scapy学习笔记

Ringo Yungpo Kao

December 12, 2017

# scrapy学习笔记

## 学习环境

|  |  |
| --- | --- |
| 类别 | 环境 |
| 系统 | Windows 10 64bit |
| Shell | PowerShell |

## 环境安装

### Python安装

1. 到官网的[下载链接](https://www.python.org/ftp/python/3.5.3/python-3.5.3-amd64.exe)下载python3.5.3进行python的安装，**勾选Add Python3.5 to PATH**，避免手动添加，然后一路下一步。
2. 在PowerShell中输入python获取版本号无误即安装成功。

PS C:\Users\Ringo> python  
Python 3.5.3 (v3.5.3:1880cb95a742, Jan 16 2017, 16:02:32) [MSC v.1900 64 bit (AMD64)] on win32  
Type "help", "copyright", "credits" or "license" for more information.

### Scrapy安装

1. 在windows下安装Scrapy前必须要安装[pywin32](https://nchc.dl.sourceforge.net/project/pywin32/pywin32/Build%20220/pywin32-220.win-amd64-py3.5.exe)。
2. 下载[Lxml](https://download.lfd.uci.edu/pythonlibs/yjwkc9i2/lxml-4.1.1-cp35-cp35m-win_amd64.whl)，并在相应目录下在PowerShell输入pip install .-4.1.1-cp35-cp35m-win\_amd64.whl
3. 输入pip install Scrapy安装Scrapy。

## 初探Scrapy爬虫数据抓取

### Scrapy命令式交互模式

本次现在命令行下面进行对于数据的抓取 抓取的目标网站是[赶集网]

#### 获得任一个价格的xpath

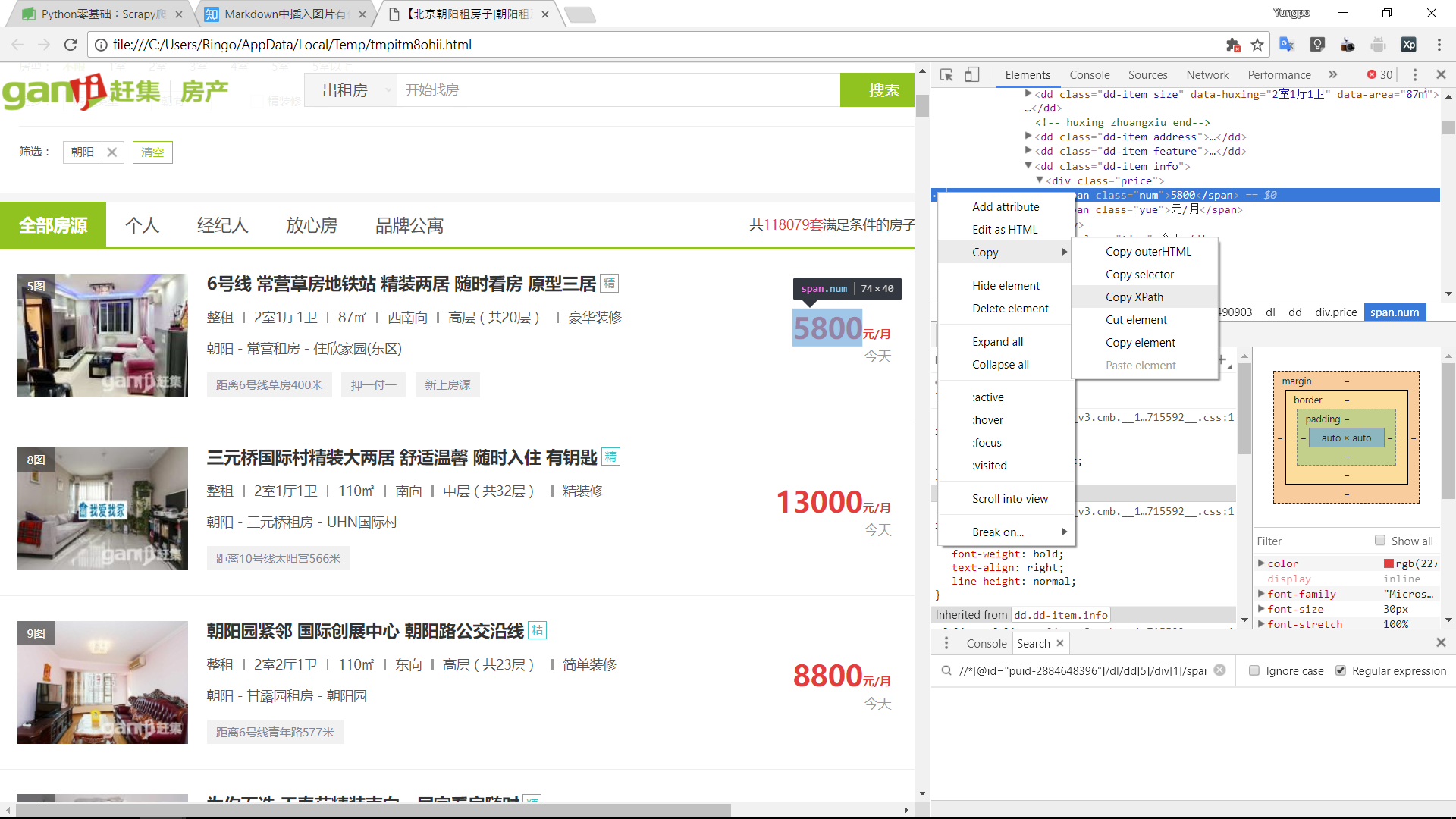
(http://bj.ganji.com/fang1/chaoyang/) 在PowerShell下输入

scrapy shell http://bj.ganji.com/fang1/chaoyang/

打开链接获取变量的数据保存的变量是一个response 通过response查看获取数据为200以及相应的地址

>>> response  
<200 http://bj.ganji.com/fang1/chaoyang/>

通过view(response)查看获取的地址 

按F12进入控制台获取价格的xpath  获得xpath为 > //\*[@id="puid-2892490903"]/dl/dd[5]/div[1]/span[1]

通过以下两个语句获取相应的标签以及文本

>>> response.xpath('//\*[@id="puid-2892490903"]/dl/dd[5]/div[1]/span[1]').extract()  
['<span class="num">5800</span>']  
>>> response.xpath('//\*[@id="puid-2892490903"]/dl/dd[5]/div[1]/span[1]/text()').extract()  
['5800']

**验证成功**

#### 抓取所有价格的xpath

通过修改单个价格xpath获取所有的价格文本 其中单个价格的xpath为 > //\*[@id="puid-2892490903"]/dl/dd[5]/div[1]/span[1]/text()

将其修改为 > //div[@class="f-list-item ershoufang-list"]/dl/dd [5]/div[1]/span[1]/text()

由于每个租房信息的标签为div 其中且通过id查找仅能查找到单个租房信息，所以通过class进行查找。

>>> response.xpath('//div[@class="f-list-item ershoufang-list"]/dl/dd[5]/div[1]/span[1]/text()').extract()  
['5800', '13000', '8800', '4800', '22000', '7200', '7000', '12000', '6000', '7000', '11000', '5500', '5700', '6000', '63  
00', '22000', '8500', '5300', '5500', '8500', '6000', '6500', '6090', '5300', '5000', '9200', '2900', '6800', '1000', '9  
000', '6700', '12800', '9200', '900', '5500', '4900', '5000', '5750', '6500', '4500', '7800', '8500', '8500', '8000', '5  
200', '6500', '8700', '6100', '2350', '8500', '5200', '6500', '40000', '7200', '7000', '9600', '5600', '6800', '1300', '  
5900', '3000', '6700', '5700', '6500', '4600', '5500', '11000', '5500', '8800', '3800', '4600', '5500', '1890', '2500',  
'4800', '7500', '6500', '2680', '3300', '12000', '1500', '1740', '5500', '1700', '2750', '1780', '8500', '6700', '12000'  
, '3000', '3400', '9000', '6800', '7600', '7300', '7300', '700', '7999', '3100', '9000', '4900', '2950', '1300', '2550',  
 '1750', '2750', '2250', '3500', '900', '1900', '2300', '5500', '14000', '7500', '9000', '899', '4500', '4200', '9500',  
'2200']  
>>> len(response.xpath('//div[@class="f-list-item ershoufang-list"]/dl/dd[5]/div[1]/span[1]/text()').extract())  
120

在命令行捕捉到了所有的价格信息，同理也可以捕捉到所有标题信息、地址信息、详情信息等 其中标题信息集合的xpath为 > //div[@class="f-list-item ershoufang-list"]/dl/dd[1]/a/text()

### Scrapy爬虫数据抓取

#### 项目初始化

执行建项命令

Scrapy startproject zufang

建立项目后用Pycharm打开项目看到目录结构

.  
+-- zufang  
| +-- spiders  
| +-- \_\_init\_\_.py  
| +-- \_\_init\_\_.py  
| +-- items.py  
| +-- middlewares.py  
| +-- pipelines.py  
| +-- setting.py  
+-- scrapy.cfg

在spiders目录下添加ganji.py爬虫文件

ganji.py

import scrapy  
   
class GanjiSpider(scrapy.Spider):  
 name = "zufang"  
 start\_urls = ['http://bj.ganji.com/fang1/chaoyang/']  
   
 def parse(self, response):  
 print(response)  
 title\_list = response.xpath('//div[@class="f-list-item ershoufang-list"]/dl/dd[1]/a/text()').extract()  
 money\_list = response.xpath('//div[@class="f-list-item ershoufang-list"]/dl/dd[5]/div[1]/span[1]/text()').extract()  
   
 for i,j in zip(title\_list, money\_list):  
 print(i,":",j)

运行

scrapy crawl zufang  
   
2017-12-17 22:33:16 [scrapy.utils.log] INFO: Scrapy 1.4.0 started (bot: zufang)  
2017-12-17 22:33:16 [scrapy.utils.log] INFO: Overridden settings: {'SPIDER\_MODULES': ['zufang.spiders'], 'ROBOTSTXT\_OBEY': True, 'BOT\_NAME': 'zufang', 'NEWSPIDER\_MODULE': 'zufang.spiders'}  
...  
2017-12-17 22:33:17 [scrapy.core.engine] INFO: Spider opened  
2017-12-17 22:33:17 [scrapy.extensions.logstats] INFO: Crawled 0 pages (at 0 pages/min), scraped 0 items (at 0 items/min)  
2017-12-17 22:33:17 [scrapy.extensions.telnet] DEBUG: Telnet console listening on 127.0.0.1:6024  
2017-12-17 22:33:17 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://bj.ganji.com/robots.txt> (referer: None)  
2017-12-17 22:33:18 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://bj.ganji.com/fang1/chaoyang/> (referer: None)  
<200 http://bj.ganji.com/fang1/chaoyang/>  
安贞安华西里精装二居 和业主签合同 能长租 南北通透卧室朝南 : 6300  
地铁5.10号线，朝南一居室出租，采光好，有客厅，家电齐全 : 4500  
牡丹园 华严北里小区 精装三居室出租 寻找居家客户紧邻10号 : 7200  
凯德锦绣租售中心东四环豪华社区精装两居室出租 : 11000  
为你而选 尚东阁精装开间随时看房入住集中供暖近地铁电梯房 : 3800  
双桥地铁站东北方向 朝阳路周家井公交车站下车就到 金真子后面 : 2400  
为你而选 风格派 1室0厅 51平 : 5600  
...  
月底活动周+来电立减现金+房源有限+无+我们是认真的 : 2700  
无附加费 地铁6号线青年路站 精致装修 押一付一 : 1500  
房租差个两三百+以后多付两三千+不是个例+住点好的蛋壳 : 1550  
泰福苑近地铁温馨舒适 好住不贵 比公寓更便宜 欢迎来电 : 4100  
酒仙桥卡布其诺南北通透三居室精装婚房看房方便配套齐全 : 11000  
卡夫卡精装大一居 : 5500  
温馨港湾, : 900  
2017-12-17 22:33:18 [scrapy.core.engine] INFO: Closing spider (finished)  
2017-12-17 22:33:18 [scrapy.statscollectors] INFO: Dumping Scrapy stats:  
{'downloader/request\_bytes': 445,  
 'downloader/request\_count': 2,  
 'downloader/request\_method\_count/GET': 2,  
...  
 'scheduler/enqueued/memory': 1,  
 'start\_time': datetime.datetime(2017, 12, 17, 14, 33, 17, 115954)}  
2017-12-17 22:33:18 [scrapy.core.engine] INFO: Spider closed (finished)