勞動部產業新尖兵計畫

人工智慧金融應用與實務培訓班

課程模組： AI 金融科技課程 - 網路爬蟲技術

1. 認識網路爬蟲

葉建華 (Yeh, Jian-hua)

tdi.jhyeh@tdi.edu.tw
au4290@gmail.com

講次內容

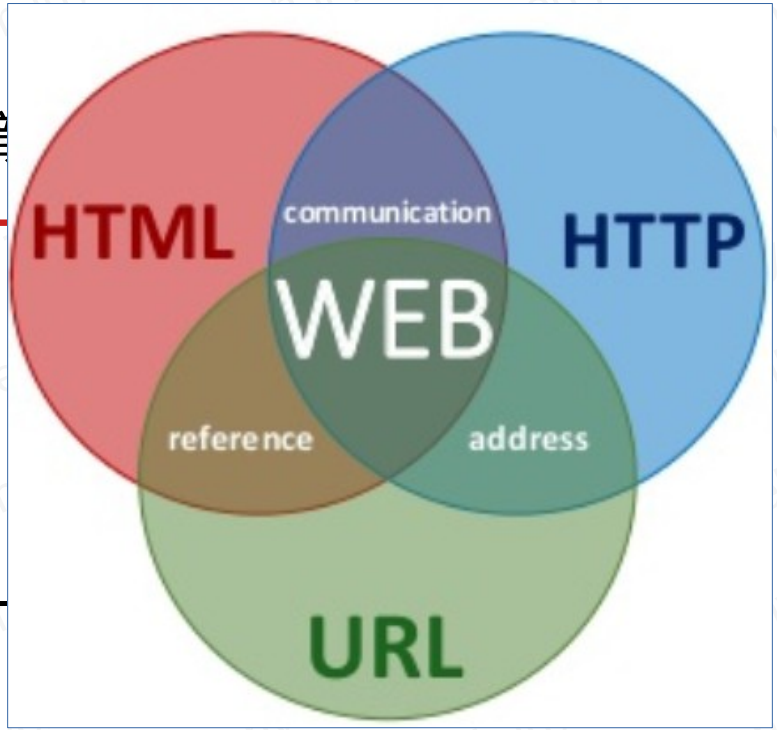
- 認識 Web、 HTML、 URL
- 何謂網路爬蟲
- 搜尋引擎介紹
- 認識 HTTP 與 HTTPS

什麼是 World Wide Web

- World Wide Web 簡稱 WWW
  - 將資訊內容以 HTML 形式進行組織呈現
  - 將 HTML 形式的資訊內容透過 HTTP 協定對外提供存取
  - 電腦與網際網路交換資訊的途徑
  - 這類文件被稱為「Web 文件」，具有唯一的位址，稱為 URL(Uniform Resource Locator, 一致性資源定位器)

什麼是 World Wide Web

- World Wide Web 簡稱 WWW
  - 將資訊內容以 HTML 形式進行組織
  - 將 HTML 形式的資訊內容透過 HTTP
  - 電腦與網際網路交換資訊的途徑
  - 這類文件被稱為「Web 文件」，
    URL(Uniform Resource Locator, -



講次內容

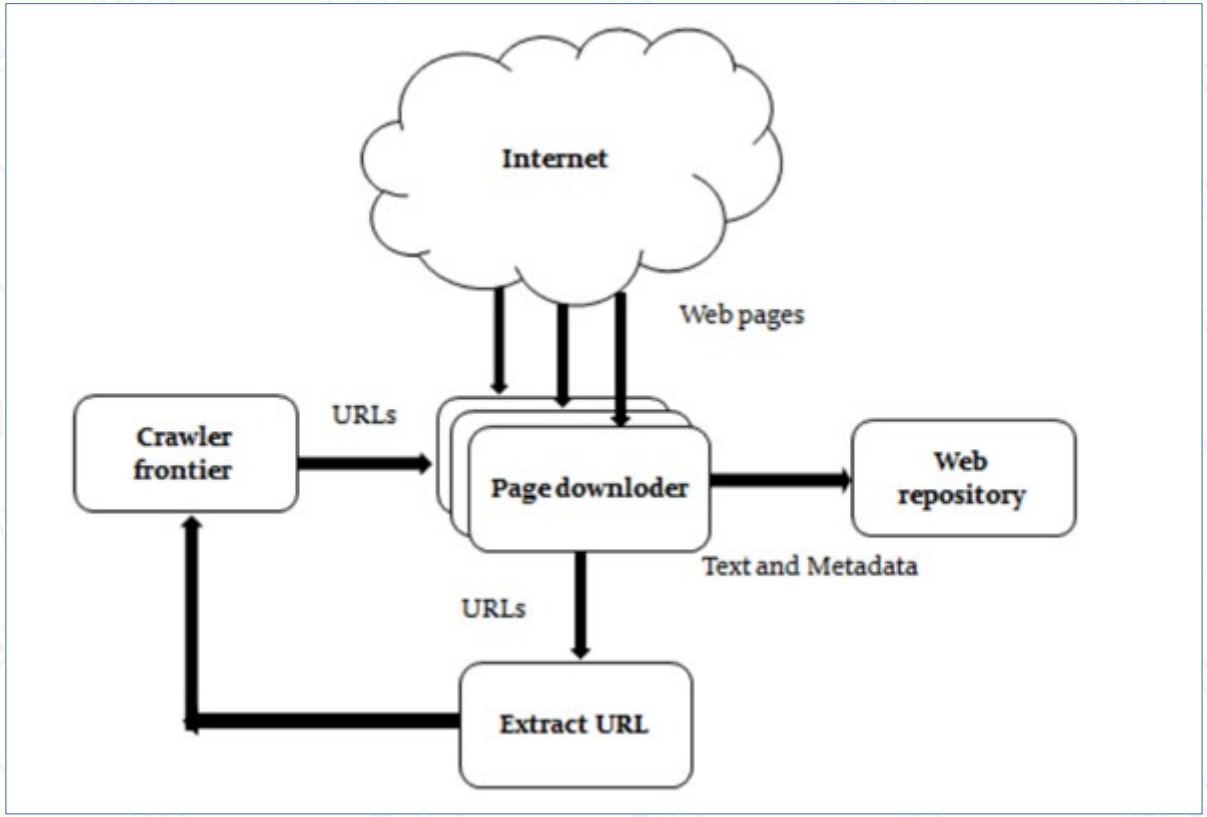
- 認識 Web、 HTML、 URL
- 何謂網路爬蟲
- 搜尋引擎介紹
- 認識 HTTP 與 HTTPS

網路爬蟲?

- 就是自動化**抓取網頁內容**的程式
  - 進入到某個頁面
  - 取出頁面中所需要的欄位資訊
  - 將資訊轉出處理（後處理、儲存等等）
  - 循環直到所有頁面都被處理完畢

通常就是讀檔、爬檔、寫檔的循環

網路爬蟲!


Internet
Web pages
Crawler frontier
URLs
Page downloder
Web repository
Text and Metadata
URLs
Extract URL

講次內容

- 認識 Web 、 HTML 、 URL
- 何謂網路爬蟲
- 搜尋引擎介紹
- 認識 HTTP 與 HTTPS

搜尋引擎?

- 我知道你都叫它 Google，可是並不精確
- 搜尋引擎是一種資訊檢索系統，讓你找資料用的
- 關鍵詞（一或多） =>

  相關結果清單（相關性排列）

搜尋引擎歷史

- 1990+: Archie, Gopher
- 1994: Yahoo Web Directory, WebCrawler, Lycos (robot), Infoseek
- 1995: Metacrawler, Excite, AltaVista (NLP, and/or/not)
- 1996: HotBot, Inktomi
- 1997: Yandex
- 1998: Google, MSN Search
- 2000: Baidu
- 2006: Ask.com
- 2008: DuckDuckGo
- 2009: Bing

然後？然後就全看 Google 了啊！


THE WORLD'S
Top 50 Websites
Below, we show the key players—from Google to Twitter—that currently dominate the Internet.
HOW TO READ THIS
Global rank
Circle size based on monthly traffic volume
Shared parent company
VISUALCAPITALIST.COM ← URL
BREAKDOWN BY CATEGORIES (GLOBAL, NOV 2020)
Social Networks and Online Communities 11
News and Media 9
Search Engines 7
Marketplace 7
Adult 4
Programming and Developer Software 3
TV Movies and Streaming 2
Email 2
Others 5
Amazon-owned Twitch attracts 1 billion monthly visitors as the most popular gaming site worldwide.
With explosive growth, newcomer Zoom surpasses Netflix, LinkedIn, and Reddit in monthly visitors.
Since 2019, Instagram's monthly visitors have jumped from 3.2 billion to 6.1 billion.
Top 3 most followed Twitter accounts
Barack Obama 125M
Justin Bieber 113M
Katy Perry 109M
Facebook, along with nearly every Big Tech company rank in the top 20.
Together, Google and subsidiary YouTube attract over 127 billion visits monthly—more than the next 48 websites combined.
China's search giant Baidu has also ventured into AI and autonomous vehicles.
Google 92.5B U.S. GOOGLE.COM
facebook 25.5B U.S. FACEBOOK.COM
34.6B U.S. YOUTUBE.COM
SOURCE: SIMILARWEB
VISUAL CAPITALIST
/visualcapitalist @visualcap visualcapitalist.com
COLLABORATORS RESEARCH + WRITING Dorothy Neufeld | ART DIRECTION + DESIGN Joyce Ma

阿就都給它玩就好了 ...

搜尋引擎!

- 為什麼談搜尋引擎?
  - 因為搜尋引擎就是靠網路爬蟲來取得 www 上的資訊

    所以我們也要玩!

講次內容

- 認識 Web、 HTML、 URL
- 何謂網路爬蟲
- 搜尋引擎介紹
- 認識 HTTP 與 HTTPS

HTTP

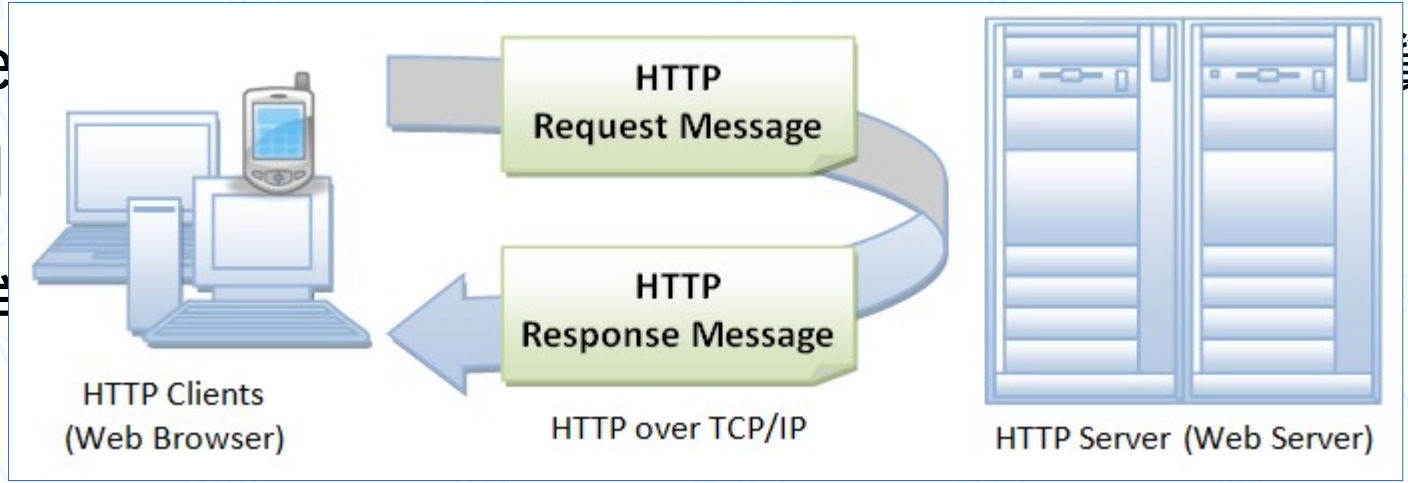
- HTTP, HyperText Transfer Protocol
  - 伺服端伺服器和用戶端瀏覽器之間溝通的標準協定
  - 每個物件從伺服器中獲取都需建立一個 TCP 連接，通訊埠 (port)80 來傳輸網頁的 HTTP 服務
  - 主要版本有 HTTP/1.0 、 HTTP/1.1 、 HTTP/2.0

HTTP，再細一點

- Client 端透過網址、超連結向 Server 下達 HTTP 請求 (Request)
- Server 處理請求：虛擬目錄 (virtual directory) 的資源 (HTML 文檔、影像檔或動態生成的結果)
- 處理完畢後，使用 MIME 格式回應 (Respond) 回 Client 端

HTTP，再細一點

- Client 端透過網址、超連結向 Server 下達 HTTP 請求 (Request)
- Serve 資源 (HTM
- 處理完 Client 端



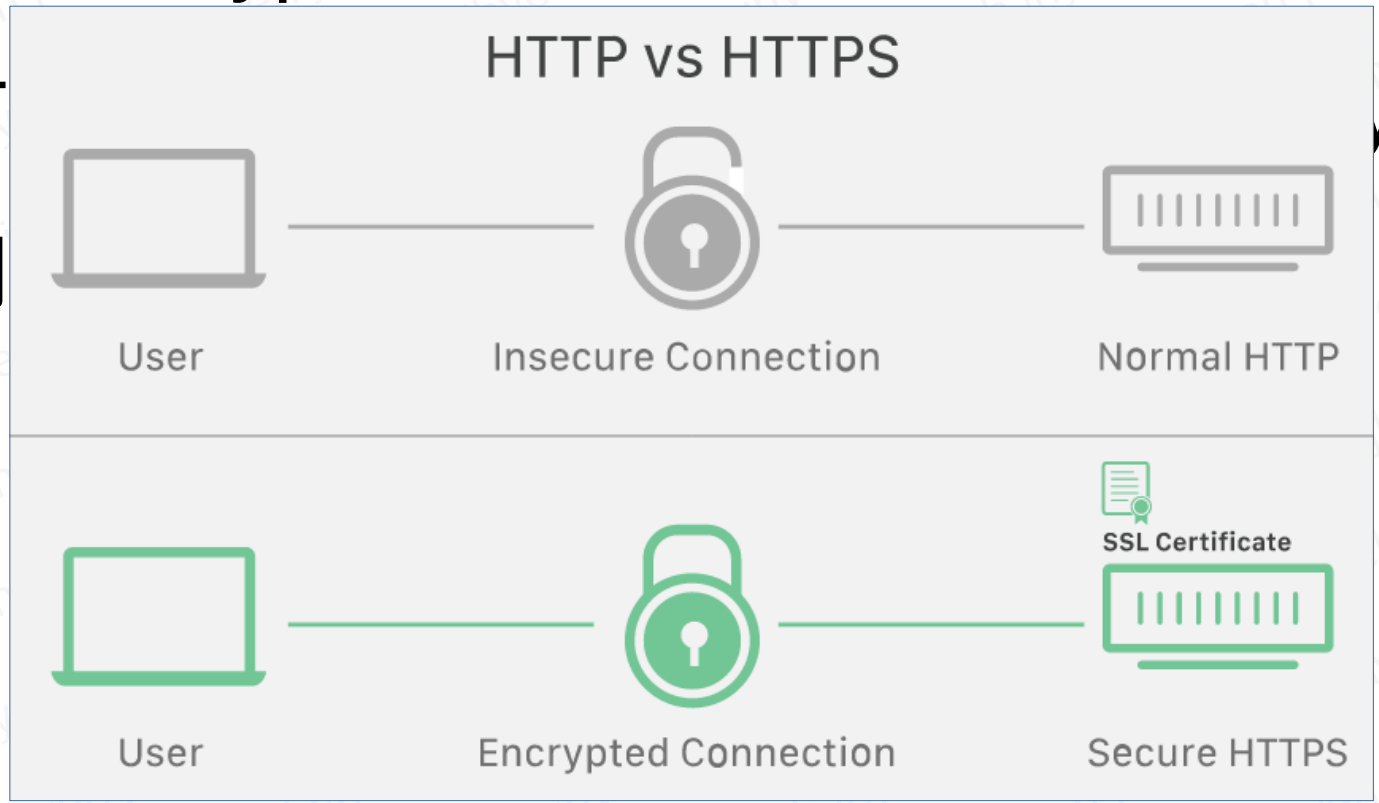
HTTPS(?)

- HTTPS, HyperText Transfer Protocol Security
- HTTP 中間插入 SSL(Secure Sockets Layer) 層
- 使用密碼對雙方的連線加密

攔截網路資訊者難以解讀!

HTTPS(?)

- HTTPS, HyperText Transfer Protocol Security
- HTT ...yer) 層
- 使用 ...



這個講次中，你應該學到了 ...

- 什麼是 Web、 HTML、 URL
- 什麼是網路爬蟲
- 什麼是搜尋引擎
- 什麼是 HTTP？ 什麼是 HTTPS？