# 中華大學資訊工程學系

# 109 學年度專題製作期初報告

題目

**網路爬蟲與資料整合**


指導教授：　　周智勳 教授

專題團隊：B10602034 顏　寬 學生

B10602006 徐世鈞 學生

B10602082 徐聖凱 學生

B10602072 高詠傑 學生



中華民國 110 年 8 月 版本 2.0 版

# 摘要

近年來「網路爬蟲」成為一種流行的趨勢，電腦相關產業多少都有耳聞，簡單來說就是一個自動化抓取網頁內容的程式，最簡單的方法不過就是一筆一筆的資料複製貼上，但如果資料成千上萬時該怎麼辦，這時就是網路爬蟲發揮功用的時候。

每所學校都有自己的官方網站，然而這些網頁內總是伴隨著龐大的資訊量，在許多情況下，學生很可能只是要搜尋某個處室的電話或位置，但對於不熟悉網頁排版或者不熟悉電腦的人，很可能因此迷失方向。因此我們將打造一個桌面應用程式，它能過濾網頁內的資訊，並自動化尋找使用者所需的內容，將其展現出來。捨去大量不相關的雜訊，讓使用者快速找到解答。

目前研究方向以「中華大學」為主題，將對其進行爬蟲分析並使用「Python」為主架構打造專屬網站。未來如果要強化的話，可以朝全國學校官方網站邁進，那將更能發揮此功用。

**關鍵字：網路爬蟲、Python、網頁、中華大學**

# 目錄

# 圖目錄

# 表目錄

# Chapter 01 專題導論

## 1.1 動機

在這個只要打開身邊的裝置就可以連上網路的時代，你可以很輕鬆地接收來自世界各地的消息與資訊，但在這些資訊爆炸的搜尋下，你可能會發現許多網頁呈現出一樣的內容，這要歸功於一種近年來新流行的技術「網路爬蟲」，在具有特定的搜尋條件下，它可以自動獲取網站內的資訊，並從中獲取使用者需要的資料。

以此為前提下，我們使用此技術探索學校的官方網站，將不重要的要素排除，接著彙整出學生比較想看的資訊，並將其整理放入應用程式中，供使用者快速找到所需的資訊。

## 1.2 範圍與目標

將結合四年在大學所學的知識，以專題的形式呈現出來。

運用「Python」為主架構打造桌面應用程式介面和網路爬蟲，讓使用者快速地尋找校園網站內的資訊。結合過去的課程內容，以此專題模擬一個小型專案，撰寫需求規格書等相關資源。

## 1.3 專題應用簡介

依照學生查詢資訊的習慣，將開發出以下的內容：

- 校園網頁搜尋：提供跨網站搜尋的可能性。
- 快速單位連結：使用者快速連結至常用單位。
- 熱門網站列表：前 50 名網站展示給使用者。

# Chapter 02 專題規劃

## 2.1 團隊工作規劃

| 工作內容 | 負責成員 |
|---|---|
| 資料搜尋、資料整理 | 徐聖凱、高詠傑 |
| 文件撰寫、問卷設計 | 徐世鈞 |
| 程式設計 | 顏寬 |
| 版型設計、美工處理 | 顏寬、徐世鈞 |
| 整合測試、程式除蟲 | 顏寬、徐聖凱、高詠傑 |
| 報告討論 | 顏寬、徐聖凱、高詠傑 |

表 2.1 團隊工作規劃表

## 2.2 進度時程安排

| | 三月 | 四月 | 五月 | 六月 | 七月 | 八月 |
|---|---|---|---|---|---|---|
| **前置討論** | | | | | | |
| Python GUI 設計 | | | | | | |
| **網路爬蟲測試** | | | | | | |
| GUI 功能與爬蟲測試 | | | | | | |
| **中期討論** | | | | | | |
| **全功能測試階段** | | | | | | |
| **規模實際測驗** | | | | | | |
| **後期討論** | | | | | | |
| **結果分析與最終報告** | | | | | | |

表 2.2 進度時程安排表

# Chapter 03 應用程式開發

## 3.1 開發環境

- Python 主程式碼

一種廣泛使用的直譯式、進階和通用的程式語言。Python 支援多種程式設計範式，包括函數式、指令式、結構化、物件導向和反射式程式。它擁有動態型別系統和垃圾回收功能，能夠自動管理記憶體使用，並且其本身擁有一個巨大而廣泛的標準庫。

Python 由吉多·范羅蘇姆創造於 1991 年，它是 ABC 語言的後繼者。

圖 3.1 Python 圖標

Python 的設計哲學強調代碼的可讀性和簡潔的語法，尤其是使用空格縮排劃分代碼塊。相比於 C 或 Java，Python 讓開發者能夠用更少的代碼表達想法。不管是小型還是大型程式，該語言都試圖讓程式的結構清晰明瞭。

Python 直譯器本身幾乎可以在所有的作業系統中執行。目前由 Python 軟體基金會管理。

--出自< 維基百科-Python >

- Asp.net GUI 介面

由微軟在.NET Framework 框架中所提供，開發 Web 應用程式的類別庫，封裝在 System.Web.dll 檔案中，顯露出 System.Web 命名空間，並提供 ASP.NET 網頁處理、擴充以及 HTTP 通道的應用程式與通訊處理等工作，以及 Web Service 的基礎架構。ASP.NET 是 ASP 技術的後繼者，但它的發展性要比 ASP 技術要強大許多。

ASP.NET 可以運行在安裝了.NET Framework 的 IIS 伺服器上，若要在非微軟的平台上執行，則需要使用 Mono 平台，ASP.NET 在 2.0 版本已經定型，在.NET Framework 3.5 上則加上了許多功能，像是 ASP.NET AJAX、ASP.NET MVC Framework、ASP.NET Dynamic Data 與 Microsoft Silverlight 的伺服器控制項等。


圖 3.1 ASP.NET 圖標

由微軟在.NET Framework 框架中所提供，開發 Web 應用程式的類別庫，封裝在 System.Web.dll 檔案中，顯露出 System.Web 命名空間，並提供 ASP.NET 網頁處理、擴充以及 HTTP 通道的應用程式與通訊處理等工作，以及 Web Service 的基礎架構。ASP.NET 是 ASP 技術的後繼者，但它的發展性要比 ASP 技術要強大許多。

--出自< 維基百科- ASP.NET >

## 3.2 應用程式開發工具

- PyCharm

  一個用於計算機編程的集成開發環境(IDE)，主要用於 Python 語言開發，由捷克公司 JetBrains 開發，提供代碼分析、圖形化調試器，集成測試器、集成版本控制系統，並支持使用 Django 進行網頁開發。

- Anaconda

  一個免費開源的 Python 和 R 語言的發行版本，用於計算科學，Anaconda 致力於簡化軟體套件管理系統和部署。

- Asp.net

  一款

- Beautiful Soup

  一個 Python 包，功能包括解析 HTML、XML 文件、修復含有未閉合標籤等錯誤的文件。這個擴充包為待解析的頁面建立一棵樹，以便提取其中的資料，這在網路資料採集時非常有用。

- Requests

  一個 Python HTTP 庫，在 Apache License 2.0 下發布。該項目的目標是使 HTTP 請求更簡單，更人性化。

- SQLite

  遵守 ACID 的關聯式資料庫管理系統，SQLite 不是一個客戶端/伺服器結構的資料庫引擎，而是被整合在使用者程式中。

- Excel

  直觀的介面、出色的計算功能和圖表工具，再加上成功的市場行銷，使 Excel 成為最流行的個人電腦資料處理軟體。

## 3.3 應用程式架構

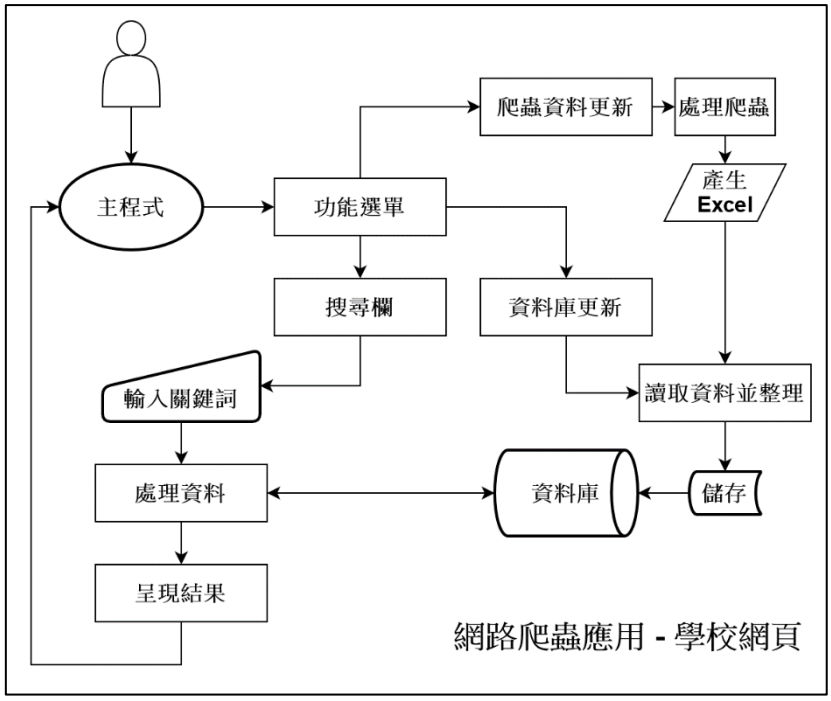


圖 3.3 應用程式架構

- 校園網路爬蟲網頁

  使用者將程序開啟後，將可直接執行以下三大動作：

  1. 校園資訊：校園網頁搜尋：提供跨網站搜尋的可能性。
  2. 快速單位連結：使用者快速連結至常用單位。
  3. 熱門網站列表：前 50 名網站展示給使用者。

- 中華大學 校園官方網站

  此專題將針對中華大學的網站進行數據撈取。

## 3.4 儀器設備需求表

| 設備名稱 | 數量 | 備註 |
|---|---|---|
| 電腦 | 1 | - |

表 3.4 儀器設備需求表

# Chapter 04 應用程式實作

## 4.1 程式首頁


圖 4.1 程式首頁

## 4.2 程式搜尋


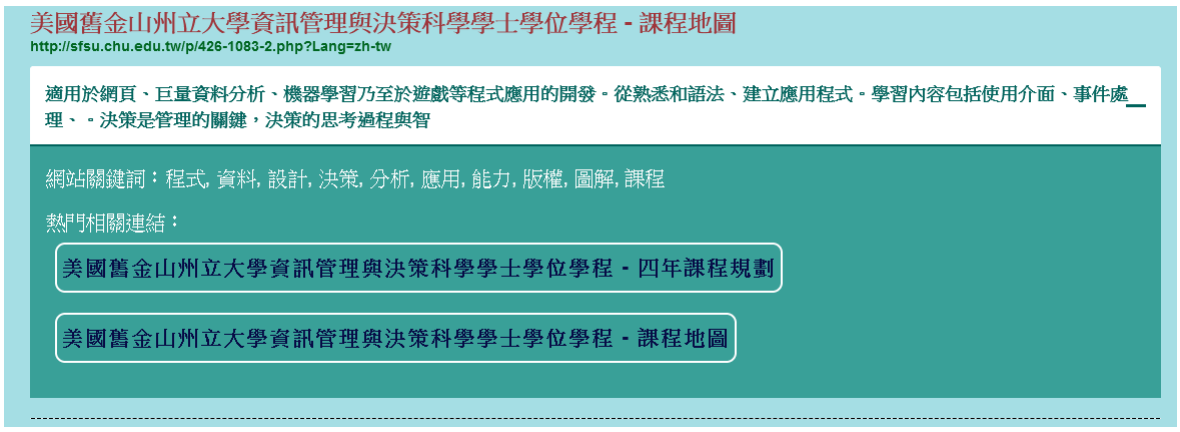


圖 4.2 程式搜尋

# 4.3 附屬功能

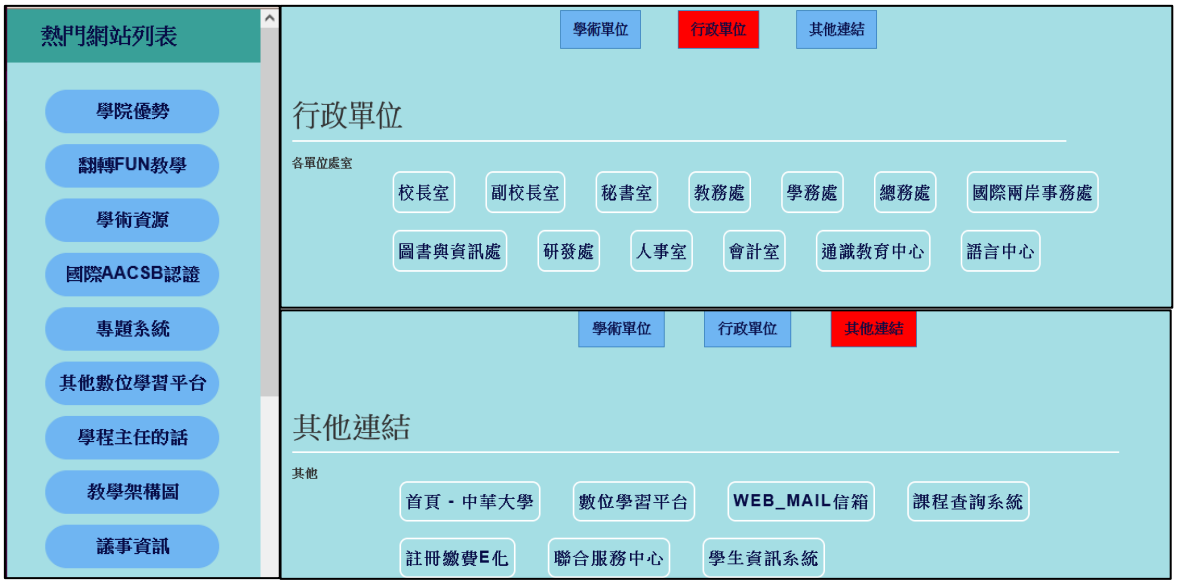


圖 4.3 附屬功能

# 4.4 爬蟲關鍵程式

```
Academic = requests.get(URL_acad)
BS_2_3 = BeautifulSoup(Academic.text, 'html.parser')
# 資訊電機學院
Data_2_5_1 = BS_2_3.find('div', id='Dyn_1_3')
Txt_2_1_3 = Data_2_5_1.find_all('a')
for i, Txt_2_1_3 in enumerate(Txt_2_1_3):
    tmp: str = Txt_2_1_3.get_text().strip()
    tmp01: str = Txt_2_1_3.get('href')
    word4 = re.findall(r"^/p", tmp01)
    word5 = re.findall(r"tw/", tmp01)
    word6 = re.findall(r"^/var", tmp01)
    if word6:
        tmp01 = 'http://www1.chu.edu.tw' + tmp01
    if word4:
        tmp01 = 'http://www1.chu.edu.tw' + tmp01
    if word5:
        tmp01 = tmp01.rstrip('index.php?Lang=zh-tw')
    for j in enumerate(Txt_2_1_3):
        Ac_List.append(tmp)
        Ac_WEB_List.append(tmp01)
        tmp = ''
        tmp01 = ''
```

```
# ================學術各網頁文字搜尋==========
web = ["main", "mainbody", "maincontent", "a", "h1", "h2", "h3"]
# ==========學術總單位==========
for n in range(len(Ac_WEB_List)):
    gg = Ac_WEB_List[n]
    if gg:
        try:
            u_url = requests.get(gg, headers=headers)
            u_url.encoding = 'utf8'
        except requests.exceptions.ConnectionError:
            point_ac_all_text.append('')
            point_ac_all_freq.append('')
            ac_all_sentence.append('')
        else:
            ur_tx = BeautifulSoup(u_url.text, 'html.parser')
            tmp02 = ur_tx.find_all('div', attrs={"class": web})
            if bool(tmp02) == 1:
                for g, tmp02 in enumerate(tmp02):
                    ww: str = tmp02.get_text().strip()
                    if ww:
                        # 去除數字、英文、底線
                        ww = re.sub('[a-zA-Z0-9_]', '', ww)
                        ww = ww.replace("_", '')
                        # 替代換行，並去除其他非文字符號
                        ww = ww.replace("\n", '1')
                        ww = ''.join(ch for ch, _ in itertools.groupby(ww))
                        ww = re.sub('\s+', '', ww)
                        # 以逗號區隔句子，並儲存至串列中
                        tmp11 = ww.split('1')
                        tmp11 = [x for x in tmp11 if x != '']
                        tmp12 = ','.join(str(d) for d in tmp11)
                        # 去除多餘的數字、標點符號等
                        ww = re.sub('\d', '', ww)
                        ww = re.sub('\W+', '', ww)
                        # 彙整文章，重製串列
                        tmp03 += ww
                        tmp11 = []
                ac_all_sentence.append(tmp12)
```

```
# 計算次數比例
for i, row in enumerate(point_ac_all_freq):
    if bool(point_ac_all_freq[i]) == 1:
        y = point_ac_all_freq[i].split(',')
        t = list(map(int, y))
        x = sum(t)
        for j in t:
            f = round(j / x, 3)
            tmp06.append(f)
    else:
        tmp06.append('')
    gg = ','.join(str(d) for d in tmp06)
    ratio_ac_all_freq.append(gg)
    tmp06 = []
# ============================
```

圖 4.4 爬蟲關鍵程式

# 4.5 資料庫

| | id | keywords | depart | subpart | website | sentence | ratio | econd_ratio |
|---|---|---|---|---|---|---|---|---|
| | 過濾 | 過濾 | 過濾 | 過濾 | 過濾 | 過濾 | 過濾 | 過濾 |
| 1 | 1 | 大學 | 資訊電機學院辦公室 | 學術 | http://csee.chu.edu.tw/ | 企業滿意度連續三年全國私立大學...企業滿意... | 0.106 | 0.005 |
| 2 | 2 | 薪資 | 資訊電機學院辦公室 | 學術 | http://csee.chu.edu.tw/ | 就業薪資高於私立大學平均薪資...畢就業薪資... | 0.106 | 0.017 |
| 3 | 3 | 私立 | 資訊電機學院辦公室 | 學術 | http://csee.chu.edu.tw/ | 企業滿意度連續三年全國私立大學...企業滿意... | 0.085 | 0.017 |
| 4 | 4 | 企業 | 資訊電機學院辦公室 | 學術 | http://csee.chu.edu.tw/ | 企業工讀實習...企業滿意度連續三年全國私立... | 0.064 | 0.006 |
| 5 | 5 | 消息 | 資訊電機學院辦公室 | 學術 | http://csee.chu.edu.tw/ | 最新消息...兩岸合作最新消息... | 0.043 | 0.029 |
| 6 | 6 | 學院 | 資訊電機學院辦公室 | 學術 | http://csee.chu.edu.tw/ | 學院公告...選擇資訊電機學院的理由... | 0.043 | 0.004 |
| 7 | 7 | 資訊 | 資訊電機學院辦公室 | 學術 | http://csee.chu.edu.tw/ | 選擇資訊電機學院的理由...交通資訊... | 0.043 | 0.004 |
| 8 | 8 | 滿意度 | 資訊電機學院辦公室 | 學術 | http://csee.chu.edu.tw/ | 企業滿意度連續三年全國私立大學...企業滿意... | 0.043 | 0.037 |
| 9 | 9 | 連續 | 資訊電機學院辦公室 | 學術 | http://csee.chu.edu.tw/ | 企業滿意度連續三年全國私立大學...企業滿意... | 0.043 | 0.062 |
| 10 | 10 | 三年 | 資訊電機學院辦公室 | 學術 | http://csee.chu.edu.tw/ | 企業滿意度連續三年全國私立大學...企業滿意... | 0.043 | 0.071 |
| 11 | 11 | 全國 | 資訊電機學院辦公室 | 學術 | http://csee.chu.edu.tw/ | 企業滿意度連續三年全國私立大學...企業滿意... | 0.043 | 0.014 |
| 12 | 12 | 就業 | 資訊電機學院辦公室 | 學術 | http://csee.chu.edu.tw/ | 就業薪資高於私立大學平均薪資...畢就業薪資... | 0.043 | 0.004 |
| 13 | 13 | 高於 | 資訊電機學院辦公室 | 學術 | http://csee.chu.edu.tw/ | 就業薪資高於私立大學平均薪資...畢就業薪資... | 0.043 | 0.14 |
| 14 | 14 | 平均 | 資訊電機學院辦公室 | 學術 | http://csee.chu.edu.tw/ | 就業薪資高於私立大學平均薪資...畢就業薪資... | 0.043 | 0.02 |
| 15 | 15 | 畢業生 | 資訊電機學院辦公室 | 學術 | http://csee.chu.edu.tw/ | 畢業生就業率高、待業率低...畢業生就業率... | 0.043 | 0.007 |
| 16 | 16 | 就業率 | 資訊電機學院辦公室 | 學術 | http://csee.chu.edu.tw/ | 畢業生就業率高、待業率低...畢業生就業率... | 0.043 | 0.013 |
| 17 | 17 | 待業 | 資訊電機學院辦公室 | 學術 | http://csee.chu.edu.tw/ | 畢業生就業率高、待業率低...畢業生就業率... | 0.043 | 0.073 |
| 18 | 18 | 率低 | 資訊電機學院辦公室 | 學術 | http://csee.chu.edu.tw/ | 畢業生就業率高、待業率低...畢業生就業率... | 0.043 | 0.073 |
| 19 | 19 | 公告 | 資訊電機學院辦公室 | 學術 | http://csee.chu.edu.tw/ | 學院公告... | 0.021 | 0.002 |
| 20 | 20 | 招生 | 資訊電機學院辦公室 | 學術 | http://csee.chu.edu.tw/ | 招生訊息... | 0.021 | 0.002 |
| 21 | 21 | 大學 | 電機工程學系 | 學術 | http://ee.chu.edu.tw/ | 國立交通大學電子研究所助理研究員...中華大... | 0.098 | 0.004 |
| 22 | 22 | 畢業生 | 電機工程學系 | 學術 | http://ee.chu.edu.tw/ | 電機系的畢業生有許多人進入台積電 聯發科... | 0.088 | 0.014 |
| 23 | 23 | 企業 | 電機工程學系 | 學術 | http://ee.chu.edu.tw/ | 企業實習與產業接軌, 提供多元專業的企業實... | 0.069 | 0.007 |
| 24 | 24 | 中華大學 | 電機工程學系 | 學術 | http://ee.chu.edu.tw/ | 中華大學不僅為一所畢業生「就業率高、薪資... | 0.069 | 0.003 |
| 25 | 25 | 薪資 | 電機工程學系 | 學術 | http://ee.chu.edu.tw/ | 中華大學不僅為一所畢業生「就業率高、薪資... | 0.069 | 0.011 |

圖 4.5 資料庫

# Chapter 05 系統介面

## 5.1 實際運行情況

- 前端

  1. Asp.net 在大部分時間都是穩定運作，只有少數情況會閃爍延遲新。
  2. 資料在存取上，受限於資料庫本身特性，讀取資料會受到些微延遲。

- 後端

  1. 爬蟲模組會針對我們提供的標籤，指向性的搜尋頁面內容。
  2. 龐大的網頁數量，進而導致一次執行，動輒要一個小時起跳。

## 5.2 程式評估

- 關鍵詞比例值計算

| | |
|---|---|
| 第一階段排序 | 計算每個詞彙出現頻率高低進行，並針對同網頁的文章所有的關鍵詞，取前 20 名。 |
| 第二階段排序 | 針對第一排序，再次進行排序。通過相同關鍵詞出現在不同網頁的比例值進行整理計算。 |
| 第三階段排序 | 彙整第二排序，將結果帶回同網頁底下的排序，這將導致與第一階段不同的排序結果。 |

**表 5.2 關鍵詞比例值計算表**

# Chapter 06 未來展望

## 6.1 結論

畢業專題之所以選擇做爬蟲，是因為當初很看好 Python 的可彈性，豐富的套件，快速開發沒有問題。語言選好了，接下來就是主題，正在苦惱之時，意外的發現學校網頁有許多弊端的操作問題，團隊決定以此為目標，設法做出一個類搜尋引擎，供學生在查詢上快速指向目標網頁。

專題完成時已達到以下目標：

1. 快速導向目標
2. 爬蟲分析大量文章
3. 關鍵詞比例值權重計算
4. 自動摘要檢視網頁內容
5. 使用者互動體驗

這次是發揮了這四年所學的東西放入其中，整合前後端，可以說是一個大工程。尤其是爬蟲程式最具有挑戰性，需要不斷測試以及正規化資料，學到了不少東西，因為算是人生第一份作品，還是蠻有挑戰性的。

在這邊對這一年來參與的團隊以及我們指導教授，熱心的付出與努力，獻上最誠摯的感謝與祝福。

## 6.2 未來展望

- 善用爬蟲，可以做更多有特色性的功能。
- 擴增更多模塊以利於使用者搜尋，例如整合 moddle 或其他常用連結。

# 參考資料

- 維基百科
- 相關知識網頁列表

| 編號 | 網頁名稱(含連結) | 網頁介紹 |
|---|---|---|
| 01 | towardsdatascience | How to build a smart search engine |
| 02 | bootdey | create this bootstrap snippet |
| 03 | androidcss | CSS3 floating button design tutorial |
| 04 | codepen | 向下滾動會出現 TOP 按鈕 |
| 05 | dotblogs | SQL-刪除重複資料 |
| 06 | hotexamples | C# to SQL |
| 07 | 爬蟲學習筆記 | Requests, Beautiful Soup, 正規表達式 |
| 08 | codepen | fixed to top navbar work. |
| 09 | codesandbox | toggle |
| 10 | asp.net | Autocomplete |

表 參考資料 相關知識網頁列表