**实习七 网页数据采集**

1. **实习目的**

练习网页数据采集，掌握利用Selenium包采集动态网页数据。

1. **实习内容**

在链家网上采集吴泾地区的二手房数据，包括位置信息、房屋信息和价格信息等数据。网页如图1所示，网址为：https://sh.lianjia.com/ershoufang/rs吴泾/，网页上包含每个二手房的位置信息（如图中的双柏小区-吴泾）、房屋信息（如图中的2室1厅|67.28平米|……）和价格信息（如图中的285万，单价42361元/平米）。采集的数据存储到Excel文件（xlsx）中（图2）。

****

图1

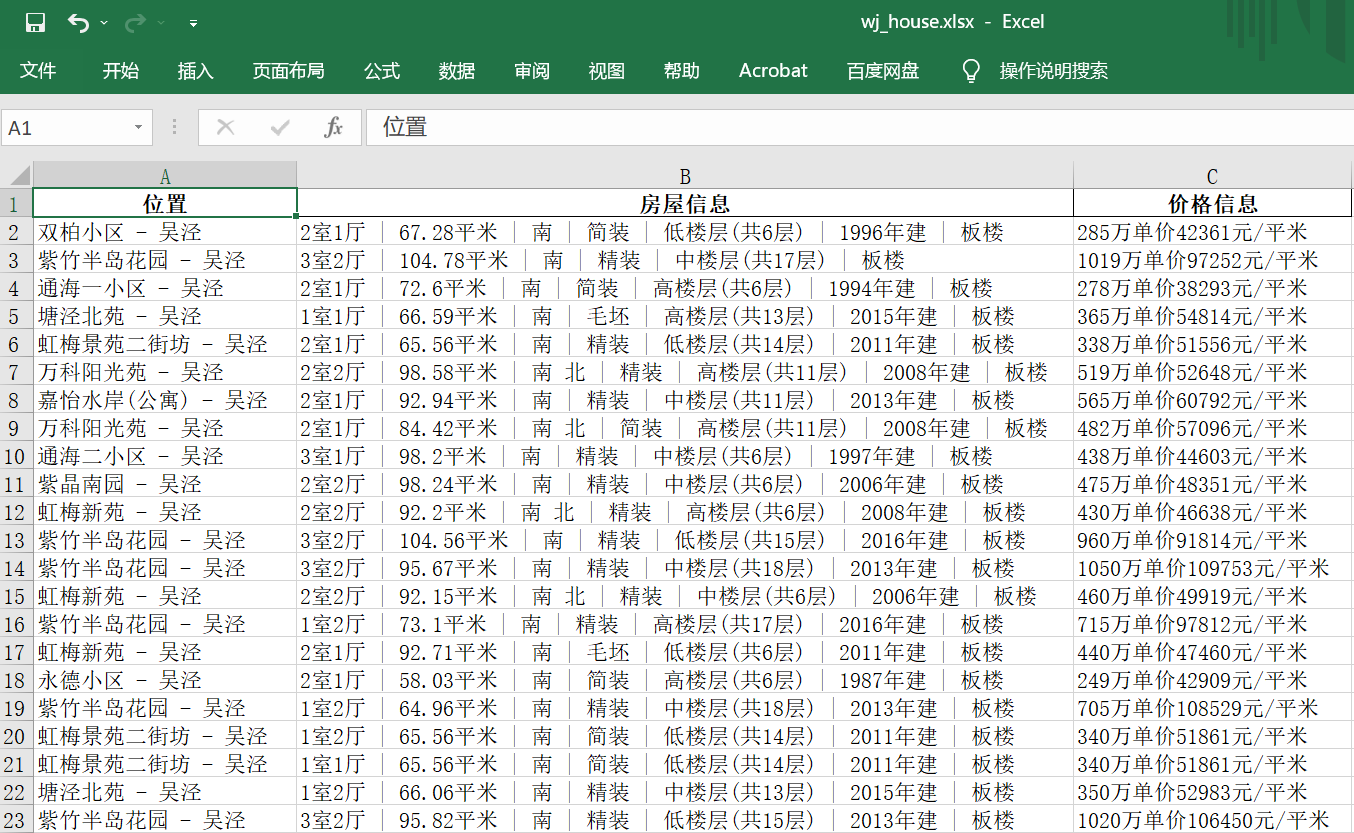


图2

1. **注意点**

1）查找到的二手房数据会分布在多个网页中，如图3所示，查找到的数据分布在7个页面，需要通过模拟点击元素（图中的1、2、3…7）操作返回每个页面。



图3

2）由于页面的返回会有一个延迟，另外有些网站对非正常速度的网页浏览会拒绝访问，因此在采集每个页面的数据时，之间应有一个时间停顿，可利用time模块中的sleep()函数，如time.sleep(2)表示停顿2秒。

3）为提高采集速度，在实例化webdriver对象时，可以设置在加载网页时不加载CSS、图片以及JavaScript代码，以提高网页加载速度。测试显示设置和没有设置的采集时间分别为1分7秒和2分10秒。设置不加载CSS、图片以及JavaScript代码的参考代码如下：

fp = webdriver.FirefoxProfile()

# 限制CSS加载

fp.set\_preference("permissions.default.stylesheet", 2)

# 限制图片加载

fp.set\_preference('permissions.default.image', 2)

# 限制JavaScript执行

fp.set\_preference('javascript.enabled', False)

driver = webdriver.Firefox(executable\_path = r'c:\Anaconda3\Scripts\geckodriver.exe',

firefox\_profile=fp)

4）写文件使用相对路径（文件名前不加路径）。

1. **作业提交**

提交ipynb文件（放在之前作业的同个ipynb文件中）。