回顾一下上次的爬虫

1.上次的爬虫我们是在豆瓣读书上面《悠悠我心》的短评里面爬取用户的信息，并且吧用户的头像下载到本地中

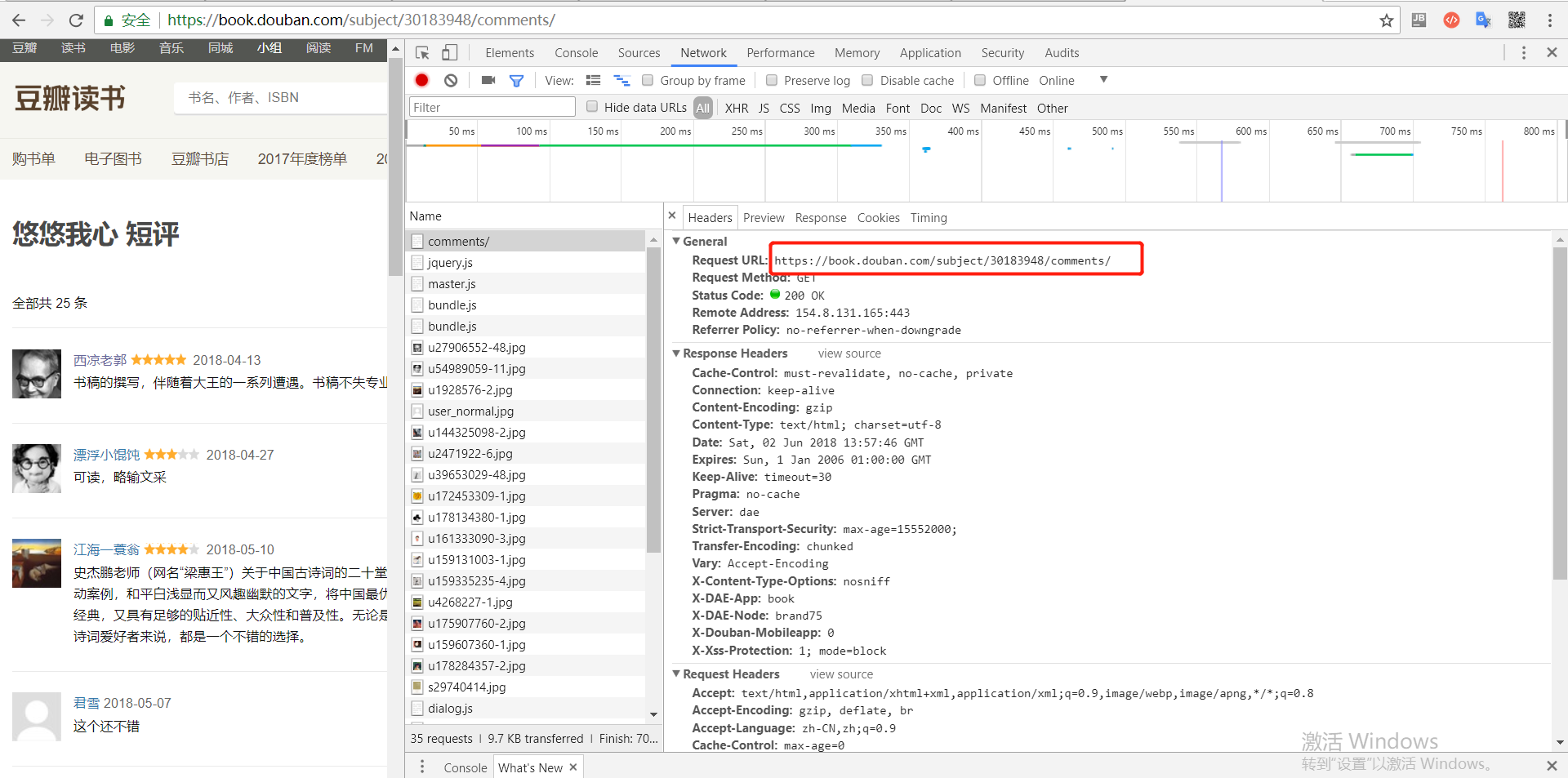
（https://book.douban.com/subject/30183948/comments/）首先打开网页，分析一下我们想要获取的东西是在哪里的，



看到浏览器的控制台的网络选项中，打开这个网页的时候会对服务器发送很多个资源请求

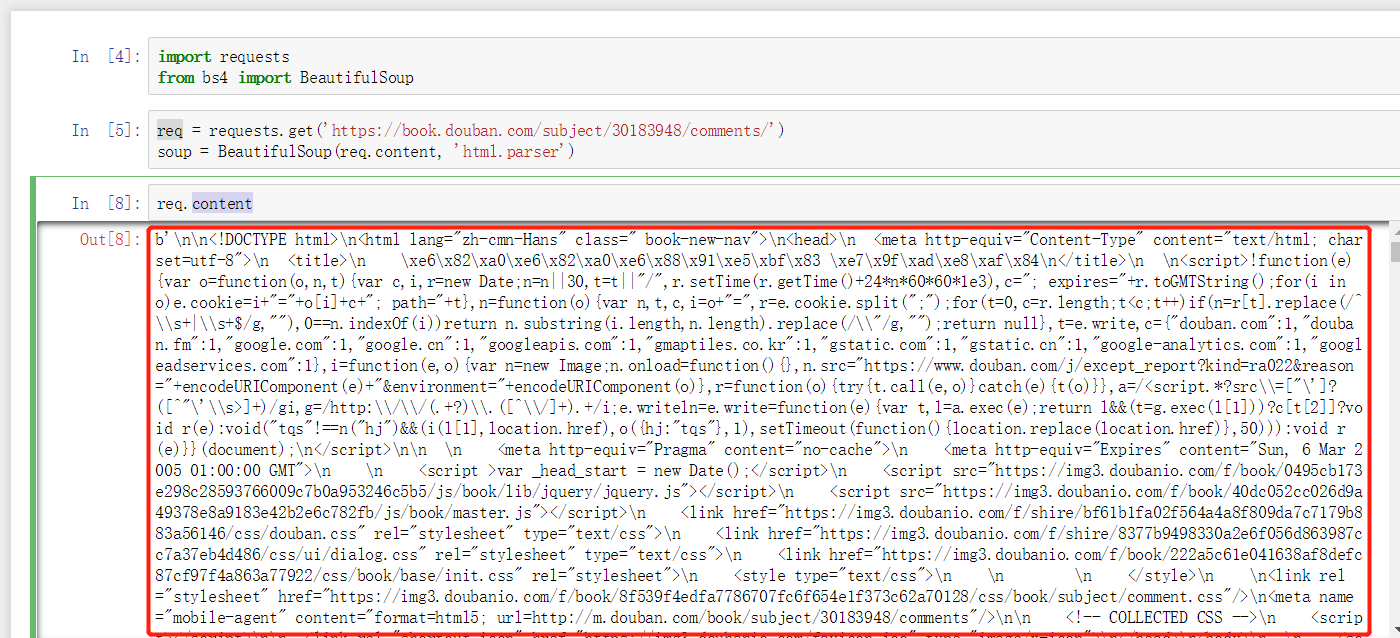
（服务器中一般会安装服务的，把一些资源通过特定的url暴露给用户，用户通过url可以获取到相应的资源），我们拿到请求返回给我们的数据，就可以对数据进行处理，

2.经过查看之后看到网页中显示的评论的内容是在<https://book.douban.com/subject/30183948/comments/>这个接口返回的数据里面的



数据是保存在html的结构里面的，（html是浏览器的能识别的结构，浏览器能通过html的文本标记吧页面渲染出结构出来），

3.我们这次爬虫用到的库是 request（用于网络请求）跟BeautifulSoup（用于解析html结构的字符串，解析之后BeautifulSoup会把html类型的字符串实例化成一个特定的对象，这个对象有BeautifulSoup提供的方法）



为什么要使用到BeautifulSoup库呢？呐先看请求获取到的数据是字符串类型（用html标准写的字符串，为什么说是字符串呢？因为电脑的系统保存文件的时候都是用二进制来保存的，通常是8位是组成一个码，通过这些不同码的组合来表示不同的内容，平常我们看到的html也是这样的，应该可以说所有的文件都是这样的，都是通过二进制来保存的，只是不同的应用程序对于不同的文件有不同的处理，例如我把文件excel用office的软件打开，office就会把文件显示出来内容，但是我改了文件的后缀之后，注意哦我只是改了文件后缀，但实际上文件的本质内容【编码】还是没有变化的，office不会会把内容按照excel的处理程序来处理这个文件，就会显示不出内容的）的，

4.我们在这段字符串里面获取到我们想要的东西呢？我们就会想到通过python的字符串的内建方法来进行匹配和截取，但是呀对于内容很复杂的字符串，我们进行匹配和截取就要写很多复杂的逻辑了，所以我们就用到了BeautifulSoup，他会把html的结构解析好，我们通过对html节点的操作，就更方便的去处理啦（这段呢，我解析的不太好，你不清楚的话我跟你讨论一下啦）



最后，通过BeautifulSoup解析好的对象，分析好我们想要的内容在结构中的那个位置，就可以获取到啦(注意哦，是先要先获取到相应的节点，再对节点的属性，内容进行操作的哦)

