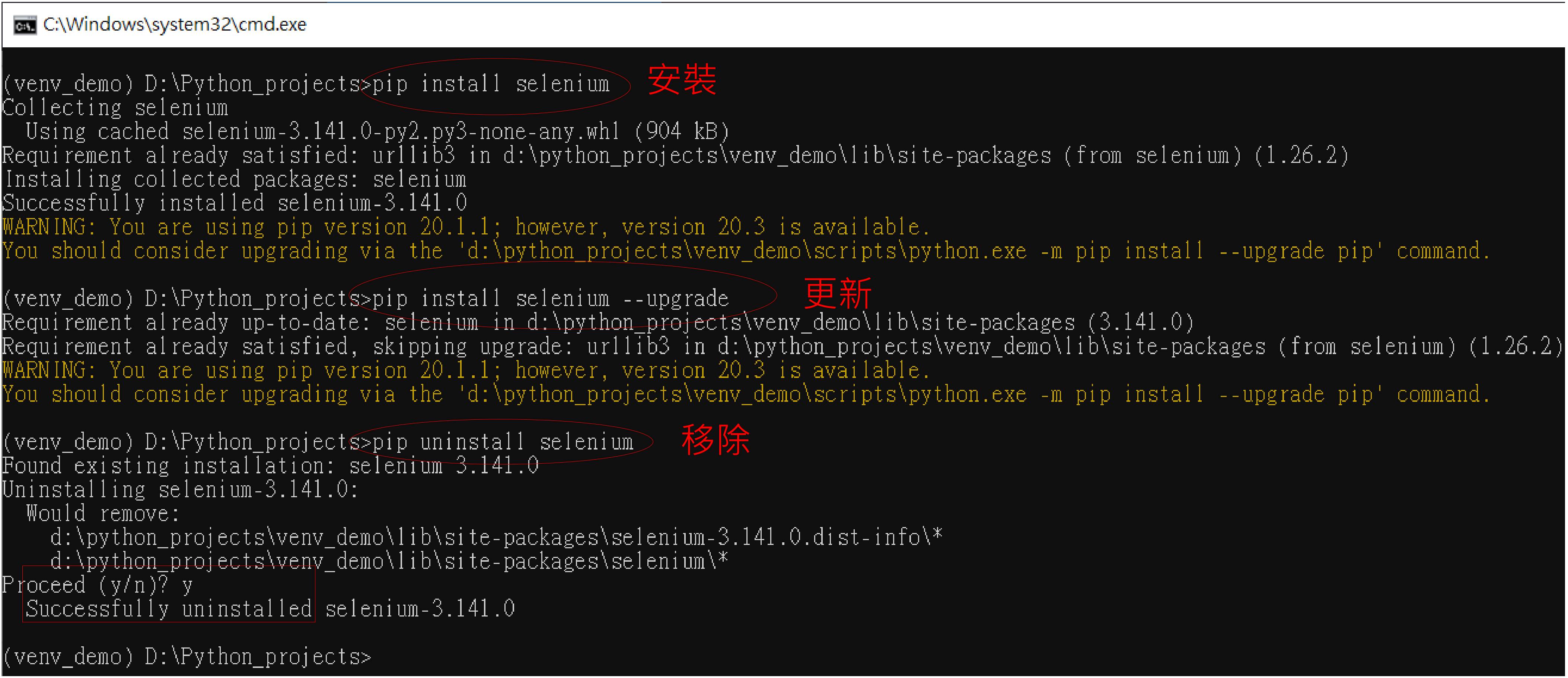
安裝 Selenium 和 WebDriver 來操作瀏覽器

網路爬蟲要應付的三種類型網站

- API (應用程式介面， Application Programming Interface)
  - Public : 中央氣象局
    - https://opendata.cwb.gov.tw/fileapi/v1/opendataapi/{}?Authorization={}&format={}
  - Private : 喜馬拉雅音頻分享平台
- 靜態網頁
  - 一般的小說網站：例如「愛下電子書」 https://tw.aixdzs.com/
- 動態網頁 (一般會安裝 python 的 Selenium 模組和 Chrome WebDriver)
  - 一般會牽扯到 java script 程式語言：例如 windy.com

安裝和更新 Selenium



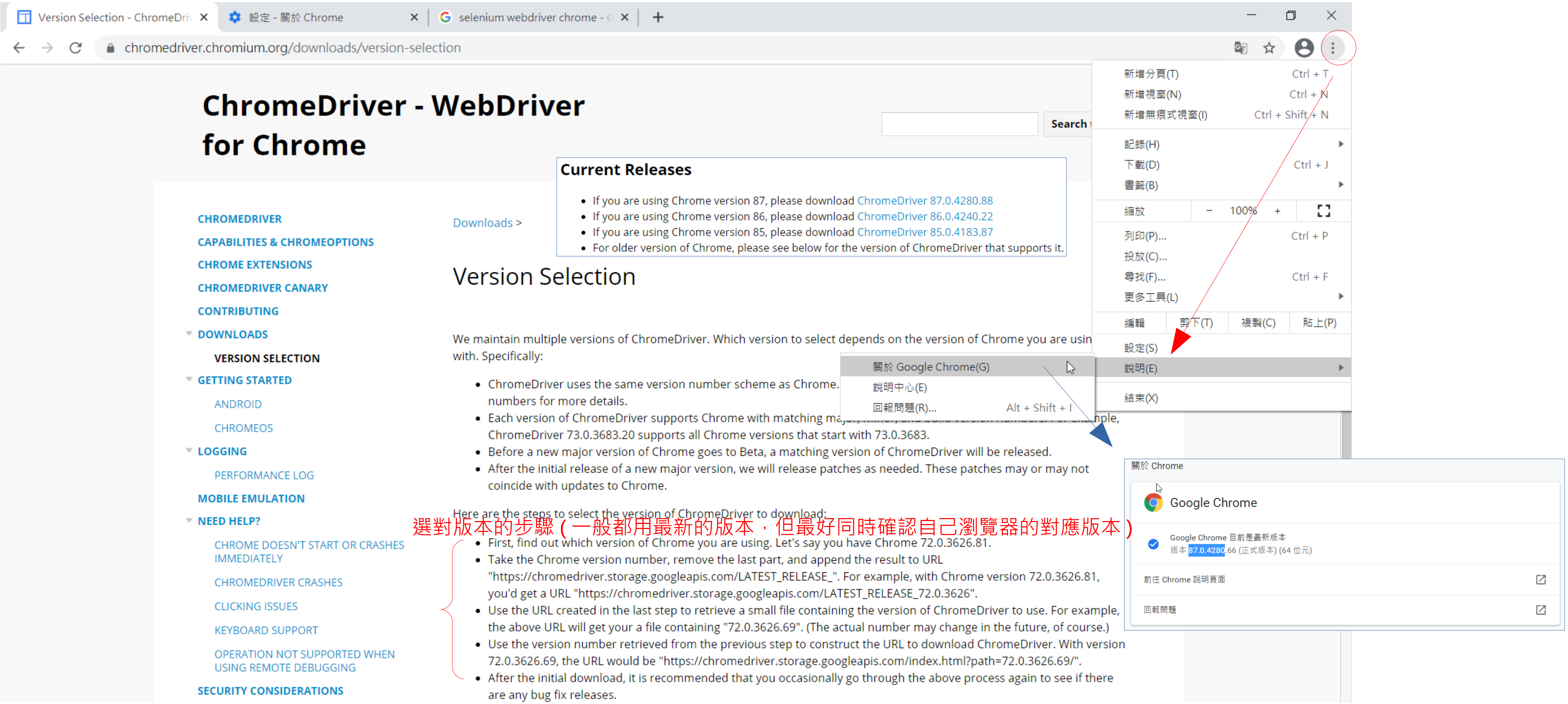
```
(venv_demo) D:\Python_projects>pip install selenium
Collecting selenium
  Using cached selenium-3.141.0-py2.py3-none-any.whl (904 kB)
Requirement already satisfied: urllib3 in d:\python_projects\venv_demo\lib\site-packages (from selenium) (1.26.2)
Installing collected packages: selenium
Successfully installed selenium-3.141.0
WARNING: You are using pip version 20.1.1; however, version 20.3 is available.
You should consider upgrading via the 'd:\python_projects\venv_demo\scripts\python.exe -m pip install --upgrade pip' command.

(venv_demo) D:\Python_projects>pip install selenium --upgrade
Requirement already up-to-date: selenium in d:\python_projects\venv_demo\lib\site-packages (3.141.0)
Requirement already satisfied, skipping upgrade: urllib3 in d:\python_projects\venv_demo\lib\site-packages (from selenium) (1.26.2)
WARNING: You are using pip version 20.1.1; however, version 20.3 is available.
You should consider upgrading via the 'd:\python_projects\venv_demo\scripts\python.exe -m pip install --upgrade pip' command.

(venv_demo) D:\Python_projects>pip uninstall selenium
Found existing installation: selenium 3.141.0
Uninstalling selenium-3.141.0:
  Would remove:
    d:\python_projects\venv_demo\lib\site-packages\selenium-3.141.0.dist-info\*
    d:\python_projects\venv_demo\lib\site-packages\selenium\*
Proceed (y/n)? y
  Successfully uninstalled selenium-3.141.0

(venv_demo) D:\Python_projects>
```

安裝

更新

移除

下載 WebDriver (Google Chrome)



ChromeDriver - WebDriver for Chrome

Current Releases

- If you are using Chrome version 87, please download ChromeDriver 87.0.4280.88
- If you are using Chrome version 86, please download ChromeDriver 86.0.4240.22
- If you are using Chrome version 85, please download ChromeDriver 85.0.4183.87
- For older version of Chrome, please see below for the version of ChromeDriver that supports it.

Version Selection

We maintain multiple versions of ChromeDriver. Which version to select depends on the version of Chrome you are usin with. Specifically:

- ChromeDriver uses the same version number scheme as Chrome. numbers for more details.
- Each version of ChromeDriver supports Chrome with matching ma nple, ChromeDriver 73.0.3683.20 supports all Chrome versions that start with 73.0.3683.
- Before a new major version of Chrome goes to Beta, a matching version of ChromeDriver will be released.
- After the initial release of a new major version, we will release patches as needed. These patches may or may not coincide with updates to Chrome.

Here are the steps to select the version of ChromeDriver to download:

選對版本的步驟（一般都用最新的版本，但最好同時確認自己瀏覽器的對應版本）

- First, find out which version of Chrome you are using. Let's say you have Chrome 72.0.3626.81.
- Take the Chrome version number, remove the last part, and append the result to URL "https://chromedriver.storage.googleapis.com/LATEST_RELEASE_". For example, with Chrome version 72.0.3626.81, you'd get a URL "https://chromedriver.storage.googleapis.com/LATEST_RELEASE_72.0.3626".
- Use the URL created in the last step to retrieve a small file containing the version of ChromeDriver to use. For example, the above URL will get your a file containing "72.0.3626.69". (The actual number may change in the future, of course.)
- Use the version number retrieved from the previous step to construct the URL to download ChromeDriver. With version 72.0.3626.69, the URL would be "https://chromedriver.storage.googleapis.com/index.html?path=72.0.3626.69/".
- After the initial download, it is recommended that you occasionally go through the above process again to see if there are any bug fix releases.

關於 Chrome

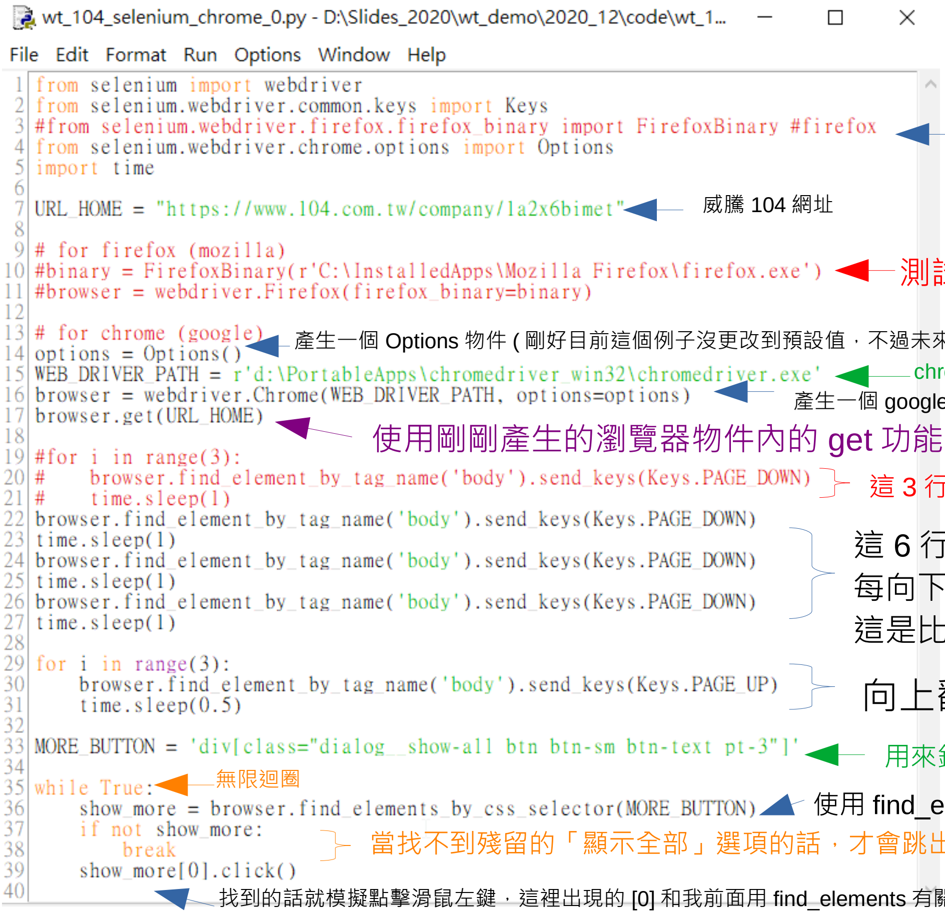
Google Chrome

Google Chrome 目前是最新版本
版本 87.0.4280.66 (正式版本) (64 位元)

操控威騰 104 網頁的程式碼

<span style="color:red">挑戰</span>：
註解的地方有 3 行是 Mozilla Firefox 瀏覽器的。試著還原並做些修改，再搭配 firefox 版本的 WebDriver 來操作。

程式開啟網頁後會模擬鍵盤的 PgDn(PAGE_DOWN)，操作 3 次，間隔 1 秒，之後 PgUp 也是 3 次，間隔 0.5 秒。最後，程式會尋找網頁的「顯示全部」選項，找到後模擬滑鼠左鍵的執行 (click)


wt_104_selenium_chrome_0.py - D:\Slides_2020\wt_demo\2020_12\code\wt_1...
File Edit Format Run Options Window Help

```python
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
#from selenium.webdriver.firefox.firefox_binary import FirefoxBinary #firefox
from selenium.webdriver.chrome.options import Options
import time

URL_HOME = "https://www.104.com.tw/company/1a2x6bimet"

# for firefox (mozilla)
#binary = FirefoxBinary(r'C:\InstalledApps\Mozilla Firefox\firefox.exe')
#browser = webdriver.Firefox(firefox_binary=binary)

# for chrome (google)
options = Options()
WEB_DRIVER_PATH = r'd:\PortableApps\chromedriver_win32\chromedriver.exe'
browser = webdriver.Chrome(WEB_DRIVER_PATH, options=options)
browser.get(URL_HOME)

#for i in range(3):
#    browser.find_element_by_tag_name('body').send_keys(Keys.PAGE_DOWN)
#    time.sleep(1)
browser.find_element_by_tag_name('body').send_keys(Keys.PAGE_DOWN)
time.sleep(1)
browser.find_element_by_tag_name('body').send_keys(Keys.PAGE_DOWN)
time.sleep(1)
browser.find_element_by_tag_name('body').send_keys(Keys.PAGE_DOWN)
time.sleep(1)

for i in range(3):
    browser.find_element_by_tag_name('body').send_keys(Keys.PAGE_UP)
    time.sleep(0.5)

MORE_BUTTON = 'div[class="dialog__show-all btn btn-sm btn-text pt-3"]'

while True:
    show_more = browser.find_elements_by_css_selector(MORE_BUTTON)
    if not show_more:
        break
    show_more[0].click()

```

載入 4 個「模組」(?)，其中
# 開頭那行代表 python 不會執行

威騰 104 網址

測試 firefox 瀏覽器用的，暫時被「註解」掉

產生一個 Options 物件 (剛好目前這個例子沒更改到預設值，不過未來會設定其他值)

chrome webdriver 在我電腦內的路徑，大家的位置應該都不一樣

產生一個 google chrome 瀏覽器物件

使用剛剛產生的瀏覽器物件內的 get 功能去開啟威騰的 104 網頁

這 3 行被註解掉，用迴圈的方式做 PAGE_DOWN 三次

這 6 行和上面被拿掉的 3 行做相同的事，每向下翻頁一次，就等 1 秒，如此重複 3 次。這是比較「笨」的寫法。

向上翻頁 3 次，每次等 0.5 秒 (用到縮排)

用來鎖定網頁「顯示全部」選項的資訊 (下一頁詳解)

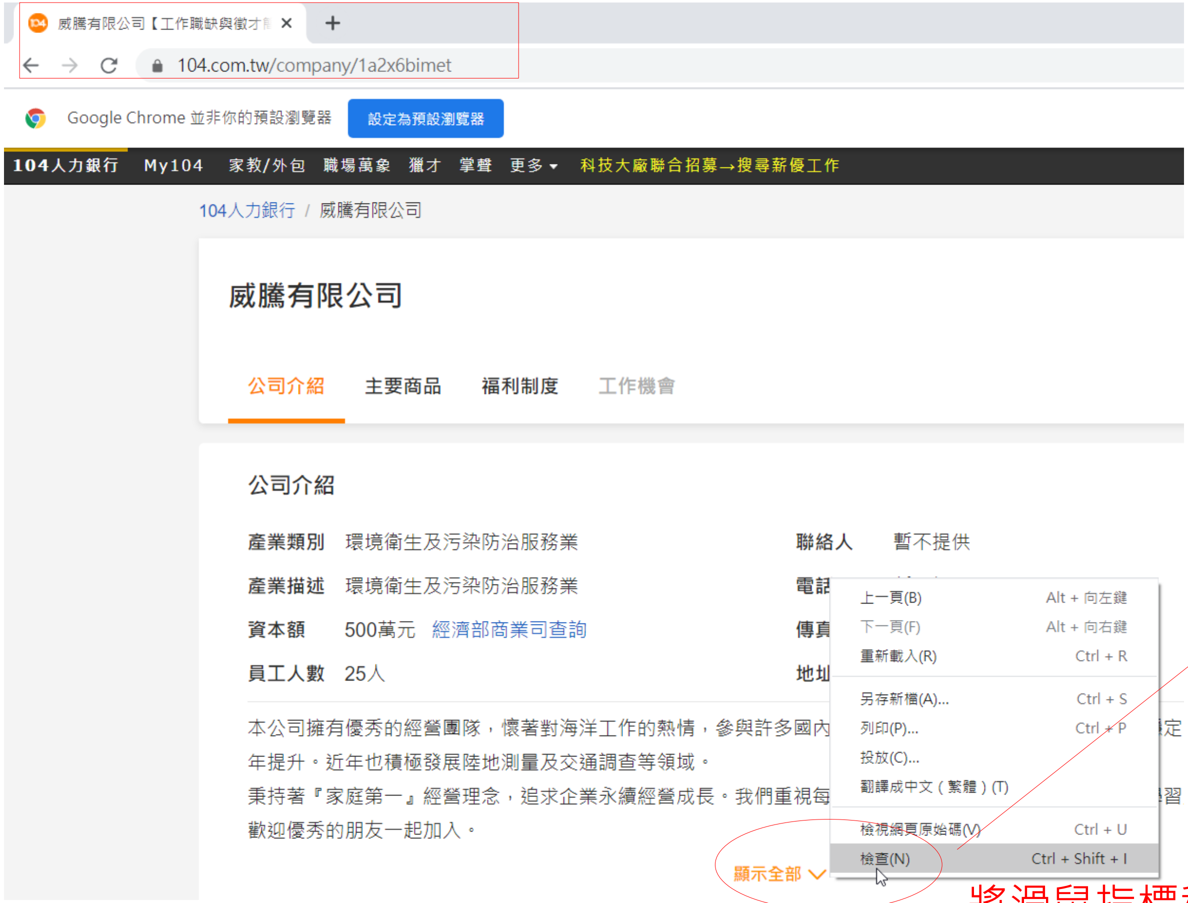
無限迴圈

使用 find_elements_by_css_selector 功能來尋找「顯示全部」

當找不到殘留的「顯示全部」選項的話，才會跳出永無止盡的迴圈

找到的話就模擬點擊滑鼠左鍵，這裡出現的 [0] 和我前面用 find_elements 有關，改成 find_element 單數才合理，有興趣人把它改掉吧～

進入瀏覽器的開發者模式



威騰有限公司【工作職缺與徵才
104.com.tw/company/1a2x6bimet

Google Chrome 並非你的預設瀏覽器 設定為預設瀏覽器

104人力銀行 My104 家教/外包 職場萬象 獵才 掌聲 更多 科技大廠聯合招募→搜尋薪優工作

104人力銀行 / 威騰有限公司

威騰有限公司

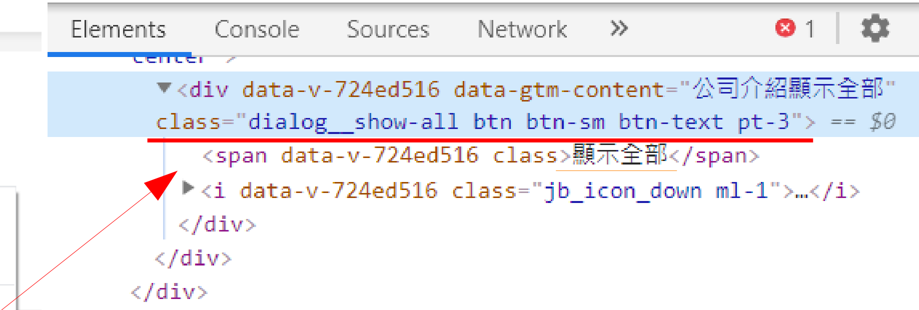
公司介紹 主要商品 福利制度 工作機會

公司介紹

產業類別	環境衛生及污染防治服務業	聯絡人	暫不提供
產業描述	環境衛生及污染防治服務業	電話	
資本額	500萬元 經濟部商業司查詢	傳真	
員工人數	25人	地址	

本公司擁有優秀的經營團隊，懷著對海洋工作的熱情，參與許多國內 定
年提升。近年也積極發展陸地測量及交通調查等領域。
秉持著『家庭第一』經營理念，追求企業永續經營成長。我們重視每 習
歡迎優秀的朋友一起加入。

顯示全部



```
  ▼<div data-v-724ed516 data-gtm-content="公司介紹顯示全部"
  class="dialog__show-all btn btn-sm btn-text pt-3"> == $0
      <span data-v-724ed516 class>顯示全部</span>
    ▶<i data-v-724ed516 class="jb_icon_down ml-1">…</i>
    </div>
  </div>
</div>
```

將滑鼠指標移到「顯示全部」上方，之後按右鍵