**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**

摘要-----------------------------------------------------------------01

目錄-----------------------------------------------------------------02

圖目錄---------------------------------------------------------------03

表目錄---------------------------------------------------------------04

Chapter 01 專題導論--------------------------------------------------05

1.1 動機

1.2 範圍與目標

1.3 專題應用簡介

1.4 遭遇問題與解決方案

Chapter 02 專題規劃--------------------------------------------------06

* 1. 團隊工作規劃
  2. 進度時程安排

Chapter 03 應用程式開發----------------------------------------------07

3.1 開發環境

3.2 應用程式開發工具

3.3 應用程式架構

3.4 儀表設備需求表

Chapter 04 應用實作--------------------------------------------------10

4.1 程式首頁

4.2 程式搜尋

4.3 附屬功能

4.4 爬蟲關鍵程式

4.5 資料庫

Chapter 05 系統介面--------------------------------------------------13

5.1 實際運行情況

5.2 程式評估

Chapter 06 未來展望--------------------------------------------------14

6.1 結論

6.2 未來展望

參考文獻-------------------------------------------------------------15

**圖目錄**

**圖3.1 Python圖標**---------------------------------------------------07

**圖3.1 ASP.NET圖標**--------------------------------------------------07

**圖3.3 應用程式架構**--------------------------------------------------09

**圖4.1 程式首頁**------------------------------------------------------10

**圖4.2 程式搜尋**------------------------------------------------------10

**圖4.3 附屬程式**------------------------------------------------------11

**圖4.4 關鍵爬蟲程式**--------------------------------------------------11

**圖4.5 資料庫**--------------------------------------------------------12

**表目錄**

**表2.1 團隊工作規劃表**------------------------------------------------06

**表2.2 進度時程安排表**------------------------------------------------06

**表3.4 儀器設備需求表**------------------------------------------------09

**表5.2** 關鍵詞**比例值計算表**--------------------------------------------13

**表 參考資料 相關知識網頁列表**-------------------------------------------15

**Chapter 01 專題導論**

* 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. **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. **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**

* 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**

* 1. **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**

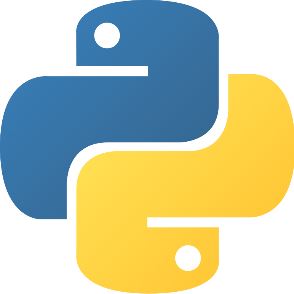
**Chapter 03 應用程式開發**

**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。

　　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。

**圖3.1 Python圖標**

　　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。

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。

**圖3.1 ASP.NET圖標**

--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**

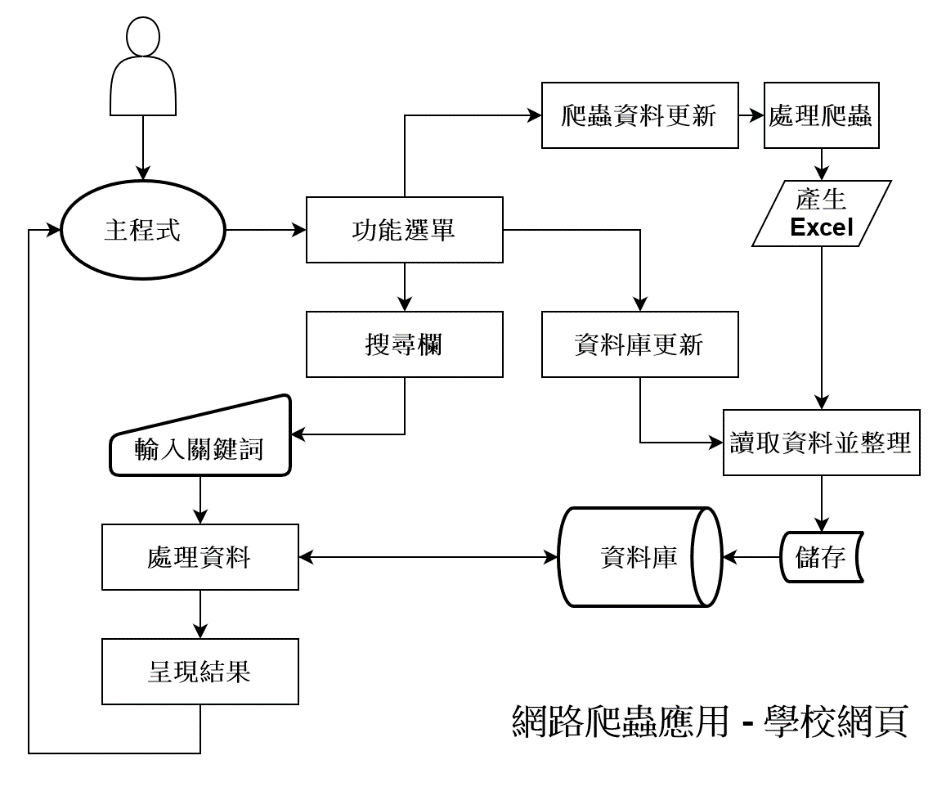
　　一個Python HTTP庫，在Apache License 2.0下發布。該項目的目標是使HTTP請求更簡單，更人性化。

* **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**

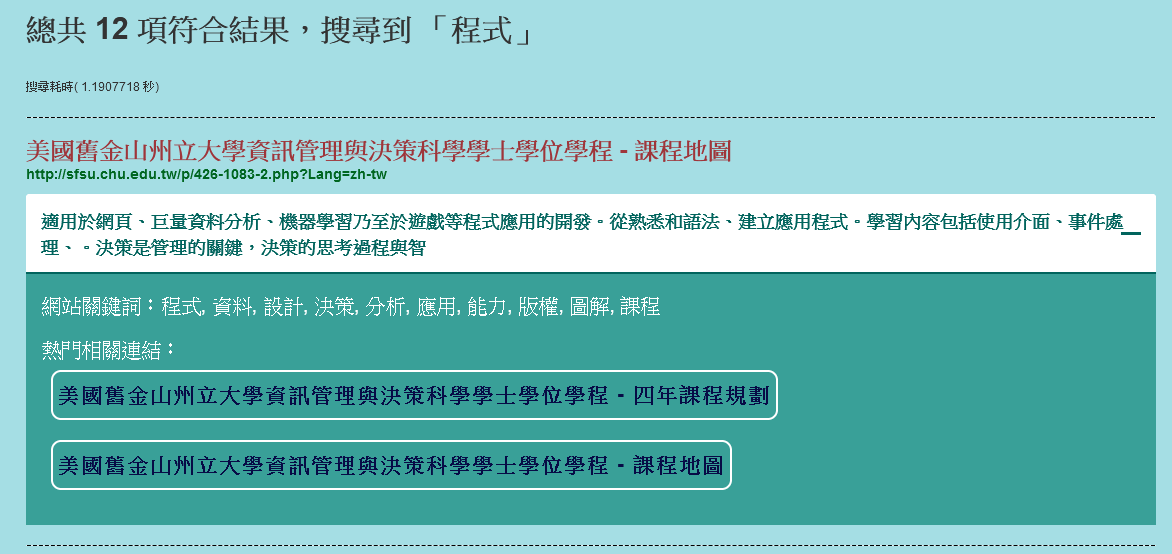


**圖4.1 程式首頁**

**4.2 程式搜尋**

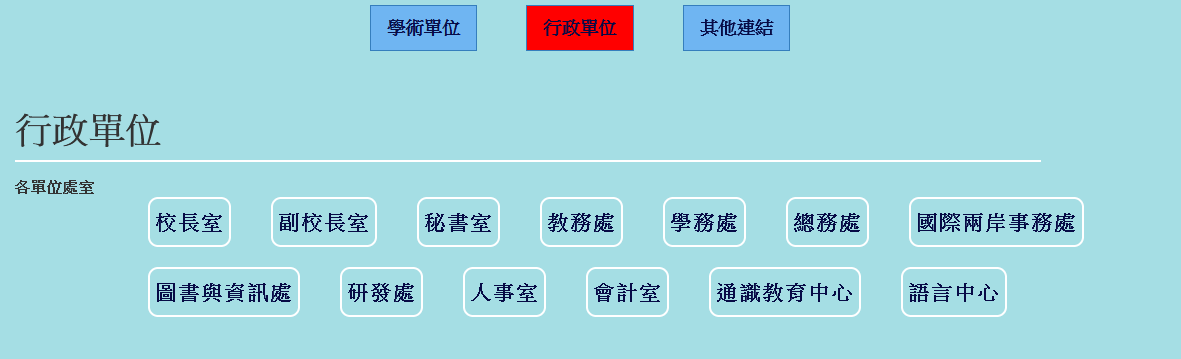
**一張含有 文字 的圖片

自動產生的描述**

****

**圖4.2 程式搜尋**

**一張含有 文字 的圖片

自動產生的描述4.3 附屬功能**

**圖4.3 附屬功能**

**4.4 爬蟲關鍵程式**

**一張含有 文字 的圖片

自動產生的描述一張含有 文字 的圖片

自動產生的描述**

**一張含有 文字 的圖片

自動產生的描述**

**一張含有 文字 的圖片

自動產生的描述**

**圖4.4 爬蟲關鍵程式**

**4.5 資料庫**

**一張含有 文字, 電腦, 電子用品, 合成器 的圖片

自動產生的描述**

**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**](https://zh.wikipedia.org/wiki/Wikipedia:%E9%A6%96%E9%A1%B5)
* **相關知識網頁列表**

|  |  |  |
| --- | --- | --- |
| **編號** | **網頁名稱(含連結)** | **網頁介紹** |
| **01** | [**towardsdatascience**](https://towardsdatascience.com/how-to-build-a-smart-search-engine-a86fca0d0795) | **How to build a smart search engine** |
| **02** | [**bootdey**](https://www.bootdey.com/snippets/view/Search-Results#preview) | **create this bootstrap snippet** |
| **03** | [**androidcss**](http://androidcss.com/css/css3-floating-button-tutorial/) | **CSS3 floating button design tutorial** |
| **04** | [**codepen**](https://codepen.io/itrong/pen/MoxypL) | **向下滾動會出現 TOP 按鈕** |
| **05** | [**dotblogs**](https://dotblogs.com.tw/lastsecret/2010/07/13/16532) | **SQL-刪除重複資料** |
| **06** | [**hotexamples**](https://csharp.hotexamples.com/site/file?hash=0x59d54812374022737a0b857bef7605a915148fe33f39187befa44ff4a0a6779e&fullName=SendAlertForm.cs&project=starboxs/GPS_Service) | **C# to SQL** |
| **07** | [**爬蟲學習筆記**](https://yanwei-liu.medium.com/python%E7%88%AC%E8%9F%B2%E5%AD%B8%E7%BF%92%E7%AD%86%E8%A8%98-%E4%B8%80-beautifulsoup-1ee011df8768) | **Requests, Beautiful Soup, 正規表達式** |
| **08** | [**codepen**](https://codepen.io/daanvankerkom/pen/bRbKEL) | **fixed to top navbar work.** |
| **09** | [**codesandbox**](https://codesandbox.io/s/trusting-shaw-wcd4d?file=/index.html:557-572) | **toggle** |
| **10** | [**asp.net**](https://forums.asp.net/t/1226079.aspx?Autocomplete+extender+and+space+symbol) | **Autocomplete** |

**表 參考資料 相關知識網頁列表**