動態網頁擷取

朱克剛

何謂DHTML

- DHTML ➔ Dynamic HTML
- 瀏覽器上看到的畫面是透過 JavaScript 動態產生的，而不是事先透過 HTML 標籤寫好的
  - *例如：使用者按下按鈕後改變瀏覽器上顯示的內容*
  - *例如：透過 AJAX 技術抓回資料後顯示在瀏覽器上*

範例

```
<html>
<script language="javascript">
function onload(){
   document.getElementById("main").innerHTML ="Hello World";
}
</script>
<body onload="onload()">
   <div id="main"></div>
</body>
</html>
```

第 1 步：安裝 Selenium 模組

```
$ pip install Selenium
```

第 2 步：下載瀏覽器驅動程式

- Firefox
  - https://github.com/mozilla/geckodriver/releases
- Chrome
  - https://sites.google.com/a/chromium.org/chromedriver/downloads
- PhontomJS
  - http://phantomjs.org
  - *Selenium 已不支援*

第3步：程式碼

```
from selenium import webdriver

## Firefox
options = webdriver.firefox.options.Options()
options.add_argument("--headless")
driver_path = "/Users/ckk/Downloads/geckodriver"
driver = webdriver.Firefox(executable_path=driver_path, options=options)

## Chrome
#options = webdriver.chrome.options.Options()
#options.add_argument("--headless")
#driver_path = "/Users/ckk/Downloads/chromedriver"
#driver = webdriver.Chrome(executable_path=driver_path, options=options)

## PhantomJS
#driver_path = "/Users/ckk/Downloads/phantomjs"
#driver = webdriver.PhantomJS(executable_path=driver_path)

driver.get("http://localhost/test.html")
pageSource = driver.page_source
print(pageSource)
driver.close()
```

第４步：處理後的網頁內容

```
<html>
<head>
<script language="javascript">
function onload(){
   document.getElementById("main").innerHTML ="Hello World";
}
</script>
</head>
<body onload="onload()">
   <div id="main">Hello World</div>
</body>
</html>
```

AJAX 處理

- 使用 AJAX 時，資料回來比較慢，所以需要 sleep 一小段時間

```
from selenium import webdriver
import time

…

driver.get("http://...")
time.sleep(3)
pageSource = driver.page_source
```