**爬虫说明文档**

**2021211338 郭柏彤**

**一、功能描述：**

本程序为基于scrapy框架的python数据爬取程序，可爬取链家官网上的新房以及二手房数据，爬取网址分别为：

新房（第3-第7页）：<https://bj.fang.lianjia.com/loupan/pg3/>

二手房(第3-第7页)：<https://bj.lianjia.com/ershoufang/pg3/>

爬取新房数据的爬虫名称为lianjia1，爬取二手房数据的爬虫名称为lianjia2，对于新房。分别爬取楼盘名称，地址，类型，房屋类型，面积，单价，总价。对于二手房，分别爬取小区名称，地址，类型，单价，总价。所有爬取到的数据保存于json文件中，其中，新房的数据保存于scrapy-test-firsthand.json，二手房的数据保存于scrapy-test-secondhand.json

**二、指令说明：**

使用指令 scrapy crawl lianjia1 开始爬取新房数据。

使用指令 scrapy crawl lianjia2 开始爬取二手房数据。

**三、item文件项目说明：**

class firsthanditem(scrapy.Item): #新房的项目

name = scrapy.Field() #名称

position = scrapy.Field() #地址

types = scrapy.Field() #类型

houseType = scrapy.Field() #房屋类型

space = scrapy.Field() #面积

unitPrice = scrapy.Field() #单价

totalPrice = scrapy.Field() #总价

class secondhanditem(scrapy.Item): #二手房的项目

name = scrapy.Field() #小区名称

position = scrapy.Field() #地址

types = scrapy.Field() #类型

unitPrice = scrapy.Field() #单价

totalPrice = scrapy.Field() #总价

示例：新房：{"name": "北辰墅院1900", "types": "住宅", "position": "顺兴街11号院望尊园", "houseType": "3室", "space": "建面 83-135㎡", "unitPrice": "36000", "totalPrice": "总价430(万/套)"}二手房：{"name": "中建国际港 ", "position": "枣园", "types": "2室1厅 | 86.96平米 | 南 | 简装 | 低楼层(共33层) | 板楼", "unitPrice": "49,219元/平", "totalPrice": "428"}

**四、项目包含文件说明：**

本项目包含文件：spiders文件夹：内含\_init\_.py,spider1.py,spider2.py

\_init\_.py, begin.py, items.py, middlewares.py, pipelines.py settings.py 其中，spider1和spider2分别为两个爬虫的主体文件，其中包含了网页的解析和获取数据功能，settings文件为爬虫的设置，可在其中修改爬虫的爬取间隔等参数。