#### 2019211301班 池纪君 2019213688

```
实验要求
实验环境
实验分析
  Scrapy爬虫原理简述
  普通Scrapy爬虫存在的问题
     动态网页
     网页端反爬虫机制
  Scrapy with Selenium
     在Scrapy中描述中间件
     Selenium
     Scrapy框架下Selenium的使用
代码分析
  代码文件树
  lianjia spider.py
  http.py
  items.py
  middlewares.py
  pipelines.py
  settings.py
  begin.py
  学堂在线爬虫核心代码
     xtzx spider.py
     items.py
     pipelines.py
实验结果
  学堂在线 (.csv)
  链家二手房(.json)
参考
```

## 实验要求

- 1. 爬取<u>学堂在线</u>的计算机类课程页面内容。 要求将课程名称、老师、所属学校和选课人数信息,保存到一个csv文件中。
- 2. 爬取链家官网二手房的数据

要求爬取北京市东城、西城、海淀和朝阳四个城区的数据(每个区爬取5页),将楼盘名称、总价、平米数、单价保存到json文件中。

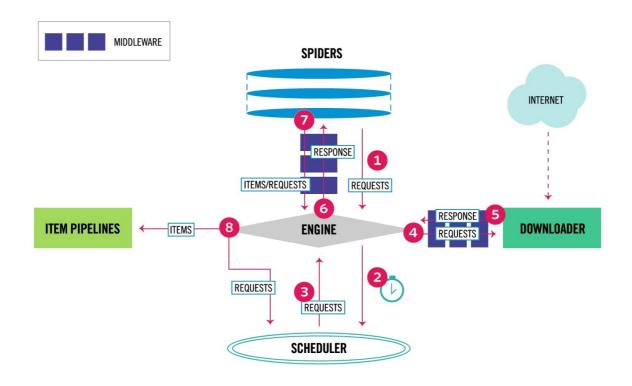
## 实验环境

- Windows 19043.1348
- Anaconda(Conda 4.10.3)
- Python 3.9.7 64-bit

```
Scrapy==2.5.1
selenium==4.1.0
```

## 实验分析

## Scrapy爬虫原理简述



• **Scrapy**是一个用Python编写的自由且开源的网络爬虫框架。它在设计上的初衷是用于爬取网络数据,但也可用作使用API来提取数据,或作为生成目的的网络爬虫。该框架目前由网络抓取的开发与服务公司Scrapinghub公司维护。

#### • 组成介绍:

1. 引擎(EGINE)

引擎负责控制系统所有组件之间的数据流,并在某些动作发生时触发事件。

2. 调度器(SCHEDULER)

用来接受引擎发过来的请求, 压入队列中, 并在引擎再次请求的时候返回. 可以想像成一个URL的优先级队列, 由它来决定下一个要抓取的网址是什么, 同时去除重复的网址。

3. 下载器(DOWLOADER)

用于下载网页内容,并将网页内容返回给EGINE,下载器是建立在twisted这个高效的异步模型上的。

4. 爬虫(SPIDERS)

SPIDERS是开发人员自定义的类,用来解析responses,并且提取items,或者发送新的请求。

5. 项目管道(ITEM PIPLINES)

在items被提取后负责处理它们,主要包括清理、验证、持久化(比如存到数据库)等操作。

- 6. 下载器中间件(Downloader Middlewares)位于Scrapy引擎和下载器之间,主要用来处理从EGINE传到DOWLOADER的请求request,已经从DOWNLOADER传到EGINE的响应response。
- 7. 爬虫中间件(Spider Middlewares) 位于EGINE和SPIDERS之间,主要工作是处理SPIDERS的输入(即responses)和输出(即requests)。

### 普通Scrapy爬虫存在的问题

#### 动态网页

问题描述:

在通过scrapy框架进行某些网站数据爬取的时候,往往会碰到页面动态数据加载的情况发生,如果直接使用scrapy对其url发请求,是绝对获取不到那部分动态加载出来的数据值。但是通过观察我们会发现,通过浏览器进行url请求发送则会进

行js的渲染,加载出对应的动态加载出的数据。那么如果我们也必须获取动态加载的数据值。

#### 解决:

使用selenium创建浏览器对象,然后通过该浏览器对象进行请求发送,获取动态加载的数据值。

#### 网页端反爬虫机制

#### • 问题描述:

在编写爬虫的过程中,有些网站会设置反爬机制,对于非浏览器的访问拒绝响应;或短时间频繁爬取会触发网站的反爬机制,导致 ip 被封无法爬取网页。这就需要在爬虫程序中修改请求的 headers 伪装浏览器访问,或者使用代理发起请求。从而绕开网站的反爬机制获取正确的页面。Python的requests发送的请求所带的请求头中,User-Agent 会标识为 python 程序发送的请求,对于有反爬的网站会识别其 headers 从而拒绝返回正确的网页。此时需要对发送的请求伪装成浏览器的 headers。

#### • 解决:

- 1. (本实验中选用此方法)使用自动化测试工具 selenium 可以模拟使用浏览器 访问网站。要使用浏览器需下载对应版本的 driver。浏览器 driver 也能伪装 User-Agent ,只需在创建 webdriver 浏览器对象的时候传入设置即可。
- 2. 当一个 ip 短时间访问太过频繁时,网站的反爬机制会被触发,将会要求输入验证码甚至封锁 ip 禁止访问。此时需要使用代理 ip 发起请求才可获取正确的网页。

#### **Scrapy with Selenium**

#### 在Scrapy中描述中间件

Scrapy的中间件主要分为下载中间件和爬虫中间件两种。

- 1. Downloader Middleware即下载中间件,它是处于Scrapy的Request和Response 之间的处理模块。作用有:
  - a. Scheduler从队列中拿出Request发送给Downloader执行下载,这个过程会经过下载中间件的处理。在Scheduler调度出队列的Request发送给Doanloader下载之前,也就是我们可以在Request执行下载之前对其进行修改。
  - b. 另外,当Downloader将Request下载完成得到Response返回给Spider时会再次经过Downloader Middleware处理。在下载后生成的Response发送给

- Spider之前,也就是在生成Resposne被Spider解析之前对其进行修改。
- Downloader Middleware的功能十分强大,修改User-Agent、处理重定向、设置代理、失败重试、设置Cookies等功能都需要借助它来实现。
- 2. Spider Middleware是介入到Scrapy的Spider处理机制的钩子框架。Spider Middleware有如下三个作用:
  - a. 在Downloader生成的Response发送给Spider之前,也就是在Response发送给Spider之前对Response进行处理。
  - b. 我们可以在Spider生成的Request发送给Scheduler之前,也就是在Request发送给Scheduler之前对Request进行处理。
  - c. 我们可以在Spider生成的Item发送给Item Pipeline之前,也就是在Item发送给Item Pipeline之前对Item进行处理。

#### Selenium

- selenium最初是一个自动化web测试工具,通过代码模拟人使用浏览器自动访问目标站点并操作,比如跳转、输入、点击、下拉等。由于开发者的不断完善,目前的功能越来越强大,基本支持各种交互操作。同时,不止支持有界浏览,还支持无界浏览。
- Selenium 测试直接在浏览器中运行,就像真实用户所做的一样,所以利用这点,可以更好的进行数据采集。

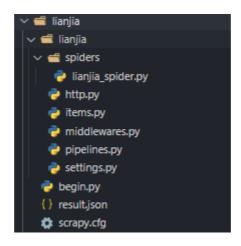
### Scrapy框架下Selenium的使用

- 使用流程
  - 1. 重写爬虫文件的\_\_init\_\_()构造方法,在该方法中使用selenium实例化一个浏览器对象(因为浏览器对象只需要被实例化一次)。
  - 2. 重写爬虫文件的closed(self,spider)方法,在其内部关闭浏览器对象,该方法是在爬虫结束时被调用。
  - 3. 重写下载中间件的process\_response方法,让该方法对响应对象进行拦截, 并篡改response中存储的页面数据。
  - 4. 在settings配置文件中开启下载中间件
- Scrapy\_selenium库将Selenium库进行底层的封装,将Selenium中创建 Webdriver,获取网页并运行网页对应的Javascript脚本这一操作进行模块化,在 爬虫Request网页的时候调用,起到中间件的作用。
- 这样,只需要调用 SeleniumRequest() 函数,并将 parse() 作为回调函数,即可实现

## 代码分析

以链家二手房的数据爬取为例:

### 代码文件树



### lianjia\_spider.py

该文件为爬虫类的定义文件。

- 在执行 scrapy crawl 后, lianjiaSpider 类被声明,并以 start\_requests() 开始执行。该函数产生5个初始的自定义请求 Myrequest。
- 在下载器将服务器response返回后,进入回调函数 parse()。该函数将对response 进行解析,并产生结构化的房屋数据项目: LianjiaItem ,该项目将被返回给 pipeline。若需要继续爬取下一页的信息,需要继续产生自定义请求 MyRequest 。

```
import scrapy
from lianjia.http import MyRequest
from lianjia.items import LianjiaItem
class lianjiaSpider(scrapy.Spider):
    name = "lianjia"
    max_page = 5
    start_urls = [
        "https://bj.lianjia.com/ershoufang/xicheng/",
        "https://bj.lianjia.com/ershoufang/dongcheng/",
        "https://bj.lianjia.com/ershoufang/haidian/",
        "https://bj.lianjia.com/ershoufang/chaoyang/"
    ]
    def start_requests(self):
        for url in self.start_urls:
            yield MyRequest(url=url,
                          wait_time=1,
```

```
callback=self.parse,
                      dont_filter=True,
                      cb_kwargs={'page': 1})
def parse(self, response, **kwargs):
    print(kwargs["page"])
    houses = response.xpath("/html/body/div[4]/div[1]/ul/li")
    for house in houses:
        item = LianjiaItem()
        item['name'] = house.xpath("./div[1]/div[1]/a/text()").get()
        price = house.xpath("./div[1]/div[6]/div[1]/span/text()").get(
        ) + house.xpath("./div[1]/div[6]/div[1]/i[last()]/text()").get()
        item['price'] = price
        item['size'] = house.xpath(
            "./div[1]/div[3]/div/text()").get().split('|')[1].strip()
        item['perunit'] = house.xpath(
            "./div[1]/div[6]/div[2]/span/text()").get()
        item['region'] = response.url.split('/')[4]
        yield item
    if kwargs["page"] + 1 < self.max_page:</pre>
        nxt_page_url = response.xpath(
            "/html/body/div[4]/div[1]/div[7]/div[2]/div/a[last()]/@href"
        ).get()
        yield MyRequest(url=response.urljoin(nxt_page_url),
                      wait_time=1,
                      callback=self.parse,
                      dont_filter=True,
                      cb_kwargs={'page': kwargs["page"] + 1})
```

### http.py

该文件重定义了 scrapy 的请求 request 操作,使其支持等待这一操作。

需要注意的是,由于 scrapy.request 中使用了 cb\_kwargs 这一名称,故在本类的实现中

```
from scrapy import Request

class MyRequest(Request):
    def __init__(self, wait_time=None, wait_until=None, *args, **cb_kwargs):
        self.wait_time = wait_time
        self.wait_until = wait_until
        super().__init__(*args, **cb_kwargs)
```

### items.py

该文件定义了房屋数据的结构体。

```
# Define here the models for your scraped items
#
# See documentation in:
# https://docs.scrapy.org/en/latest/topics/items.html
import scrapy

class LianjiaItem(scrapy.Item):
    # define the fields for your item here like:
    # name = scrapy.Field()
    name = scrapy.Field()
    price = scrapy.Field()
    size = scrapy.Field()
    perunit = scrapy.Field()
    region = scrapy.Field()
```

### middlewares.py

该文件定义了爬虫的下载中间件 LianjiaDownloaderMiddleware。

- 在初始化该类的时候,初始定义了 Webdriver 。
- 在 from\_crawler 中,可以用爬虫的配置对中间件进行初始化,本项目中并不需要。
- 在 process\_request 中,该中间件会收到爬虫的请求,并将该请求发送给远端网页,再将结果作为 HtmlResponse 返还给爬虫。在该函数中用了隐式等待,以保证 Webdriver 加载完网页。
- 在 spider\_closed 中,中间件会关掉浏览器驱动 Webdriver 。

```
from scrapy import signals
from selenium import webdriver
from selenium.webdriver.support.ui import WebDriverWait
from scrapy.http import HtmlResponse
# from selenium.webdriver.support import expected_conditions as Expect
# from selenium.webdriver.common.by import By
class LianjiaDownloaderMiddleware:
    """Scrapy middleware handling the requests using selenium"""
    def __init__(self):
        driver_executable_path = 'V:\Code\Py_projects\spiders\chromedriver.exe'
        driver_options = webdriver.ChromeOptions()
        driver_options.add_argument('--headless')
        driver_options.add_argument('--ignore-certificate-errors-spki-list')
        driver_options.add_argument('log-level=3')
        self.driver = webdriver.Chrome(executable_path = driver_executable_path, optio
ns = driver_options)
    @classmethod
```

```
def from_crawler(cls, crawler):
       """Initialize the middleware with the crawler settings"""
       middleware = cls()
       crawler.signals.connect(middleware.spider_closed, signals.spider_closed)
       return middleware
   def process_request(self, request, spider):
       self.driver.implicitly_wait(30)
        self.driver.get(request.url)
       self.driver.find_elements_by_class_name('sellListContent')
       # x = WebDriverWait(self.driver, 30).until(Expect.presence_of_element_located
((By.CLASS_NAME, "result")))
       for cookie_name, cookie_value in request.cookies.items():
           self.driver.add_cookie(
               {
                    'name': cookie_name,
                    'value': cookie_value
               }
       if request.wait_until:
           WebDriverWait(self.driver, request.wait_time).until(
                request.wait_until
            )
       body = str.encode(self.driver.page_source)
       # Expose the driver via the "meta" attribute
        request.meta.update({'driver': self.driver})
        return HtmlResponse(
           self.driver.current_url,
           body=body,
           encoding='utf-8',
           request=request
        )
   def spider_closed(self):
        """Shutdown the driver when spider is closed"""
       self.driver.quit()
```

#### pipelines.py

本文定义了 LianjiaPipeline 中间件。

- process\_item 为接收到爬虫产生的完整房屋数据结构后,对该数据进行的操作(一般是存储)
- close\_spider 为爬虫结束后进行的操作。此处为存储到 json 中。

```
import json
```

```
class LianjiaPipeline:
    items = [[{"region":"西城"}],[{"region":"东城"}],[{"region": "海淀"}],[{"region":
    "朝阳"}]]
    regid = {"xicheng": 0, "dongcheng": 1, "haidian": 2, "chaoyang": 3}

def process_item(self, item, spider):
    print(item['region'], type(item['region']))
    self.items[self.regid[item['region']]].append(dict(item))
    return item

def close_spider(self, spider):
    try:
        self.file = open('result.json', "w", encoding="utf-8")
    except Exception as err:
        print(err)

    json_str = json.dumps(self.items, ensure_ascii=False, indent=4)

    self.file.write(json_str)
    self.file.close()
```

### settings.py

本文件定义了爬虫项目的参数和中间件使用。

```
BOT_NAME = 'lianjia'
SPIDER_MODULES = ['lianjia.spiders']
NEWSPIDER_MODULE = 'lianjia.spiders'
# Obey robots.txt rules
ROBOTSTXT\_OBEY = False
# Configure maximum concurrent requests performed by Scrapy (default: 16)
CONCURRENT_REQUESTS = 32
# Disable cookies (enabled by default)
COOKIES ENABLED = False
# Enable or disable downloader middlewares
# See https://docs.scrapy.org/en/latest/topics/downloader-middleware.html
DOWNLOADER_MIDDLEWARES = {
   'lianjia.middlewares.LianjiaDownloaderMiddleware': 543,
}
# Configure item pipelines
# See https://docs.scrapy.org/en/latest/topics/item-pipeline.html
ITEM_PIPELINES = {
   'lianjia.pipelines.LianjiaPipeline': 300,
}
```

### begin.py

本文件为脚本文件,作用是执行 scrapy crawl lianjia。

```
from scrapy import cmdline
cmdline.execute("scrapy crawl lianjia".split())
```

### 学堂在线爬虫核心代码

#### xtzx spider.py

```
import scrapy
from lianjia.http import MyRequest
from lianjia.items import LianjiaItem
class lianjiaSpider(scrapy.Spider):
    name = "lianjia"
    max_page = 1
    id = 1
    start_urls = [
        "https://bj.lianjia.com/ershoufang/xicheng/",
        "https://bj.lianjia.com/ershoufang/dongcheng/",
        "https://bj.lianjia.com/ershoufang/haidian/",
        "https://bj.lianjia.com/ershoufang/chaoyang/"
    ]
    def start_requests(self):
        for url in self.start_urls:
            yield MyRequest(url=url,
                            callback=self.parse,
                            dont_filter=True,
                            cb_kwargs={'page': 1})
    def parse(self, response, **kwargs):
        houses = response.xpath("/html/body/div[4]/div[1]/ul/li")
        for house in houses:
            item = LianjiaItem()
            item['name'] = house.xpath("./div[1]/div[1]/a/text()").get()
            price = house.xpath("./div[1]/div[6]/div[1]/span/text()").get(
            ) + house.xpath("./div[1]/div[6]/div[1]/i[last()]/text()").get()
            item['price'] = price
            item['size'] = house.xpath(
                "./div[1]/div[3]/div/text()").get().split('|')[1].strip()
            item['perunit'] = house.xpath(
                "./div[1]/div[6]/div[2]/span/text()").get()
            item['region'] = response.url.split('/')[4]
            item['id'] = self.id
            self.id += 1
            yield item
```

#### items.py

```
import scrapy

class LianjiaItem(scrapy.Item):
    # define the fields for your item here like:
    # name = scrapy.Field()
    name = scrapy.Field()
    price = scrapy.Field()
    size = scrapy.Field()
    perunit = scrapy.Field()
    region = scrapy.Field()
    id = scrapy.Field()
```

### pipelines.py

```
import json
class LianjiaPipeline:
    items = [[{"region":"西城"}],[{"region":"东城"}],[{"region": "海淀"}],[{"region":
 "朝阳"}]]
    regid = {"xicheng": 0, "dongcheng": 1, "haidian": 2, "chaoyang": 3}
    def process_item(self, item, spider):
        print(item['region'], type(item['region']))
        self.items[self.regid[item['region']]].append(dict(item))
        return item
    def close_spider(self, spider):
            self.file = open('result.json', "w", encoding="utf-8")
        except Exception as err:
            print(err)
        json_str = json.dumps(self.items, ensure_ascii=False, indent=4)
        self.file.write(json_str)
        self.file.close()
```

## 实验结果

### 学堂在线(.csv)

```
id, page, course_name, teachers, school, number
1,1,C++语言程序设计基础,"['郑莉', '李超']",清华大学,466884 人
2,1,数据结构(上),['邓俊辉'],清华大学,456221 人
3,1,数据结构(下),['邓俊辉'],清华大学,381149人
4,1,操作系统,"['向勇', '陈渝']",清华大学,218892 人
5,1,Java程序设计,['郑莉'],清华大学,213223 人
6,1,网络技术与应用,"['沈鑫剡', '俞海英']",中国人民解放军陆军工程大学,199238 人
7,1,C++语言程序设计进阶,"['郑莉', '李超']",清华大学,135101 人
8,1,C程序设计案例教程(基础),['张莉'],中国农业大学,117460 人
9,1,C程序设计案例教程(进阶),['张莉'],中国农业大学,112999 人
10,1,大数据技术与应用,['李军'],清华大学,104728 人
11,2,软件工程,"['刘强', '刘璘']",清华大学,103433 人
12,2,计算机文化基础,"['李秀', '姚瑞霞']",清华大学,91338 人
13,2,程序设计基础,"['徐明星', '王瑀屏']",清华大学,89903 人
14,2,人工智能原理,['王文敏'],北京大学,79359 人
15,2,组合数学,['马昱春'],清华大学,78961 人
16,2,大数据系统基础,"['王建民', '徐葳']",清华大学,78893 人
17,2,0ffice办公软件应用,"['史巧硕', '朱怀忠']",河北工业大学,75083 人
18, 2, 算法设计与分析, ['王振波'], 清华大学, 73176 人
19,2,VC++面向对象与可视化程序设计(上):Windows编程基础,['黄维通'],清华大学,71678 人
20,2,大数据平台核心技术,"['武永卫', '姚文辉']",清华大学,69717 人
21,3,VC++面向对象与可视化程序设计(下):MFC编程基础,['黄维通'],清华大学,69490 人
22,3,Web前端攻城狮,"['刘强', '吴亮']",清华大学,64233 人
23,3,大学计算机教程,"['张莉', '马钦']",中国农业大学,63595 人
24,3,计算机应用基础,"['宋承继', '王坤']",陕西工业职业技术学院,62522 人
25,3,汇编语言程序设计,"['张悠慧', '翟季冬']",清华大学,60110 人
26,3,大学计算机基础,['卫春芳'],湖北大学,59128 人
27,3,学做小程序——基础篇,"['刘强', '小程序慕课讲师']",清华大学,57109 人
28,3,基于Linux的C++,['乔林'],清华大学,56120 人
29,3,微软亚洲研究院大数据系列讲座,"['洪小文', '宋睿华']",Microsoft,54240 人
30,3,计算机网络,"['袁华', '杜广龙']",华南理工大学,53106 人
31,4,大学计算机基础,"['徐红云', '刘欣欣']",华南理工大学,52760 人
32,4,JAVA程序设计进阶,['许斌'],清华大学,52680 人
33,4,大学计算机—计算思维的视角,['郝兴伟'],山东大学,52341 人
34,4,面向对象程序设计(C++),"['黄震春', '徐明星']",清华大学,51814 人
35,4,计算几何,['邓俊辉'],清华大学,47959 人
36,4,R语言数据分析,['艾新波'],北京邮电大学,45128 人
37, 4, Python程序设计基础, ['许志良'], 深圳信息职业技术学院, 44720 人
38,4,大数据机器学习,['袁春'],清华大学,42246 人
39,4,数据库技术与程序设计,"['高裴裴', '康介恢']",南开大学,41800 人
40,4,程序设计基础(上),"['赵宏', '闫晓玉']",南开大学,41221 人
41,5,单片机原理及应用,"['杨居义', '王颖丽']",绵阳职业技术学院,40847 人
42,5,Web开发技术,"['王成良', '陈静']",重庆大学,39557 人
43,5,ARM微控制器与嵌入式系统,"['曾鸣', '薛涛']",清华大学,39448 人
44,5,Linux 内核分析与应用,"['陈莉君', '谢宝友']",西安邮电大学,39226 人
45,5,区块链和加密数字货币,['罗玫'],清华大学,38381 人
46,5,程序设计基础(下),"['赵宏', '李妍']",南开大学,37441 人
47,5,大学计算机基础,"['李敏', '高裴裴']",南开大学,36824 人
48,5,C语言程序设计(上),"['李凤霞', '薛静锋']",北京理工大学,33570 人
```

```
49,5,数据结构与算法设计,"['张小艳', '史晓楠']",西安科技大学,33449 人
50,5,高级大数据系统,['王智'],清华大学,33088 人
```

## 链家二手房(.json)

由于网络的随机性和多个连接的并发请求,地区中的id不连续。

```
[
   [
          "region": "西城"
       },
           "name": "汽北小区小区 南北通透两居室 诚心出售",
           "price": "650万",
           "size": "52.8平米",
           "perunit": "123,107元/平",
           "region": "xicheng",
           "id": 1
       },
           "name": "南北双通透,中间层采光好,层高2.7米,有钥匙随时看",
           "price": "1290万",
           "size": "83.3平米",
           "perunit": "154,862元/平",
           "region": "xicheng",
           "id": 2
       },
           "name": "侨办大院 三室一厅 有电梯 满五年唯一",
           "price": "1400万",
           "size": "93.9平米",
           "perunit": "149,095元/平",
           "region": "xicheng",
           "id": 3
       },
           "name": "红居南街 2室1厅 西北",
           "price": "469万",
           "size": "45.77平米",
           "perunit": "102,469元/平",
           "region": "xicheng",
           "id": 4
       },
           "name": "红莲晴园 三居双卫 阳台大 中间楼层 采光好 无遮挡",
           "price": "1200万",
           "size": "117.67平米",
           "perunit": "101,981元/平",
           "region": "xicheng",
           "id": 5
       },
           "name": "三里河南二巷双南两居,采光好,无遮挡",
```

```
"price": "789万",
   "size": "54.9平米",
    "perunit": "143,716元/平",
   "region": "xicheng",
   "id": 6
},
   "name": "小区车位充足,四居室,中组部分房,阜成门,月坛公园",
    "price": "1218万",
   "size": "117.7平米",
   "perunit": "103,484元/平",
   "region": "xicheng",
   "id": 7
},
   "name": "德胜里一区 双南向2居室 业主诚售",
   "price": "810万",
   "size": "50.1平米",
   "perunit": "161,677元/平",
   "region": "xicheng",
   "id": 9
},
   "name": "1997北礼士路 高楼层 带电梯正规大客厅 房龄新 精装修",
   "price": "760万",
   "size": "60.1平米",
   "perunit": "126,456元/平",
   "region": "xicheng",
   "id": 10
},
   "name": "西城区 南北三居 明厨明卫 低楼层 采光充足 满五年",
   "price": "780万",
   "size": "77.78平米",
   "perunit": "100,283元/平",
   "region": "xicheng",
   "id": 11
},
   "name": "丰汇园小区 3室2厅 东 西",
   "price": "2450万",
   "size": "149.2平米",
   "perunit": "164,210元/平",
   "region": "xicheng",
   "id": 12
},
   "name": "新德街35号院2000年建成钢混结构南北向两居中间层",
   "price": "1340万",
   "size": "89.5平米",
   "perunit": "149,721元/平",
   "region": "xicheng",
   "id": 21
},
   "name": "阜成门南大街 带电梯两居室 视野好",
   "price": "615万",
   "size": "40.1平米",
```

```
"perunit": "153,367元/平",
   "region": "xicheng",
   "id": 22
},
{
   "name": "广安门 天宁寺 广华轩 电梯 复式",
   "price": "1480万",
   "size": "193.46平米",
   "perunit": "76,502元/平",
   "region": "xicheng",
   "id": 23
},
   "name": "阳光丽景南北通透 经典大三居 有产权车位 高层视野",
   "price": "2390万",
   "size": "146.71平米"
   "perunit": "162,907元/平",
   "region": "xicheng",
   "id": 24
},
   "name": "车公庄大街南北两居室 满五年住房 随时看房和签约",
   "price": "720万",
   "size": "56.7平米",
   "perunit": "126,985元/平",
   "region": "xicheng",
   "id": 25
},
   "name": "牛街西里西南向两居 无遮挡采光好看房方便 诚意出售",
   "price": "763万",
   "size": "74.93平米",
   "perunit": "101,829元/平",
   "region": "xicheng",
   "id": 35
},
   "name": "长椿街 广安门 朝南小两居 满二唯一",
   "price": "560万",
   "size": "50.22平米",
   "perunit": "111,510元/平",
   "region": "xicheng",
   "id": 36
},
   "name": "长椿街 满五南北通透明厨明卫中间楼层有电梯",
   "price": "720万",
   "size": "60.73平米",
   "perunit": "118,558元/平",
   "region": "xicheng",
   "id": 37
},
   "name": "陶然亭平原里小区南北通透一居室采光无遮挡",
   "price": "595万",
   "size": "50.12平米",
   "perunit": "118,716元/平",
   "region": "xicheng",
```

```
"id": 38
},
{
   "name": "裕中西里 南北通透三居室 正南正北 采光好",
   "price": "1250万",
   "size": "89.1平米",
   "perunit": "140,292元/平",
   "region": "xicheng",
   "id": 39
},
   "name": "西城安德路优质三居室,田字格户型格局好采光好",
   "price": "885万",
   "size": "68.3平米",
   "perunit": "129,576元/平",
   "region": "xicheng",
   "id": 40
},
{
   "name": "98年钢混,电梯直达,满五年唯一,精装,业主诚心出售",
   "price": "660万",
   "size": "73.96平米",
   "perunit": "89,238元/平",
   "region": "xicheng",
   "id": 50
},
   "name": "黄瓜园南北通透三居室,中间楼层,1992年楼龄新",
   "price": "1200万",
   "size": "97.1平米",
   "perunit": "123,584元/平",
   "region": "xicheng",
   "id": 51
},
   "name": "西城两居,南北无遮挡,视野好,小区管理好,湾子站。",
   "price": "609万",
   "size": "58.86平米",
   "perunit": "103,466元/平",
   "region": "xicheng",
   "id": 52
},
   "name": "满五年家庭唯一住房,中间楼层,南北朝向通透 采光好",
   "price": "930万",
   "size": "61.6平米",
   "perunit": "150,975元/平",
   "region": "xicheng",
   "id": 53
},
   "name": "马甸南村 南北通透两居 把边户型 临近公园",
   "price": "990万",
   "size": "65.11平米",
   "perunit": "152,051元/平",
   "region": "xicheng",
   "id": 54
},
```

```
"name": "西直门南大街四居室业主诚心出售,看房方便",
       "price": "1300万",
       "size": "94.2平米",
       "perunit": "138,005元/平",
       "region": "xicheng",
       "id": 55
   },
       "name": "裕中东里南北三居 全明格局 可签约",
       "price": "998万",
       "size": "77.8平米",
       "perunit": "128,278元/平",
       "region": "xicheng",
       "id": 66
   },
       "name": "双南向,电梯房,采光好, 满五年住房, 随时看房",
       "price": "730万",
       "size": "50.7平米",
       "perunit": "143,985元/平",
       "region": "xicheng",
       "id": 67
   }
],
[
   {
       "region": "东城"
   },
       "name": "东二环 阳光都市 紧挨工体三里屯",
       "price": "2200万",
       "size": "240.55平米",
       "perunit": "91,458元/平",
       "region": "dongcheng",
       "id": 8
   },
       "name": "北河沿大街 中间楼层东向两居室",
       "price": "666万",
       "size": "53.03平米",
       "perunit": "125,590元/平",
       "region": "dongcheng",
       "id": 13
   },
       "name": "东城区 悠胜美苑 03年新楼 商品房 视野好",
       "price": "545万",
       "size": "70.17平米",
       "perunit": "77,669元/平",
       "region": "dongcheng",
       "id": 14
   },
       "name": "忠实里东区 中间层 带电梯 商品房",
       "price": "632万",
       "size": "57.71平米",
       "perunit": "109,514元/平",
```

```
"region": "dongcheng",
   "id": 15
},
{
   "name": "京城仁合朝东向两居室没有遮挡视野开阔。",
   "price": "1100万",
   "size": "100.14平米",
   "perunit": "109,847元/平",
   "region": "dongcheng",
   "id": 16
},
   "name": "龙潭湖 龙潭北里小区 规整社区 配套齐全全南向三居室",
   "price": "630万",
   "size": "65.37平米",
   "perunit": "96,375元/平",
   "region": "dongcheng",
   "id": 17
},
   "name": "东城区 2001年商品房 南向两居 双阳台 满五唯一",
   "price": "913万",
   "size": "114.17平米",
   "perunit": "79,969元/平",
   "region": "dongcheng",
   "id": 18
},
   "name": "菊儿胡同 4室1厅 南 北",
   "price": "1430万",
   "size": "92.35平米",
   "perunit": "154,846元/平",
   "region": "dongcheng",
   "id": 26
},
   "name": "中海紫御公馆明厨明卫双南向大卧室两居",
   "price": "1028万",
   "size": "90.08平米",
   "perunit": "114,121元/平",
   "region": "dongcheng",
   "id": 27
},
   "name": "商务部家属院 东交民巷 东单 全南两居室,满五年唯一",
    "price": "636万",
   "size": "49.79平米",
   "perunit": "127,737元/平",
   "region": "dongcheng",
   "id": 28
},
   "name": "富贵园1居室中间楼层,小区人车分流距离地铁100米,",
   "price": "648万",
   "size": "63.74平米",
   "perunit": "101,664元/平",
   "region": "dongcheng",
   "id": 29
```

```
},
{
   "name": "广渠门忠实里 三居室 东南向 有电梯采光好",
   "price": "720万",
   "size": "82.2平米",
   "perunit": "87,592元/平",
   "region": "dongcheng",
   "id": 41
},
   "name": "广渠门内大街 不临街 交通方便",
   "price": "655万",
   "size": "63.93平米",
   "perunit": "102,456元/平",
   "region": "dongcheng",
   "id": 42
},
   "name": "东城二环里新改造小区靠公园地铁",
   "price": "430万",
   "size": "45.59平米",
   "perunit": "94,319元/平",
   "region": "dongcheng",
   "id": 43
},
   "name": "国瑞城高层南北三居室 地铁崇文门站",
   "price": "1699万",
   "size": "157.42平米",
   "perunit": "107,928元/平",
   "region": "dongcheng",
   "id": 44
},
   "name": "崇文门国瑞城中区南北三居,前后观景,双明卫,有钥匙",
   "price": "2230万",
   "size": "186.63平米",
   "perunit": "119,488元/平",
   "region": "dongcheng",
   "id": 56
},
   "name": "沙滩后街55号院 南北通透 精装修两居室 满五唯一",
   "price": "780万",
   "size": "56.91平米",
   "perunit": "137,059元/平",
   "region": "dongcheng",
   "id": 57
},
   "name": "华城滨河世家+高层大一居+两居格局+满五唯一+钥匙房",
   "price": "830万",
   "size": "66.1平米",
   "perunit": "125,568元/平",
   "region": "dongcheng",
   "id": 58
},
```

```
"name": "本家润园C区 视野好 临地铁 两居室 配套成熟",
   "price": "720万",
   "size": "70.22平米",
   "perunit": "102,535元/平",
   "region": "dongcheng",
   "id": 59
},
   "name": "东城,赵堂子胡同小区,精装自住,中粮和商务部家属楼",
   "price": "599万",
   "size": "41.66平米",
   "perunit": "143,784元/平",
   "region": "dongcheng",
   "id": 68
},
   "name": "海运仓规整社区 三室两卫复式带电梯 南北朝向采光好",
   "price": "1150万",
   "size": "95.18平米",
   "perunit": "120,824元/平",
   "region": "dongcheng",
   "id": 69
},
   "name": "广渠门桥 金世纪嘉园西南两居室出售",
   "price": "850万",
   "size": "108.11平米",
   "perunit": "78,624元/平",
   "region": "dongcheng",
   "id": 70
},
   "name": "朝内大街 精装修 南北东通透两居室 中间楼层随时签约",
   "price": "690万",
   "size": "63.88平米",
   "perunit": "108,016元/平",
   "region": "dongcheng",
   "id": 71
},
   "name": "兴隆都市馨园全南两居,双阳台,得房率92%",
   "price": "798万",
   "size": "72.73平米",
   "perunit": "109,721元/平",
   "region": "dongcheng",
   "id": 81
},
   "name": "板厂南里 低楼层 南北通透 满五年唯一 不临街",
   "price": "590万",
   "size": "50.04平米",
   "perunit": "117,906元/平",
   "region": "dongcheng",
   "id": 82
},
   "name": "人车分流,环境好、高层,楼层好、东西不临街,景观好",
   "price": "1600万",
```

```
"size": "139.17平米",
       "perunit": "114,968元/平",
       "region": "dongcheng",
       "id": 83
   },
       "name": "房子南北通透,满五年唯一住房,中间楼层视野开阔。",
       "price": "899万",
       "size": "77.02平米",
       "perunit": "116,723元/平",
       "region": "dongcheng",
       "id": 84
   },
       "name": "中海紫御公馆 开发商精装两居 对小区花园",
       "price": "1150万",
       "size": "91.86平米",
       "perunit": "125,191元/平",
       "region": "dongcheng",
       "id": 85
   },
       "name": "新景家园东区 商品房 东西三居室 满五年唯一",
       "price": "1160万",
       "size": "88.45平米",
       "perunit": "131,148元/平",
       "region": "dongcheng",
       "id": 97
   },
       "name": "琉璃井东街 2室1厅 南 北",
       "price": "486万",
       "size": "54.65平米",
       "perunit": "88,930元/平",
       "region": "dongcheng",
       "id": 98
   }
],
Γ
   {
       "region": "海淀"
   },
       "name": "韦伯豪家园大复式 南北通透,全明卫 诚心出售",
       "price": "2500万",
       "size": "231.2平米",
       "perunit": "108,132元/平",
       "region": "haidian",
       "id": 19
   },
       "name": "复兴路79号院 南北向 两居室 满五年 采光好",
       "price": "526万",
       "size": "66.4平米",
       "perunit": "79,217元/平",
       "region": "haidian",
       "id": 20
   },
```

```
"name": "金隅山墅 南北通透 毛坯 上叠拼别墅",
    "price": "1600万",
   "size": "224.81平米",
   "perunit": "71,172元/平",
   "region": "haidian",
   "id": 30
},
   "name": "前沙涧路7号院 2室1厅 南 北",
   "price": "360万",
   "size": "86.55平米",
   "perunit": "41,595元/平",
   "region": "haidian",
   "id": 31
},
   "name": "中监所西区 南北通透三居带客厅 楼龄新",
   "price": "1314万",
   "size": "95.54平米",
   "perunit": "137,535元/平",
   "region": "haidian",
   "id": 32
},
   "name": "房子是东西通透的2居室,带电梯,小区管理好",
   "price": "585万",
   "size": "62平米",
   "perunit": "94,355元/平",
   "region": "haidian",
   "id": 33
},
   "name": "翠微南里1层南北2室2厅2卫 适合老年人诚心卖多14平",
   "price": "580万",
   "size": "54.29平米",
   "perunit": "106,834元/平",
   "region": "haidian",
   "id": 34
},
   "name": "水科院 西南两居 带电梯 满五唯一 看房方便",
   "price": "850万",
   "size": "73.2平米",
   "perunit": "116,121元/平",
   "region": "haidian",
   "id": 45
},
   "name": "恩济里1层大两居室 大客厅 看房随时",
   "price": "585万",
   "size": "63.4平米",
   "perunit": "92,272元/平",
   "region": "haidian",
   "id": 46
},
   "name": "昌运宫三环内带电梯,闹中取静,",
```

```
"price": "680万",
   "size": "72.6平米",
    "perunit": "93,664元/平",
   "region": "haidian",
   "id": 47
},
   "name": "南北通透双阳台,业主诚意出售,小区停车方便,视野好",
    "price": "698万",
   "size": "77.7平米",
   "perunit": "89,833元/平",
   "region": "haidian",
   "id": 48
},
   "name": "学知园 三居两卫 精装修 全明格局 满五唯一 有电梯",
   "price": "938万",
   "size": "111.9平米",
   "perunit": "83,825元/平",
   "region": "haidian",
   "id": 49
},
   "name": "复兴路77号院 2室1厅 南",
   "price": "468万",
   "size": "52.5平米",
   "perunit": "89,143元/平",
   "region": "haidian",
   "id": 60
},
   "name": "飞达社区 南北 双卧朝南 厅朝东 明卫 精装 中间楼层",
   "price": "1320万",
   "size": "104.5平米",
   "perunit": "126,316元/平",
   "region": "haidian",
   "id": 61
},
   "name": "精装修两居室 采光充足 满五唯一 随时可签约",
   "price": "552万",
   "size": "55.4平米",
   "perunit": "99,639元/平",
   "region": "haidian",
   "id": 62
},
   "name": "电梯房 两居 中间层 采光好 满五唯一",
   "price": "690万",
   "size": "76.78平米",
   "perunit": "89,868元/平",
   "region": "haidian",
   "id": 63
},
   "name": "金沟河2号院 海淀五棵松 正规三居 南北通透购房成本低",
   "price": "475万",
   "size": "56.4平米",
```

```
"perunit": "84,220元/平",
   "region": "haidian",
   "id": 64
},
{
   "name": "玉海园三里新上两居室+精装修+南北通透+满五年",
   "price": "593万",
   "size": "72.65平米",
   "perunit": "81,625元/平",
   "region": "haidian",
   "id": 72
},
   "name": "梅园甲低楼层,满五唯一,性质好。",
   "price": "820万",
   "size": "105.32平米",
   "perunit": "77,858元/平",
   "region": "haidian",
   "id": 73
},
   "name": "满庭芳园东南向三居室 经典户型 有钥匙看房方便",
   "price": "1420万",
   "size": "132.71平米",
   "perunit": "107,001元/平",
   "region": "haidian",
   "id": 74
},
   "name": "北三环 蓟门桥 大钟寺 带电梯 两居室",
   "price": "570万",
   "size": "61.4平米",
   "perunit": "92,834元/平",
   "region": "haidian",
   "id": 75
},
   "name": "门头馨村北二区两居 南北朝向 二层 稳定出售",
    "price": "520万",
   "size": "92.89平米",
   "perunit": "55,981元/平",
   "region": "haidian",
   "id": 86
},
   "name": "北太平庄路2号院 2室1厅 南 北",
   "price": "930万",
   "size": "97.65平米",
   "perunit": "95,239元/平",
   "region": "haidian",
   "id": 87
},
   "name": "金雅园南区南北三居室一梯两户居家装修有车位",
   "price": "1360万",
   "size": "141.85平米",
   "perunit": "95,876元/平",
   "region": "haidian",
```

```
"id": 88
   },
   {
       "name": "西直门明光村小区东南向三居双卧朝南采光好安静不临街",
       "price": "830万",
       "size": "79.5平米",
       "perunit": "104,403元/平",
       "region": "haidian",
       "id": 89
   },
       "name": "铁东精装两居室 中间楼层 配套齐全",
       "price": "510万",
       "size": "51.1平米",
       "perunit": "99,805元/平",
       "region": "haidian",
       "id": 90
   },
   {
       "name": "强Y·府学上院南北通透三居室满五唯一",
       "price": "1165万",
       "size": "139.58平米",
       "perunit": "83,465元/平",
       "region": "haidian",
       "id": 99
   },
       "name": "燕北园 中间层 不临街 南北向 三居",
       "price": "820万",
       "size": "81.6平米",
       "perunit": "100,491元/平",
       "region": "haidian",
       "id": 100
   },
       "name": "西三环 厂洼 全明格局 南北通透精装 近地铁",
       "price": "705万",
       "size": "67.4平米",
       "perunit": "104,600元/平",
       "region": "haidian",
       "id": 101
   },
       "name": "索家坟、南北通透、精装修、诚心出售、随时配合签约",
       "price": "550万",
       "size": "54.2平米",
       "perunit": "101,477元/平",
       "region": "haidian",
       "id": 102
   }
],
   {
       "region": "朝阳"
   },
       "name": "华腾园南向三居,正看花园,精装修,有固定车位",
       "price": "985万",
```

```
"size": "126.14平米",
   "perunit": "78,088元/平",
   "region": "chaoyang",
   "id": 65
},
   "name": "晨光B区 南北三居 满五唯一 诚意出售 随时签约",
   "price": "880万",
   "size": "133.96平米",
   "perunit": "65,692元/平",
   "region": "chaoyang",
   "id": 76
},
   "name": "嘉美风尚跃层复式 东南向 高层 无遮挡落地窗 随时可看",
   "price": "1320万",
   "size": "193.17平米"
   "perunit": "68,334元/平",
   "region": "chaoyang",
   "id": 77
},
   "name": "户型方正,三个方向采光,高层无遮挡",
   "price": "440万",
   "size": "109.77平米",
   "perunit": "40,084元/平",
   "region": "chaoyang",
   "id": 78
},
   "name": "苹果派 南北通透三居 满五唯一 诚心卖 无遮挡",
   "price": "699万",
   "size": "112.77平米",
   "perunit": "61,985元/平",
   "region": "chaoyang",
   "id": 79
},
   "name": "满五年唯一,楼层好,采光好,无遮挡,户型方正大厨房",
   "price": "429万",
   "size": "66.64平米",
   "perunit": "64,376元/平",
   "region": "chaoyang",
   "id": 80
},
   "name": "10号院低楼层三居室 南北通透 无遮挡",
   "price": "400万",
   "size": "68.86平米",
   "perunit": "58,089元/平",
   "region": "chaoyang",
   "id": 91
},
   "name": "农光里小区 精装修 户型方正 近地铁",
   "price": "430万",
   "size": "66.78平米",
   "perunit": "64,391元/平",
```

```
"region": "chaoyang",
   "id": 92
},
{
   "name": "亚运花园 满五唯一 东南大两居 双卫",
   "price": "782万",
   "size": "113.66平米",
   "perunit": "68,802元/平",
   "region": "chaoyang",
   "id": 93
},
   "name": "满五年唯一,中间层,通透两居室,户型好",
   "price": "455万",
   "size": "60.85平米",
   "perunit": "74,775元/平",
   "region": "chaoyang",
   "id": 94
},
   "name": "将府家园 2004年楼龄 不临街 看房方便",
   "price": "725万",
   "size": "125.32平米",
   "perunit": "57,852元/平",
   "region": "chaoyang",
   "id": 95
},
   "name": "6号线 十里堡 南北通透两居 交通便利 看房方便",
   "price": "308万",
   "size": "58.26平米",
   "perunit": "52,867元/平",
   "region": "chaoyang",
   "id": 96
},
   "name": "瑞和国际 东向三居室 两梯六户 封闭管理",
   "price": "760万",
   "size": "143.84平米",
   "perunit": "52,837元/平",
   "region": "chaoyang",
   "id": 103
},
   "name": "京通苑南向三居 高层无遮挡 看房方便",
    "price": "750万",
   "size": "195.78平米",
   "perunit": "38,309元/平",
   "region": "chaoyang",
   "id": 104
},
   "name": "远洋万和城A区 全新房未住 刷卡入户 只有契税 诚售",
   "price": "3350万",
   "size": "286.8平米",
   "perunit": "116,807元/平",
   "region": "chaoyang",
   "id": 105
```

```
},
{
   "name": "华瀚国际 3室2厅 南 北",
   "price": "1400万",
   "size": "164.98平米",
   "perunit": "84,859元/平",
   "region": "chaoyang",
   "id": 106
},
   "name": "安华里社区 精装修 南北田字格三居室 满五唯一",
   "price": "730万",
   "size": "85.48平米",
   "perunit": "85,401元/平",
   "region": "chaoyang",
   "id": 107
},
   "name": "劲松七区 三居 光明桥 龙潭湖 南北通透",
   "price": "360万",
   "size": "60.32平米",
   "perunit": "59,682元/平",
   "region": "chaoyang",
   "id": 108
},
   "name": "富力又一城C区 南北通透三居室 中间楼层 前后观景",
   "price": "840万",
   "size": "155.14平米",
   "perunit": "54,145元/平",
   "region": "chaoyang",
   "id": 109
},
   "name": "西坝河北里 3室1厅 东南",
   "price": "1065万",
   "size": "142.84平米",
    "perunit": "74,559元/平",
   "region": "chaoyang",
   "id": 110
},
   "name": "华威北里满五唯一电梯房卧室朝南",
    "price": "299万",
    "size": "45.98平米",
    "perunit": "65,029元/平",
    "region": "chaoyang",
   "id": 111
},
   "name": "店长推荐 高楼层西南向的两居 景观萧太后河 采光好",
    "price": "370万",
    "size": "78.32平米",
   "perunit": "47,243元/平",
   "region": "chaoyang",
   "id": 112
},
```

```
"name": "朝阳区,立水桥,商办类,全款购买,可注册公司",
   "price": "280万",
   "size": "84.62平米",
   "perunit": "33,090元/平",
   "region": "chaoyang",
   "id": 113
},
   "name": "福怡苑南北通透正规三居室,两个卧室朝南。",
   "price": "420万",
   "size": "83.9平米",
   "perunit": "50,060元/平",
   "region": "chaoyang",
   "id": 114
},
   "name": "南沙滩小区,高层视野开阔,配套齐全,交通便利。",
   "price": "436万",
   "size": "52.9平米",
   "perunit": "82,420元/平",
   "region": "chaoyang",
   "id": 115
},
   "name": "万豪国际公寓 北向 正规一居室",
   "price": "475万",
   "size": "69.22平米",
   "perunit": "68,622元/平",
   "region": "chaoyang",
   "id": 116
},
   "name": "华美橡树岭 南北两居 满五年唯一 有钥匙",
   "price": "399万",
   "size": "91.52平米",
   "perunit": "43,598元/平",
   "region": "chaoyang",
   "id": 117
},
   "name": "潘家园 铁道宿舍 西南向明厨明卫两居室 诚意出售",
   "price": "430万",
   "size": "64.11平米",
   "perunit": "67,073元/平",
   "region": "chaoyang",
   "id": 118
},
   "name": "东一时区满五年唯一,全明格局2居有地铁看房方便",
   "price": "449万",
   "size": "85.11平米",
   "perunit": "52,756元/平",
   "region": "chaoyang",
   "id": 119
},
   "name": "和平街十三区高层西南两居室,满5年唯一随时签约",
   "price": "668万",
```

## 参考

https://github.com/clemfromspace/scrapy-selenium