國立台北商業大學

資訊管理系

112．資訊系統專案設計

**系統手冊**



**組別：第112510組**

**題目：網路爬蟲機器人**

**指導老師：陳信宏**

**組長：10856040鍾皓年**

**組員：10856010郭宗翰 10856030彭鈺程**

**10856031彭鈺達 10856032陳永祥**

**中華民國112年5月10日**

目錄

[第一章 前言 5](#_Toc134191273)

[1-1背景介紹 5](#_Toc134191274)

[1-2動機 5](#_Toc134191275)

[1-3系統目的與目標 5](#_Toc134191276)

[1-4預期成果 5](#_Toc134191277)

[第二章 營運計畫 6](#_Toc134191278)

[2-1 可行性分析 6](#_Toc134191279)

[2-2 商業模式－Business model 8](#_Toc134191280)

[2-3 市場分析－STP 9](#_Toc134191281)

[2-4競爭力分析－SWOT 10](#_Toc134191282)

[第三章 系統規格 11](#_Toc134191283)

[3-1 系統架構 11](#_Toc134191284)

[3-2 系統軟、硬體需求與技術平台 11](#_Toc134191285)

[3-3 開發標準與使用工具 12](#_Toc134191286)

[第四章 專案時程與組織分工 13](#_Toc134191287)

[4-1 專案時程 13](#_Toc134191288)

[4-2 專案組織與分工 14](#_Toc134191289)

[第五章 需求模型 16](#_Toc134191290)

[5-1 功能分解圖（Functional decomposition diagram） 16](#_Toc134191291)

[5-2 需求清單： 17](#_Toc134191292)

[第六章 程序模型 18](#_Toc134191293)

[6-1 資料流程圖(Data flow diagram)。 18](#_Toc134191294)

[6-2 程序規格書(Process specification)。 19](#_Toc134191295)

表目錄

[第一章 前言 5](#_Toc134300323)

[第二章 營運計畫 6](#_Toc134300324)

[表 3-3-1開發標準與使用工具 10](#_Toc134300325)

[第三章 系統規格 11](#_Toc134300326)

[表 3-2-1系統軟、硬體需求 11](#_Toc134300327)

[表 3-3-1開發標準與使用工具 12](#_Toc134300328)

[第四章 專案時程與組織分工 13](#_Toc134300329)

[表 4-1-1、專案時程 13](#_Toc134300330)

[表4-2-1、專案組織與分工 14](#_Toc134300331)

[第五章 需求模型 16](#_Toc134300332)

[第六章 程序模型 18](#_Toc134300333)

圖目錄

[第一章 前言 6](#_Toc134301105)

[第二章 營運計畫 7](#_Toc134301106)

[第三章 系統規格 12](#_Toc134301107)

[第四章 專案時程與組織分工 14](#_Toc134301108)

[第五章 需求模型 17](#_Toc134301109)

[圖 5-1、功能分解圖 17](#_Toc134301110)

[第六章 程序模型 19](#_Toc134301111)

[圖 6-1-1、系統環境圖 19](#_Toc134301112)

[圖 6-1-2、圖0 19](#_Toc134301113)

第一章 前言

1-1背景介紹

在現在這個大數據時代，科技的進步日新月異，不管是5G、AI、大數據等科技都在不斷的成熟。而網路上的資訊量非常龐大，因此需要一種自動化的方式來收集、分析和整理這些資料。

1-2動機

隨著互聯網的迅速發展，人們需要大量的資訊來滿足日常生活和工作需求。如果是用人工的方式蒐集資料，搜集上萬筆（或者更多）的資料可不是一天就能完成的，而爬蟲卻可以自動從網路上獲取大量的資訊，可以有效地提高人們的資訊搜尋效率和準確性。

1-3系統目的與目標

主要利用爬蟲去更快地抓取資料，讓爬蟲程式自動化，利用AI技術，能自動分析網頁架構

1-4預期成果

透過培養機器人，讓USER在任何網站都可以不受任何限制的進行爬蟲，透過自動化分析，得出想要的結論

第二章 營運計畫

2-1 可行性分析

市場可行性: 在市場方面，網頁爬蟲是一個非常有應用價值的技術，能夠幫助使用者自動化地瀏覽和擷取網頁上的資訊。因此，將這項技術應用於提供使用者自由選取網站並爬取資料的網頁介面，有著許多應用場景和市場潛力。 首先，對於需要搜集大量網頁資訊的公司或組織而言，這個專題提供了一個自動化的方式，能夠減少人力和時間成本。例如，市場調查公司需要搜集大量網站上的商品資訊，以做出更準確的市場分析報告；或是網路安全公司需要監控特定網站上的漏洞和風險，以提高網路安全性等。此外，對於個人使用者而言，這個專題也提供了一個便捷的工具，讓使用者可以自由選取網站並爬取想要的資訊。例如，一些學術研究人員需要搜集網站上的文獻和資料，以便進行進一步的研究；或是一些愛好者需要搜集網站上的影片和圖片，以滿足自己的興趣和需求。 總結來說，這個專題在市場上有著廣泛的應用價值，能夠滿足不同用戶的需求，且有著不斷擴大的應用範圍和市場潛力。

財務可行性: 財務可行性是這個專題實施的一個重要方面。考慮到實施這個專題所需的資源和成本，以及預期的營收和利潤，以下是一些財務可行性方面的考慮： 首先，對於實施這個專題所需的資源和成本，主要包括人力、硬體、軟體和營運費用等方面。這些成本都需要事先估算和預算，以便進一步確定這個專題的財務可行性。 其次，對於預期的營收和利潤，主要來自於使用者的付費訂閱和廣告收入等方面。使用者可以按月或按年付費，以使用網頁介面提供的爬蟲服務。此外，也可以通過與廣告商合作，通過網頁介面上的廣告投放獲得收入。這些收入預測需要根據市場調查和市場需求來進行預測和評估。最後，在考慮財務可行性時，需要考慮到競爭對手和市場趨勢等方面。確定專題在市場上的差異化優勢，以便應對市場競爭和變化，進一步確定財務可行性。 綜合考慮以上方面，可以進一步進行財務模型的建立和評估，以確定這個專題的財務可行性。如果預期的收益和利潤能夠超過成本和開支，那麼這個專題就是財務可行的。

2-2 商業模式－Business model

商業模式: 訂閱服務模式：用戶訂閱服務，按照數據爬取的次數、爬取的數據量等參數進行計費。這個商業模式可以通過提供不同的訂閱方案，讓用戶根據自己的需求選擇適合自己的方案。

付費數據模式：提供用戶爬取數據的服務，收取一定的費用。這個商業模式可以通過對不同的數據進行定價，讓用戶根據自己的需求選擇要爬取哪些數據，以及爬取的數量。 廣告模式：通過向用戶推送廣告，獲取收益。這個商業模式可以通過分析用戶的爬蟲需求，向用戶展示相關廣告，從中獲取收益。 咨詢服務模式：為需要特定數據的客戶提供定制化的咨詢服務，協助客戶分析數據並提供專業的建議。這個商業模式可以通過對用戶需求的了解，向客戶提供定制化的咨詢服務。

2-3 市場分析－STP

S： 地域區分：爬蟲機器人可以面向不同國家和地區的市場，全世界都可以用。 統計變數：根據不同收入水平，必須有電腦和網路才能使用 生活型態：常接觸到電腦、會使用電腦來查詢資料 人生階段：主要為上班族

T： 集中行銷：針對特定的利基市場，如大型企業、大型研究機構等。 市場專業化：提供一種爬蟲機器人產品，針對不同的目標市場。 無差異行銷：針對所有的市場提供相同的爬蟲機器人產品。

P： 功能利益面：爬蟲機器人的功能利益包括自動化網路數據收集、處理與分析，提高工作效率，減少人力成本，並且能夠檢測網路上的變化，提供即時資訊。 價格比較面：在爬蟲機器人市場中，價格比較是非常重要的因素，因為許多競爭對手也提供類似的產品。因此，要考慮定價策略與價格彈性，確保價格合理且具有競爭力。 適用情境面：爬蟲機器人適用於各種行業，例如金融、電子商務、新聞媒體等，可以應用於大量數據的收集、分析、處理等工作，也可以為企業提供即時的市場情報和競爭情報。 文化象徵面：爬蟲機器人可以被看作是一個現代化和高效率的工具，這符合現代社會對於科技與效率的追求。此外，爬蟲機器人還有助於促進數據的共享和流通，這與現代社會中重視信息共享和開放的價值觀相符。 模仿競爭者：在爬蟲機器人市場中，模仿競爭者的行為是常見的，但同時也需要保持創新和差異化，以吸引更多的消費者和擴大市場佔有率。

2-4競爭力分析－SWOT

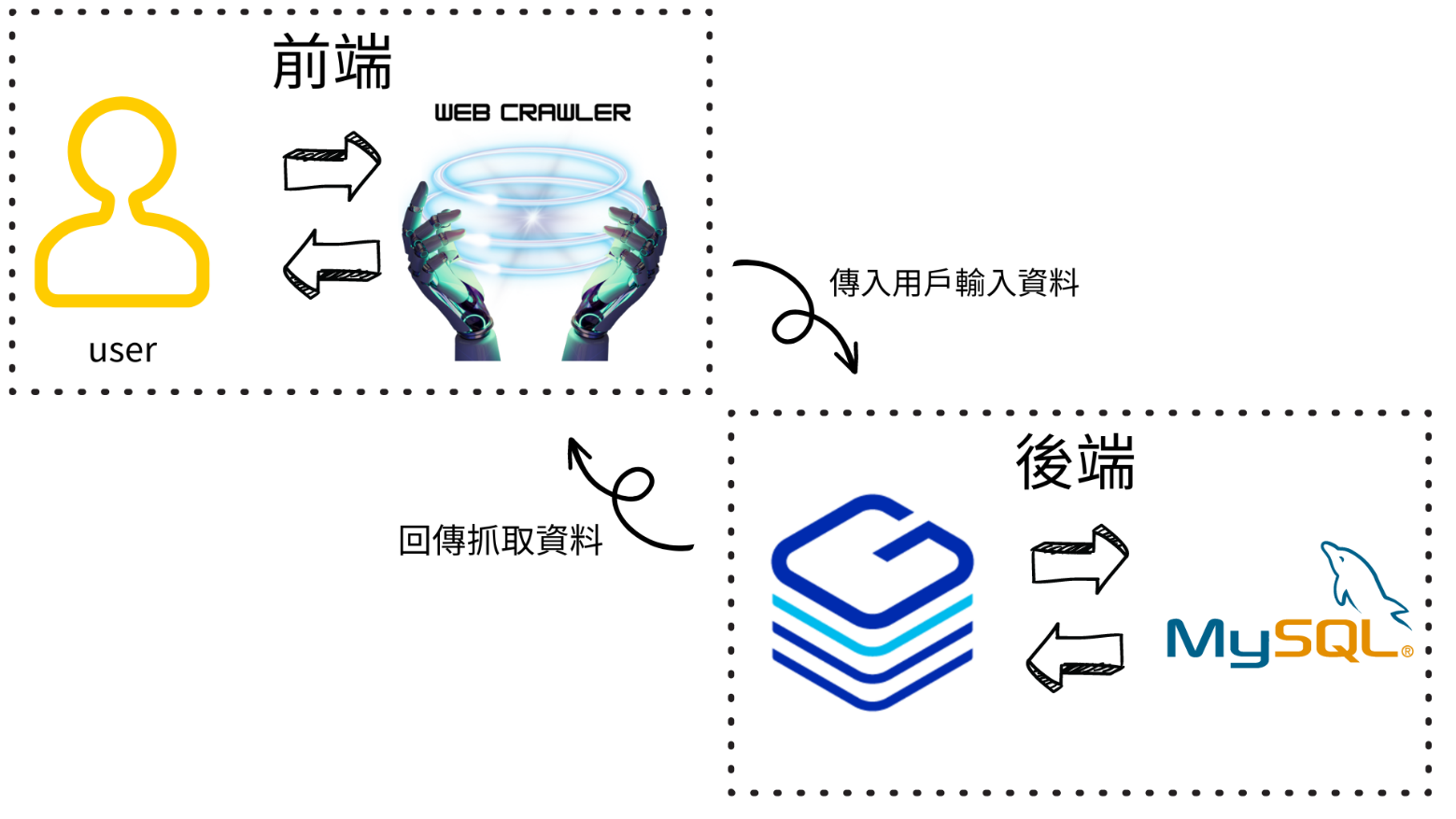
▼表 3-3-1開發標準與使用工具

|  |  |  |
| --- | --- | --- |
| 內部因素  外部因素 | 優勢 (Strengths) | 劣勢 (Weaknesses) |
| 自動化程度高：爬蟲機器人可以自動收集資料，節省人力成本且提高工作效率。 精確性高：爬蟲機器人可以快速且準確地收集資料，避免人為錯誤。 彈性大：爬蟲機器人可以針對不同網站進行自定義設置，擁有較高的彈性度。 | 需要技術支持：開發和維護爬蟲機器人需要具備較高的技術水平，這也增加了使用成本。 可能受限於網站規則：部分網站可能設置了反爬蟲機制，導致爬蟲機器人無法正常收集資料。 安全風險：爬蟲機器人可能會被惡意使用，導致資料泄漏等安全問題。 |

|  |  |  |
| --- | --- | --- |
| 機會 (Opportunities) | SO  (Strengths-Opportunities) | WO  (Weaknesses-Opportunities) |
| 巨大市場：隨著大數據和人工智慧的發展，對於自動化資料收集的需求逐漸增加，為爬蟲機器人帶來了市場機會。 新興行業的需求：部分新興行業如社交媒體監測、競爭情報收集等，需要大量的資料收集和分析，爬蟲機器人可以為這些行業提供解決方案。 整合其他技術：爬蟲機器人可以與其他技術如自然語言處理、機器學習等進行整合，為使用者提供更全面的解決方案。 | 基於爬蟲機器人的自動化和精確性高的優勢，可以針對新興行業的需求，如社交媒體監測、競爭情報收集等，提供解決方案，擴大市場佔有率。 | 基於爬蟲機器人的可定制化和彈性大的優勢，可以與其他技術如自然語言處理、機器學習等進行整合，為使用者提供更全面的解決方案，增加附加價值。 |
| 威脅 (Threats)： | ST  (Strengths-Threats) | WT  (Weaknesses-Threats) |
| 法律風險：部分網站可能設置了反爬蟲機制，這可能會涉及到法律風險。 技術風險：爬蟲機器人的技術不斷更新 | 基於爬蟲機器人的自動化和精確性高的優勢，可以針對市場競爭激烈的威脅，提供更優秀的產品和服務，增強競爭力。 | 基於爬蟲機器人的技術需要不斷更新的劣勢，可以通過合作或收購其他技術公司，補充技術優勢，增強產品的創新能力。 |

第三章 系統規格

3-1 系統架構



3-2 系統軟、硬體需求與技術平台

▼表 3-2-1系統軟、硬體需求

|  |  |  |  |
| --- | --- | --- | --- |
| 軟、硬體需求 | | | |
| 作業系統版本 | 最低系統需求 | 建議系統需求 |
| Windows 7 以上版本 | Windows 10 以上版本 |
| 處理器磁碟可用空間 | 雙核心以上 | 四核心以上 |
| 1 GB以上可用空間 | 2 GB以上可用空間 |
| RAM、網路 | 4 GB以上可用記憶體 | 8 GB以上可用記憶體 |
| 4G行動網路、Wi-Fi | |

3-3 開發標準與使用工具

▼表 3-3-1開發標準與使用工具

|  |  |
| --- | --- |
| 系統開發環境 | |
| 作業系統 | Win10、Win11 |
| 開發平台 | Python |
| 應用程式 | Visual Studio Code、 |
| 程式開發工具 | |
| 前端 | HTML、CSS、JavaScript |
| 後端 | Python |
| 文件美工工具 | |
| 文件 | Microsoft word |
| 簡報 | Microsoft PowerPoint |
| 圖樣 |  |
| 專案管理平台 | |
| 專案管理 | GitHub、GitKraken |
| 檔案存放 | MySQL、GitHub |

第四章 專案時程與組織分工

4-1 專案時程

▼表 4-1-1、專案時程

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **年** | **111年** | | | | | | | | | | |
| **月** | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 系統發想 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| 工具學習 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| 前端設計 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| 前端開發 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| 資料庫設計 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| 資料庫建置 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| 後端設計 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| 後端開發 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| 系統整合 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| 系統測試 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| 手冊製作 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| Logo製作 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| PPT製作 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

4-2 專案組織與分工

▼表4-2-1、專案組織與分工

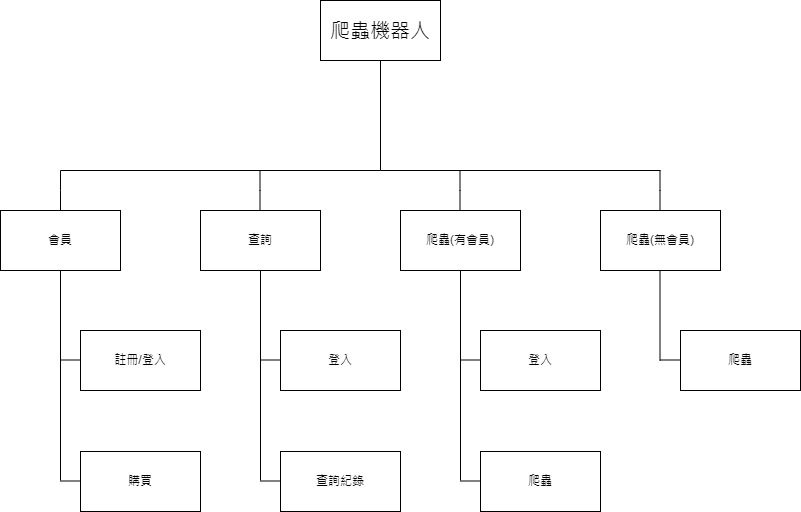
●主要負責人 〇次要負責人

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 項目/組員 | | 10856041  鍾皓年 | | 10856010  郭宗翰 | | 10856030  彭鈺程 | | 10856031  彭鈺達 | 10856032  陳永祥 | |
| 後端開發 | 資料庫建置 |  | | ● | |  | | 〇 |  | |
| 伺服器架設 | ● | | 〇 | | 〇 | |  |  | |
| Flask架構 | ● | |  | | 〇 | |  |  | |
| 爬蟲功能 | ● | | 〇 | |  | |  |  | |
| 會員系統 |  | | 〇 | | 〇 | | ● |  | |
| 前端開發 | HTML設計 | 〇 | |  | | ● | |  |  | |
| JavaScript設計 | ● | | 〇 | |  | |  |  | |
| jQuery設計 | ● | |  | |  | |  |  | |
| CSS設計 | 〇 | |  | | ● | | 〇 |  | |
| 美術設計 | UI/ UX | 〇 | |  | |  | |  | ● | |
| Web/APP介面設計 | ● | |  | |  | |  | 〇 | |
| 色彩設計 |  | | 〇 | |  | |  | ● | |
| Logo設計 |  | |  | | 〇 | |  | ● | |
| 素材設計 |  | |  | |  | | 〇 | ● | |
| 文件撰寫 | 統整 | ● | |  | |  | |  |  | |
| 第1章 前言 |  | | ● | |  | |  |  | |
| 第2章 營運計畫 |  | |  | | 〇 | | ● |  | |
| 第3章 系統規格 | 〇 |  | | ● | | |  | |  |
| 第4章專題時程與組織分工 | ● |  | | 〇 | | 〇 | |  | |
| 第5章 需求模型 |  |  | | 〇 | | ● | |  | |
| 第6章 程序模型或設計模型 |  | ● | |  | | 〇 | |  | |
| 第7章 資料模型或實作模型 |  |  | | ● | |  | |  | |
| 第8章 資料庫設計 |  | ● | |  | |  | |  | |
| 第9章 程式 | ● | 〇 | |  | | 〇 | |  | |
| 第10章 測試模型 |  |  | | 〇 | | ● | |  | |
| 第11章 操作手冊 | ● |  | |  | | 〇 | | 〇 | |
| 第12章 使用手冊 | ● | 〇 | |  | |  | | 〇 | |
| 報告 | 簡報製作 |  | 〇 | |  | |  | | ● | |
| 影片製作 | 〇 |  | |  | | 〇 | | ● | |
| 海報製作 |  |  | |  | |  | | ● | |

第五章 需求模型

5-1 功能分解圖（Functional decomposition diagram）

▼圖 5-1、功能分解圖



5-2 需求清單：

需求清單: 功能需求

1.1 支持用戶自主選擇爬取網站和版面

1.2 支持用戶自主設定爬取的時間間隔和範圍

1.3 支持對爬取的數據進行過濾、清洗、轉換等操作

1.4 提供結構化的數據格式，便於用戶進行數據分析和應用

1.5 支持對爬取過程進行監控和管理

1.6 支持用戶自主設定爬取的代理IP和其他配置

非功能需求

2.1 可靠性：爬取過程中出現錯誤時，需要能夠自動恢復或提示用戶處理

2.2 效率：需要支持高效的爬取速度，並能夠應對高併發的情況

2.3 安全性：需要對用戶數據進行加密和保護，防止數據外洩或損壞

2.4 可擴展性：需要能夠方便地擴展爬取的網站和功能

限制條件

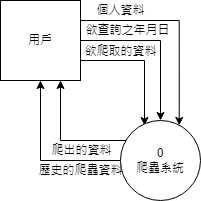
3.1 法律法規：必須遵守當地和國家的相關法律法規，如保護個人隱私等

3.2 技術限制：需要適應不同網站的技術架構和防盜採取的技術手段

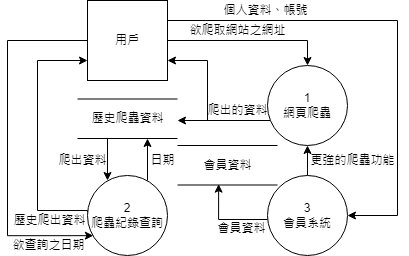
第六章 程序模型

6-1 資料流程圖(Data flow diagram)。

▼圖 6-1-1、系統環境圖



▼圖 6-1-2、圖0



6-2 程序規格書(Process specification)。

|  |  |  |  |
| --- | --- | --- | --- |
| **編號** | **1** | **功能名稱** | 網頁爬蟲 |
| **操作說明** | 利用下拉式選單來選擇爬取的網站，選擇要爬取的版面和爬取的頁數 | | |
| **輸入值** | 欲爬取的網站 | | |
| **輸出值** | 網站的資料 | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **編號** | **2** | **功能名稱** | 爬蟲記錄查詢 |
| **操作說明** | 登入會員帳號，選擇要查詢的日期或關鍵字 | | |
| **輸入值** | 欲查詢的日期或關鍵字 | | |
| **輸出值** | 歷史爬出的資料 | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **編號** | **3** | **功能名稱** | 會員系統 |
| **操作說明** | 用帳號、密碼來註冊或登入 | | |
| **輸入值** | 帳號與密碼 | | |
| **輸出值** | 登入成功、失敗 | | |

**附錄**

|  |  |
| --- | --- |
| **評審建議事項** | **修正情形** |
|  |  |