Data Scraping

2022-11-15

Brief Introduction

使用 **`Scrapy`** 框架抓取链家官网新房和二手房 `3-7` 页的数据。抓取新房的 *楼盘名称*、*类型*、*地点*、*房型*、*面积*、*单价*、*总价*; 抓取二手房的 *小区名称*、*地点*、*类型*、*单价*、*总价*。数据保存至 **`json`** 文件.

Primary Method

- **Scrapy** - `main frame for data scraping`
- **Beautiful Soup 4** - `html element filter`

Special Design

- **Runner**
  通过**Runner**在一个项目中配置两个爬虫任务，分别爬取新房数据和二手房数据.

  ```
  runner.crawl(NewHouseSpider) # 新房
  runner.crawl(SecHandHouseSpider) # 二手房
  runner.join().addBoth(lambda _:reactor.stop())
  reactor.run()
  ```

- **BS4**
  通过 `BS4` 进行元素过滤及提取, 主要运用 `select()` , `find()` 及 `find_all()` 方法.

  ```
  document = BeautifulSoup(response.text,features="lxml")
  for node in document.select('.resblock-desc-wrapper'):
      item=NewHouseData()
      item['type'] = node.find('span',class_ = 'resblock-type').text
  ```

- **Exception**
  在爬取过程中发现部分数据没有 **`总价`** 节点，对其进行异常处理.

  ```
  try:
      item['total_price'] = price_node.find(class_ = 'second').text[2:] # 过滤'总价'前缀
  except Exception:
      item['total_price'] = '' # 没有该节点，设为空
  ```

Data Storage

`task1/data/new_house.json` ：新房数据

`task1/data/sec_hand_house.json` ：二手房数据