臺灣股市每日資料自動擷取系統開發

以上市與上櫃股票為例

組員：C112151113 張勝鈞

組員：C112151133 蔡長浩

目錄

[一、 摘要 2](#_Toc201959222)

[二、 研究背景與動機 2](#_Toc201959223)

[三、 文獻探討 2](#_Toc201959224)

[四、 研究方法 3](#_Toc201959225)

[1. 系統輸入： 3](#_Toc201959226)

[2. 系統流程與模組 3](#_Toc201959227)

[3. 技術細節 3](#_Toc201959228)

[五、 結論與展望 3](#_Toc201959229)

[六、 參考資料 4](#_Toc201959230)

## 摘要

本專題目標是設計一套能自動擷取臺灣股市中上市（TWSE）與上櫃（TPEx）公司每日收盤價格資料的系統。系統使用 Python 語言，結合 requests、pandas、BeautifulSoup 等工具，搭配多執行緒提升整體效率。使用者可以透過系統自動抓取近半年內的歷史交易資料，並將結果儲存成各自的 CSV 檔案。另外，系統會自動排除上櫃的 ETF 類別，以避免非標準股票影響資料處理。此系統將可協助投資人、研究者有效蒐集與分析臺灣股票市場資料。

## 研究背景與動機

臺灣股票市場涵蓋上千家上市與上櫃公司，其每日交易資料對於投資研究與量化分析非常重要。雖然交易所官方網站都提供歷史資料查詢功能，但多數需要透過手動操作、複雜選單與篩選後下載，效率不彰且容易出錯。此外，ETF 類商品雖亦具代號，但其價格與成分股、追蹤指數密切相關，並不適合與一般股票一同分析。這個專題希望透過程式設計實現：

1. 自動化擷取每日股價資料。
2. 支援多檔股票批次查詢。
3. 排除 ETF 資料，確保分析資料的一致性。
4. 支援近半年歷史資料，供後續回測與分析使用。

## 文獻探討

在相關領域中，網路爬蟲技術已被廣泛應用於金融資料擷取（如使用 requests + BeautifulSoup 搭配 API 提取結構化資料）。TWSE 與 TPEx 均有提供 JSON 格式資料，可透過程式模擬請求加以擷取；而資料處理與儲存則普遍使用 pandas 進行欄位處理與 CSV 存檔。多執行緒（ThreadPoolExecutor）與非同步技術在高頻資料抓取中被證實可大幅降低總耗時。對 ETF 的分類與識別，也可透過簡單的欄位判別實作排除規則。

## 研究方法

1. 系統輸入：

股票清單為三份 CSV 檔：

1. list\_company\_number.csv（上市公司）
2. over\_the\_counter\_number.csv（上櫃公司）
3. emerging\_stock\_market.csv（興櫃公司）

2. 系統流程與模組

|  |  |
| --- | --- |
| 模組 | 功能 |
| 股票代號擷取 | 從 CSV 檔中讀取第一欄代號，轉為字串陣列 |
| ETF 過濾 | 若該股票資料欄位數不足（通常無產業別），即判定為 ETF 並略過 |
| 資料來源 | 上市：TWSE🡺API（GET） 上櫃：TPEx🡺POST 請求 |
| 多執行緒 | 使用 ThreadPoolExecutor 同時查詢多支股票 |
| 檔案儲存 | 每支股票建立獨立 CSV 檔案，存於 daily\_stock\_data 資料夾中 |

3. 技術細節

* 1. 使用 datetime 與 timedelta 判定半年範圍
  2. 每支股票資料包含：「日期」、「收盤價」
  3. 網址動態拼接並處理查詢參數（上市用 GET；上櫃用 POST）
  4. 使用 headers 模擬正常瀏覽器請求，避免遭封鎖

## 結論與展望

本專題目前已完成以下項目：

1. 自動讀取上市與上櫃股票代號
2. 自動擷取半年內每日股價資料（例如：上市：twse API ）
3. 排除 ETF 股票（以欄位資料進行過濾）
4. 多執行緒提升擷取效率（10 執行緒）
5. 每支股票儲存為獨立檔案（CSV 格式）

未來展望：

1. 擴充支援興櫃公司收盤價爬取（已取得來源頁面與格式）。

2. 加入自動錯誤重試與日誌記錄機制，提升穩定性。

3. 加入簡易視覺化工具，協助資料初步分析（折線圖、趨勢圖等）。

4. 讓程式自動執行，每日定時更新資料。

## 參考資料

1. 臺灣證券交易所（TWSE）：<https://www.twse.com.tw/>
2. 臺灣櫃檯買賣中心（TPEx）：<https://www.tpex.org.tw/>
3. Python 官方文件：<https://docs.python.org/3/library/concurrent.futures.html>