9.3 裁判文书网数据挖掘实战

裁判文书网：<http://wenshu.court.gov.cn/Index>是最权威的生效裁判文书公示网站，对于金融行业内部风控和舆情监控都有较大的作用，用常规手段来对该网站进行数据挖掘难度很大，而如果通过selenium方法的话则可以较轻易的进行信息的获取。



以在该网站搜索和关键词“房地产”相关案件为例，我们需要做的就是在上图的搜索框里输入“房地产”这个词并点击搜索按钮，这个和上一章讲解模拟在百度首页上搜索python是类似的，简单回顾下之前在百度上搜索python的代码：

from selenium import webdriver

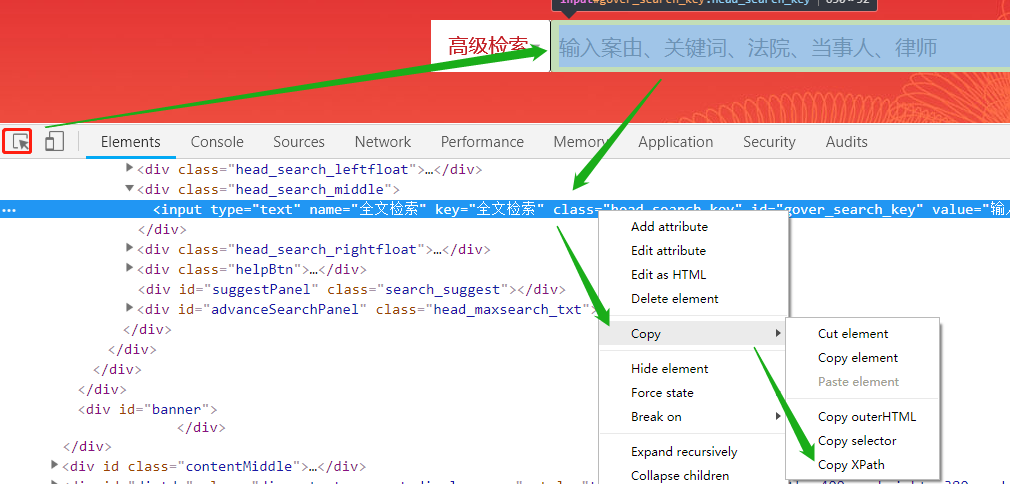
browser = webdriver.Chrome()

browser.get("https://www.baidu.com/")

browser.find\_element\_by\_xpath('//\*[@id="kw"]').send\_keys('python')

browser.find\_element\_by\_xpath('//\*[@id="su"]').click()

这边采用类似的方法，只要找到裁判文书网的搜索框和搜索按钮的xpath或css\_selector的内容即可。以找xpath内容为例，用F12在网页上找到该元素，然后鼠标右击，选择Copy，然后选择Copy Xpath即可，如下图所示：



获取到xpath内容之后，在裁判文书网搜索“房地产”和在百度首页搜索“python”的原理是一致的，唯一要注意的的就是在裁判文书网上，其搜索框里是有一些默认文本内容的，所以在输入“房地产”这个关键词之前，要把原来的搜索框用clear()方法清空一下，完整代码如下：

from selenium import webdriver

import time

browser = webdriver.Chrome()

browser.get('http://wenshu.court.gov.cn/')

browser.maximize\_window()

browser.find\_element\_by\_xpath('//\*[@id="\_view\_1540966814000"]/div/div[1]/div[2]/input').clear() # 清空原搜索框

browser.find\_element\_by\_xpath('//\*[@id="\_view\_1540966814000"]/div/div[1]/div[2]/input').send\_keys('房地产') # 在搜索框内模拟输入'房地产'三个字

browser.find\_element\_by\_xpath('//\*[@id="\_view\_1540966814000"]/div/div[1]/div[3]').click() # 点击搜索按钮

time.sleep(10) # 如果还是获取不到你想要的内容，你可以把这个时间再稍微延长一些

data = browser.page\_source

browser.quit()

print(data)

这边为了美观，还加了一行窗口最大化的代码：browser.maximize\_window()。运行代码之后，可以看到它已经能够自动访问裁判文书网，并搜索房地产三个字，然后跳转到与房地产相关案件的页面了。成功获取的网页源代码里也包含我们想要的案件信息。

这里注意一定要记得引入time模块，在通过data = brower.page\_source获取源代码之前，得先利用time.sleep()来稍微等待一段时间，因为裁判文书网在搜索“房地产”三个字的时候其实是有一个页面加载等待的过程，有的时候甚至要等待十多秒才能把网页刷新出来。如果不通过time.sleep()等待一段时间的话，可能获取的内容就不是最终想要的内容。上述代码熟悉了之后，也可以设置无界面浏览器形式让程序在后台运行。

获取源码之后，需要什么信息便可通过编写正则表达式来进行提取了。这节的重点不在于教大家如何写正则表达式，而是教大家如何通过selenium来模拟鼠标键盘点击访问裁判文书网。

此外注意，2019年8月份之后裁判文书网改版，其反爬非常强，所以模拟键盘鼠标操作后等待很久也等不到刷新，所以这里主要给大家练习下如何通过selenium库模拟键盘鼠标操作。

9.4 巨潮资讯网数据挖掘实战

巨潮资讯网：<http://www.cninfo.com.cn/new/index>是中国证券监督管理委员会指定的上市公司信息披露网站，创建于1995年，是国内最早的证券信息专业网站，同时亦是国内首家全面披露深沪3000多家上市公司公告信息和市场数据的大型证券专业网站。



这里讲解下巨潮资讯网数据挖掘的项目背景：这个是为了通过数据挖掘在巨潮资讯网上获取各个上市公司的理财公告的下载链接并进行下载，从而通过分析这些理财公告判断是否存在合适的资金方（也即合适的投资者们）。我们也可以从上面搜索别的内容，比如公司年报，从而进行年报分析，下一章会讲解如何进行PDF解析来分析这些公告。本章节先演示如何获取下图上市公司理财报告的标题、下载链接和发布日期。



**9.4.1 获取网页源码**

首先第一步还是来获取网页源码，在巨潮上搜索“理财”两个字，能够看到很多公告，其网址为：<http://www.cninfo.com.cn/new/fulltextSearch?notautosubmit=&keyWord=%E7%90%86%E8%B4%A2>。你也可以把里面百分号的内容换成“理财”两个字：[http://www.cninfo.com.cn/new/fulltextSearch?notautosubmit=&keyWord=理财](http://www.cninfo.com.cn/new/fulltextSearch?notautosubmit=&keyWord=%E7%90%86%E8%B4%A2)，然后用selenium来进行网页源码的获取（传统的方法获取不了源码）。

from selenium import webdriver

browser = webdriver.Chrome()

url = 'http://www.cninfo.com.cn/new/fulltextSearch?notautosubmit=&keyWord=理财'

browser.get(url)

data = browser.page\_source

print(data)

这边为了演示之后章节要讲的理财公告pdf下载，所以采用有界面模式，运行上面的代码便可获取到网页源代码了，并且可以在其中找到我们想要的内容。如果没有获取到想要的源码，可以试着在data = browser.page\_source前加个time.sleep()用以防止页面加载刷新时间过长导致源码获取失败，括号里填入你想等待的时间即可，单位为秒。在有界面的模式下可以看到它并没有一个页面加载的过程，所以目前来说不需要加time.sleep()进行等待。

**9.4.2 正则表达式编写**

首先通过F12来观察下这些公告的标题、网址的规律：



然后我们再通过在爬取到的源代码里观察，发现其标题有如下规律：

p\_title = '<span title="" class="r-title">标题</span>'

其网址规律如下：

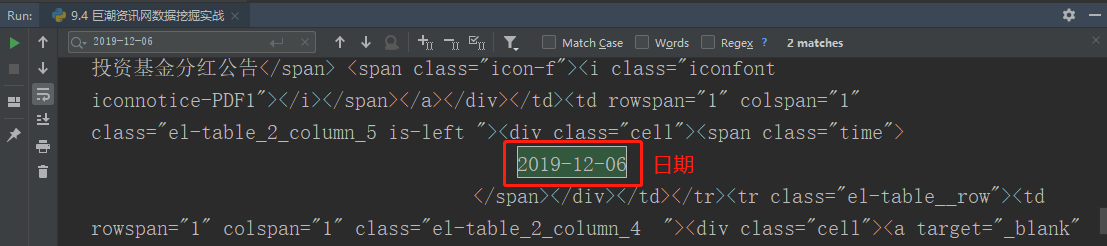
p\_href = '<a target="\_blank" href="网址data-id='

所以标题和网址的正则编写如下所示：

p\_title = '<span title="" class="r-title">(.\*?)</span>'

p\_href = '<a target="\_blank" href="(.\*?)data-id='

再在源代码里观察观察日期的规律：



其日期则有如下规律：

p\_date = '<span class="time">日期</span>'

所以日期的正则编写如下所示：

p\_date = '<span class="time">(.\*?)</span>'

此外，注意在<span class="time">和</span>间存在换行，但是在第2章讲正则表达式的时候我们提到过，(.\*?)不能直接匹配换行，因此需要通过第2章讲过的re.S的方法来取消换行的影响，代码如下：

date = re.findall(p\_date, data, re.S)

完整代码如下所示：

p\_title = '<td class="sub-title"><a href=".\*?" target="\_blank">(.\*?)</td>'

p\_href = '<td class="sub-title"><a href="(.\*?)" target="\_blank">.\*?</td>'

p\_date = '<div class="sub-time-time">(.\*?)</div>'

title = re.findall(p\_title, data)

href = re.findall(p\_href, data)

date = re.findall(p\_date, data, re.S)

数据提炼完后也可以把这些内容打印出来，看是否把想要的内容都爬取到了。

**9.4.3 数据清洗及打印输出**

在标题里含有一些<em>等<>包围的内容，我们通过re.sub()方法来进行处理。爬取到的网址缺少一个前缀：<http://www.cninfo.com.cn>，可以使用字符串拼接解决这个问题。还有个不太容易发现的问题，就是爬取到的网址里还有多这么一串字母：“amp;”，可以利用re.sub方法将其替换为空值。因为获取的日期中有换行符(/n)和一些空格，所以需要通过strip()函数来清除空格和换行符；最后一个小问题是爬取到的日期里面有的含有具体的时分秒信息，比如说：2013-09-27 16:15，我们只需要前面的年月日信息，两者之间以空格分开，那么可以通过split()函数提取前者，具体代码如下：

for i in range(len(title)):

title[i] = re.sub('<.\*?>', '', title[i])

href[i] = 'http://www.cninfo.com.cn' + href[i]

href[i] = re.sub('amp;', '', href[i])

date[i] = date[i].strip() # 清除空格和换行符

date[i] = date[i].split(' ')[0] # 只取“年月日”信息，不用“时分秒”信息

print(str(i + 1) + '.' + title[i] + ' - ' + date[i])

print(href[i])

**9.4.4 定义和调用函数**

定义和调用函数和之前所学类似，把原来url中的理财换成keyword即可：

def juchao(keyword):

url = 'http://www.cninfo.com.cn/new/fulltextSearch?notautosubmit=&keyWord=' + keyword

完整代码如下：

from selenium import webdriver

import re

def juchao(keyword):

browser = webdriver.Chrome()

url = 'http://www.cninfo.com.cn/new/fulltextSearch?notautosubmit=&keyWord=' + keyword

browser.get(url)

data = browser.page\_source

# print(data)

browser.quit()

# 巨潮网网页结构调整后的正则表达式代码（2019-12-31日改）

p\_title = '<span title="" class="r-title">(.\*?)</span>'

p\_href = '<a target="\_blank" href="(.\*?)data-id='

p\_date = '<span class="time">(.\*?)</span>'

title = re.findall(p\_title, data)

href = re.findall(p\_href, data)

date = re.findall(p\_date, data, re.S)

for i in range(len(title)):

title[i] = re.sub(r'<.\*?>', '', title[i])

href[i] = 'http://www.cninfo.com.cn' + href[i]

href[i] = re.sub('amp;', '', href[i])

date[i] = date[i].strip() # 清除空格和换行符

date[i] = date[i].split(' ')[0] # 只取“年月日”信息

print(str(i + 1) + '.' + title[i] + ' - ' + date[i])

print(href[i])

keywords = ['理财', '现金管理', '纾困']

for i in keywords:

juchao(i)

这边还可以加入异常处理函数和无界面模式设置。如果有的时候出现错误了，可能是因为访问的页面还没有缓冲出来，这时你可以引入time模块，然后在data = browser.page\_source前面加一行：time.sleep()，在括号里填入你想等待的时间即可，单位为秒。

巨潮资讯网还有很多知识点可以挖掘，比如有的读者就已经发现了，这个关于理财的公告有好几万条，但是它第一页只显示十条，那么该如何把这些信息都获取了呢？以及如果只想要2018年的数据该怎么处理呢？获得这些理财公告的下载网址之后，如何进行把这些理财公告pdf批量下载下来呢？这些问题下一章将进行详细讲解。