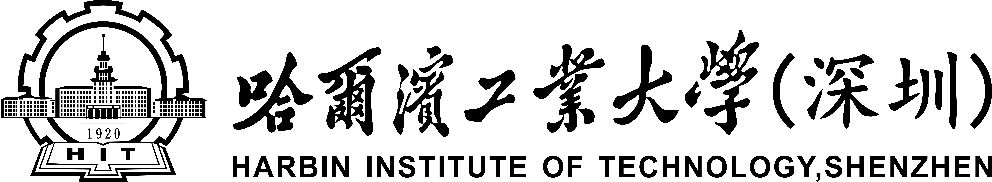
****

**大数据导论课程作业**

**作业一 爬虫的解析与存储**

**专业： 计算机科学与技术**

**学号： 190110509**

**姓名： 王铭**

**一、作业内容**

使用python的scrapy爬虫框架爬取“安居客”上的一万个租房的房源描述信息以及月租价格。

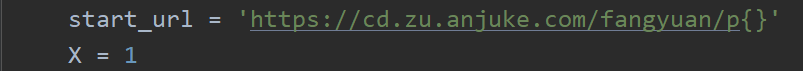
**二、程序实现过程分析**

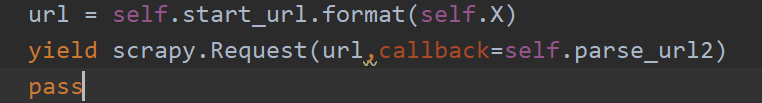


观察网页，可以发现每个地区的房源信息含有50页，一页含有60个房源的url，即一个地区可爬取3000个房源url，从每个房子的具体描述网页中解析描述房子的信息和价格，选取5个地区，深圳、广州、天津、成都、北京进行爬取（手动更换5次起始url即可）。

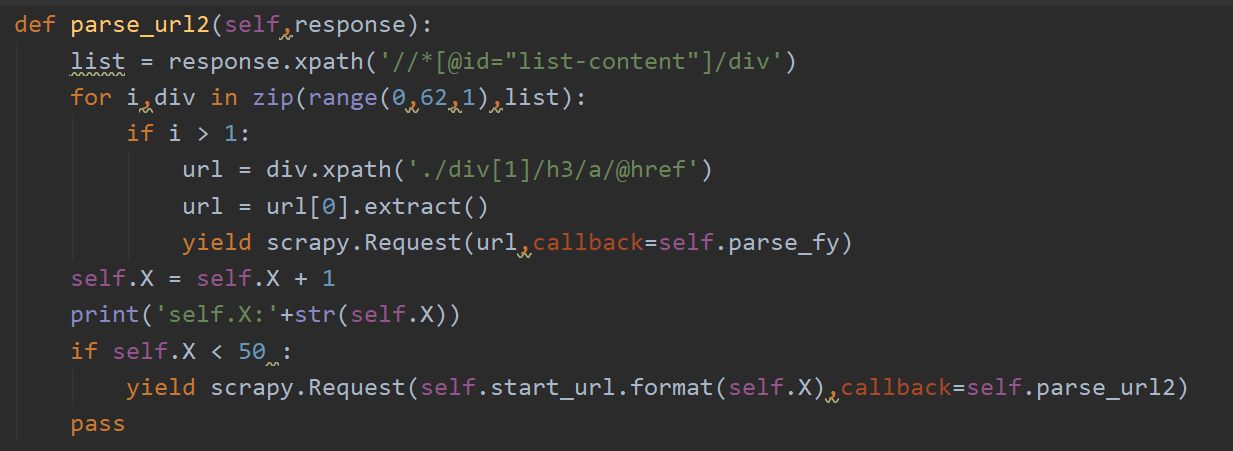
以成都为例：

①设置起始url，页面索引初始化为1，在start\_requests中格式化起始url，X可从1增加到50，发起请求，用parse\_url2解析。



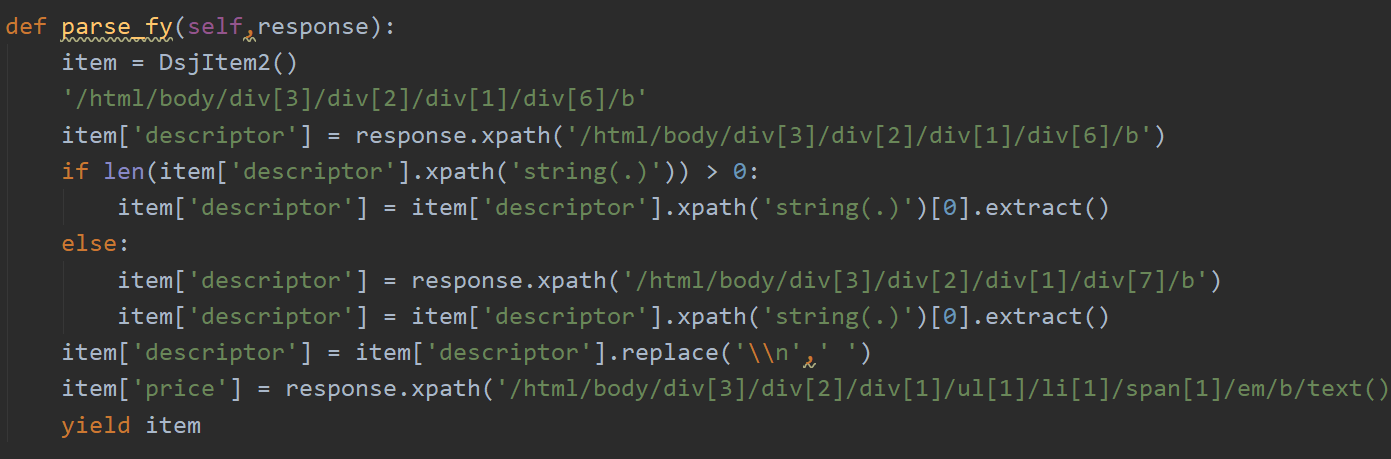


②解析并获取每个房源具体的url的函数parse\_url2：



从页面中提取60个具体标题的url，提出请求后用parse\_fy解析。提取完后，进行下一个页面解析，还是用parse\_url2解析。

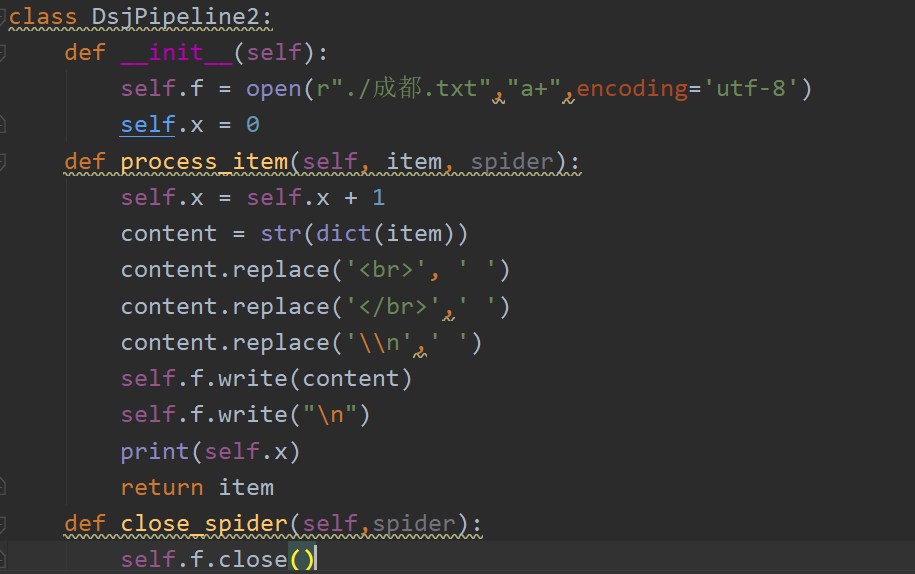
③解析房源的描述信息和价格的函数parse\_fy：



descriptor是描述信息，price是价格，用yield交还给管道存储

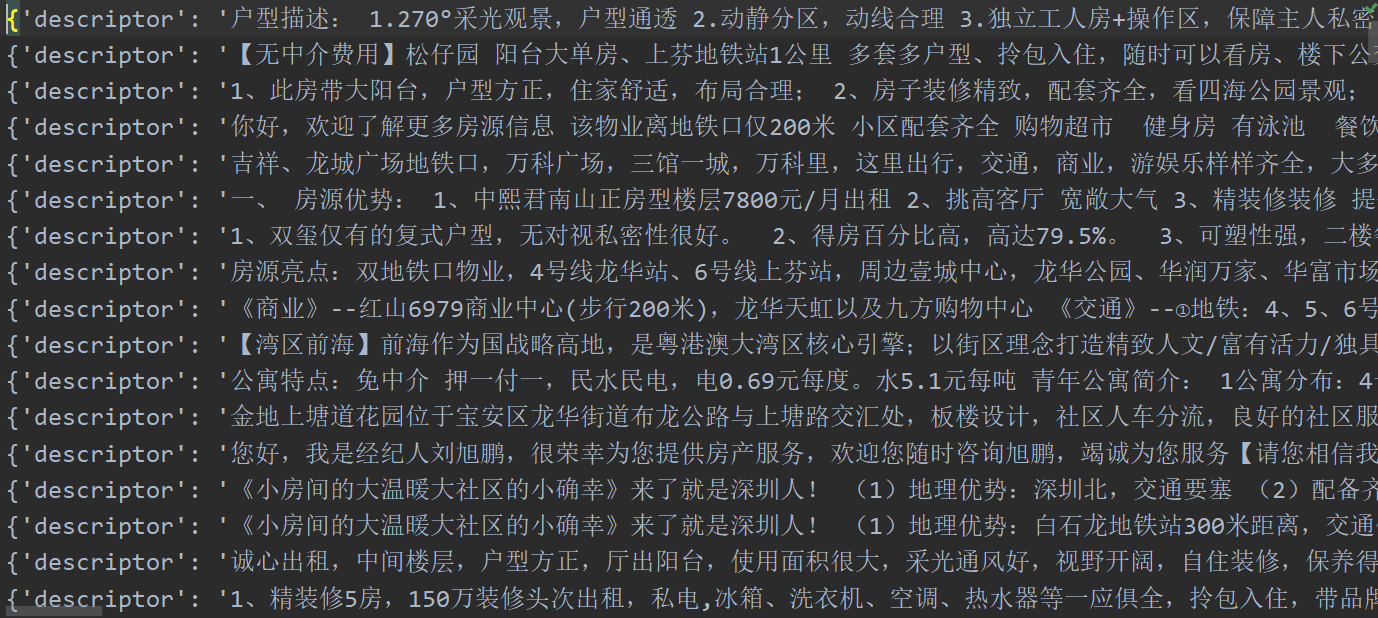
④管道处理：

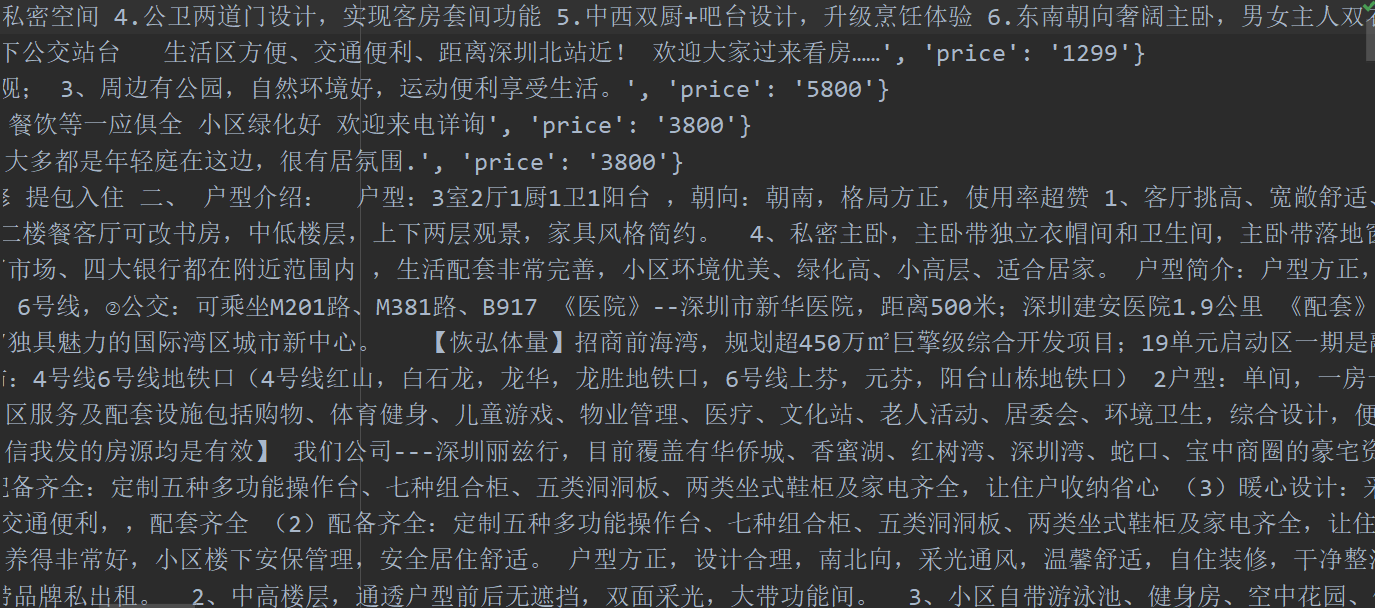
写入具体的txt文件。



三、结果截图（作业一可以截图部分TXT文件）

将5个地区的所有信息整和在一个txt文件中，爬虫解析后的txt部分内容如下：。





利用python的jieba库，进一步处理并分词：

