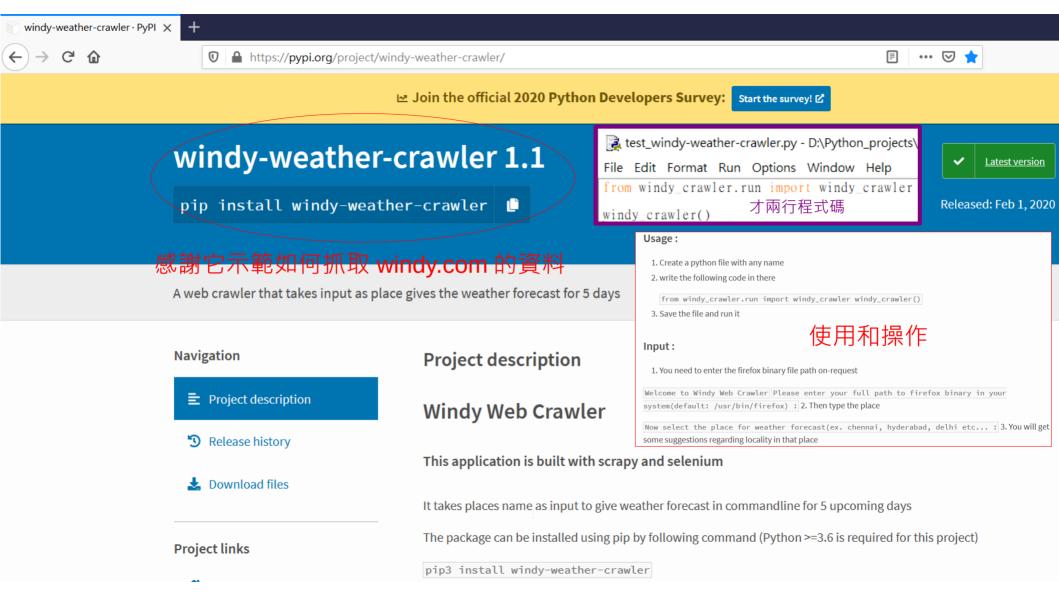
windy.com Webscraper

嘗試下載 windy.com 的氣象預報

- 無法直接解析網頁的 html source code 來獲得資料
 - 例如「港灣環境資訊網」的做法
- 也無法透過 web api 下載資料
 - 沒有公開的api,尚弄不清楚私人付費api的操作
- 最後只能透過 selenium 來模擬「人操作瀏覽器」
 - 第三方模組 windy-weather-crawler (需要修改才能正常)
 - 參考windy-weaher-crawler的程式碼,並實現全新的爬蟲
 - 下載並修改 windy.com的預報資料,最後生成 excel 檔案
 - 直接輸入gps座標來準確選定地點
 - 透過 docker 移植程式到群暉主機跑,尚不穩定



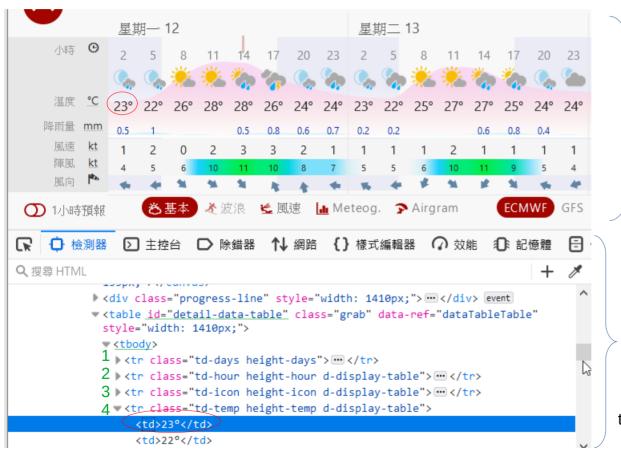
windy-weather-crawler 執行過程

```
C:\Windows\system32\cmd.exe
(venv windy2) D:\Python projects\windy>python test windy-weather-crawler.py
Welcome to Windy Web Crawler
Please enter your full path to firefox binary in your system(default: /usr/bin/firefox) :
c:/InstalledApps/Mozilla Firefox/firefox.exe
Path taken: c:/InstalledApps/Mozilla Firefox/firefox.exe
Now select the place for weather forecast(ex. chennai, hyderabad, delhi etc... :
kaohsiung
2020-10-13 10:23:40 [py.warnings] WARNING: D:\venv windy2\lib\site-packages\scrapy\spidermiddlewares\offsite.py:65: URLW
arning: allowed domains accepts only domains, not URLs. Ignoring URL entry https://www.windy.com in allowed domains.
 warnings.warn(message, URLWarning)
                                                         020–10–13 10:23:52 [selenium.webdriver.remote.remote connection] DEBUG: Finished Request
   Place Name
                                                        Time of Day | Temp(C)
                                                                               Wind Speed(Kt)
   Kaohsiung Sanmin District
   Kaohsiung International Airport RCKH
   Kaohsiung Lingva District
   Kaohsiung
   Kaohsiung Sanmin District
   Hafen Kaohsiung Qianzhen District
   Kaohsiung International Airport Xiaogang District
   Kaohsiung Arena Gushan District
                                                        >星期三 14
   Kaohsiung Museum of History Yancheng District
                                                          Time of Day | Temp(C)
                                                                               Wind Speed(Kt)
   Kaohsiung Museum of Fine Arts Gushan District
   Select the Index from above table: 3
```

修改 windy-weather-crawler 後才能動 (windy.com 可能改變了網頁寫法)

```
forecast spider.py - D:\venv windy2\Lib\site-packages\windy crawler\spiders\forecast spider.py (3.8.5)
File Edit Format Run Options Window Help
              if airport:
                  max temp, min temp = str(els.text).split('\n')
                  temps list.append(min temp + '-' + max temp)
49
50
51
52
53
54
55
56
57
58
59
60
              else:
                  temps_list.append(els.text)
          for els in winds:
              if airport:
                  trash, min_speed, max_speed = str(els.text).split('\n')
                  winds list.append(min speed + '-' + max speed)
              else:
                  winds list.append(els.text)
          for els in times:
              times list.append(els.text)
          for day in days list:
61
              print('=>' + day)
                                                   OAM 這值在目前的 windy.com 不存在,
62
              #ind = times_list.index('OAM')
63
              #if ind == 0:
                                                   造成 library 跑不起來 (若不改程式的話)
64
                   ind = 8
65
              ind = 8 # 上面3行拿掉,直接將ind設8
66
              current_times, times_list = times_list[:ind], times_list[ind:]
67
              current temps, temps list = temps list[:ind], temps list[ind:]
68
              current winds, winds list = winds list[:ind], winds list[ind:]
69
              display_list = []
70
71
              for i in range(0, len(current_times)):
                  display_list.append([current_times[i], current_temps[i], current_winds[i]])
```

用 xpath 鎖定資料 (以溫度為例)



瀏覽器的「官方」畫面

瀏覽器的開發者模式下的視窗

temps_xpath='//*[@id="detail-data-table"]/tbody/tr[4]/*'

擷取瀏覽器沒顯示的資料

```
小時 😉
     温度
         °C
   降雨量 mm
     風速
     陣風
     風向
                                                                           days xpath='//*[@id="detail-data-table"]/tbody/tr[1]/*'
                                                                           days and dates = browser.find elements by xpath(xpath=days xpath)
                                       Meteog. ➤ Airgram
                                                               ECMWF )
 ○ 1小時預報
                                                                           for els in days and dates:
    → 檢測器
                                                              ብ:記憶體
                                                                              data = els.find element by tag name('div')
Q 搜尋 HTML
                                                                              print(data.get_attribute('data-daydiv'))
          ▼ <table id="detail-data-table" class="grab" data-ref="dataTableTable"
           style="width: 1410px;">

▼ 

            ▼
              ▼ 

               \d>
              > 
                ---
```

```
days list, hours list, temps list, winds list = [], [], [], []
                                                                                             54 dates list, rains list = [], []
File Edit Format Run Options Window Help
  from selenium import webdriver
                                                                                               for els in days and dates:
  I from selenium.webdriver.firefox.firefox binary import FirefoxBinary
                                                                                                   data = els.find element by tag name('div')
  3 from selenium.webdriver.common.kevs import Keys
                                                                                                   print('data:', data.text)
                                                                                                   days list.append(data.text)
  5 from selenium.webdriver.firefox.options import Options
                                                                                                   print(data.get attribute('data-daydiv'))
                                                                                                   dates list.append(data.get attribute('data-daydiy'))
   import time
                                                                                            62 for els in temperatures:
                                                参考 windy-weather-crawler
                                                                                                   temps list.append(els.text)
 9 from json excel converter import Converter
                                                                                             64 for els in winds:
10 from json excel converter.xlsx import Writer
                                                                                                   winds list.append(els.text)
11 import ison
                                                                                            66 for els in hours:
                                                後全新改寫程式
                                                                                                   hours_list.append(els.text) 因為網頁儲存時間的原始格式
13 URL HOME = "https://www.windv.com"
                                                                                            69 for els in rains:
                                                                                                   els in rains:
rains_list.append(els.text) 不適合表格的產生,所以新增
15 #options = Options()
16 #options.add_argument('-headless')
17 binary = FirefoxBinary(r'C:\InstalledApps\Mozilla Firefox\firefox.exe')
                                                                                             72 five days = []
18 browser = webdriver.Firefox(firefox binary=binary)
                                                                                             73 last hour = 0
19 #browser = webdriver.Firefox(firefox binary=binary, options=options) #headless
                                                                                             74 date index = 0
20 browser.get(URL_HOME)
                                                                                               for i, hour in enumerate(hours list):
                                                                                                   item = \{\}
22 search = browser.find element by xpath(xpath='//*[@id="q"]')
                                                                                                   # combine 'date' and 'hour' to generate 'time'
23 search.send keys('kaohsiung')
24 print('Sending Search Data and retrieving suggestions, Please wait...')
                                                                                                   this hour = int(hour)
                                                                                                   if this hour < last hour:
25 time.sleep(2)
26 | results = browser.find_elements_by_xpath(xpath='//*[@id="search"]/div[4]/div[1]/*')
                                                                                                       date index += 1
                                                                                                   this date = dates list[date index] + '-{:02d}'.format(this hour)
                                                                                                   last hour = this hour
 28 for i, place in enumerate(results):
                                           尋找輸入框並填入地點
                                                                                                   item['time'] = this date
       print(i, place.text)
                                                                                                   item['day'] = days list[date index]
31 # select index
                                                                                                   item['hour'] = hour
32 search.send keys(Keys.ARROW_DOWN)
                                                                                                   item['temp'] = temps_list[i]
                                          模擬按下「方向鍵」四次
33 search.send keys(Keys.ARROW DOWN)
                                                                                                   item['rain'] = rains_list[i]
34 search.send keys(Keys.ARROW DOWN)
                                                                                                   item['wind'] = winds list[i]
35 search.send keys(Keys.ARROW DOWN)
                                          和最後按下「 Enter 」
36 time.sleep(\overline{2})
                                                                                                   five days.append(item)
37 search.send_keys(Keys.ENTER)
38 time.sleep(2)
                                                                                            94| #meteog xpath = '/html/body/div[4]/div[2]/div[6]/div[2]/div[4]' # 語法較死,會掛
                                                                 測試改變選單
                                                                                             95 meteog xpath='//*[@id="detail-box"]/div[2]/div[4]' # 語法較不特定
40 days_xpath='//*[@id="detail-data-table"]/tbody/tr[1]/*'
                                                                                             96 meteog = browser.find_element_by_xpath(xpath=meteog_xpath)
41 hours xpath='//*[@id="detail-data-table"]/tbody/tr[2]/*'
                                                                                             97 meteog.click()
42 temps_xpath='//*[@id="detail-data-table"]/tbody/tr[4]/*'
43 winds xpath='//*[@id="detail-data-table"]/tbody/tr[6]/*'
                                                                                            99 with open('test4.json', mode='wt', encoding='utf-8') as f:
                                                                       產生 json~
                                                                                                   #f.write(json.dumps(five_days, indent=2)) # generate \uxxxx instead of 中文
45 days_and_dates = browser.find_elements_by_xpath(xpath=days_xpath)
                                                                                                   json.dump(five days, fp=f, indent=2, ensure ascii=False)
46 hours = browser.find elements by xpath(xpath=hours xpath)
47 temperatures = browser.find_elements_by_xpath(xpath=temps_xpath)
                                                                                            103 filename excel = 'test4.xlsx'
                                                                    產生 excel
48 winds = browser.find elements by xpath(xpath=winds xpath)
                                                                                            104 conv = Converter()
                                                                                            105 conv.convert(five_days, Writer(file=filename_excel))
50 rains xpath='//*[@id="detail-data-table"]/tbody/tr[5]/*'
51 rains = browser.find_elements_by_xpath(xpath=rains_xpath)
                                                                                            107 | #browser.close()
```

windy firefox selenium SimulateBrowser.py - D:\Python projects\windy\windy firefox selenium SimulateBrowser.py

	全 ▼		В	С	D	E	F
1	tin		day	hour	temp	rain	wind
2	20:	02	星期一 12	2	23°	0.5	1
3	2020-10-12-	05	星期一 12	5	22°	1	2
4	2020-10-12-	-08	星期一 12	8	26°		0
5	2020-10-12-	11	星期一 12	11	28°		2
6	2020-10-12-	14	星期一 12	14	28°	0.5	3
7	2020-10-12-	17	星期一 12	17	26°	0.8	3
8	2020-10-12-	20	星期一 12	20	24°	0.6	2
9	2020-10-12-	23	星期一 12	23	24°	0.7	1
10	2020-10-13-	02	星期二 13	2	23°	0.2	1
11	2020-10-13-	05	星期二 13	5	22°	0.2	1
12	2020-10-13-	-08	星期二 13	8	25°		1
13	2020-10-13-	11	星期二 13	11	27°		2
14	2020-10-13-	14	星期二 13	14	27°	0.6	1
15	2020-10-13-	17	星期二 13	17	25°	0.8	1
16	2020-10-13-	20	星期二 13	20	24°	0.4	1
17	2020-10-13-	23	星期二 13	23	24°		1
18	2020-10-14-	02	星期三 14	2	24°	0.5	1
19	2020-10-14-	05	星期三 14	5	23°	0.4	2
20	2020-10-14-	-08	星期三 14	8	25°		0
21	2020-10-14-	11	星期三 14	11	27°	0.9	1
22	2020-10-14-	14	星期三 14	14	25°	3.9	3
23	2020-10-14-	17	星期三 14	17	24°	1.4	1
24	2020-10-14-	20	星期三 14	20	23°	0.2	2
25	2020-10-14-	23	星期三 14	23	22°	0.2	1
26	2020-10-15-	02	星期四 15	2	22°		2
27	2020-10-15-	05	星期四 15	5	21°		2
28	2020-10-15-	-08	星期四 15	8	24°		0
29	2020-10-15-	11	星期四 15	11	27°		2
30	2020-10-15-	14	星期四 15	14	27°	0.2	2
31	2020-10-15-	17	星期四 15	17	25°	0.2	1
32	2020-10-15-	20	星期四 15	20	23°	0.2	3
33	2020-10-15-	23	星期四 15	23	23°	0.7	1
34	2020-10-16-	02	星期五 16	2	22°	0.7	2
35	2020-10-16-	05	星期五 16	5	21°		1
36	2020-10-16-	-08	星期五 16	8	25°		1
37	2020-10-16-	11	星期五 16	11	26°		1
38	2020-10-16-	14	星期五 16	14	26°	0.6	2
39	2020-10-16-	17	星期五 16	17	24°	1.6	1
40	2020-10-16-	20	星期五 16	20	23°	2.4	2

整理成表格,方便

程式進一步處理



星期一

直接輸入 GPS 座標

```
windy_firefox_selenium_InputGpsLocation.py - D:\Python_projects\windy\wi... —

File Edit Format Run Options Window Help

browser = webdriver.Firefox(firefox_binary=binary)
#browser = webdriver.Firefox(firefox_binary=binary, options=options) #h
browser.get(URL_HOME)

gps_pos = '22.6874, 120.3074' # 22.6874° N, 120.3074° E 左營高鐵
search = browser.find_element_by_xpath(xpath='//*[@id="q"]')
search.send_keys(gps_pos)
print('Sending Search Data and retrieving suggestions, Please wait...')
time.sleep(2)

search.send_keys(Keys.ARROW_DOWN)
time.sleep(2)
search.send_keys(Keys.ENTER)
time.sleep(2)
```

windy.com 除了可以輸入地名查詢外,也可以 直接輸入 GPS 座標,不過不是所有格式都支援

輸入帳號和密碼以獲得每小時資料

1. login

```
browser.get(UKL HOME)
time.sleep(2)
LOGIN_REGISTER = 'div[id="login"]'
USERNAME ENTRY = 'input[id="username"]'
USERNAME = 'xx'
PASSWORD ENTRY = 'input[id="password"]'
PASSWORD = 'XXX'
LOGIN BUTTON = 'button[id="login"]'
register login = browser.find element by css selector(LOGIN REGISTER)
if register login:
    register_login.click()
    time.sleep(2)
    username = browser.find element by css selector(USERNAME ENTRY)
    time.sleep(1)
    username.send keys(USERNAME)
    username = browser.find element by css selector(PASSWORD ENTRY)
    time.sleep(1)
    username.send keys(PASSWORD)
    login = browser.find element by css selector(LOGIN BUTTON)
    login.click()
    time.sleep(2)
```

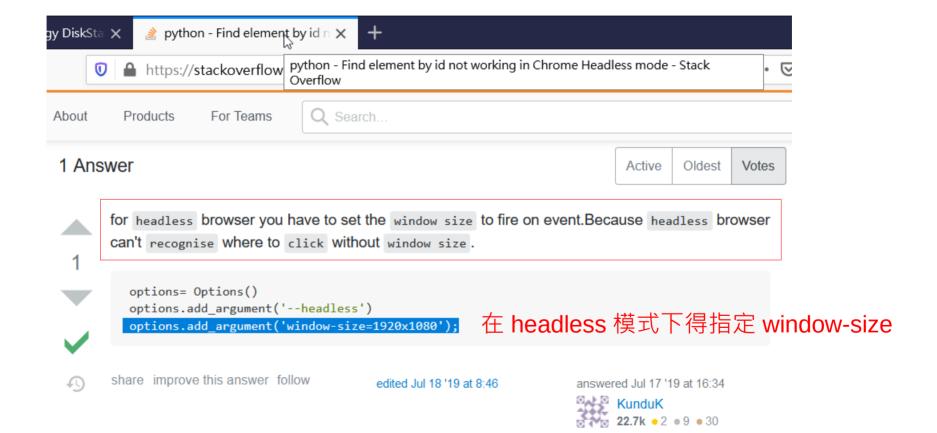
2. switch mode

```
HOUR_SWITCH = 'div[data-t="D_1H_FORECAST"]'
hour_switch = browser.find_element_by_css_selector(HOUR_SWITCH)
hour_switch.click()
time.sleep(3) # for browser to load data
```

3. logout

```
IMG_ICON = 'img[class="avatar"]'
img_icon = browser.find_element_by_css_selector(IMG_ICON)
img_icon.click()
time.sleep(1)
LOGOUT = 'a[href="#logout"]'
logout = browser.find_element_by_css_selector(LOGOUT)
logout.click()
time.sleep(1)
```

Chrome 下的 headless Mode



讓程式在群暉主機上執行成為可能



joyzoursky/python-chromedriver &

By joyzoursky • Updated 3 months ago

Python with Chromedriver, for running automated tests

Container

感謝作者

Overview

Tags

Dockerfile

Builds

docker-python-chr@medriver

Python with Chromedriver, for running automated tests

Quick Try

 $\$ \ \, \text{git clone https://github.com/joyzoursky/docker-python-chromedriver.git}$

\$ cd docker-python-chromedriver

/usr/wrokspace# python test_script.py

 $\$ docker run -it -w /usr/workspace -v $\$ (pwd):/usr/workspace joyzoursky/python-chromedriver:3.7 bash /usr/wrokspace# pip install selenium

Image includes

需要的工具都包裝在一起了

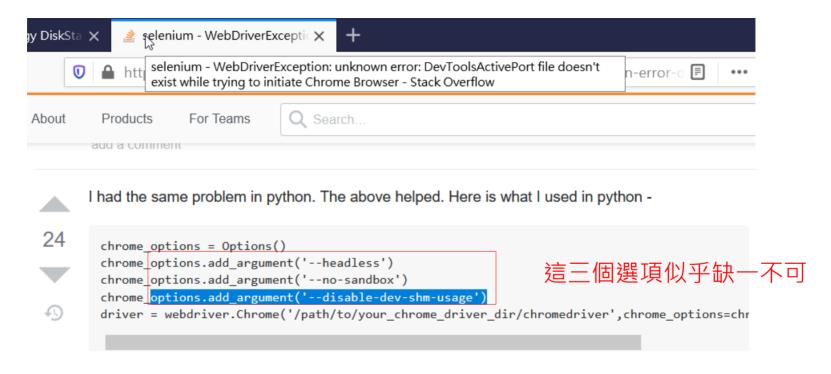
- Python (Debian or Alpine based)
- Google Chrome
- Chromedriver
- Selenium (in some versions)
- Xvfb (in some versions)

Versions

Python 3 (Debian-based)

- 3.8, latest (based on python:3.8, Dockerfile)
- 3.8-selenium (based on python:3.8 with selenium installed, Dockerfile)

在群暉的 docker 環境下的限制



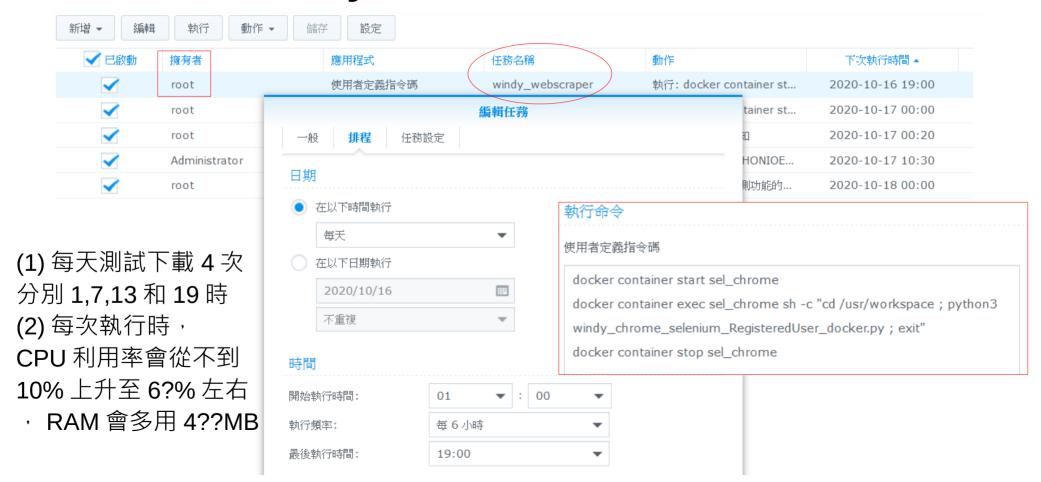
share improve this answer follow

answered Jun 13 '18 at 1:15

Vinodh Periyasamy

241 • 1 • 2

將 Windy 爬蟲納入任務排程表













首頁 / Frequently Asked Questions / When and how often are the weather data updated?



When and how often are the weather data updated?

公開的每3小時間隔資料似乎是一天兩次 更新,付費的預計 2020 底每天 4 次更新

TomSlavkovsky 2016年12月2日 下午11:29 ☑

The table below describes when and how often are the weather models updated. All times are in UTC.

Model / Runtime	Update interval	0:00	6:00	12:00	18:00
ECMWF 0.1°	12h	07:15	-	19:15	-
GFS	6h &	06:15	12:15	18:15	00:15
NESDIS	5 days	-	-	-	
WaveWatch3	6h	06:00	12:00	18:00	00:00
NEMS Europe	12h	08:30	-	20:30	-
NAM CONUS	6h	04:15	10:15	16:15	22:15
NAM Alaska	6h	03:15	09:15	15:15	21:15
NAM Hawaii	6h	04:15	10:15	16:15	22:15
ICON-EU	6h	04:30	10:30	16:30	22:30

無法完全掌控 loading time

120 waves_list = []
121 * for els in waves:

122

waves list.append(els.text)

```
63 #gps pos = '22.6874, 120.3074' # 22.6874° N, 120.3074° E 左營高鐵
search = browser.find element by xpath(xpath='//*[@id="q"]')
65 search.send keys(gps pos)
66 print('Sending Search Data and retrieving suggestions, Please wait...')
67 time.sleep(2)
68
69 search.send keys(Keys.ARROW DOWN)
70 time.sleep(2)
71 search.send keys(Keys.ENTER)
72 time.sleep(10) # 2 may fail, 5 seems ok
                                          1. 輸入 GPS 座標後,第一次載入資料
74 HOUR SWITCH = 'div[data-t="D 1H FORECAST"]'
75 hour switch = browser.find element by css selector(HOUR SWITCH)
76 hour switch.click()
77 time.sleep(5) # for browser to load data, 3 seems ok
                                                     2. 從「3小時預報」切換成1小時
113 # switch to waves mode and extract forecast records
114 waves select xpath='//*[@id="detail-box"]/div[2]/div[2]'
115 waves select = browser.find element by xpath(xpath=waves select xpath)
116 waves select.click()
117 <u>time.sleep(5) # 2 secs failed</u> 3. 從「基本」模式切到「波浪」模式
118 waves xpath='//*[@id="detail-data-table"]/tbody/tr[4]/*' # same as temps
119 waves = browser.find elements by xpath(xpath=waves xpath)
```