## 二手房房源抓取&展示程序说明文档

### 总体功能：

1. 多线程抓取五个公司二手房源基本信息（完成）
2. 聚合房源（完成）
3. 结合前端页面进行相似房源列表展示（完成）
4. 定时周期抓取（未完成）
5. 抓取详细的房源信息（未完成）
6. 房屋列表页筛选功能，房屋详情页的实现（未完成）

#TODO：

1. 日志Logger没写，部署到服务器端需要写Logger
2. 在抓取二手房时，每一遍抓取只能向里面存取新的房源，但是没法判断哪些房源已经下架，这样会造成数据库里的房源数量大于网站上的房源数量

### 抓取目标网站

**public** **final** **static** String ***LIANJIA\_HOMEPAGE*** = "http://bj.lianjia.com/ershoufang/";

**public** **final** **static** String ***KUFANG\_HOMEPAGE*** = "http://beijing.koofang.com/sale/";

**public** **final** **static** String ***FANGTIANXIA\_HOMEPAGE*** = "http://esf.fang.com/house/a21/";

**public** **final** **static** String ***WOAIWOJIA\_HOMEPAGE*** = "http://bj.5i5j.com/exchange/n1";

**public** **final** **static** String ***MAITIAN\_HOMEPAGE*** = "http://bj.maitian.cn/esfall";

注意：

1. 房天下抓取的是自营房源
2. 在抓取的时候有几个网站需要分商业区抓取，因为房源列表最多给出100页
3. 各个网站大概房源总数：

* 链家：30000+
* 我爱我家：19000+
* 麦田：4000+
* 酷房：29000+（网站上写这么多，但是有大量重复的）
* 房天下：7000+

1. 各个网站页面不完全规范，jsoup抓取的时候注意抓取空指针
2. 有时候加载页面会出现READ TIME OUT错误，可以适当延长time参数，也可以存到list里进行重复抓取
3. 链家的网站有ip 访问限制，需要动态代理
4. 已有代码的抓取某个房源详情页部分代码不太合理，页面稍有改动就抓不到了。

### 数据：

使用数据库：MongoDB，因为数据字段数量不确定

存储原始数据的collection

"lianjia\_origin\_data",

"fangtianxia\_origin\_data",

"kufang\_origin\_data",

"maitian\_origin\_data",

"woaiwojia\_origin\_data"

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名称 | 类型 | Key | 描述 |
| houseUrl | String |  | 房源的URL地址 |
| buildingName | String |  | 小区名称 |
| buildingTile | String |  | 房屋标题 |
| totalPrice | Double |  | 总价 |
| unitPrice | Int |  | 单价 |
| Area | Double |  | 面积 |
| houseType | String |  | 房屋类型 |
| orientation | String |  | 朝向 |
| floor | String |  | 楼层 |
| buildingTags | String |  | 房屋特点标签 |
| decoration | String |  | 装修 |
| buildingAge | String |  | 建造年代 |
| companyName | String |  | 代理公司名称 |
| buidingNum | String | Primary | 房源编号 |
| List<Review> | Arraylist() |  | 存储经纪人评论的列表 |

存储示例：略

聚合后相似房源collection：

“similar\_house\_data”

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名称 | 类型 | Key | 描述 |
| Size | int |  | 同一特征相似房源数量 |
| List<HouseSource> similarHouseList | List |  | 相似房源列表 |
| houseSource | HouseSource |  | 代表性房源 |

存储举例(BSON&JSON)

Document{

{

\_id=58b7cfd6782aed0d42ef5517,

houseSource=Document{

{

area=60.0,

buildingAge=2000年建/塔楼,

buildingName=宝隆温泉公寓,

buildingNum=101101182933,

buildingTags=近地铁,

buildingTitle=宝隆温泉公寓2室1厅345万,

companyName=链家,

decoration=精装,

floor=中楼层/共31层,

houseType=2室,

houseUrl=http: //bj.lianjia.com/ershoufang/101101182933.html,

orientation=南,

reviews=[

Document{

{

agentName=廖露漫,

agentPhone=4008896039转4015,

date=2017/02/25,

review=【装修描述】：房子精装修，都是环保材料，现在改成两居室。布局合理！【交通出行】：地铁9号线科怡路站就在小区东边100米处，地铁9号线；【周边配套】：小区属于北京第二中关村的丰台科技园区，总部基地以北。小区方圆两公里有丰台万达广场，永旺梦乐城，资和信百货，欧尚超市。

}

},

Document{

{

agentName=高岩,

agentPhone=4008891675转9266,

date=2017/02/23,

review=房子南向，中间楼层，采光好。新装修一年左右，都是环保材料，有孩子，一居改成两居室，布居合理。地铁9号线科怡路站直线距离100米左右，也是公交站点，交通便利。小区方圆两公里有丰台万达广场，永旺梦乐城，资和信百货，欧尚超市。即将使用天坛医院总院，丰台医院。配套齐全，居民生活便利。

}

}

],

totalPrice=345.0,

unitPrice=57072

}

},

similarHouseList=[

Document{

{

area=60.0,

buildingAge=2000年,

buildingName=宝隆温泉公寓,

buildingNum=4514129849,

buildingTitle=五小学区地铁房宝隆公寓周边总价两居室你还在等什么,

companyName=酷房网,

decoration=中等装修,

floor=中楼层/共31层,

houseType=2室,

houseUrl=http: //beijing.koofang.com/sale/d-1006189186.html,

orientation=南,

reviews=[

Document{

{

agentName=张翔宇,

agentPhone=18600539689,

date=,

review=

}

}

],

totalPrice=345.0,

unitPrice=57500

}

}

],

size=2

}

}

### 类

|  |  |  |
| --- | --- | --- |
| 类 | 所属层次 | 业务逻辑 |
| CrawlController | Controller |  |
| ClassifyController | Controller |  |
| MongoDao | DAO |  |
| HouseSource | Model | 抽象一个房源的类 |
| Review | Model | 抽象一条经纪人评论 |
| SimilarHouseSource | Model | 存储相似房源 |
| ClassifyService | Service | 实现聚合的类 |
| {CompanyName}CrawlService | Service | 不同公司的爬虫服务 |
| ClassifyMongoService | Service | 封装好的Mongodb服务 |
| MongoService | Service | 封装好的Mongodb服务 |
| threads.area包中的类 |  | 实现多线程按商区抓取二手房URL |
| threads.house包中的类 |  | 实现多线程抓取具体所有房源的详细信息 |
| ConfigurationUtil |  | 配置类 |
| CrawlerWithProxy |  | 动态IP代理 |
| strDealUtil |  | 字符串处理 |

SimilarHouseSource类转化成JSON&BSON结构说明：

{ HouseSource house1（代表性房源），List[ house2, house3, house4, house 5…..], int size }

### 接口：

注意：程序中数据库collection名称写死，可以通过参数传递，不必写死

**接口名称**

/crawl/{companyName} /crawl/crawlAll

**功能与业务逻辑描述**

抓取某一个公司所有二手房源/抓取所有五个公司二手房源信息

**调用方式：**

**Get**

**输入参数：**

**公司名称：{“lianjia”，“woaiwojia”，“maitian”，“kufang”, “fangtianxia”}**

**输出参数：**

**无**

**接口名称**

/classify/{companyName} /classify/run

**功能与业务逻辑描述**

将指定数据库中的原始房源信息添加到已经分类的数据库中。添加过程会按照房屋面积，朝向，小区名称，几室这四个元素严格相等进行划分。（可以有更优算法）

**调用方式：**

**输入参数：**

**公司名称：{“lianjia”，“woaiwojia”，“maitian”，“kufang”, “fangtianxia”}**

**输出参数：**

**接口名称**

/read/esf-list/{page} /read/index

**功能与业务逻辑描述**

读取分类之后的房源列表。可以进一步查看相似房源的URL

**调用方式：**

**Get**

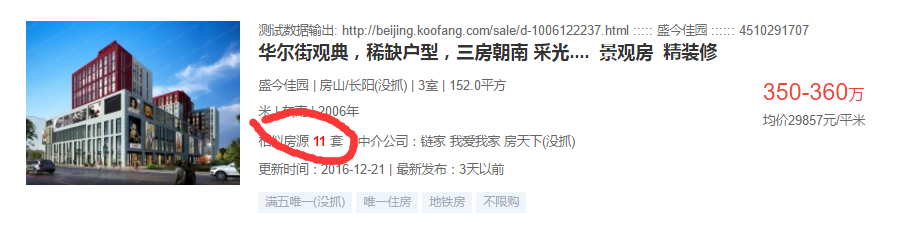
**输入参数：**

**page页数**

**index重定向到第一页**

**（页面只做了简单的翻页，需要进一步完善）**

**输出参数：**



点击可以看到一个简单的列表

