**BS4**

1、bs4：BeautifulSoup：

作用：用来解析网页，提取指定数据，一般和bs4使用的是正则，虽然用的比较多，但是bs4只能解析html格式的数据。

2、我们首先要安装bs4：

pip install bs4（注：这里可能有依赖关系，第一次我没装上，在代码里还是不识别，后来我去官网下载对应版本的，又安装一个lxml就好了）。

3、整体使用步骤：

我们在这里使用的类是bs4里面的BeautifulSoup，然后通过这个类将网页html格式字符串生成一个对象，然后通过对象的方法来进行查找指定元素。

1. 将本地的html文件转化为对象

soup = BeautuifulSoup（open（‘xxx.html’,encoding=’utf8’）,’lxml’）

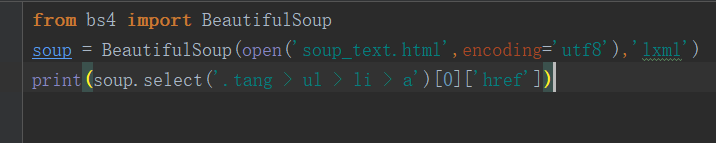
4、练习：  
 因为我是第一次接触bs4，所以在本地写了一个html文件，来熟悉以下bs4

5、练习过程：

（1）准备本地html文件（如图）



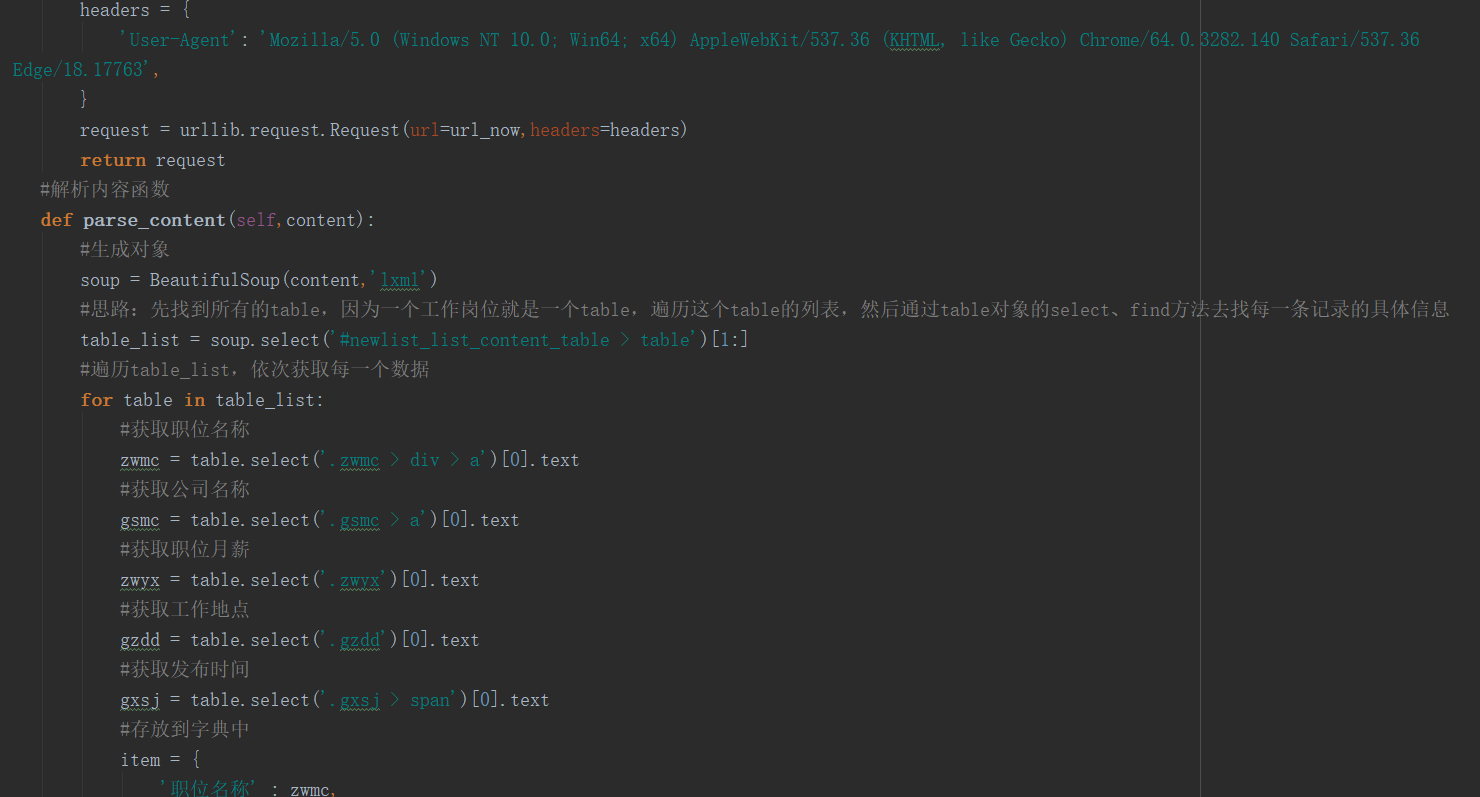
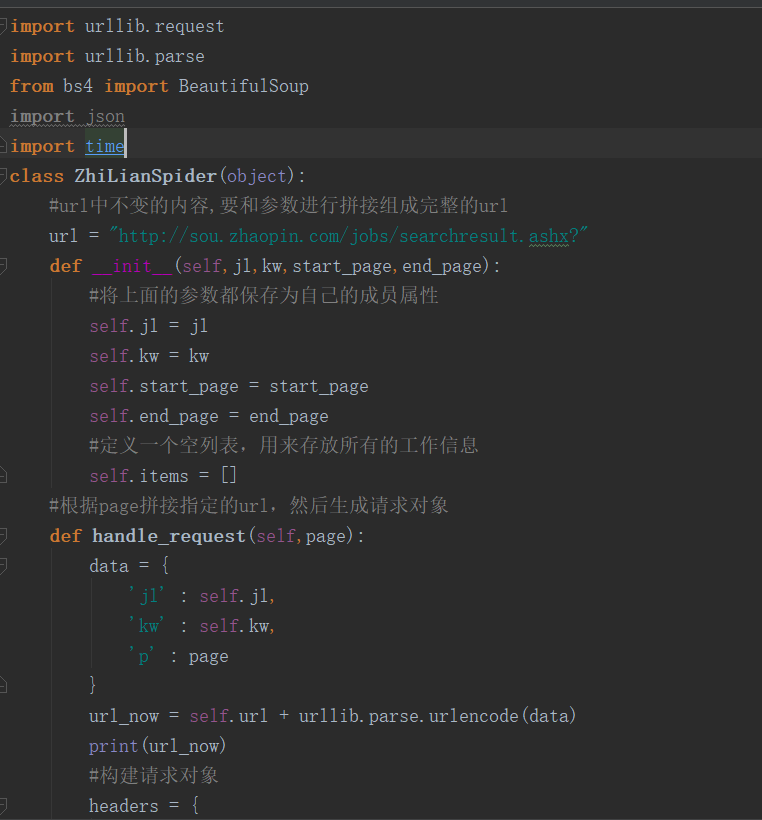
1. 写代码读取内容（代码如下）：

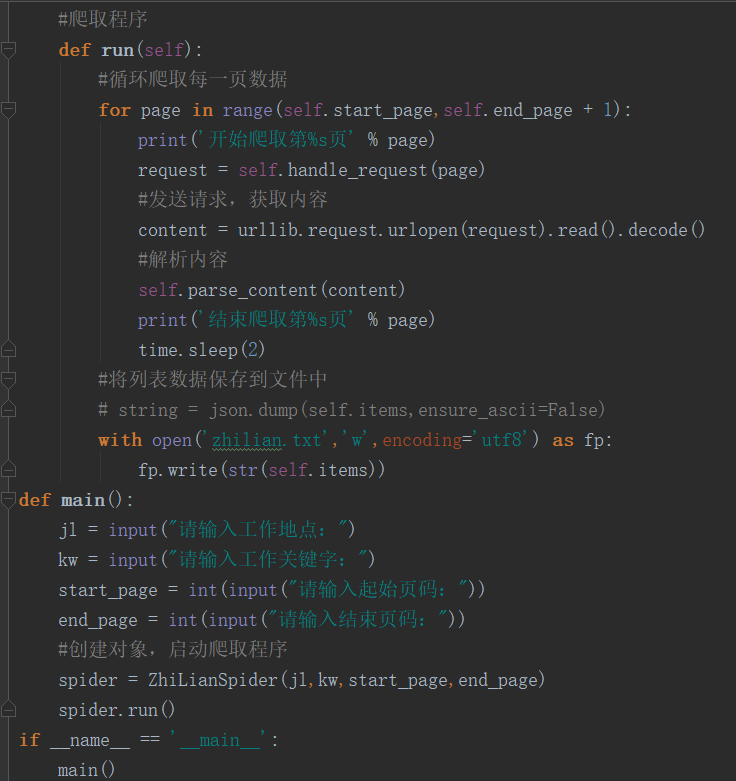


1. 结果如下：



再熟悉bs4之后，我写了一个关于智联招聘的爬虫，应用的东西比较多，直接上代码然后分析，代码如下：





分析代码：

因为爬取的内容比较多，所以我采用的面向对象编程的思想，把不同功能的代码封装在不同的模块里

* 首先还是老规矩，导入响应的库，这里导入bs4库和以往不一样，from bs4 improt BeautifulSoup，另外我们还需要导入time库，在爬虫的时候有个等待时间，要不给太快了给网站爬崩了就不好了
* 然后构件主函数 def mian():，这里是爬虫的主程序，控制各个模块
* 经过分析url我们知道，jl就是工作地点，kw就是工作关键字，还是老规矩，经过我们输入这些内容然后爬取响应的内容，另外还是起始页码和结束页码
* 然后就是下一步，构建对象，启动爬虫程序，在这里我们写一个类叫做ZhiLianZhaoPin我们先把jl，kw，start\_page,end\_page当作参数传过去
* 在这里我们先写爬取程序，def run(self):还是老规矩写一个for循环，循环爬取每一页内容，和以往不一样的是，我们在这里爬取之前print一下开始爬取第几页内容，和结束爬取第几页内容，更清晰，然后就是发送请求，获取内容，在这里我们把循环后的page传过去，传到之后写的handle\_request类里面去，后面会介绍，就是发送请求，获取内容

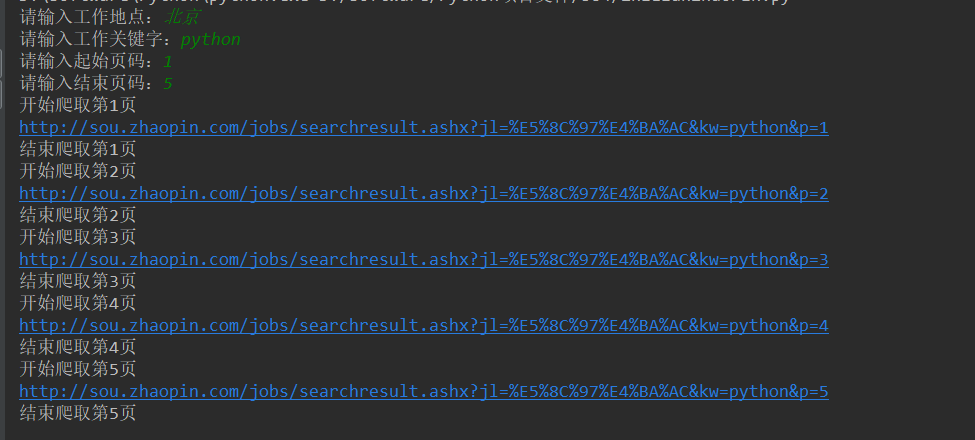


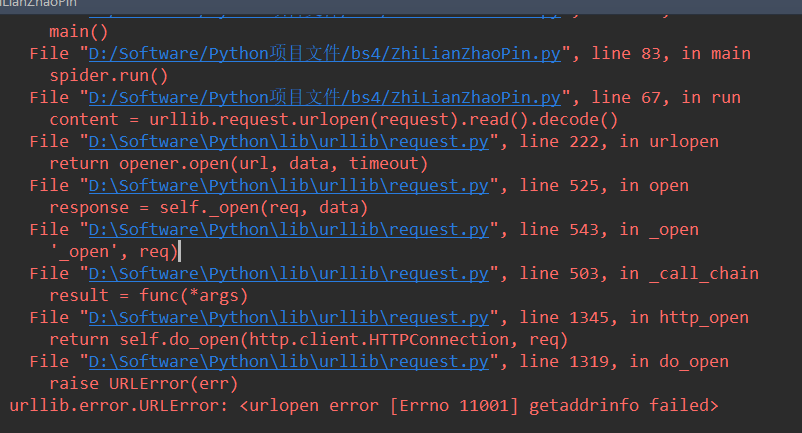
* 在这个类里面，我们先写上url，但是这里url有不变的内容，要和参数进行拼接组成完整的url
* 在这个类里面，先定义一个函数def \_\_init\_\_(self,jl,kw,start\_page,end\_page):先接收刚才传进来的参数，然后将上面的参数都保存为自己的成员属性，例如：self.jl=jl
* 然后定义一个空列表，用来存放所有的工作信息
* 再根据page拼接指定的url，然后生成请求对象，在这里所以我们还要定义另一个函数handle\_request，用来构建请求对象，我们还是在这里定义一个data字典，这里的jl和kw不做介绍，就是刚才参数，这里的p就是页码，page
* 然后拼接url，生成url\_now，在这里打印一下url
* 然后构建请求对象，老规矩放一个请求头，然后发送请求，最后return返回请求对象
* 构建请求之后，在定义一个解析内容函数，def paese\_content(self,content)：
* 这里我们就要先生成对象，利用bs4，和刚才练习的是一样的，这次我们解析请求过来的数据
* 这里我们要分析怎么解析，分析后有一个思路：先找到所有的table,因为一个工作岗位就是一个table，遍历这个table的列表，然后通过table对象的select、find方法去找每一条记录的具体信息

table\_list = soup.select('#newlist\_list\_content\_table > table')[1:]

* 然后我们遍历table\_list，依次获取每一个数据，职位名称，公司名称，职位月薪，工作地点，发布时间，然后存放到字典里，在存放到我们之前定义的列表里
* 在我们的run函数里面，发送完请求，解析完内容，将内容存放到item列表里面，这个爬虫程序相当于告一段落了，刚才我们写的time还没用上，在这里sleep2秒钟
* 然后就是将列表数据保存到文件中，大功告成

运行结果如下：



总结：虽然最后成功爬取到了内容，但是我第一次接触，这之中我遇到不少错误，首先就是bs4下载的问题，这个问题困扰我整整一天，下载成功了但是在代码里不执行，最后下载了相关依赖解决了，还有一个问题就是下面这个错误截图  


首先我以为是我decode()那里出了问题，换了好几个编码都不行，最后我百度之后，竟然是在url那里不能用单引号，用双引号，也算是学到了，后来又遇到一次，我蒙了，我寻思这引号也没问题啊，最后一看没联网（尴尬），还有就是bs4的使用格式我还不是特别熟练，后来又单独练了bs4的用法，熟练多了，这还是一个很有力的助手，继续学习。