该任务为爬取西安二手房价数据，首先确定需要爬取的网站和数据，这是十分重要的。

经过筛选，选取了**链家**网站的数据，搜索结果如下（部分截图）：

|  |  |  |
| --- | --- | --- |
|  |  |  |

该网站有如下优点：

1.该网站涵盖的数据量巨大，一共含有111774个数据，足够满足建模分析。

2.该网站对于房子的各个特征归纳总结到位，便于分析。

3.该网站便于数据爬取。

爬取分析：

该网站已经整理完毕，并且分类完毕，因此可以选取部分数据爬取。

1.进行网站构成分析：

一共有100页数据，其中网页链接构成如下：

|  |  |  |
| --- | --- | --- |
|  | https://xa.lianjia.com/ershoufang/pg2co32/ |  |

构成十分有规律，当进行翻页时，pg后面的数字加1即可，十分利于爬取，利用一个变量表示页数即可，细节不再赘述。

2.对网页构成分析：发现包含数据的网页不是动态网页，直接调用request库即可进行访问爬取。

3.对数据包含分析：调用F12查看元素可以发现：

|  |  |  |
| --- | --- | --- |
|  |  |  |

所需的数据主要包含在中，可以采取几种方式：

1.使用re正则匹配式进行查找匹配。

2.对于html语法树进行匹配，找到对应容器，提取文字即可。

4.优化：考虑到有一百页数据，数据量大，因此采用线程池来进行爬取，使用到了方法，具体代码附录给出。

5.爬取结果：不出意外地由于高强度的爬取，我的ip被网页禁止了。最终只爬取到了不足200条数据。最后，使用缓解了一下我和网页的关系，多爬取了一点数据。