01 環境設定與網頁爬蟲初探

工欲善其事，必先利其器。第一章的內容主要分為三部份，首先是開發環境的建置，有鑑於網頁爬蟲程式通常只是資料分析的第一步，取得資料之後，常常需要資料清理與分析，或使用機器學習模型進行預測等工作，因此 1-1 節會介紹如何安裝 Anaconda，一個開源且適合資料科學工作的套件管理系統，並透過其操作介面建立爬蟲程式所需的開發環境，透過 Anaconda 建置開發環境也是我們認為對初學者最友善的方式。接著，1-2 節與 1-3 節介紹常見的整合開發環境（Integrated Development Environment，IDE），PyCharm 與 Jupyter Notebook。IDE 可作為「使用純文字編輯器撰寫程式碼，並透過命令列執行程式」的替代方案，以提昇開發者的產能。最後，1-4 節將講解 HTML 網頁文件的基本架構，並帶領讀者實作第一隻網頁爬蟲程式。

1-1 環境設定及套件安裝：Anaconda

Anaconda 是適合資料科學工作者的開源套件管理系統，針對 Python 使用者提供簡單的安裝與豐富的套件。Anaconda 包含了 Python 的實作環境，以及與網路爬蟲、資料分析、資料視覺化、機器學習相關的多種套件，如網路爬蟲最常使用到的套件 Requests 與 BeautifulSoup，資料分析與視覺化套件 Pandas 與 Matplotlib，自然語言處理套件 nltk 與 spaCy，機器學習與深度學習套件 scikit-learn 與 TensorFlow 等，都可以透過 Anaconda 的圖形環境來安裝。其自帶的 IDE 如 Jupyter Notebook 與 Sypder 等也提供了整合開發環境，讓開發者可以更簡便地逐行執行或分享程式碼內容。本節將介紹如何在 Windows 10 內安裝 Anaconda，建立一個虛擬環境，安裝所需套件並執行 Python 程式（本書使用 Python 3）。需要了解如何在 Mac 安裝 Anaconda 的讀者請參考附錄。