# 腾讯房产V1.1版本改动

|  |  |  |  |
| --- | --- | --- | --- |
| 版本 | 作者 | 版本描述 | 完成日期 |
| 1.1 | S | 基本实现对腾讯房产网站楼盘数据抓取及去重 | 2017.5.10 |

因为项目的需要同样是爬取安居客的网络数据，所以大体的结构还是用的原版本的结构，没有做大的改动，只是一些小点。内容如下：

## 抓取IP代理

由于代理网站的整体框架添加了JS，所以之前的提取规则没有数据，从网站返回的JS中提取代理IP信息，再用正则匹配IP和端口。

## 中间件

添加了重定向中间件house/downloadermiddlewares/redirect\_with\_captcha.py中把出现重定向的URL，以及是哪个URL出现302和用哪个代理IP访问的这三点记录到文件302\_error.txt中。

## Spiders

之前的程序爬取时是按在售、代售、售馨和在租四个分类点进行分类爬取的。新 的需求是直接爬取总的楼盘，按页数爬取确定总量。所以就不需要参数ND了。

在原基础的代码上添加了提取手机页面数据的函数，parsefirst、parseModdle、 parseTraffic、\_parseTraffic以及mapTrafficDetails和mapModdleDetails几个函 数来处 理提取手机端页面的数据。

parsefirst匹配的是移动页面中的基本信息，parseModdle和mapModdleDetails 匹 配的是移动页面中的详情信息，parseTraffic、\_parseTraffic以及 mapTrafficDetails 这三个函数是处理移动页面中的学校、生活、交通、餐饮等相关详情。