**網路爬蟲學習指南**

作者：楊德倫，日期：2022/04/03

# 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>

# 三、教學影片

有了Python、HTML、CSS的基礎之後，就可以開始學習網路爬蟲相關的技術：

|  |  |
| --- | --- |
| **說明** | **連結** |
| 練習參考專案網址  （近期已改版，與影片稍有不同） | <https://github.com/telunyang/python_web_scraping> |

影片當中的Python執行平台，以Jupyter為主，可以安裝Anaconda來建立執行環境，抑或自行google了解Jupyter的安裝與使用；進階使用者則可自訂環境，不受開發環境拘束：

|  |  |
| --- | --- |
| **說明** | **連結** |
| 上課前的環境說明與基礎練習（一） | <https://www.youtube.com/watch?v=vKwIV-5hvkQ> |
| 基礎練習（二） 正規表達式說明（一） | <https://www.youtube.com/watch?v=4RwSZ5_VnW0> |
| * 正規表達式說明（二） * 網路請求與回應 （Request and Response）說明 * HTML、CSS Selector 基礎說明 * Requests 套件操作說明 * BeautifulSoup 說明 | <https://www.youtube.com/watch?v=WdSdrsyrxJk> |
| Requests 套件實務練習（一） | <https://www.youtube.com/watch?v=FrcWpT3NUgo> |
| Requests 套件實務練習（二） | <https://www.youtube.com/watch?v=gYRsJD2FtVM> |
| Requests 套件實務練習（三） 從 00:07:47 開始看。 | <https://www.youtube.com/watch?v=857lgp8KVGE> |

練習結束後，你將可以學到:

1. 將網頁中結構性資料加以整理，將 list 或 dict 變數儲存成 .json 檔案。
2. 讀取前述 .json 檔案，轉成 list 或 dict 變數，繼續使用。
3. 取得 LINE 官方貼圖(真正貼圖檔，高畫質)。
4. 取得所有金庸小說，共 977 (或 976) 回。

接下來，我們將使用瀏覽器自動化工具（selenium & web driver），**模擬我們人類操作瀏覽器**，例如自動訂高鐵票**搶票**、自動預約室內運動場域、還有如**搶訂**疫苗殘劑。

使用selenium & web driver 可以成為 requests 的替代方案，但建議兩個都深入了解，對面不同的案例和情境，使用不同工具。

|  |  |
| --- | --- |
| **說明** | **連結** |
| 1. Selenium 介紹與基礎練習  2. Excel 操作套件 openpyxl | <https://www.youtube.com/watch?v=4mqlEV_PsSk> |
| Selenium 套件實務練習 - YouTube | <https://www.youtube.com/watch?v=FyHv7t2-g9o> |
| Selenium 套件實務練習 - 臺灣證券交易所  (時間: 00:58:33 到 01:29:25) | <https://www.youtube.com/watch?v=TmNftYQjiCk> |
| 額外補充: yt-dlp  (時間: 01:32:03到影片結束) | <https://www.youtube.com/watch?v=TmNftYQjiCk> |
| Selenium 套件實務練習 - Instagram  （第一段: 00:00:00 到 00:58:22） | <https://www.youtube.com/watch?v=TmNftYQjiCk> |
| Selenium 套件實務練習 - Instagram  （第二段: 00:00:00 到 00:55:42） | <https://www.youtube.com/watch?v=kauePzw9uUk> |

在練習中，你將學會:

1. 如何用 Python 操作 Excel 檔，有時候並不一定要把資料存在 json 當中，Excel 也是一種選擇。
2. 如何取得 YouTube 和 Instagram 的重要資訊。
3. 如何取得臺灣證券交易所「外資及陸資買賣超彙總表」的 csv 檔，並且建立、儲存 快照(screenshot)。

# 四、延伸應用【optional】

專案資料夾中，17.ipynb、18.ipynb 是後來新增的，先前有自學過網路爬蟲的同學，有空也可以研究看看，17 說明 ActionChains，18 說明 PyAutoGUI。

# 五、參考資料

[1] 認識網路爬蟲

<https://www.webscrapingpro.tw/what-is-web-scraping/>

[2] 网络爬虫：Data Scraping vs Data Crawling

<https://reurl.cc/vel63e>