**Ajax学习**

1. 我通过资料了解到的Ajax：

Ajax，全称为Asynchronous JavaSript andXML，即异步的JavaScript和XML，是利用Java在保证页面不断刷新，页面链接不改变的情况下与服务器交换数据并更新部分网页的技术。

利用这个技术的有豆瓣电影之类的，我通过学习爬取了豆瓣电影，肯德基和百度贴吧。

1. 基本原理
2. 发送请求：

新建Request对象，然后调用onreadystatechangge属性设置了监听，然后调用open()和send()方法向服务器发送了一个请求，的掉服务器返回响应，并解析。

1. 解析内容：

得到相应之后，监听会被触发，利用response可以得到相应内容，例如HTML和Json。

1. 渲染网页：

解析完响应内之后，就可以调用JavaScript来针对解析完的内容对网页进行下一步的处理了，可以保存到本地等。

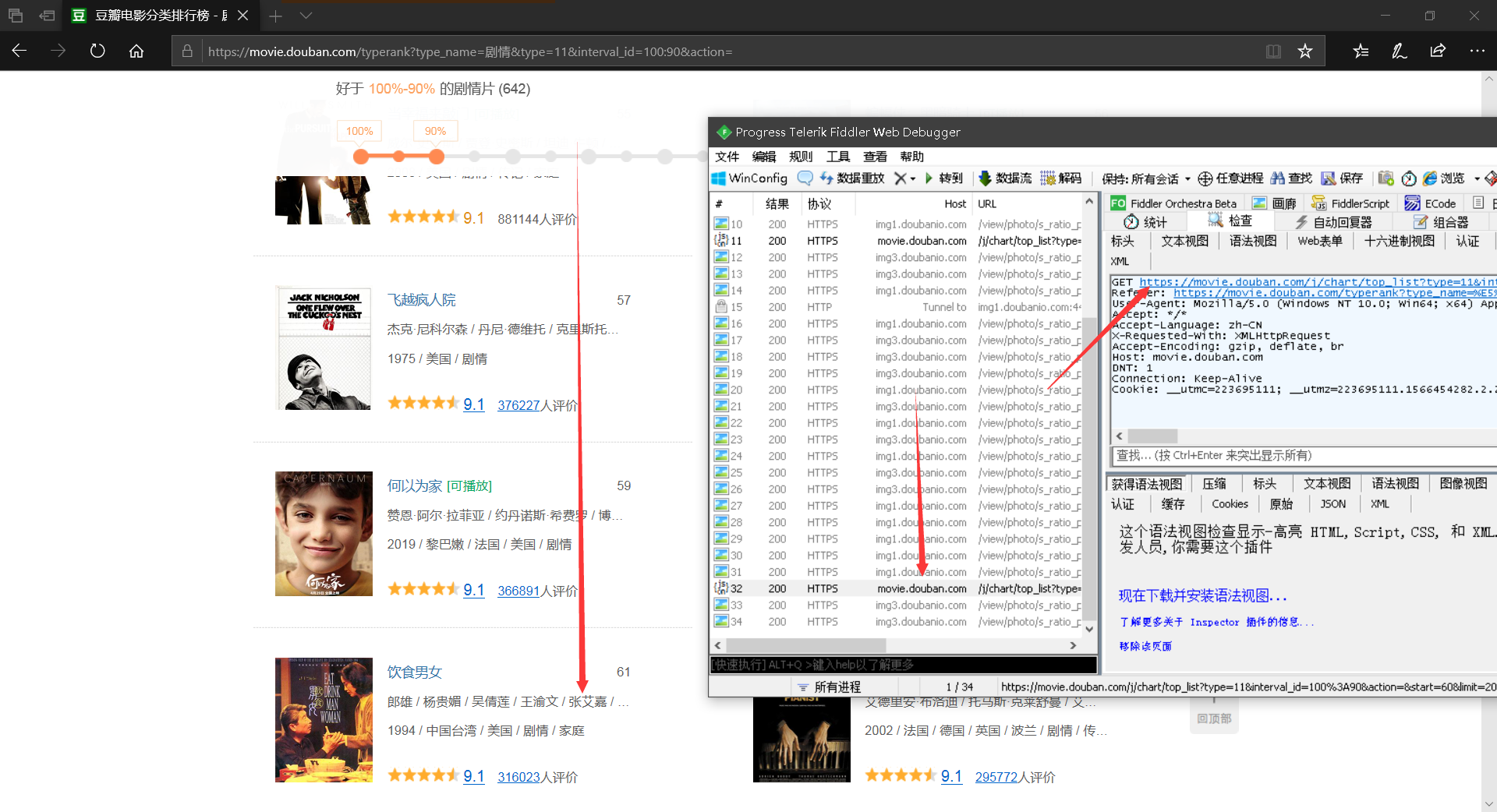
3、我爬取的实例：  
 1、豆瓣电影：

（1）网站分析：

* 打开豆瓣电影网站：<https://movie.douban.com/>，选择排行榜，随便点击一个分类（如剧情）。
* 一直往下滑，可以看到到底不断出现新的内容，也就是往下没完。
* 基本可以判定使用了ajax请求，进行异步的加载

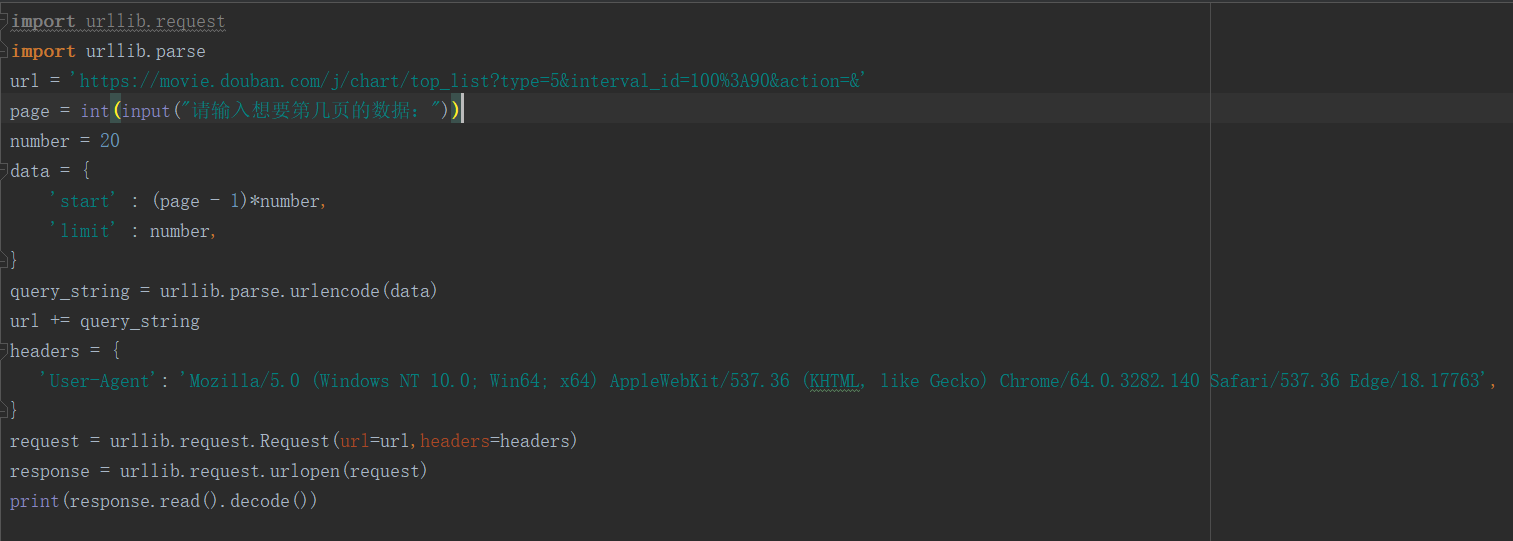
注：爬取ajax网站我们要获取请求信息，有一个接口，谷歌浏览器可以自己抓包，不是谷歌浏览器就要通过Fiddler软件抓包，我在这里使用的是Fiddler

抓包信息如下图：



可以看到请求在不断更新，点击一个请求，就可以看到请求的信息。

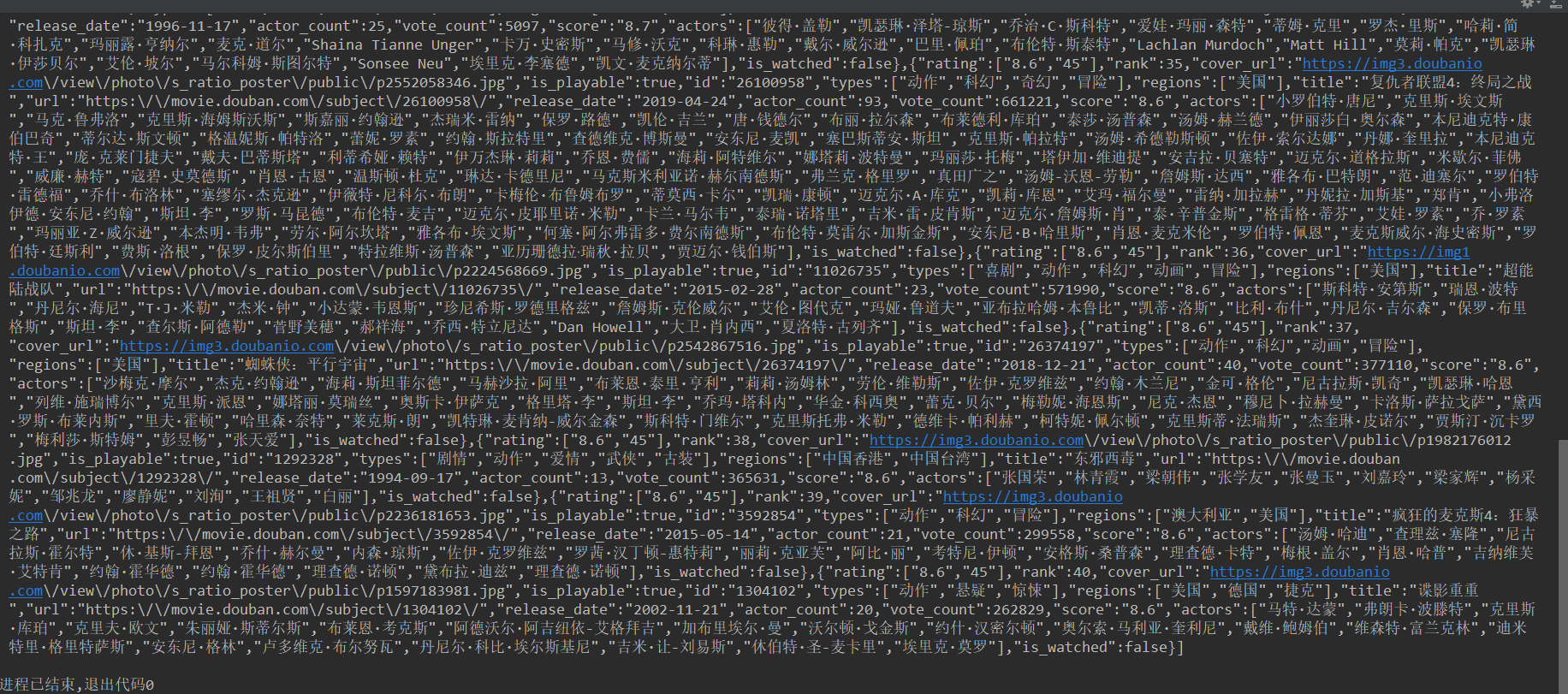
所以写代码如下：



代码分析：

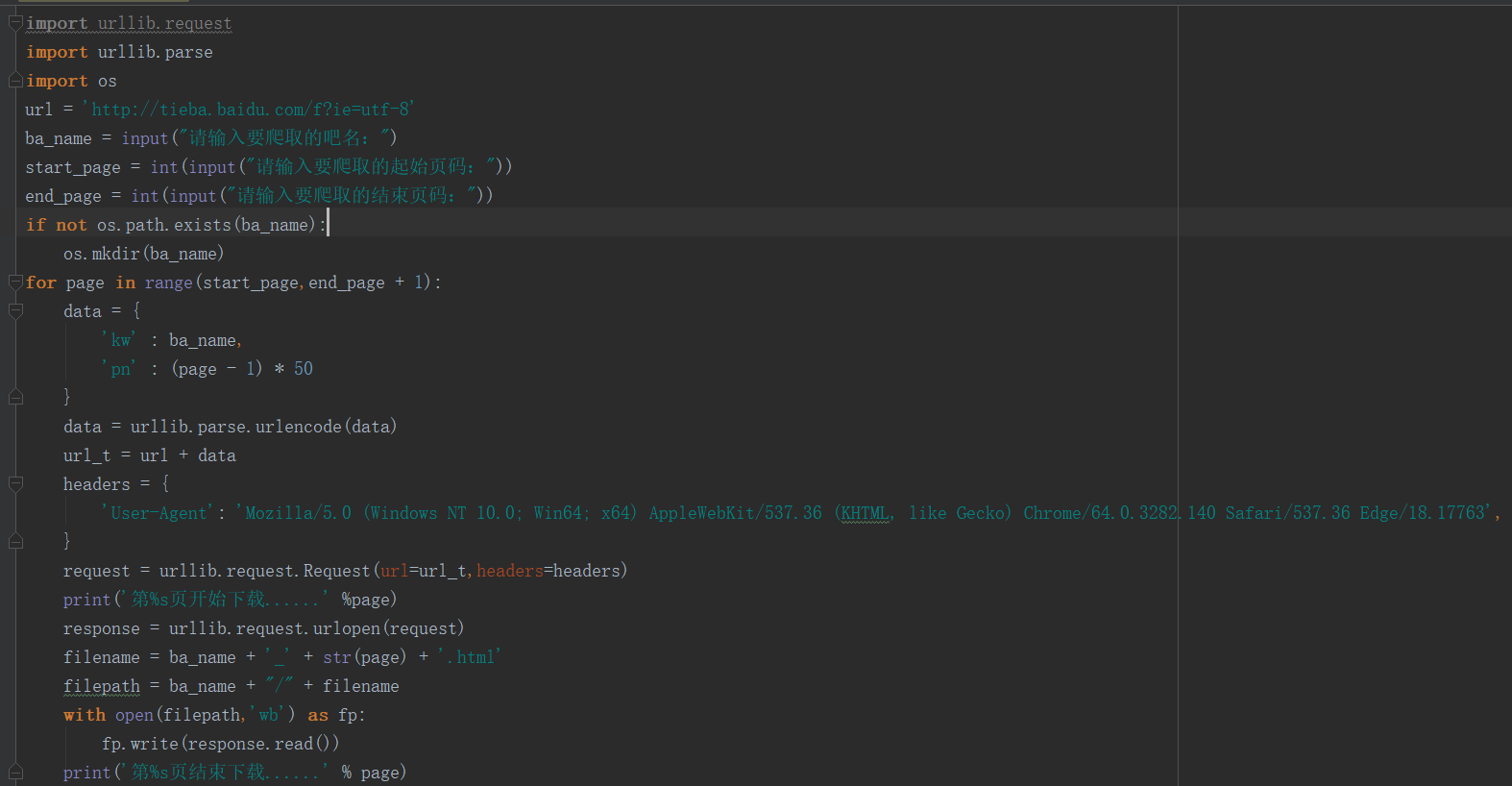
* 我们需要导入urllib的request和parse库
* 写入url地址，因为我们不能只爬取一夜的内容，所以我们要输入页数，经过分析limit是每一页的内容，我们暂定一页20个内容，所以写一个data字典，用来存放页数和数量
* 然后构建请求，在这里有一个要注意的，url在这里要拼接
* 然后写请求头，这里我用的是微软浏览器的请求头
* 然后发送请求，最后打印内容

下面是运行结果：



2、百度贴吧：  
 百度贴吧分析后和豆瓣电影一样，虽然有翻页功能，但是页面的url不变，所以判断是ajax，这样的话步骤和豆瓣电影一样，只是我增加了可以自己选吧名称，爬取页数，最后保存到本地的文件夹，并且文件夹是通过爬取的吧名称命名的。

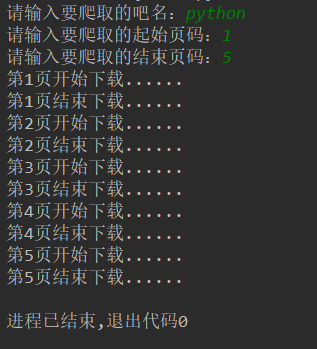
代码如下：



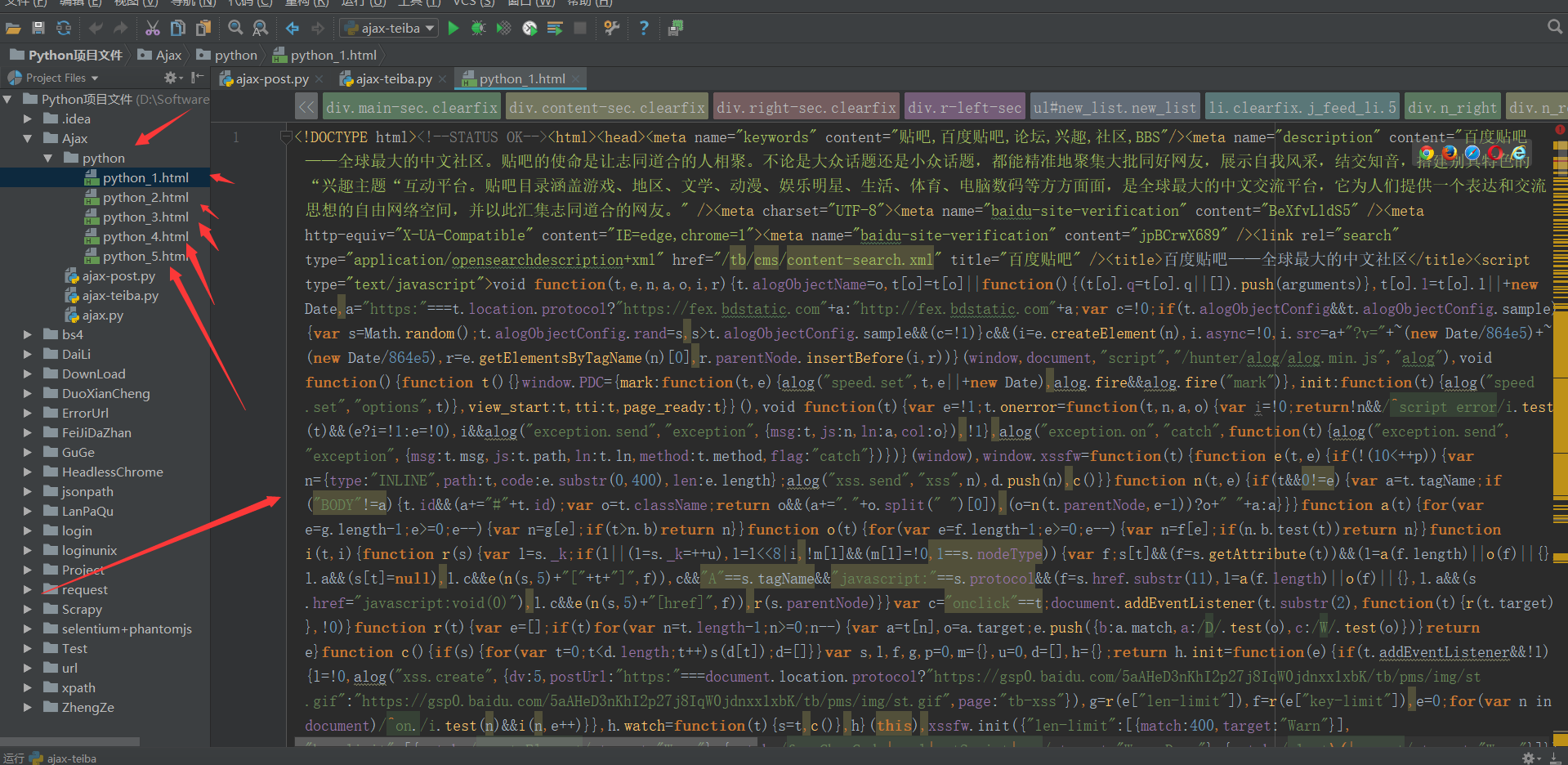
代码分析：

* 首先还是导入库，这次需要导入os库，因为要实现本地保存
* 然后写入url，分析url，kw是吧的名字，pn是页码，所以需要我们手动获取，首先创建一个文件夹，用爬取的吧名字命名
* 在这里写一个循环，因为我们起始页码都是我们手动输入的，然后在循环里写一个data字典，分别存放kw和pn
* 然后还是老规矩，拼接url，把kw和pn放进去。还是同一个请求头
* 然后就是发送请求
* 最后保存文件到本地，这里因为要用吧的名字命名文件夹，所以filename=ba\_name=’\_’+str(page)+’.html使用名字和页码命名

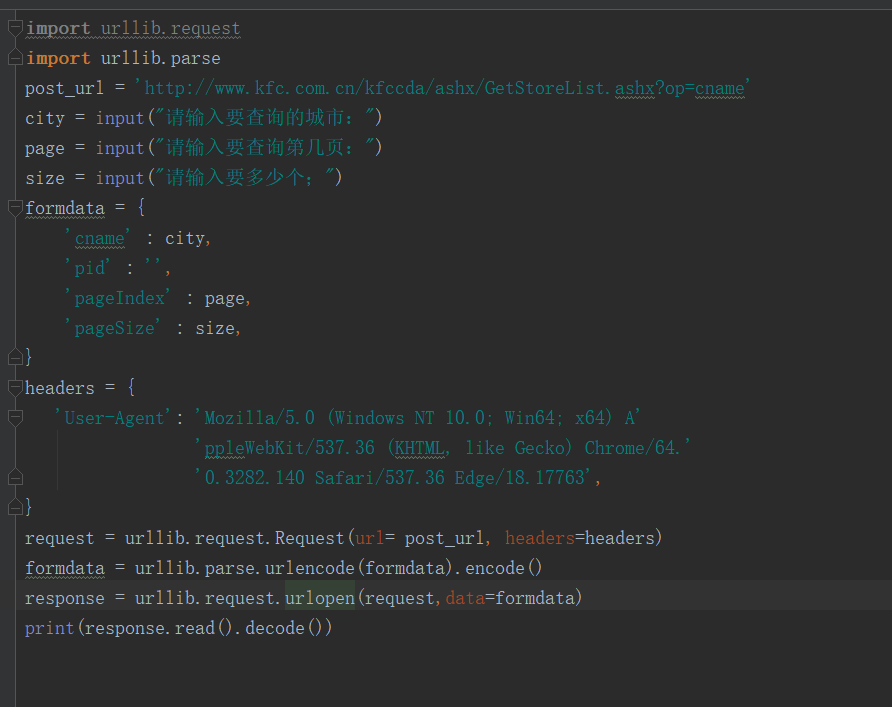
以下是代码运行过程：



运行后结果：

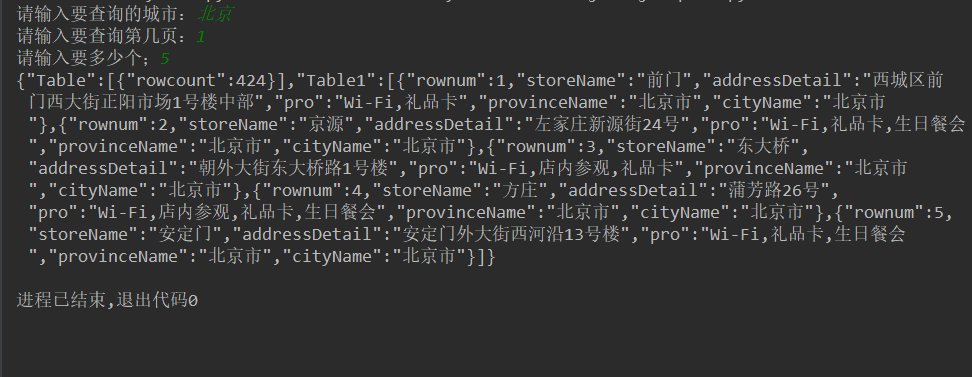


1. 肯德基店铺查询：

基本步骤和上面两个一样，这个拿来巩固练习，话不多说，代码如下：  


代码分析：

* 老规矩，导入需要的库，然后写入url
* url中的city就是我们手动输入的城市名字（cname），pid不知道是啥，直接写成空字符，pageIndex就是我们查询的第几页（page），pageSize就是我们需要查询多少个（size）
* 然后拼接url，写请求头，构建请求，最后打印内容

运行结果如下：  


Ajax爬虫总结：经过查阅资料视频学习，通过三个小项目对Ajax有了更多的了解，我也写出几个在我学习中的问题，首先，刚开始并不知道什么情况该用Ajax，后来了解到，在页面链接不改变的情况下页面的内容发生了变化，这样的情况下就可以用，其次有的时候在浏览器和抓包的request里的内容不一样，因为request获取的都是原始 的HTML文档，而浏览器中的页面则是经过js处理数据后生成的结果，如果遇到这样的页面，直接利用requests等库来抓取原始页面，是无法获取到有效数据的，这是需要分析网页后台向接口发送的Ajax请求，如果用requests来模拟Ajax请求，那么就可以成功抓取了，最后，在抓包的时候一定要过滤好内容，因为东西太多了，ajax一般都在js文件里面，点击解析内容，最后点击raw或者json就能找到所需内容了。