租房数据分析

> 要求抓取链家官网北上广深 4 个一线城市，再加上一个离你家乡最近的一个非一线城市/或者你最感兴趣的一个城市的数据（选择老家湖南省常德市）。

- 爬虫代码文件见 `lianjia` 文件夹
- `Python` 画图及数据处理脚本见 `code` 文件夹
- 全部数据见 `data` 文件夹
- 全部可视化图片见 `images` 文件夹
- 数据分析实验报告见 实验报告.pdf

技术栈

- 爬虫框架：`Scrapy 2.12.0`
- Python 项目管理：`uv 0.5.7 (Homebrew 2024-12-06)`
- 可视化库：`matplotlib 3.9.4`，`seaborn 0.13.2`，`geopandas 1.0.1`
- 数学库：`numpy 2.2.0`，`pandas 2.2.3`

快速开始

1. 克隆本项目到本地

```
git clone https://github.com/Word2VecT/Lianjia-Spider-Demo
cd Lianjia-Spider-Demo
```

2. 安装 `uv` Python 项目管理器（需要 Python 环境以及 `pip`）

```
pip install uv
```

3. 使用 `uv` 安装虚拟环境

```
uv venv
```

4. 根据提示,激活虚拟环境 (以 `fish` 为例)

```
source .venv/bin/activate.fish
```

5. 安装项目依赖

```
uv pip install -r pyproject.toml
```

6. 进入爬虫代码文件夹

```
cd lianjia/spider
```

7. 修改 `rent_house_spider.py` 中的城市代码（bj，sh，gz，sz，changde 已测试，其他城市可能可以，未测试）
8. 运行爬虫

```
cd ..
scrapy crawl rent_house
```

9. 数据被保存在 `rent_house_data.json` 中
10. Enjoy it!

由于直接访问只能爬取到 100 页总共 3000 条数据，因此采用分类爬取的策略。首先不添加分类限制条件，判断总页数是否是 100，是则进行分类，分区域依次进行爬取，如果页数仍然是 100，则在分区域的基础上，分价格进行爬取，同理依次选择区域、价格、房型、朝向、楼高作为条件进行爬取。通过以上的限制条件分类爬取，实现了租房数据的全部获取。

反爬虫机制

链家具有反爬机制，短时间多次访问会被直接 `302` 重定向至验证码页面，我们采用了以下的反爬策略：

- 采用随机 `User-Agent`，模拟不同的浏览器访问，见 `settings.py` 相关设置。
- 采用 快代理 提供的隧道代理服务，并设置每次请求更换 IP，值得注意的是，由于已经采用代理，**无需设置 `cookie`，设置后反而会被封**，见 `settings.py` 相关设置以及 `middlewares.py` 中相关实现。
- 采用延时 `DELAY` 策略，每次请求后延时 0.5 秒，避免短时间内多次访问，见 `settings.py` 相关设置。
- 设置检测重试 `RETRY` 机制，由于未设置 `cookie`，未登录状态下可能会有小程序弹窗，刷新后消失，因此设置中间件 `middleware` 当未检测到数据时，自动重试，

见`settings.py`相关设置以及`middlewares.py`中相关实现。

- 设置自定义重定向重试`RETRY`机制，由于某些代理 ip 可能因为多次访问被链家封禁，这个时候换个 ip 刷新即可正常访问，而`Scarpy`的默认重定向的重试机制是，跟随重定向页面后刷新重试，设置中间件`middleware`，取消跟随，在原始页面重试，见`settings.py`相关设置以及`middlewares.py`中相关实现。
- 为避免个性化推荐引起不同 ip 爬取内容重复，设置`COOKIES\_ENABLED = False`，禁用`cookie`，见`settings.py`相关设置。