

Python3使用selenium库爬取安居客

前言

某些网站为了反爬虫，可能会采取多种方式。比如，使用ajax技术异步加载数据、检测HTTP协议头信息等。当然，我们可以通过制作头信息，并提前找到实际数据的请求的URL路径。但这样就会存在许多问题，如：程序中会包含许多与程序本身逻辑的头信息，并且如果当每次需要寻找的数据的真实URL路径不同时，可能还需要再进行头信息的构造。这样，就会使得程序本身被大部分的头信息占据。所以，为了解决这种问题，并且使得反爬取更加容易，selenium库出现了。它能够自动化操作浏览器进行点击、输入等人类一切可以在网页上完成的操作。由于selenium本身就是模拟人类去利用浏览器访问，所以使用selenium爬取数据时，就完全不用担心爬虫没有添加协议头信息以及实际请求数据与请求页面不相同的问题。本文基于Chrome浏览器、Python3、pycharm、Chrome浏览器以及selenium，爬取了安居客在成都锦江区，价格在150-200万的部分二手房的信息。

安装selenium库以及Chrome driver

1. Chrome driver是用于selenium库自动化操作浏览器。不同浏览器，需要下载对应浏览器的driver。如Firefox需要下载Firefox driver。
2. selenium库安装

```
pip3 install selenium
```

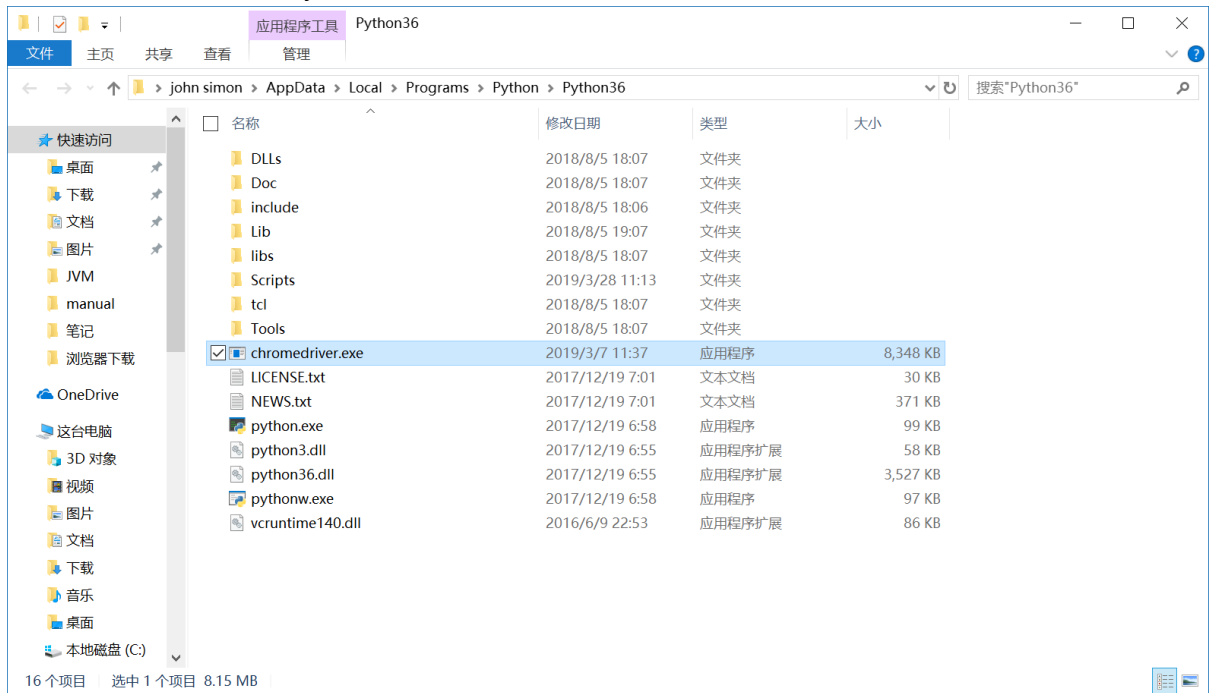
3. Chrome driver安装

1. 首先下载Chrome driver的解压包,下载地址: [点这里](#) 注意: chromedriver的版本要与你使用的chrome版本对应, 对应关系如下:

chromedriver版本	支持的Chrome版本
v2.33	v60-62
v2.32	v59-61
v2.31	v58-60
v2.30	v58-60
v2.29	v56-58
v2.28	v55-57
v2.27	v54-56
v2.26	v53-55
v2.25	v53-55
v2.24	v52-54
v2.23	v51-53

chromedriver版本	支持的Chrome版本
v2.22	v49-52
v2.21	v46-50
v2.20	v43-48
v2.19	v43-47
v2.18	v43-46
v2.17	v42-43
v2.13	v42-45
v2.15	v40-43
v2.14	v39-42
v2.13	v38-41
v2.12	v36-40
v2.11	v36-40
v2.10	v33-36
v2.9	v31-34
v2.8	v30-33
v2.7	v30-33
v2.6	v29-32
v2.5	v29-32
v2.4	v29-32

2. 将driver的exe文件放到Python的安装目录下，即安装成功



关于seleniumAPI简单总结

selenium归根结底还是基于driver对网页进行爬取。所以，首先要做的就是获取到对应浏览器的driver对象。之后，在利用获取到的driver对象，进行元素选择、信息输入、表单提交、frame移动等复杂操作。本例只使用到了元素选择方面API，对于其他方面，在此不做讲解。毕竟，API还是要通过自己的阅读和使用才能熟练掌握。

1. 获取driver 使用Chrome浏览器就使用Chrome()方法，对应的，如果使用Firefox，就使用Firefox()
2. 元素选择 selenium的元素选择API类似于js的元素选择函数。如果读者对于利用js对DOM树中元素进行操作比较熟练的话，那么应该很容易理解这方面的API。只不过，selenium的元素选择API在选择的方式进行更多的扩展。如，可以使用css样式、xpath等方式进行选择。而选择器的写法，也与css选择器、xpath选择语法完全相同。读者完全可以放心使用。
3. 其他方面 既然是想要使用selenium库进行网页爬取的开发者，那么自然自身对于前端还是比较了解。比如类似表单如何完成，如何提交等。其实，另外这些方面的API也就是将用户实际的网页操作封装成了方法提供出来。所以，完全可以做到见文知义。这也是我没有再赘述的原因之一。如果是不清楚selenium库中有哪些API，有两种解决方法可供选择。一种是通过网上查找相关博客，有很多这方面的总结。另一种则是查看源码，直接查看库中API。

使用pycharm进行代码编写

```
import csv
import re

from selenium import webdriver

browser = webdriver.Chrome()
browser.get('https://chengdu.anjuke.com/sale/jinjiang/m58/')

house_links = browser.find_elements_by_css_selector('#houselist-mod-new a')
# 预存各个链接网址
hrefs = []
```

```
for link in house_links:
    hrefs.append(link.get_attribute('href'))

row_label = ['introduction', 'tags', 'saler']
data = []
for i in range(0, len(hrefs)):
    browser.get(hrefs[i])
    row = []

    if i == 0:
        labels = browser.find_elements_by_class_name('houseInfo-label')
        # 添加房屋细节label
        for label in labels:
            row_label.append(label.text)

    # 添加介绍列数据
    long_title = browser.find_element_by_class_name('long-title').text
    row.append(long_title)

    # 添加标签列数据
    tags = browser.find_elements_by_class_name('info-tag')
    tag_text = ''
    for tag in tags:
        # 去除HTML标签, 合并tag
        tag_text += re.sub('<.*?>', '', tag.text)
        # 两个空白字符分隔不同tag
        tag_text += ' '

    row.append(tag_text)

    # 添加售卖者列数据
    owner = browser.find_element_by_class_name('brokercard-name').text
    row.append(owner)

    # 添加房屋所有详细列数据
    contents = browser.find_elements_by_class_name('houseInfo-content')
    for content in contents:
        row.append(content.text)

    data.append(row)
    print('第' + str(i+1) + '条房屋信息爬取完成')

print('开始写入文件')
# utf-8-sig避免中文乱码
with open('anjuke.csv', 'w', encoding='utf-8-sig', newline='') as csvfile:
    writer = csv.writer(csvfile)
    writer.writerow(row_label)
    writer.writerows(data)

print('文件写入完成')
browser.close()
```

实际爬取测试

1. 爬取效果：在程序中，没有采用多线程进行爬取，所以爬取网页速度较慢。

【多图】甩卖！！精装标准套三带书房，正读盐小，随时看房，急售

https://chengdu.anjuke.com/prop/view/A1666582857?from=filter&spread=filtersearch&invalid=1®ion_ids=904&position=1&kwtype=filter&now_time=1556079434

Chrome 正受到自动测试软件的控制。

安居客 成都 · 首页 新房 二手房 租房 商铺写字楼 海外地产 楼讯 房价 问答

登录 注册 下载APP

成都房产网 > 成都二手房 > 锦江二手房 > 三圣乡二手房 > 锦江城市花园三期

甩卖！！精装标准套三带书房，正读盐小，随时看房，急售

150万 | 3室2厅 | 85平方米 | 房贷计算

安选 真实在售 假赔百元

下载app早报 | 安选假赔百元细则


室内图(9)

户型图(1)

环境图片

环境图(10)

周边地图



室内图片

杨海

查看TA的店铺

等级: ★★★★★

得分: 房源: 1.9 服务: 4.0 评价: 5.0

勋章:

公司: 买房无忧

门店: 买房无忧三圣乡店

公司执照编码 91510100350576595E

点击查看电话

房屋信息

房屋编码: 928836060404737, 发布时间: 2019年04月24日

所属小区: 锦江城市花园三期

房屋户型: 3室 2厅 1卫

房屋单价: 17647 元/m²

所在位置: 锦江 - 三圣乡 - 喜树街618号

建筑面积: 85平方米

参考首付: 45.00万

建造年代: 2012年

房屋朝向: 南

参考月供: 8834元

微信扫一扫，在线聊

【多图】双桥子家乐福旁边蜀都花园18楼套三双卫107.19平170万

https://chengdu.anjuke.com/prop/view/A1664525949?from=filter-saleMetro-salesqx&spread=filtersearch&invalid=1®ion_ids=904&position=3&kwtype=filter&now_time=1556079434

Chrome 正受到自动测试软件的控制。

安居客 成都 · 首页 新房 二手房 租房 商铺写字楼 海外地产 楼讯 房价 问答

登录 注册 下载APP

成都房产网 > 成都二手房 > 锦江二手房 > 牛市口二手房 > 蜀都花园

双桥子家乐福旁边蜀都花园18楼套三双卫107.19平170万

170万 | 3室2厅 | 107.2平方米 | 房贷计算

安选 真实在售 假赔百元

下载app早报 | 安选假赔百元细则


室内图(11)

户型图(1)

环境图片

环境图(10)

周边地图



室内图片

孙玲

查看TA的店铺

等级: ★★★

得分: 房源: 2.1 服务: 4.1 评价: 4.3

勋章:

公司: 成都优优好房

门店: 好房YOU双桥店

公司执照编码 91510105094650476Q

点击查看电话

房屋信息

房屋编码: 926386770534400, 发布时间: 2019年04月22日

所属小区: 蜀都花园

房屋户型: 3室 2厅 2卫

房屋单价: 15860 元/m²

所在位置: 锦江 - 牛市口 - 水碾河路14号

建筑面积: 107.2平方米

参考首付: 51.00万

建造年代: 2012年

房屋朝向: 南北

参考月供: 10012元

微信扫一扫，在线聊

2. 数据展示（部分）：

introductio	tags	saler	所属小区	房屋户型	房屋单价	所在位置	建筑面积	参考首付	建造年代	房屋朝向	参考月供	房屋类型	所在楼层	装修程度	产权年限	配套电梯	房本年限	产权性质	唯一住房
甩卖! ! 有	150万	3室杨海	锦江城市7	3室 2厅 1	17647 元/	锦江 - 三圣乡 - 喜树街618号	85平方米	45.00万	2012年	南	8834元	普通住宅	高层(共34)	精装修	70年	有	满五年	商品房	否
0.5个点费	150万	2室陈保利	卓锦城6期	2室 1厅 1	19711 元/	锦江 - 三圣乡 - 国香街780号	76.1平方	45万	2014年	东南	8834元	普通住宅	低层(共30)	简单装修	70年	有	不满二年	商品房	否
双桥子家	170万	3室孙玲	蜀都花园	3室 2厅 2	15860 元/	锦江 - 牛市口 - 水碾河路14号	107.2平方	51.00万	2003年	南北	10012元	普通住宅	高层(共18)	精装修	70年	有	满五年	商品房	否
6000佣金)	175万	2室龙丽	卓锦城3期	2室 2厅 2	18122 元/	锦江 - 三圣乡 - 国香街333号	96.6平方	52.50万	2010年	南	10307元	普通住宅	中层(共24)	精装修	70年	有	满五年	商品房	是
亚洲伊藤	150万	3室粟艳	东洪广厦	3室 2厅 1	17241 元/	锦江 - 三圣乡 - 樱花街383号	87平方米	45.00万	2012年	西	8834元	普通住宅	中层(共32)	精装修	70年	有	满二年	商品房	是
绿地468	185万	3室李伟	绿地468公	3室 2厅 1	21023 元/	锦江 - 三圣乡 - 芙蓉西路707号	88平方米	56万	2015年	西南	10896元	普通住宅	低层(共25)	毛坯	70年	有	满五年	商品房	否
甩卖! ! 有	150万	3室杨海	锦江城市7	3室 2厅 1	17647 元/	锦江 - 三圣乡 - 喜树街618号	85平方米	45.00万	2012年	南	8834元	普通住宅	高层(共34)	精装修	70年	有	满五年	商品房	否
						锦江 - 沙													