



Day 20

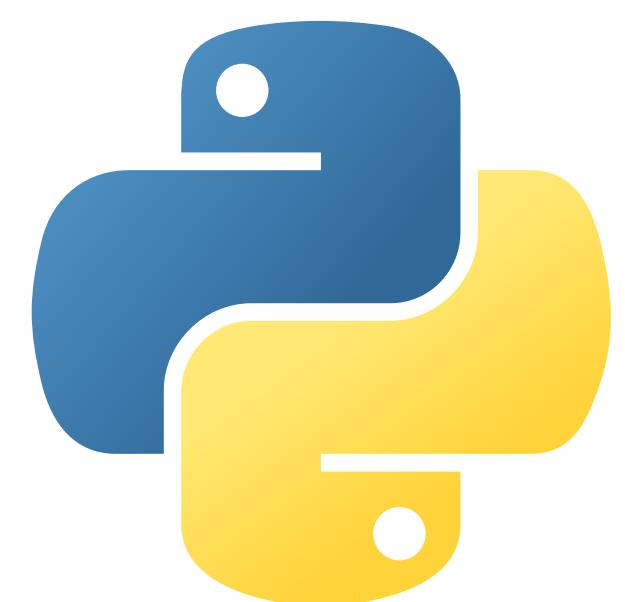
cupay

動態網頁資料爬蟲

動態網頁爬蟲 - API Request



出題教練：張維元



python

本日知識點目標

- 了解 API Request 用於動態網頁爬蟲的原理
- 能夠使用 API Request 撰寫動態網頁爬蟲

第二種動態網頁爬蟲策略

關於這種利用到 JavaScript 的非同步特性載入更多資料的網頁稱為動態網頁。而爬蟲程式也會因為沒有執行到 JavaScript 導致資料不完全的現象。

第二種解法透過利用 Python 模仿 JavaScript 呼叫 API 這個動作，也可以達到動態取得資料的目的。



利用 開發者工具 觀察 JavaScript 行為

可以從瀏覽器的開發中工具中提供的「Network」功能，去查看所有網頁中的傳輸行為。這種透過 JavaScript 動態載入的請求，就可以從中觀察到。接著就可以利用我們前面談到的 API 爬蟲方式進行。



本署新版空氣品質監測網（網址：<https://airtw.epa.gov.tw>）即日起試行上線，歡迎瀏覽並提供使用意見。

[環保署](#) \ [空氣品質監測網](#) \ [資料查詢與服務](#) \ [空氣品質測站](#) \ [月平均值查詢](#)

月平均值查詢

- 一、初步統計，未經驗証，僅供參考
二、無效值定義：月平均值計算需日平均有效值 ≥ 20 天，方為有效月平均值，未達 20 天者，則視為無效月平均值。

測站：高雄市-三民 年度：2015

查詢

[TOP](#)

- [空氣品質監測網](#)
- [空氣品質監測](#)
- [沙塵網站](#)
- [河川揚塵監測](#)
- [紫外線監測](#)
- [品質保證作業](#)
- [資料查詢與服務](#)
- [空氣品質測站](#)
- [即時值查詢](#)
- [不良日數月報表\(PSI\)](#)
- [不良日數月報表\(AQI\)](#)
- [污染物測值月報表](#)

The screenshot shows the Network tab in the Chrome DevTools Performance panel. At the top, there are icons for Elements, Console, Sources, Network (which is underlined), Performance, and Memory. Below the tabs are various controls: a red circle, a grey circle with a slash, a video camera icon, a funnel icon, a magnifying glass icon, and a 'View:' dropdown with grid and list icons. There are also checkboxes for 'Group by frame', 'Preserve log', and 'D'. A 'Filter' input field is followed by a row of buttons: 'Hide data URLs' (unchecked), 'All' (disabled), 'XHR' (selected and highlighted in grey), 'JS', 'CSS', 'Img', 'Media', and 'For'. The main area displays a horizontal timeline with six network requests. The first request is a green bar labeled '500 ms'. The second is a light blue bar labeled '1000 ms'. The third is a light orange bar labeled '1500 ms'. The fourth is a light purple bar labeled '2000 ms'. The fifth is a light green bar labeled '2500 ms'. The sixth is a light pink bar labeled '3000 ms'. Each bar has a small vertical line extending downwards.

Name	Status	Type	Ir

0 / 2 requests | 0 B / 7.6 KB transferred | 0 B / 7.1 KB resources

在網頁上叫出 Console 切換到 Network Tab，中間選 XHR，這裡會記載所有網頁中的 API 呼叫。

本署新版空氣品質監測網（網址：<https://airtw.epa.gov.tw>）即日起試行上線，歡迎瀏覽並提供使用意見。

環保署\空氣品質監測網\資料查詢與服務\空氣品質測站\月平均值查詢

月平均值查詢

一、初步統計，未經驗証，僅供參考
二、無效值定義：月平均值計算需日平均有效值 ≥ 20 天，方為有效月平均值，未達 20 天者，則視為無效月平均值。

測站：高雄市-三民 年度：2015 檢索

監測項目	單位	監測日期	監測值	標註
SO ₂	ppb			無此測項
CO	ppm			無此測項
O ₃	ppb			
PM10	$\mu\text{g}/\text{m}^3$			
NO _x	ppb			

此時點選重新整理，會發現網址沒有動（表示沒有發送新的 HTML 網頁請求），但畫面有新的內容出現，左下角也多了一次新的 API 呼叫。

Network Performance Memory

View: Group by frame Preserve log

Filter Hide data URLs All XHR JS CSS Img Media Font

Name	Status	Type
MonthlyAverage.aspx	200	xhr

1 / 14 requests | 13.1 KB / 25.9 KB transferred | 12.9 KB / 27.8 KB resources



CUPOY

Name	Headers	Preview	Response	Cookies	Timing
MonthlyAverage....	General Request URL: https://taqm.epa.gov.tw/taqm/tw/MonthlyAverage.aspx Request Method: POST Status Code: 200 OK Remote Address: 223.200.80.179:443 Referrer Policy: no-referrer-when-downgrade				
MonthlyAverage....	Response Headers view source Name Headers Preview Response Cookies Timing MonthlyAverage.... User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4453.125 Safari/537.36 X-MicrosoftAjax: Delta=true				
	Form Data view source view URL encoded ScriptManager1: ctl05\$UpdatePanel1 ctl05\$btnQuery __EVENTTARGET: __EVENTARGUMENT: __VIEWSTATE: gu5vTMr3FLvk... zvYna0pQXFR9g7j/JFAz1x1dzjQTTWP4rU+q/1X44qy4qhQ4evHTjnRnLZliTRxHtkR+jwIgl2miw7uYE2GaCfkrJkxNNMz4cXBxvCKBu984iV1MqVb3WUPWz1q6UA... oSBIy9sEYrnLNW8c5RVPtzgbfoYF00Qrzdv5UKe952zrbW9Yd2uFI2eXQxqWLUBj9Uuca/id4ml7lFwNh/6iQf8pMaXMqi0kKE+vqDzz5NhAPmwGM0E3L5ZRG				

點開就可以得到完整的 API 呼叫內容，包含網址，Headers 和資料。

利用 開發者工具 觀察 JavaScript 行為



簡單來說，在一個動態網頁中，我們可以簡單地透過開發者的工具的觀察，知道 JavaScript 發送了哪些請求。換句話說，我們就可以模仿這部分，改用 Python 來發出 API，將原問題簡化為前面講過的 API 存取。

重要知識點複習

- 了解 API Request 用於動態網頁爬蟲的原理
- 能夠使用 API Request 撰寫動態網頁爬蟲



解題時間

LET'S CRACK IT

請跳出 PDF 至官網 Sample Code & 作業

開始解題

