基于 Scrapy 框架二手房数据爬虫程序的设计与实现

汪宗伟

(江西信息应用职业技术学院软件工程系 江西南昌 330043)

摘 要:当今社会,互联网技术发展成熟,各个行业都通过各种平台对接到互联网。这种方式给各个行业积累了大量的数据,这部分数据通常堆积在各个企业或者平台之中,作为研究者需要一套完整的理论框架和可行的技术手段,从网络平台种爬取海量数据,同时对数据进行整理和分析。爬虫技术作为一种最快获取信息的方式,被广泛应用在各行各业,利用 Scrapy 爬虫技术获取二手房的房源信息数据,给后续分析二手房市场整体情况提供数据支持。

关键词:爬虫;Scrapy;二手房数据

Design and Implementation of Data Crawler Program for Second-hand Housing Based on Crawler Frame

WANG Zong-wei

(Jiangxi Vocational and Technical College of Information Application 330043)

Abstract: Currently, internet technology is mature; various industries are connected to the Internet through a variety of platforms. By this way, every industry accumulated a huge amount of data. This part of the data is usually piled up in various enterprises or platforms. As researchers, we need a complete set of theoretical framework and feasible technical means to get the huge amount of data from the internet platform, then organize and analyze it. The crawler technology has been used in all walks as the fastest way to get the information, using Scrapy crawler technology to obtain the second-hand housing data, providing data support for the follow-up analysis of the overall situation.

Key Words: Crawler; Scrapy; Second-hand house data

1 引言

房屋市场是一个比较庞大的市场,也是一个非常复杂的市场。同时房屋价格受到多方面的影响,比如经济发展、城市规划和城市调控等。同时一手期房市场多数是由相关的大型房地产企业所有主导,所以受到企业和人为的客观调控因素较强。二手房市场一半认为是全自由市场,主要业务是个体房东向个体客户出售自己的房源,一般来说二手房市场更能够真实地反映该地图客观的房屋市场情况。

本文选取一个数据量相对比较充足的二手房交易平台“安居客”,以江西省南昌市青云谱区的所有二手房房源数据作为此次分析的数据源。在使用网络数据爬虫技术获取完成数据以后,从房屋数据中依次提取房屋类型、房屋位置、地产品牌、房屋年限、相关配套等方面,为后续的数据分析提供数据支撑。

2 爬虫技术概述

2.1 网络爬虫定义

网络爬虫,主要用于采集互联网上的各种数据,在模拟环境中,自己包装一个网络请求,通过请求服务器地址可以获取到一些数据,通常情况下网络爬虫是采集一个网站上的公开数据。故爬虫指的是:模拟本地的请求,向网站或者服务器发起请求,获取资源数据后对数据进行解析,解析完成后使用

作者简介:汪宗伟(1992——),男,助教,主要研究方向:大数据与人工智能。

特定的存储技术(如:数据库、文件存储)将数据有序地存储在本地的一种技术。简单来说,从技术层面来说就是通过程序模拟浏览器请求站点的行为,把站点返回的 HTML 代码、JSON 数据、图片、视频等保存到本地或服务器,当需要使用时可以随时提取使用。

2.2 爬虫技术步骤

总体上来说,网络爬虫数据采集步骤分为四个步骤:

封装网络请求,包装请求参数和设置反爬虫伪装;

发送网络请求,根据返回码判断请求情况,并提取返回的数据;

使用特定的数据解析方法将需要的数据解析出来,并封装;

使用数据库或者磁盘文件的方式对数据进行存储。

这四个步骤中,难度最高、工作量最大的集中在第三步:数据解析。其中 python 提供了很多数据解析的库,市面上很多浏览器也进行对应方法的辅助。一般情况下可以先使用浏览器进行网页源码的查看,找到需要的内容在网页中哪个位置,再使用选定的解析内容方式,同时浏览器也会提供对应元素结构的规则(结构公式),只要将其代入到方法中即可。

一般来说,一个网站的数据是分布在许多网页当中的,就如同购物网站的商品数据一样,不同的商品在不同的网页中,通过输入或者翻页可以观察到。虽然数据所在的网页不同,但它们所在板块和格式却大致是一致的,所以解析方法是不需要更改的,只需要找到网页地址变化的规律,或者传递参数的变化规律即可。当整理完成所有的网址,重复之前的爬虫步骤,就可以把所有需要的数据都爬取下来,并使用特定的方式进行存储。可选择的存储方式有 csv 文件、数据库、普通文档,在有数据库可以使用的条件下,推荐使用数据库的形式存储最为可靠、便捷。

2.3 网络爬虫框架 Scrapy

Scrapy 框架是一款基于 python 语言进行设计开发的网络爬虫框架。由于在实际的项目中,爬虫所需要采集的数据量是非常大的,解析的内容也非常多,故在实际爬虫项目中需要一个稳定高效的框架进行支撑。Scrapy 则承担了这个角色,在 Scrapy 框架中根据分工的不同设计了很多角色,每个模块负责一部分功能,在每个对应的模块写入对应的业务代码,比如解析模块加入数据解析的代码,数据管道模块写入存储代码,各模块之间进行分工协作完成爬虫任务。

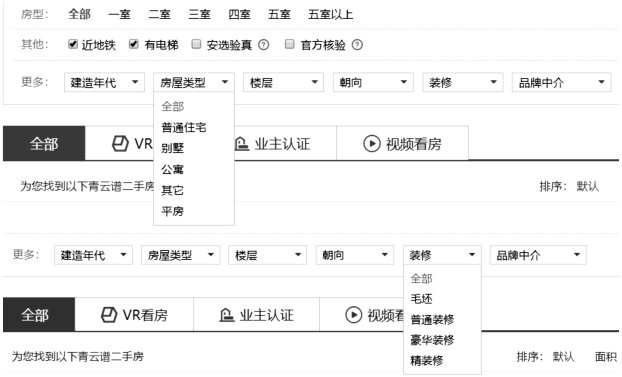
3 爬虫代码的设计与实现

3.1 网址和数据字段的确认

本文以二手房数据为示例的,故选取的字段主要是对二手房市场分析中具有参考性的字段。大部分字段都可以在网页的界面代码中直接获取,少部分隐藏的字段可能存放在网络请求的返回值中或 javascript 代码中,需要进行更复杂的网络请求才可获取,以安居客为例,可以进入网站二手房列表界面,观察每个子项的数据是否满足要求:

在此列表单元中,展示的是某个房屋的基础数据,由此可以直接得到的数据有:房屋名称、房屋规程、房屋总价、房屋单价、房屋面积、房屋楼层、建筑年限、地产品牌、具体位置、优势标签等。

除此以外有部分信息内容没有直接展示在界面中,而是需要通过筛选项来进行准备定位的,其具体的筛选项可以在网站的筛选栏中看到:



由此两个筛选项可以得到两个另外的数据维度即:房屋类型和装修情况。由此所需要用来分析的数据都已经足够,下一步就要设计代码用来爬取数据了。

3.2 找到网址变化规则

由于“爬”取完整的数据,在平台网址中一个界面一般都对应一个网址,要把所有的网页数据都采集下来,也就是说需要爬取很多个网站对应的网址,那第一步就是要把这个网址都找出来。网址的开发人员在网址设计的时候都给了它对应的规律的,所以先要分析网址找到对应的规律,找到规律以后就可以获取到完整的网址了。

提取一个网址:

https://nc.anjuke.com/sale/qingyunpuqu/p2-t71/#filtersort

其网址的规则组成为

nc.anjuke.com:表示的域名,一般不会改变

Sale:应用名,表示一个厂商下面的不同业务

Qingyunpuqu: 取名名称的拼英全称,这里指青云谱区

p2:网址页码,第一页不需要,第二页以后的页码用 P(n)表示

t71:表示筛选的房屋类型,其具体对应是{普通住宅:'t9/','别墅 ':'t2/',' 公寓 ':'t71/',' 平房 ':'t12/',' 其它 ':'t4/'}

#filtersort :结尾符号,仅在第二页以上的界面会出现,暂时未有具体的含义。

找到网址的变化规则以后,可以在后续请求爬虫的时候进行网址 url 的组装了。

3.3 编写 Spider 数据爬虫类

新建一个 python 文件,创建一个 Spider 类并继承 Scrapy.

Spider 类。继承之后需要修改三个参数，分别是：爬虫名称（name），首次的爬取网址（start_urls），爬取的网址域名（do_main）。设置完成以后运行后会自动发送请求到该网址，并在 parse(response)函数中接收到请求网页返回的数据，所有请求数据都包装在 response 对象中，其中对象的 state_code 为响应码 200 为成果返回数据，对象的 text 为具体的网页文本数据，对象的 url 为当前请求的网址。

获取的数据以后下一步就可以对数据进行解析，一般来说解析数据有多种方法，常用分别是 bs4 的标签结构选择器（select）方法，lxml 中的 xpath 网页标签路径的方法，以及 beautifulSoup 中的正则表达式字符匹配的方法。其中第一种方法最适合网页爬虫的初学者，在谷歌浏览器种也有对应辅助的选项，可以极大地降低爬虫算法的编写难度，故作为本次爬虫选取的方案。

使用谷歌浏览器打开第一个网页，使用 F12 按键或者选择查看网页源码的选择时可以进入一个新的网页，在这个网页的左半部分是网页的展示，内容右边则是内容对应的源码属性标签，爬虫的时候只需要以这种方式找到需要的标签。将鼠标光标移动至需要的标签，点击鼠标左键，会展开一个功能选择项，选择 copy selector 即可将该元素的结构拷贝到结构种，在代码种将拷贝的内容放入 select()方法种，可以到类似一下的完整方法：

houseinfos = jsonSoup.select('#houselist-mod-new > li')

执行此行代码以后，符合该规则的标签都会被解析出来，包装在参数 houseinfos 里面，由于房屋数据实在列表中的，这里解析出来的是一个数组，里面存放的该网页上所有的房屋数据，但这个数据还是非结构的，需要进一步解析里面的标签字段。所以需要套用一层 for 循环，逐个解析里面的各组房屋数据，比如解析的房屋单价的数据代码如下：

danjia = int(price.select('span.unit-price')[0].text.replace('元/m2',''))

价格数据可以用数值表示，不过解析出来的字段统一是字符串类型，需要进行一次类型转换，其它的字段依次为参照进行解析即可。

如果到这个时候爬虫代码都进行顺利的话，表示已经完成单个网页的数据解析工作，现在可以尝试多个网址的数据解析。多个数据解析并不困难，由于已经把之前网址的变化规则分析清楚，并已经用函数拼接完成所有网址的数组，遍历此数组中的所有地址即可，把地址放在 yield 中需要包装的 request 对象中，yeld 会作为一个爬虫中期，以一个网页为一个爬虫周期即可，其案例代码如下：

for i in list(self.housekind.keys())[1:]:

url='https://nc.anjuke.com/sale/qingyunpuqu/'+self.housekind[i]

yield Request(url=url,callback=self.parse)

3.4 中转和存储数据

在数据解析完成后，在项目文件中找到 items.py 文件，该文件创建了一个 MyscrapyItem 类，该类是作为数据仓库用作爬取的数据的缓存容器，在该仓库类中定一个自己的子容器，当每次爬取数据的时候，可以将解析完成的数据拼装成一个元组数据，将其放入到子仓库中，子仓库的名称和变量定义的名称一致，再通过 yield item 的方式将数据输送到缓存，在日志台中也可以看到放入容器的缓存数据。

当数据放入缓存容器完成以后，找到项目中的 pipelines 文件该文件中有一个 MyscrapyPipeline 类，其中有一个 process_item(self, item, spider)，该函数有两个参数，其中 item 即是之前用于存放缓存数据的仓库。通过 item[子容器]可以提取到该子容器中存放的缓存数据。

数据存放在缓存中当然是不够的，缓存数据会随着程序的终止也会随之清空，一般会选择一个方便存储数据的方式来存储，在条件允许的情况选择用数据库来存储是最合适的。可以使用 mySql 来作为本次项目的数据存储，在安装完成 mysql 完成后，建立一个存储房屋的数据库，同时建立一张房屋数据表，注意的是表的字段要和爬取的房屋数据字段进行一一对应，这样既可以在存储过程中不会出现错乱。

最后使用 pyMysql 作为具体链接和操作数据库的第三方包，创建数据库连接时需要对应到具体的库，防止存储错误。将数据仓库的数据使用 for 循环提取，在依次使用字符串处理的方式拼接成一条完整的数据库插入语句，使用数据对象执行该条数据库语句，最后调用 commit()进行操作提交，可以将数据存储到数据库中。

所有代码编写完成以后，运行程序，当需要爬取的数量非常大、网页数量也比较多的时候，一般会需要比较长的时间，需要预先准备好足够的存储空间以及能够保持长时间工作的设备，程序执行完成后即可以获取到完整的二手房数据了，可以用数据库软件打开数据库表查看是否成功。

4 小结

近些年随着爬虫技术的使用者越来越多，很多公司开始注重自己网络数据的防爬性。很多重要的数据不再使用简单的 html 标签加文本内容的形式进行展示，转而使用 json 文件数据的读取或使用 Ajax 异步请求的形式加载数据。爬虫工程师和反爬虫工程师是两个博弈的角色，彼此都在学习和实验更先进的技术，在竞争中两种技术也在不断地更新和进步。作为爬虫工程师，我们需要加强新技术的学习、掌握多种爬虫技术和框架、关注新的反爬技术，才能让自己在竞争中保持领先不被淘汰。

参考文献：

[1] 罗安然，基于 Python 的网页数据爬虫设计与数据整理，10.16520/j.cnki.1000-8519.2020.19.035.

[2] 江吉彬，Python 网络爬虫技术，人民邮电出版社，2019 年 4 月

[3] 谢乾坤，Python 爬虫开发 - 从入门到实战，人民邮电出版社，2019 年 8 月

[4] 薛丽敏，面向专用信息获取的用户定制主题网络爬虫技术研究 [J].2017(2):12-21.1671-1122(2017)02-0012-10