Python 数据抓取作业

吴清柳 2020211597

1 作业要求

分别爬取链家官网新房和二手房的数据 (从第 3 页开始到第 5 页). 新房需要爬取以下数据: 楼盘名称 类型 地点 房型 面积 单价 总价; 二手房需要爬取以下数据: 小区名称, 地点, 房型, 单价, 总价; 爬取的数据保存在 json 文件中.

2 内容

制作了两个爬虫 new_house 和 second_hand, 分别爬取链家的新房和二手房信息.

由于 Scrapy 默认保存的 json 文件中使用了 escape-unicode 字符格式, 且一个 json 条目放在一行, 让人类无法阅读. 为了解决这个问题, 制作了 `unicode2str.py` 来将 escape-unicode 字符转为 UTF-8 格式的人类可阅读的 json 文件.

3 测试环境

Python 3.11.0, Scrapy 2.7.1, Clang 14.0.4, macOS 13.1 22C5050e arm64; 使用 neovim v8.0.0 编写.

4 使用方法

使用前可以考虑在 settings.py 中停用 `'lianjia.middlewares.ProxiesMiddleware'` 中间件, 或者在 `middlewares.py` 中的 `ProxiesMiddleware` 类中更改代理为自己的代理.

```
# Run scrapy spiders to crawl Lianjia new house and second-hand house info and
# store them in json format.
scrapy crawl new_house -O new_house.json
scrapy crawl second_hand -O second_hand.json

# The data crawled are stored in new_house.json and second_hand.json.

# Considering that they are difficult for human to read, convert them into more
# readable type.
py unicode2str.py

# Converted json files are saved as converted_new_house.json and
# converted_second_hand.json

# Converted json files are saved as converted_new_house.json and
# converted_second_hand.json.
```

5 防反爬策略

在 settings.py 中进行如下设置来避免被链家屏蔽请求.

- 随机 User-Agent: 使用 `scrapy-user-agents` 来随机切换 UA, 放入 DOWNLOADER_MIDDLEWARES 中.
- 使用代理 IP: 制作代理中间件 ProxiesMiddleware 在 middlewares.py 中, 将通往代理的本机 8001 接口用作代理, 放入 DOWNLOADER_MIDDLEWARES 中.
- 设置下载延迟 DOWNLOAD_DELAY = 3
- 设置并行请求数 CONCURRENT_REQUESTS_PER_DOMAIN = 2, CONCURRENT_REQUESTS_PER_IP = 2.
- 不使用 cookie: COOKIES_ENABLED = False.