XPath Helper助XPath爬虫一臂之力

原创：爬虫俱乐部 Stata and Python数据分析 10月16日

本文作者：王碧琪
文字编辑：宁刘莹
技术总编：张　邯

在之前的推文《解析XML文件》中我们讲了关于XPath的基本使用方法。XPath虽然好用，但是关键在于迅速正确找到合适的节点，才能提取出相应的信息。小编一开始接触XPath的时候，经常在茫茫代码中苦苦寻找XPath的目标节点，无奈有些复杂的代码总是搞得小编焦头烂额。后来，小编偶然发现一款神器—XPath Helper，妈妈再也不用担心我找不到目标节点了！

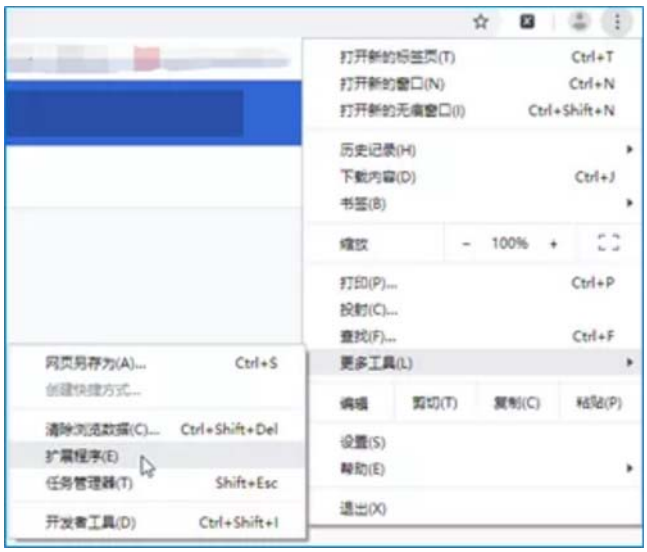
1 XPath Helper是什么?

XPath Helper是一款谷歌浏览器插件， 它支持在网页点击元素生成XPath。有了它，我们可以轻松快捷地找到目标信息对应的XPath节点，提取目标信息。

2 Xpath Helper的安装

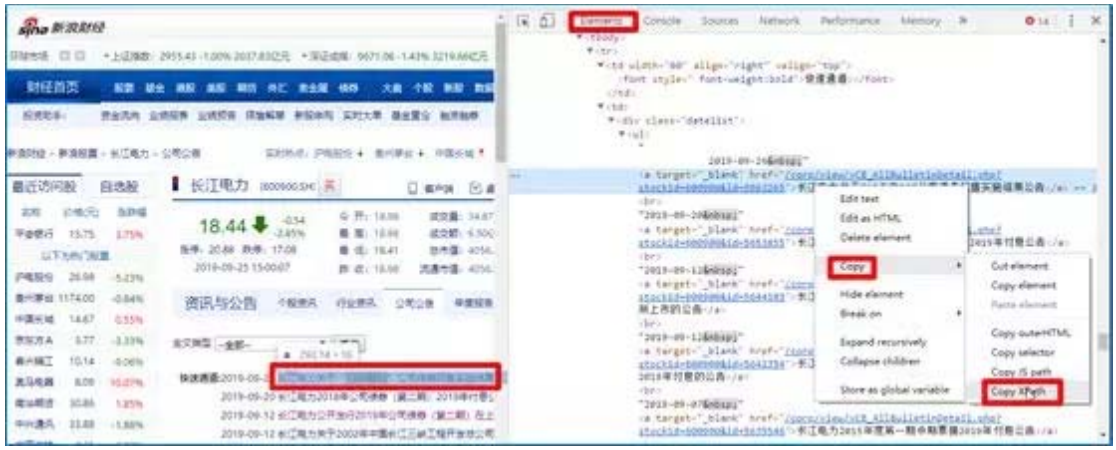
首先我们需要安装一下这个神器。在谷歌浏览器中的应用商店里（科学上网的情况下），搜索到XPath Helper插件，点击“添加至Chrome”即可。如果不能打开Chrome应用商店，可以通过网上的其他途径获取该插件，之后再手动添加至谷歌浏览器即可。手动添加方法是：打开谷歌浏览器扩展程序，并开启开发者模式，将该插件拖拽到浏览器里，如果不成功，可以选择“加载已解压的扩展程序”，将该文件夹先压缩再解压添加进去。

安装成功之后会发现在扩展程序中找到该插件，这就代表我们能用了。



1.点选copy XPath

现在我们打开一个标签页，以http://vip.stock.finance.sina.com.cn/corp/view/vCB_AllBulletin.php?stockid=600900&Page=1为例，我们按下F12键打开开发者模式，选择elements选项，将光标放在想要的源代码上，右键copy选项下的copy xpath，即可拷贝到选中的源代码。此处我们拷贝了包含第一条公告“长江电力关于“18长电02”公司债券回售实施结果公告”的这一行，如图片所示，得到的是：//*[@id="con02-7"]/table[2]/tbody/tr/td[2]/div[1]/ul/a[1]。也就意味着我们得到了它的XPath，可以直接调用了。



2.打开控制台

方法一：点击这个小图标。

方法二：快捷键 ctrl+shift+x

以上两种方式都可以调出XPath的控制台，会弹出query和result框。我们按住shift键，会发现随着光标的移动，控制台框中的内容随之变化，query中显示的是选中内容的XPath，result框中显示的对应的选中内容。比如我们把光标放在这条短评的文字上，在query中可以复制得到该XPath：/html/body/div[@class='wrapmain_wrap clearfix']/div[@class='R']/div[@id='con02-7']/table[@class='table2'][2]/tbody/tr/td[2]/div[@class='datelist']/ul/a[1]

3.小结

我们使用上述两种方法都得到了想要的公告标题内容，但是对应的XPath却不相同，使用copy XPath得到的是包含任意位置选取以及通配符的路径表达式，这种方式得到的表达式比较简短易读；而使用控制台得到的是从根节点一层一层精确定位得到的路径表达式，这种方式得到的表达式比较长但更易于理解。

4 案例实操

我们编写以下程序来提取长江电力公司的网站公告的发布日期、标题、链接：
导入第三方库：

```
import requests
import re
import json
from lxml import etree
```

获取网站代码信息：

```
url='http://vip.stock.finance.sina.com.cn/corp/view/vCB_AllBulletin.php?stocki
headers={
  'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,
  'Accept-Encoding': 'gzip, deflate',
  'Accept-Language': 'zh-CN,zh;q=0.9',
  'Cache-Control': 'max-age=0',
  'Connection': 'keep-alive',
  'Cookie': 'U_TRS1=000000f8.4cd080b9.5d5df083.84177f98; UOR=,vip.stock.financ
  'Host': 'vip.stock.finance.sina.com.cn',
  'Upgrade-Insecure-Requests': '1',
  'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36
}
html=requests.get(url,headers=headers)
html.encoding='gb18030'
tree=etree.HTML(html.text)
```

我们使用上述方法尝试获取公告标题：

```
title=tree.xpath('''/html/body/div[@class='wrap main_wrap clearfix']/div[@class
print(title)
```

发现返回结果为空。

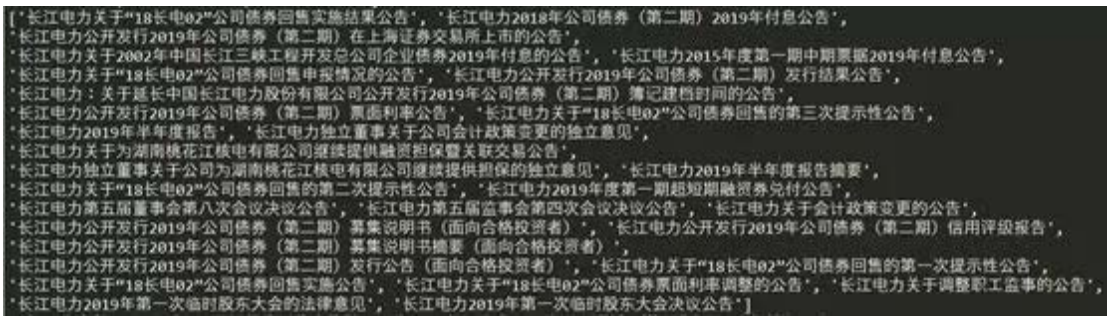
```
[]
```

这是为什么呢？小编也百思不得其解，后来发现是谷歌浏览器对网站内容进行了渲染，添加了/tbody标签。所以，只需要手动删除该标签即可。

```
title1=tree.xpath('''/html/body/div[@class='wrap main_wrap clearfix']/div[@clas
print(title1)
title2=tree.xpath('//*[@id="con02-7"]/table[2]/tr/td[2]/div[1]/ul/a/text()')
print(title2)
```

上述两种方法获得的内容结果相同。

结果如下：



类似地，我们获取公告发布日期和公告链接：

```
date1=tree.xpath('//tr/td[2]/div[1]/ul/text()') #copy方法。
datelist=[''.join(date1.split()) for date1 in date1 if ''.join(date1.split()) !=
print(datelist)
```

```
['2019-09-26', '2019-09-20', '2019-09-12', '2019-09-12',
'2019-09-07', '2019-09-06', '2019-09-05', '2019-09-02',
'2019-09-03', '2019-09-03', '2019-08-31', '2019-08-31',
'2019-08-31', '2019-08-31', '2019-08-31', '2019-08-31',
'2019-08-31', '2019-08-31', '2019-08-31', '2019-08-31',
'2019-08-30', '2019-08-30', '2019-08-30', '2019-08-30',
'2019-08-30', '2019-08-28', '2019-08-28', '2019-08-23',
'2019-08-22', '2019-08-22']
```

```
url1=tree.xpath('//*[@id="con02-7"]/table[2]/tr/td[2]/div[1]/ul/a/@href') #copy
print(url1)
```

至此我们就使用xpath helper获取了所有公告的发布日期、标题和链接了。接下来我们编写完整的程序将其保存到文件中。

完整程序如下：

```
import requests
import re
import json
from lxml import etree

url='http://vip.stock.finance.sina.com.cn/corp/view/vCB_AllBulletin.php?stocki
headers={
  'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,
  'Accept-Encoding': 'gzip, deflate',
  'Accept-Language': 'zh-CN,zh;q=0.9',
  'Cache-Control': 'max-age=0',
  'Connection': 'keep-alive',
  'Cookie': 'U_TRS1=000000f8.4cd080b9.5d5df083.84177f98; UOR=,vip.stock.financ
  'Host': 'vip.stock.finance.sina.com.cn',
  'Upgrade-Insecure-Requests': '1',
  'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36
}
html=requests.get(url,headers=headers)
html.encoding='gb18030'
tree=etree.HTML(html.text)

date=tree.xpath('//tr/td[2]/div[1]/ul/text()') #copy xpath
datelist=[''.join(date.split()) for date in date if ''.join(date.split()) !=""
titlelist=tree.xpath('''/html/body/div[@class='wrap main_wrap clearfix']/div[@
urllist=tree.xpath('//*[@id="con02-7"]/table[2]/tr/td[2]/div[1]/ul/a/@href')
with open("g:\\daydayfund.csv","w",encoding="utf8") as f:
  for date,title,url in zip(datelist,titlelist,urllist):
    info=date+','+title+','+url+'\n' #日期，标题，链接
    f.write(info)
```

结果文件中的部分内容如下：

```
2019-09-26,长江电力关于“18长电02”公司债券回售实施结果公告,/corp/view/vCB_AllBulletinDetail.php?stockid=600900&id=5663265
2019-09-20,长江电力2018年公司债券（第二期）2019年付息公告,/corp/view/vCB_AllBulletinDetail.php?stockid=600900&id=5653855
2019-09-12,长江电力公开发行2019年公司债券（第二期）在上海证券交易所上市的公告,/corp/view/vCB_AllBulletinDetail.php?stockid=600900&id=5644103
2019-09-12,长江电力关于2002年中国长江三峡工程开发总公司企业债券2019年付息的公告,/corp/view/vCB_AllBulletinDetail.php?stockid=600900&id=5642354
2019-09-07,长江电力2015年度第一期中期票据2019年付息公告,/corp/view/vCB_AllBulletinDetail.php?stockid=600900&id=5635546
2019-09-06,长江电力关于“18长电02”公司债券回售申报情况的公告,/corp/view/vCB_AllBulletinDetail.php?stockid=600900&id=5633298
2019-09-05,长江电力公开发行2019年公司债券（第二期）发行结果公告,/corp/view/vCB_AllBulletinDetail.php?stockid=600900&id=5632962
```

5 小 结

Xpath helper插件小巧好用，有了它的加持，我们在网络数据获取时更加省时省力，那么看完这编推文，你学会使用xpath helper了吗？下次获取网络数据时赶快来试试它吧！

对我们的推文累计打赏超过1000元，我们即可给您开具发票，发票类别为“咨询费”。用心做事，不负您的支持！

往期推文推荐

查找变量？用“codebook”！
distinct命令用法一览
Stata数据分析技术应用培训
玩转Python之“手把手”教你爬数据（一）
玩转Python之“手把手”教你爬数据（二）
labelsof和labelbook介绍
Statalist上的“火云邪神”
爬虫实战程序的函数封装
Zipfile(二)
利用collapse命令转化原始数据
Stata中的数值型
爬虫实战——聚募网股权众筹信息爬取
duplicates drop之前，我们要做什么？
类型内置函数-type() isinstance()
数据含义记不住？—— label“大神”来帮忙
Zipfile(一)
tabplot命令
Jupyter Notebook不为人知的秘密

关于我们

微信公众号“Stata and Python数据分析”分享实用的stata、python等软件的数据处理知识，欢迎转载、打赏。我们是由李春涛教授领导下的研究生及本科生组成的大数据处理和分析团队。

此外，欢迎大家踊跃投稿，介绍一些关于stata和python的数据处理和分析技巧。

投稿邮箱：statatraining@163.com

投稿要求：

1）必须原创，禁止抄袭；

2）必须准确，详细，有例子，有截图；

注意事项：

1）所有投稿都会经过本公众号运营团队成员的审核，审核通过才可录用，一经录用，会在该推文里为作者署名，并有赏金分成。

2）邮件请注明投稿，邮件名称为“投稿+推文名称”。

3）应广大读者要求，现开通有偿问答服务，如果大家遇到有关数据处理、分析等问题，可以在公众号中提出，只需支付少量赏金，我们会在后期的推文里给予解答。