一、 selenium简介

如果链接简单，爬虫可以通过链接用requests库提取页面信息，如爬取豆瓣top250影片信息，链接简单易懂。参考：爬取豆瓣top250影片信息

但如果遇到一些搜索之类的，基于js动态加载的网页，以上就不适合，如爬虫b站，搜索“爬虫”页面，第一页链接如下，第二页又是很长没规律的链接。很难用requests库提取页面。

```
https://search.bilibili.com/all?keyword=%E7%88%AC%E8%99%AB&from_source=webtop_search&spm_id_from=333.1007&search_source=5
```

针对以上情况，我们可以通过浏览器直接访问每个页面，然后提取页面。当然是让爬虫自己打开浏览器，输入内容访问，然后提取页面元素。这个过程就要用到 selenium 库。

selenium其实它就是一个自动化测试工具，支持各种主流的浏览器。遇到python，selenium 就变成了爬虫利器。

二、安装selenium配置环境变量

1、安装

```
pip install selenium
```

下载浏览器驱动，我用的是 Chrome 浏览器，所以下载Chrome驱动即可，当然你可以下载其他浏览器驱动。下载链接：https://chromedriver.chromium.org/，找到和自己浏览器版本一致或者最接近的。

2、配置环境变量

下载解压后，配置环境变量。

解压，然后创建一个存放浏览器驱动的目录，如： D:\Python\Driver，将下载的浏览器驱动文件（例如：chromedriver、geckodriver）丢到该目录下，我这里是chromedriver。

我的电脑–>属性–>系统设置–>高级–>环境变量–>系统变量–>Path，将“D:\Python\Driver”目录添加到Path的值中。比如：Path字段;D:\Python\Driver

关于环境变量不生效的问题：

1：尝试将浏览器驱动，直接放在python安装目录试试

2：配置环境变量后，重启电脑生效（我就是重启才生效）

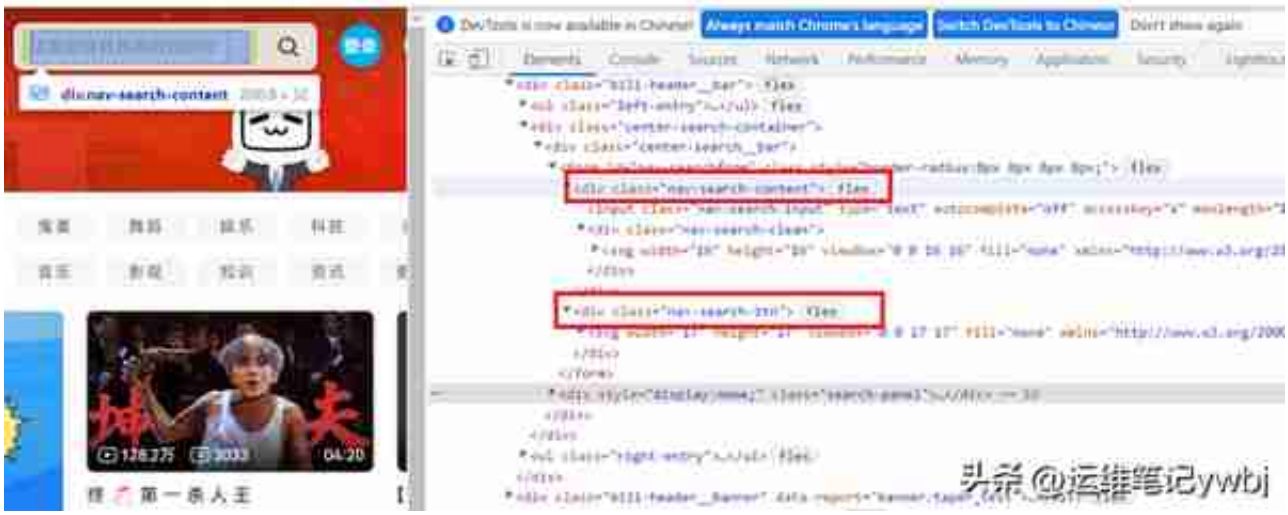
三、打开浏览器并自动搜索

1、浏览器自动访问

代码，打开浏览器，并访问。

```
#?? web ????from selenium import webdriver#????? Chrome ??dr
iver = webdriver.Chrome()#??????driver.get("https://www.bili
bili.com/")#???????????#driver.quit()
```

以上，执行代码自动打开浏览器



找到元素后，用find_element方法定位找到此元素，定位有多种方式。

```
id???find_element_by_id()name???find_element_by_name()class?
??find_element_by_class_name()link???find_element_by_link_te
xt()partial link???find_element_by_partial_link_text()tag???
```

```
find_element_by_tag_name()xpath???find_element_by_xpath()css
???find_element_by_css_selector()
```

以上也可以用 `driver.find_elements(By.ID, 'xxx')` 的方式，对应值。

```
ID = "id"XPATH = "xpath"LINK_TEXT = "link text"PARTIAL_LINK_
TEXT = "partial link text"NAME = "name"TAG_NAME = "tag name"
CLASS_NAME = "class name"CSS_SELECTOR = "css selector"
```

我这里用class值定位即可。

```
from selenium import webdriverfrom selenium.webdriver.common
.by import Bydriver = webdriver.Chrome()driver.get("https://
www.bilibili.com/")#?????input = driver.find_element(By.CLAS
S_NAME,'nav-search-input')button = driver.find_element(By.CL
ASS_NAME,'nav-search-
btn')#?????input.send_keys('??')#????button.click()
```

注：以上如果报错提示没有By这个类型，需要导入包from selenium.webdriver.common.by import By

执行效果：

最后代码为：

```
import timefrom selenium import webdriverfrom selenium.webdr
iver.common.by import Byfrom bs4 import BeautifulSoupdriver
= webdriver.Chrome()driver.get("https://www.bilibili.com/")i
nput = driver.find_element(By.CLASS_NAME,'nav-search-input')
button = driver.find_element(By.CLASS_NAME,'nav-search-btn')
input.send_keys('??')button.click()windows = driver.window_h
andles#print(windows)driver.switch_to.window(windows[-1])#??
5?time.sleep(5)#?????html=driver.page_sourcesoup = Beautiful
Soup(html,'lxml')list = soup.find(class_='video-list row').f
ind_all(class_="bili-video-card")for item in list:    #print
(item)    video_name = item.find(class_='bili-video-card__in
fo--tit').text    video_up = item.find(class_='bili-video-ca
rd__info--author').string    video_date = item.find(class_='
bili-video-card__info--date').string    video_play = item.fi
nd(class_='bili-video-card__stats--item').text    video_time
s = item.find(class_='bili-video-card__stats__duration').str
ing    video_link = item.find('a')['href'].replace('//','')
```

```
   print(video_name,video_up,video_play,video_times,video_li
nk,video_date)driver.quit()
```

执行结果为：

```
???1000????B?????????????????????????????? python???? 24.6?
29:14:25 www.bilibili.com/video/BV1bL4y1V7q1  · 4-252020?Pyt
hon????????????? ????IT 150.8? 43:30:57 www.bilibili.com/vi
deo/BV1Yh411o7Sz  · 2020-7-9Python????8?????????????????
Python??Alex 51.6? 20:39:05 www.bilibili.com/video/BV1ha4y1H
7sx  · 2020-12-7Python????????????????? python??? 80.5? 09:31
:34 www.bilibili.com/video/BV1qJ411S7F6  · 2019-11-13???????
???Python????+?????????????????????????? Python????? 3.6? 41
:15 www.bilibili.com/video/BV1gd4y1i75v  · 10-8Python?????,P
ython??+Python??+Python????5?????/Python??.Python?? IT?? 400
.8? 20:24:40 www.bilibili.com/video/BV12E411A7ZQ  · 2020-3-2
0???Python???????????python??+????? ??? 55.8? 22:19:04 www
.bilibili.com/video/BV1Db4y1m7Ho  · 2021-9-12022?Python?????
??-????+??????????????????20????????????? Python???? 6.1? 1
4:40:56 www.bilibili.com/video/BV1rv4y1K7yV  · 5-4Python????
5?????????????????? Python??Alex 7300 18:53:17 www.bilibil
i.com/video/BV1BD4y1k7Po  · 10-9?Python???????20?????Python?
??????????????????~ H??? 424 55:46:15 www.bilibili.com/video
/BV1QN4y1A7oc  · 10-10?2022???32?Python????????????????????
??????????????????????_??_??_??_??_??_?? ???? 113 40:44:18
 www.bilibili.com/video/BV1PR4y1R7Xc  · 13????????????? ???
60.5? 12:06 www.bilibili.com/video/BV1KK4y1s7iK  · 2021-1-22
???34???2022?????Python??????+?????????????????????????????
 Python???? 811 19:43:13 www.bilibili.com/video/BV1D8411x7Kx
  · 10-103???? Python?????????+????+?????? Python??? 6.6? 29
:32:59 www.bilibili.com/video/BV1hS4y1b7EJ  · 5-142021???Pyt
hon????+??????????? ????IT 68? 14:40:22 www.bilibili.com/vi
deo/BV1i54y1h75W  · 2021-3-5?Python?????9888??Python??????2
021?????????????????——??????? ??Python??? 49.2? 40:10:11 www
.bilibili.com/video/BV1ZT4y1d7JM  · 2021-10-8??1w???? ?????3
? ??????3?? ???? 20.1? 07:56 www.bilibili.com/video/BV1ih411
a7PK  · 2021-6-14Python?????????????????????52?? T????? 1.9
? 12:15:35 www.bilibili.com/video/BV1pf4y1H72c  · 2021-8-29?
????Python???????????10???????????????????? ??python 7764 39:
```

20 www.bilibili.com/video/BV1Jd4y1i7cH · 10-93??? Python??
???????+????+?????? ??IT?? 5.5? 02:16:47 www.bilibili.com/vi
deo/BV1Ry4y1V7PE · 2021-8-18????????Python?????? ?C-??? 18.
4? 01:47:55 www.bilibili.com/video/BV1wp411o7dz · 2018-5-20
?????Python?????????Scrapy???? ????? 27.9? 04:17:09 www.bili
bili.com/video/BV1jx411b7E3 · 2017-8-22???????????????? ???
?? 9.7? 05:18 www.bilibili.com/video/BV1Hr4y1w71G · 2020-11
-10????? ??????????Python?????????????????????????...????? ?
??Python 1.1? 02:52:35 www.bilibili.com/video/BV14T411P7TD
· 10-8Python??????7????2022?????Python??+?????0????????????
??????? 4.7? 51:00:54 www.bilibili.com/video/BV1Fa411q75C ·
4-11???1000????B?????????????????????????????? Python??? 14
.7? 14:40:56 www.bilibili.com/video/BV1qB4y1D77o · 6-42022?
?Python????????????????????????? ?????? 2.6? 17:38:51 www.b
ilibili.com/video/BV133411F7Cp · 7-20?python??1000????B????
????????????????????????? ?????? 73 14:40:56 www.bilibili.c
om/video/BV11N4y1A7cj · 14???????????B????Python??????????
?????99.9%???~?????????JS??/????/????/APP??/????? ???Python
2.3? 07:49:30 www.bilibili.com/video/BV1sG411n7Zf · 7-18202
2????????????????????? ??????? 3.2? 15:45:05 www.bilibili.com/
video/BV1JP4y1u72J · 3-18Python???????????? ?????????? Pyth
on?? 2.1? 04:10:56 www.bilibili.com/video/BV13e4y197vg · 7-
22???????????Python????7??????????? ?python3.8????+??? ???-?
?? 3.2? 23:57:57 www.bilibili.com/video/BV1rL411G7ep · 2021
-10-20??????????100?????????????????????????Python?????? pytho
n???? 8.4? 71:15:41 www.bilibili.com/video/BV1SA4y1976A · 4
-14??????Python?????????????Python???Python???????? IT?? 22
.2? 31:25:34 www.bilibili.com/video/BV1yY4y1w7r8 · 8-5Pytho
n???????????APP???????? ??????? 3.5? 39:07 www.bilibili.com
/video/BV18v411q76u · 2020-7-222022????????????????????? ????
??? 3.8? 13:43:02 www.bilibili.com/video/BV1iu411C79S · 4-1
6Python????????????????????????????? ??Python??? 2.5? 45:20
:00 www.bilibili.com/video/BV1Yi4y1S7Pi · 4-8????50????????
???2021???python????????????????????? ????? 2.2? 14:23:31 www
.bilibili.com/video/BV1Pg411V768 · 2021-9-1?Python?????????
Python????????????????????????????????? ?????? 2.9? 39:08:57 w
ww.bilibili.com/video/BV1R34y1h7jx · 5-16?Python??????????
??????Python ????????????????????? ?????Python 11.7? 27:51:
01 www.bilibili.com/video/BV1Hb4y167c9 · 2021-7-2010???Pyth

```
on?????????????(B???) Python_?? 12.2? 24:57:21 www.bilibili
.com/video/BV16f4y197D6  · 2020-8-11?????????????Python????
????????? ???????? 1.6? 19:41:57 www.bilibili.com/video/BV17
D4y1i7nX  · 9-20
```

这样单个页面的所有视频信息提取完成了。

五、判断网页是否加载完成

以上代码种，切换页面后，有一个等待5秒的代码。

```
time.sleep(5)
```

如果不加这行，有时候获取不到完整的信息而报错，因为代码元素没有加载完成，代码就执行完了，从而获取不到元素值而报错，所以必须加一个等待时间。报错如下：

```
Traceback (most recent call last):  File "D:\Python\test_sel
enium.py", line 42, in <module>    video_name = item.find(cl
ass_='bili-video-card__info--tit').textAttributeError: 'None
Type' object has no attribute 'text'
```

time.sleep(5)，如果临时调试，也可以用，但是在实际环境，尽量少用，因为实际环境网络差异，可能5秒也加载不出，而报错。所以需要灵活点的方式，有三种等待方法。

1、强制等待

sleep(x) x单位为s，sleep等待的是元素 。

不管你浏览器是否加载完了，程序都得等待，时间一到，继续执行下面的代码，作为调试很有用。

隐性等待和显性等待可以同时用。

注意：等待的最长时间取两者之中的大者

2、隐性等待

隐性等待的是页面，implicitly_wait(x) x单位为s。

一旦设置，这个隐式等待会在WebDriver对象实例的整个生命周期起作用，它不针对某一个元素，是全局元素等待，即在定位元素时，需要等待页面全部元素加载完成，才会执行下一个语句。

如果超出了设置时间的则抛出异常。

缺点：当页面某些js无法加载，但是想找的元素已经出来了，它还是会继续等待，直到页面加载完成（浏览器标签左上角圈圈不再转），才会执行下一句。某些情况下会影响脚本执行速度

3、显性等待

```
WebDriverWait(driver,timeout,poll_frequency=0.5,ignored_exceptions=None)
```

需要通过from selenium.webdriver.support.wait import WebDriverWait导入模块

- driver：浏览器驱动
- timeout：最长超时时间，默认以秒为单位
- poll_frequency：检测的间隔步长，默认为0.5s
- ignored_exceptions：超时后的抛出的异常信息，默认抛出NoSuchElementExeception异常。

配合该类的until()和until_not()方法

程序每隔poll_frequency秒看一眼，如果条件成立了，则执行下一步，否则继续等待，直到超过设置的最长时间，然后抛出TimeoutException。

WebDriverWait与expected_conditions结合使用，示例：

```
from selenium.webdriver.support import expected_conditions as ECfrom selenium.webdriver.common.by import Bywait = WebDriverWait(driver,10,0.5)element =wait.until(EC.presence_of_element_located((By.ID,"kw"),message="")# ???????message=“”??By.ID?????()
```

expected_conditions类提供的预期条件判断的方法：

```
?????????title????????title????????driver.titletitle_istitle
_contains???????????????????????????locator??(By.ID, ‘kw’)??
??????????????????????????????????????????????presence_of_ele
ment_locatedpresence_of_all_elements_located????????????????
????????????text????????valuetext_to_be_present_in_elementt
ext_to_be_present_in_element_value??????frame?????????locato
r????????????id?name?index?WebElementframe_to_be_available_
and_switch_to_it?????????alert??alert_is_present???????????
????locatorelement_to_be_clickable??????????????????????Web
Element????????locator???????WebElement???????????True?????F
alse?????locator?????????True?????Falseelement_to_be_selecte
delement_located_to_be_selectedelement_selection_state_to_be
element_located_selection_state_to_be??????????????DOM????
WebElement????????????staleness_of
```

4、更改代码等待方式

3个地方等待，输入关键词，点击按钮，还有切换页面。

输入搜索框，并确定。

```
wait=WebDriverWait(driver,10)#????????????input = wait.unt
il(EC.presence_of_element_located((By.CLASS_NAME,'nav-search
-input')))#????????????button = wait.until(EC.element_to
_be_clickable((By.CLASS_NAME,'nav-search-btn')))
```

调整页面等待，By.CLASS_NAME定位一个class值，并不能很准确，所以这里用By.CSS_SELECTOR定位。

对CSS_SELECTOR规则不熟的，直接查看源代码复制即可，当然也不一定非要最底层的元素值，我这里到div.video.i_wrapper.search-all-list即可。