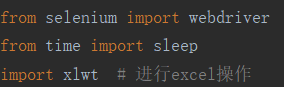
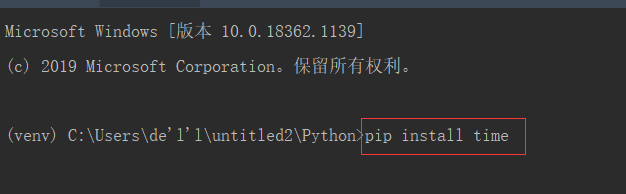
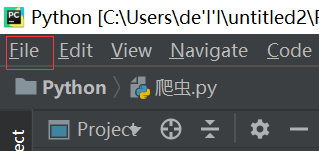
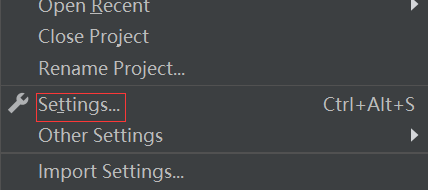
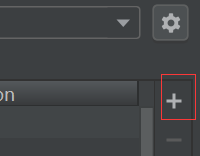
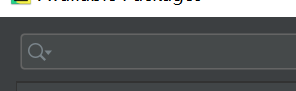
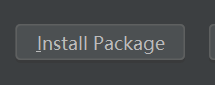
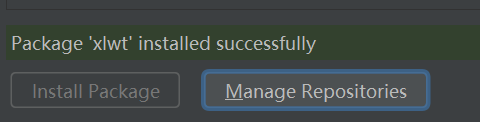
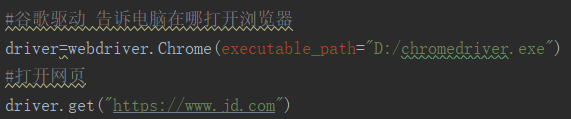
1、http://www.cae.cn/cae/html/main/col48/column\_48\_1.html



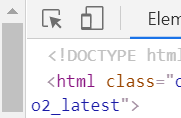
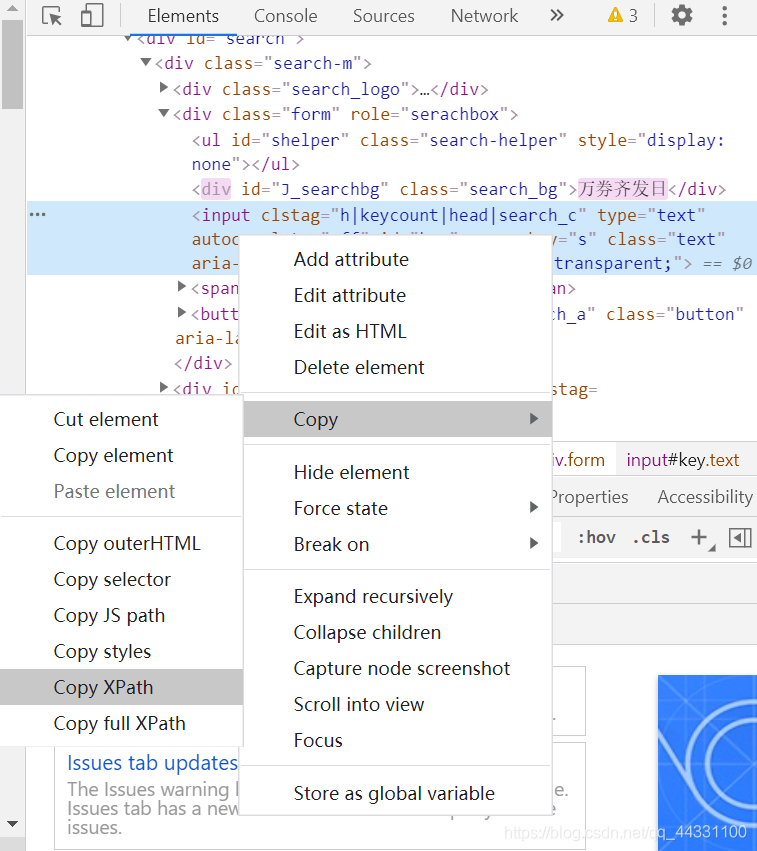
## 2、使用Python爬取京东华为手机前十页的所有商品的链接、名称、价格、评价数以及店铺名称。参考：<https://blog.csdn.net/qq_44331100/article/details/109392160>

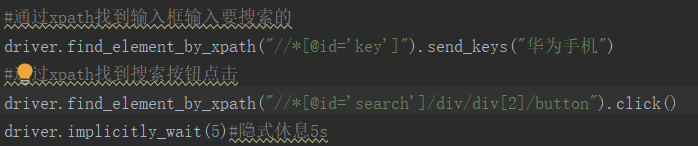
**1.前期准备**  
（1）下载驱动  
使用的是谷歌浏览器，所以要下载谷歌驱动，用来告诉电脑在哪打开浏览器。  
驱动文件下载地址：http://chromedriver.storage.googleapis.com/index.html  
（需要下载与自己的浏览器版本相一致或者相近版本的驱动，要记住驱动下载之后存放的位置，解压）  
（2）安装第三方库  
selenium,time,xlwt都是爬虫需要用到的库  
  
安装方法1：  
在pycharm最下面有一行选项，如下图，选择Terminal  
在这里插入图片描述  
选择完之后，在出来的页面输入pip install +你要安装的库，之后等待它自动安装就可以了  
例如安装time库，time不用安装。  
  
安装方法2：  
  
  
  
  
在搜索栏中搜索你要安装的库，找到之后点击左下角的Install Package  
  
之后等待安装，若成功会显示  
  
**2.开始编程**  
（1）在谷歌中打开京东  
  
executable\_path就是你下载的驱动的存放位置

（2）在京东中搜索“华为手机”

我们要电脑模拟人去在搜索框中输入“华为手机”，就要使用输入框的xpath定位到输入框的位置，那这么获取元素的xpath呢？

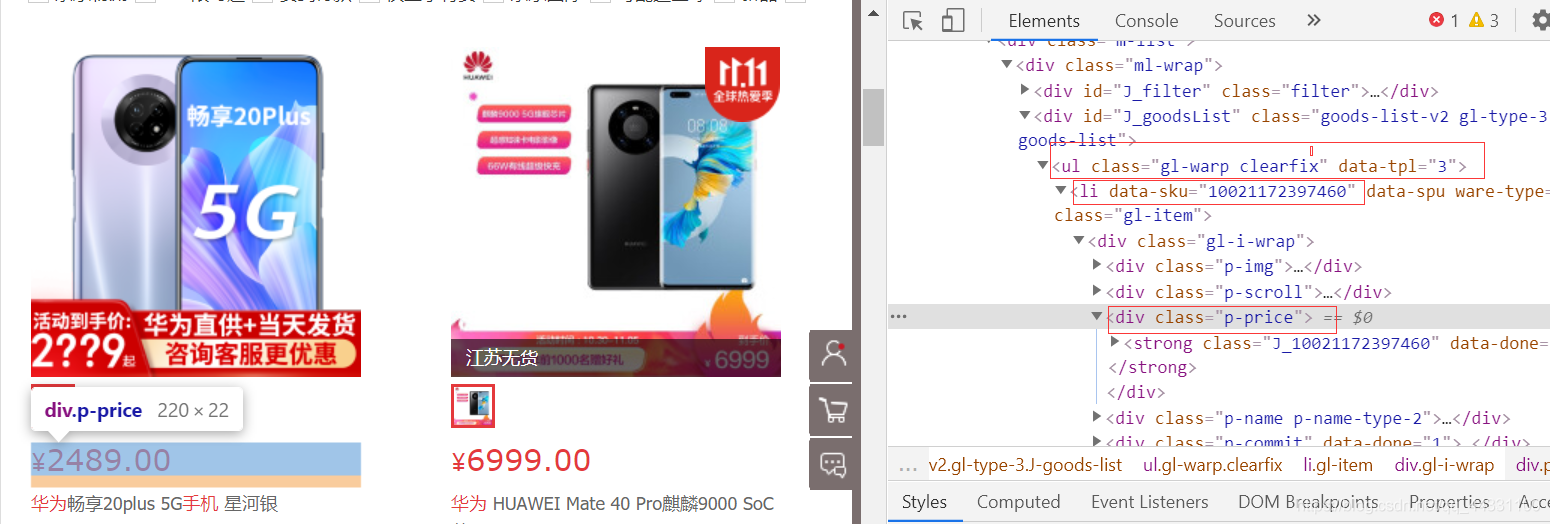
在谷歌中打开京东网页，按住Ctrl+Shift+l打开网页源码，在代码的左上角有一个箭头的图标（这个图标是你在页面滑动到哪，对应的源代码哪处高亮显示，这就表示此处对应的源代码）

  
找到源代码后右击，再点Copy\Copy XPath就得到了输入框的xpath  
  
得到xpath之后就可以写代码了,使用find\_element\_by\_xpath方法通过xpath定位输入框，在里面输入“华为手机”，再定位搜索按钮的位置，单击它进行搜索（注意：由于网络问题网页加载可能过慢，建议跳转页面要休息几秒钟，这样防止下一步进行时网页还没跳转成功）

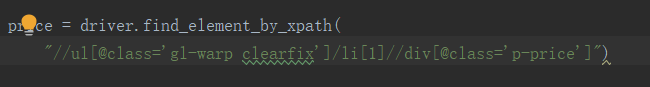
  
（3）获取单个商品信息  
进入到商品页面之后，我们先爬取第一个商品信息

我们获取商品信息同样适用xpath定位元素位置，但这个时候尽量不要使用上面的方法获得xpath，因为那样获得的xpath在定位是会根据html一层一层的寻找，花费时间太久，我们可以自己写xpath来缩短定位时间。

我们可以看到，商品对应的价格信息存放在下图红色框框出来的div中，我们就可以直接定位到对应的div中去获取价格



//表示跨级搜索，/表示逐级搜索

  
找到价格信息的位置之后要获取其中的内容，则要使用get\_attribute(“textContent”)方法

在这里插入图片描述  
获得了单个商品的价格信息后，其他信息也重复相同步骤  
（爬取一个页面多个商品在单个商品的外面加循环即可，爬取多个页面在上述基础上再加循环即可，获得的数据可以用列表存储起来，方便下面写入到表格中）  
  
（4）数据写入表格

  
**3.完整代码**

from selenium import webdriver

from time import sleep

import xlwt # 进行excel操作

#谷歌驱动 告诉电脑在哪打开浏览器

driver=webdriver.Chrome(executable\_path="D:/chromedriver.exe")

#打开网页

driver.get("https://www.jd.com")

#通过xpath找到输入框输入要搜索的

driver.find\_element\_by\_xpath("//\*[@id='key']").send\_keys("华为手机")

#通过xpath找到搜索按钮点击

driver.find\_element\_by\_xpath("//\*[@id='search']/div/div[2]/button").click()

driver.implicitly\_wait(5)#隐式休息5s

hrefs=[]

names=[]

prices=[]

commits=[]

shops=[]

for j in range(1,11):

for i in range(1, 31):

href = driver.find\_element\_by\_xpath(

"//ul[@class='gl-warp clearfix']/li[" + str(i) + "]//div[@class='p-img']/a").get\_attribute("href")#get\_attribute获得对应属性的属性值

information = driver.find\_element\_by\_xpath(

"//ul[@class='gl-warp clearfix']/li[" + str(i) + "]//div[@class='p-name p-name-type-2']")

price = driver.find\_element\_by\_xpath(

"//ul[@class='gl-warp clearfix']/li[" + str(i) + "]//div[@class='p-price']")

commit = driver.find\_element\_by\_xpath(

"//ul[@class='gl-warp clearfix']/li[" + str(i) + "]//div[@class='p-commit']/strong")

shop = driver.find\_element\_by\_xpath(

"//ul[@class='gl-warp clearfix']/li[" + str(i) + "]//div[@class='p-shop']")

hrefs.append(href)

names.append(information.get\_attribute("textContent").replace('\n', '').replace('\t', ''))

prices.append(price.get\_attribute("textContent").replace('\n', '').replace('\t', '').replace('￥',''))

commits.append(commit.get\_attribute("textContent").replace('\n', '').replace('\t', ''))

shops.append(shop.get\_attribute("textContent").replace('\n', '').replace('\t', ''))

driver.find\_element\_by\_xpath("//\*[@id='J\_bottomPage']/span[1]/a[9]").click()

sleep(2)

print("第"+str(j)+"页")

print("爬取完毕！")

#存数据

book = xlwt.Workbook(encoding="utf-8", style\_compression=0) # 创建workbook对象

sheet = book.add\_sheet('京东', cell\_overwrite\_ok=True) # 创建工作表

col = ("链接","名称","价格","评价","商店")

for i in range(0, 5):

sheet.write(0, i, col[i]) # 列名

for i in range(0,300):

sheet.write(i+1,0,hrefs[i])

for i in range(0,300):

sheet.write(i+1,1,names[i])

for i in range(0,300):

sheet.write(i+1,2,prices[i])

for i in range(0,300):

sheet.write(i+1,3,commits[i])

for i in range(0,300):

sheet.write(i+1,4,shops[i])

book.save("京东.xls") # 保存

print("关闭浏览器，保存数据")

**4.输出结果**  




3、参考代码使用scrapy爬取山东的天气预报，比较麻烦。可以自行写BeautifulSoup4或者selenium爬孝感天气预报。

**pip安装扩展库出错的解决方法**

1. python -m pip install python-docx 出错，

提示pip需要升级，用下面命令可以升级pip

python -m pip install -U pip

或 python -m pip install --upgrade pip

2.不行试试这个或许可以升级

easy\_install --upgrade pip

3、如果仍出错，使用如下命令，查看具体失败原因：

pip install --upgrade pip –vvv

4.如果发现**ReadTimeoutError 错误再用命令**

pip --default-timeout=100 install gevent

然后使用步骤2命令 升级pip

5能升级成功再安装扩展库，否则继续查找原因

python -m pip install python-docx

或者**pip3 install -U** python-docx

或 **pip3 install –U** wxPython

安装成功，就能使用

import docx 或 import wx

在pycharm里C:\Users\Hp\AppData\Roaming\Tencent\Users\79737484\QQ\WinTemp\RichOle\UOF)89}0SCPDRFGZ~C_6TS7.png

输入：pip install scrapy

C:\Users\Hp\AppData\Roaming\Tencent\Users\79737484\QQ\WinTemp\RichOle\@~L`1Q2UYF@]8@V%DYXXXF6.png

出错，提示pip要升级

输入：python -m pip install -U pip

升级成功，

输入：pip install scrapy 仍然出错，

C:\Users\Hp\AppData\Roaming\Tencent\Users\79737484\QQ\WinTemp\RichOle\$`)%}6SKIUT$4BCFAH]5P_K.png

https://blog.csdn.net/qq\_36711420/article/details/79487772  安装scrapy出现Failed building wheel for Twisted解决办法 这里有，或者在ftp 100下载 Twisted-20.3.0-cp38-cp38-win\_amd64.whl 复制在pycharm运行目录

C:\Users\Hp\AppData\Roaming\Tencent\Users\79737484\QQ\WinTemp\RichOle\)`731J69I$1CZ)LV1JQ3W$K.png

然后 安装 ： pip install Twisted-20.3.0-cp38-cp38-win\_amd64.whl

成功，继续

输入：pip install scrapy

成功！scray安装在这里：

C:\Users\Hp\PycharmProjects\pythonProject\venv\Scripts

按下shift + 鼠标右键，选择 在此处打开“powershell窗口”，

C:\Users\Hp\AppData\Roaming\Tencent\Users\79737484\QQ\WinTemp\RichOle\`L~HB(@R@SZ6G~UXA`[OZSW.png

在此路径下运行下面步骤2,即输入：scrapy startproject HbWeatherSpider或 .\scrapy startproject HbWeatherSpider创建爬虫项目，然后进入，项目目录

C:\Users\Hp\AppData\Roaming\Tencent\Users\79737484\QQ\WinTemp\RichOle\3A0_@H]@{CQAEH}2]$EQ6~K.png

把写好的everyCityinHB.py文件放在这里，在命令提示符执行scrapy crawl everyCityHB命令运行爬虫程序。

最后把结果文件截图。

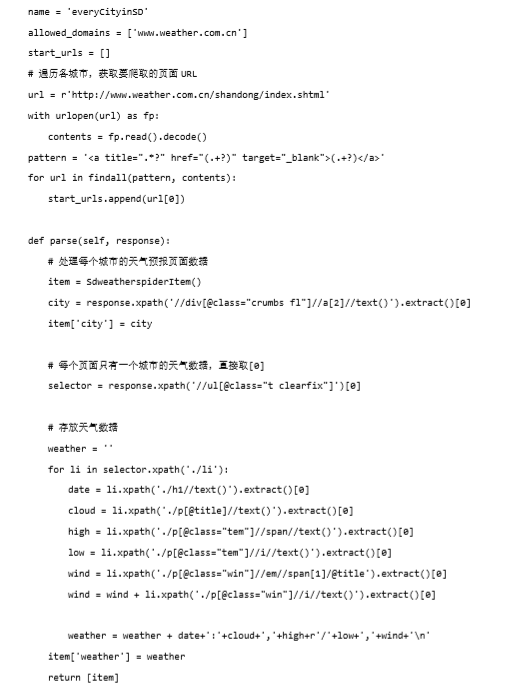
详细方法可参考网站：https://blog.csdn.net/xiaoxiaoCYG/article/details/81095843

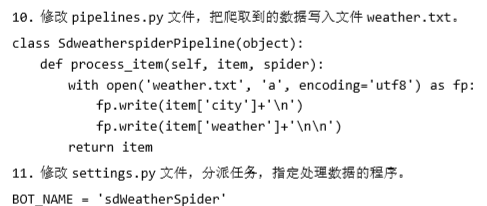


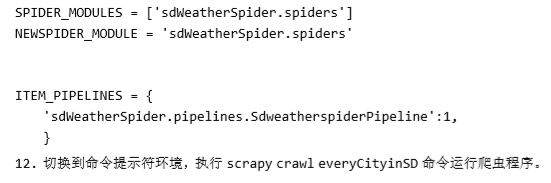




C:\Users\Hp\AppData\Roaming\Tencent\Users\79737484\QQ\WinTemp\RichOle\%T2SRAEQAP%7US@~D(83~}D.png







C:\Users\lenovo\AppData\Roaming\Tencent\Users\79737484\QQ\WinTemp\RichOle\92A7B9VQSYI({$GIK%7$}80.png

## 4、自己完成ftp爬虫或者根据自己想爬取的信息，写爬虫代码。