# 安居客V1.1版本改动

|  |  |  |  |
| --- | --- | --- | --- |
| 版本 | 作者 | 版本描述 | 完成日期 |
| 1.1 | S | 基本实现对安居客网站楼盘数据抓取及去重 | 2017.3.24 |

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：

## 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

在去重中间件ajk/downloadermiddlewares/redirect\_with\_captcha.py中添加了一段关于提取错误状态码的代码。主要根据返回的状态码302,、403、404、429、500、503这几种来记录它的URL，并写入到error.txt文件中。

在IP代理中间件ajk/downloadermiddlewares/rotate\_proxy\_every\_request.py文件中，在ProxyMiddleware下的process\_exception方法中的return response去掉了，因为在跑程序的时候总是报没有定义response，发现程序好像没有继承response。

## Spiders、items、pipelines

因为新的需求把原来提取的字段items和URL改变了，所以需要重新定义成新的字段，并根据字段定义不同的Xpath提取规则。以及在存储数据库的pipelines.py文件，创建数据库表名，字段以及对应的类型。