**國立勤益科技大學**

**資訊工程學系**

**Scripting程式設計-期末分組作業**

**PTT爬蟲系統**

**組別8**

**指導老師：朱孝國**

**組員：3B017154羅治圳**

**3B117139陳龍耀**

**中華民國　　114　年　1　月　6　日**

# 摘要

本論文旨在探討並實作一個基於Python的PTT爬蟲系統。透過運用Selenium進行網頁自動化操作，結合SQLite3資料庫進行資料存儲，以及Tkinter設計圖形化使用者介面，實現自動抓取PTT指定版面的文章標題並進行查詢與分析。本文詳細說明了系統的設計思路、實作方法、技術選型及其應用效果，並在實作過程中探討了所面臨的挑戰與解決方案。研究結果顯示，本系統能有效地自動化抓取PTT資料，並提供便利的資料查詢功能，具備實際應用價值。

**關鍵詞**：Python、網路爬蟲、Selenium、SQLite3、Tkinter、PTT

目錄

[摘要 1](#_Toc186755976)

[目錄 2](#_Toc186755977)

[第一章 緒論 3](#_Toc186755978)

[1.1 前言 3](#_Toc186755979)

[1.2 文獻探討 3](#_Toc186755980)

[第二章 作業說明 4](#_Toc186755981)

[2.1 學習目標 4](#_Toc186755982)

[2.2 預期效益 4](#_Toc186755983)

[2.3 環境介紹 4](#_Toc186755984)

[第三章 實作方法 5](#_Toc186755985)

[3.1 環境檢查 5](#_Toc186755986)

[3.2 安裝與設定 5](#_Toc186755987)

[3.2 系統設計與實作步驟 5](#_Toc186755988)

[3.2.1建立資料庫與資料表 5](#_Toc186755989)

[3.2.2設計圖形化介面 6](#_Toc186755990)

[3.2.3實作爬取功能 7](#_Toc186755991)

[3.2.4實作搜尋功能 8](#_Toc186755992)

[3.2.5錯誤處理與使用者提示 8](#_Toc186755993)

[第四章 實作心得 9](#_Toc186755994)

[4.1 心得 9](#_Toc186755995)

[4.2 結論 9](#_Toc186755996)

[參考文獻 10](#_Toc186755997)

第一章 緒論

1.1 前言

隨著網際網路的普及，網路爬蟲技術在資料收集、分析及應用領域中扮演著重要角色。PTT作為台灣最大的網路論壇之一，擁有豐富的討論內容和大量的使用者互動，是研究網路輿情、熱門話題分析的重要資料來源。然而，手動收集PTT資料不僅效率低下，且難以保證資料的及時性和完整性。為此，開發一個自動化的PTT爬蟲系統具有重要的實用價值。

本研究旨在設計並實作一個基於Python的PTT爬蟲系統，通過自動化抓取PTT指定版面的文章標題，並將其儲存至本地資料庫，進行後續的查詢與分析。系統的開發過程涵蓋了網頁自動化操作、資料庫管理及圖形化介面設計等多個技術領域，旨在提供一個完整的解決方案，滿足實際應用需求。

1.2 文獻探討

網路爬蟲是指自動化地瀏覽和抓取網頁內容的程式，廣泛應用於資料收集、搜尋引擎、價格比較等領域。常見的網路爬蟲工具包括BeautifulSoup、Scrapy和Selenium等。其中，BeautifulSoup適用於靜態網頁的解析，Scrapy則是一個功能強大的爬蟲框架，適合大規模資料抓取；而Selenium則主要用於處理動態網頁的自動化操作，模擬用戶行為，抓取動態生成的內容。

研究者常利用PTT資料進行輿情分析、社會熱點追蹤及市場調查等。例如，林（2020）利用PTT資料分析了台灣選舉期間的輿情變化，揭示了網路討論對選舉結果的影響。陳（2021）則運用PTT資料進行了產品評價分析，為企業提供了有價值的市場反饋。

資料庫是網路爬蟲系統中不可或缺的一部分，用於存儲和管理抓取的資料。SQLite3作為一個輕量級的資料庫管理系統，因其無需安裝伺服器、易於使用且與Python高度兼容，被廣泛應用於小型應用和開發測試中。相較於MySQL和PostgreSQL等大型資料庫，SQLite3在本地開發環境中具有更高的靈活性和便捷性。

使用者介面是系統與使用者互動的橋樑，良好的介面設計能提升使用者體驗。Tkinter作為Python內建的GUI庫，具有簡單易用輕量化的特點，適合快速開發桌面應用程式。透過Tkinter，開發者能夠設計直觀的使用者介面，提供友善的操作環境

第二章 作業說明

2.1 學習目標

本次作業的主要目標是透過實作PTT爬蟲系統，熟悉並掌握以下技能：

**理解網路爬蟲的基本概念與應用**：學習如何自動化抓取網頁資料，並應用於實際案例中。

**熟悉Python及其相關套件的使用**：掌握Selenium進行網頁自動化操作，運用SQLite3進行資料庫管理，利用Tkinter設計圖形化介面。

**提升問題解決能力與團隊合作能力**：在開發過程中，面對並解決各種技術挑戰，同時通過分工合作提升團隊協作能力。

2.2 預期效益

通過此次作業，預期將達成以下效益：

**程式設計能力提升**：熟練運用Python進行各項功能的實作，增強編程技巧。

**資料處理能力增強**：掌握資料抓取、儲存及查詢的完整流程，提高資料處理效率。

**應用開發能力提高**：了解並運用多種技術工具，完成一個功能完善的應用系統。

**團隊合作經驗積累**：通過分工合作，提升團隊協作能力，學會有效溝通與協調。

2.3 環境介紹

硬體環境

處理器：Intel Core i5，四核心，具備良好的運算能力。

記憶體：8GB RAM，能夠順暢運行多個開發工具與瀏覽器。

儲存設備：256GB SSD，提供快速的資料讀取與寫入速度。

軟體環境與使用工具

作業系統：Windows 10，具備良好相容性，支持多種開發工具與驅動程式。

開發工具：

編輯器：Visual Studio Code，輕量化且擴充性強，支持多種語言與插件。

Python：版本3.x，具備強大的第三方庫支援。

瀏覽器：Microsoft Edge，搭配相應版本的Edge WebDriver，實現自動化操作。

資料庫管理系統：SQLite3，輕量級且易於整合，適合本地開發與測試。

第三章 實作方法

3.1 環境檢查

在開始開發前，需確認開發環境的配置是否完整且正確

Python安裝：確認已安裝Python 3.x版本，並設置環境變數以便在命令行中直接使用python指令。

必要套件安裝：使用pip安裝所需的第三方庫，包括selenium與tkinter。Bash pip install selenium注：tkinter通常隨Python一同安裝，無需額外安裝。

WebDriver配置：下載與瀏覽器版本匹配的Edge WebDriver，並確保其路徑正確設置。下載地址：[Microsoft Edge WebDriver](https://developer.microsoft.com/en-us/microsoft-edge/tools/webdriver/" \t "_new)

3.2 安裝與設定

安裝Python套件： 除了安裝selenium，需確認tkinter已包含在Python安裝中。若未包含，可參考Python官方文檔進行安裝。

bash pip install selenium

下載Edge WebDriver： 從Microsoft官方網站下載對應版本的Edge WebDriver，C:/Users/USER/Desktop/FinalExam/FinalExam/msedgedriver.exe。確保WebDriver的版本與Microsoft Edge瀏覽器版本相匹配，以避免兼容性問題。

3.2 系統設計與實作步驟

### 3.2.1建立資料庫與資料表

使用SQLite3建立名為scrape.db的資料庫，並建立一個儲存標題的資料表titles。資料表結構如下：id：主鍵，自動遞增。cname：儲存文章標題，文本類型。

一張含有 文字, 螢幕擷取畫面, 軟體, 字型 的圖片

自動產生的描述

### 3.2.2設計圖形化介面

利用Tkinter設計主介面，包括：

URL輸入欄位：讓使用者輸入或選擇欲爬取的PTT版面URL。

文字顯示區域：顯示爬取結果，提供使用者瀏覽抓取到的標題。

操作按鈕：

開始爬取：觸發爬取動作。

搜尋：根據關鍵字篩選顯示特定標題。

刷新內容：重新爬取並更新資料。

一張含有 文字, 螢幕擷取畫面, 軟體, 作業系統 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 軟體, 作業系統 的圖片

自動產生的描述

### 3.2.3實作爬取功能

使用Selenium自動化Edge瀏覽器，打開指定URL，抓取目標標題，並將其顯示於介面及儲存至資料庫。

一張含有 文字, 螢幕擷取畫面, 軟體 的圖片

自動產生的描述

**一張含有 文字, 螢幕擷取畫面, 軟體 的圖片

自動產生的描述**

### 3.2.4實作搜尋功能

提供使用者輸入關鍵字，從已抓取的標題中篩選包含關鍵字的項目，並顯示於介面。

一張含有 文字, 螢幕擷取畫面, 軟體, 陳列 的圖片

自動產生的描述

### 3.2.5錯誤處理與使用者提示

在爬取過程中加入錯誤處理機制，確保系統的穩定性與使用者體驗。使用try-except語句捕捉可能的異常，並通過彈出提示框告知使用者錯誤原因。

一張含有 文字, 螢幕擷取畫面, 字型, 行 的圖片

自動產生的描述

第四章 實作心得

4.1 心得

在本次期末作業中，我們團隊選擇開發PTT爬蟲系統，旨在實踐網路爬蟲的基本概念並應用於實際案例中。整個開發過程中，我深刻體會到理論與實作之間的差距，尤其是在處理網頁動態內容與資料庫整合時，遇到了不少挑戰。

WebDriver兼容性問題： 初期在配置Edge WebDriver時，由於瀏覽器版本與WebDriver不匹配，導致爬取失敗，花費較多時間解決。後來通過確認瀏覽器版本並下載對應的WebDriver版本，問題得以解決。資料庫操作的同步性： 在多線程操作中，確保資料庫的同步更新是一大挑戰。由於本系統為單線程設計，這部分問題相對較少，但在未來擴展時需注意資料庫操作的同步性問題。使用者介面的友善性： 如何設計一個既簡潔又功能完善的介面，讓使用者能夠輕鬆操作，是我們反覆調整的重點。最終透過反覆測試與調整，達到了較為理想的使用者體驗。

錯誤處理與穩定性： 在爬取過程中，網頁結構的變動可能導致元素定位失敗，需要加入靈活的錯誤處理機制，以提升系統的穩定性。

在團隊合作方面，每位成員都充分發揮所長，分工明確，有效地提升了開發效率。然而，由於部分成員對Selenium不熟悉，初期在學習與應用上花費了較多時間，這也讓我們更加重視知識分享與團隊協作的重要性。透過定期的討論與協作，我們成功克服了技術上的困難，並在限定時間內完成了任務。

透過本次作業，我不僅提升了Python程式設計的能力，還深入理解了網路爬蟲的實作細節。尤其是在面對實際問題時，如何有效地進行問題分解與解決，是一個寶貴的學習經驗。此外，與團隊成員的合作也讓我體會到溝通與協調的重要性，這將對未來的學習與工作產生積極的影響。最終，這次作業不僅增強了我的技術能力，也提升了我的團隊合作與問題解決能力。

4.2 結論

本研究成功設計並實作了一個基於Python的PTT爬蟲系統，實現了自動抓取PTT指定版面的文章標題，並將其儲存至本地資料庫，提供使用者進行查詢與分析的功能。系統運用Selenium進行網頁自動化操作，結合SQLite3進行資料管理，並利用Tkinter設計圖形化介面，形成一個完整的解決方案。實作過程中，我們面對並克服了多項技術挑戰，提升了程式設計、資料處理及團隊合作能力。

未來的研究可以針對以下幾個方向進行改進與擴展：

多線程與異步操作：提升爬取效率，實現更大規模的資料抓取。

資料分析與可視化：在資料庫基礎上，加入資料分析模組，並設計可視化介面，提供更深入的數據洞察。

跨平台支援：擴展系統至其他作業系統，如macOS與Linux，提升系統的適用性。

爬取內容擴展：不僅抓取標題，還可抓取文章內容、作者信息及發佈時間等，豐富資料庫內容。

總之，本研究所開發的PTT爬蟲系統具備實用性與可擴展性，為未來相關應用的開發與研究提供了參考與基礎。

**4.3 附註工作分配**

|  |  |  |
| --- | --- | --- |
| 成員 | 負責項目 | 比例 |
| 羅治圳 | 程式設計 | 100 |
| 陳龍耀 | PPT及WORD | 100 |

備註：各成員在分工上有部分重疊，如系統整合與測試及報告由全體共同參與。

參考文獻

[1] Selenium官方文件(n.d.). Retrieved from https://www.selenium.dev/documentation/

[2] Tkinter文檔. (n.d.). Retrieved from https://docs.python.org/3/library/tkinter.html

[3] SQLite官方網站. (n.d.). Retrieved from https://www.sqlite.org/index.html

[4] MicrosoftEdgeWebDriver.(n.d.). Retrieved from https://developer.microsoft.com/en-us/microsoft-edge/tools/webdriver/

[5] PTT使用手冊. (n.d.). Retrieved from https://www.ptt.cc/bbs/index.html

[6] Python官方網站. (n.d.). Retrieved from https://www.python.org/

[7] 林, 小明. (2020). PTT輿情分析研究. 台北市: 台灣大學出版.

[8] 陳, 大華. (2021). 基於PTT的產品評價分析. 高雄市: 高雄科技大學出版社.