**網路爬蟲 - 學前先備知識**

# 一、何謂網路爬蟲？

網路爬蟲（web crawler / web scraper）是一項網路資料蒐集的技術，以程式自動化（automation）的方式，取代人工重複操作的流程，來達到「資料自動抓取」的目標。

|  |  |
| --- | --- |
|  |  |
| 圖: 網路爬蟲概念，取自https://www.instagram.com/p/Cb12imAvHVR/ | |

# 二、先備知識

以下列出的影片，為學習網路爬蟲前，必須要了解的基本知識：若是先前有一些基礎，則可滑動時間軸，來選擇需要的學習內容：

（一）Python 3.x 版本（建議3.6以上）

由於我們使用Python來開發網路爬蟲，所以必須學習 Python 語法與實作方式。可以參考以下的影片，學習時間約略為4小時4分35秒：

【python】4小時初學者Python教學

<https://www.youtube.com/watch?v=zdMUJJKFdsU>

（二）HTML（Hypertext Markup Language）

網路爬蟲最常取得網頁的內容，而特定內容（例如圖片連結、標題文字等）通常都放在HTML標籤當中，以內文（inner text）或屬性（attribute）的格式來存放，因此需要了解HTML是什麼。可以參考以下的影片，學習時間約略為1小時23分40秒：

【html】1小時初學者教學

<https://www.youtube.com/watch?v=CLUPkcLQm64>

（三）CSS（Cascading Style Sheets）

程式取得HTML後，需要了解HTML之間的階層（節點間的父子、兄弟、上下）關係，於是我們需要使用CSS的選擇器（selector）來取得單一或兩個以上的HTML元素（在某些程式語言，例如JavaScript，我們會把HTML標籤視為元素elements，每個元素都有自己的內文和屬性資料），透過程式存取元素的內文與屬性，例如圖片連結、標題文字等。

需要注意的是，我們並不關注HTML元素到底多寬、背景和文字是什麼顏色等，只要把重點放在選擇器（selector）上面，可以讓程式選到指定的HTML元素，取得指定的資料即可。

可以參考以下的影片，學習時間約略為1小時50分40秒（CSS Selector教學在時間軸1時6分20秒開始）：

【css】2小時初學者教學

<https://www.youtube.com/watch?v=Ml78vnNTBLw>

（四）SQL（Structured Query Language）【optional】

當取得資料，例如頁面標題、超連結或其它描述這筆資料的說明文字等，需要將它們儲存下來，以便記錄過去到現在擷取資料的進度，或是進一步拿來建立重要資訊，此時需要一個資料庫系統（Database Management System，DBMS），協助我們存放資料。

資料通常會存置在結構化（structured）的環境或平台當中，可以想像成類似Excel這種 N x M 的巢狀表格，有自己定義的 column（欄位）和row（每一列資料）。DBMS常見的有 MySQL、SQL Server、PostgresSQL等，為了能夠將資料存放在 DBMS 當中，需要先了解與DBMS溝通的語法，這種語法稱為SQL。可以參考以下的影片，學習時間約略為2小時46分36秒：

【資料庫】SQL 3小時初學者教學

<https://www.youtube.com/watch?v=gvRXjsrpCHw>