2022 Python網路爬虫實戰研習馬拉松-期末專題

# 目錄

[2022 Python網路爬虫實戰研習馬拉松-期末專題 1](#_Toc143959871)

[目錄 1](#_Toc143959872)

[一、專題摘要 2](#_Toc143959873)

[1.期末專題主題-PTT Stock 版 2](#_Toc143959874)

[2.期末專題基本目標 2](#_Toc143959875)

[二、實作方法介绍 2](#_Toc143959876)

[1.使用的程式碼介紹 2](#_Toc143959877)

[2.使用的模組和程式碼整合介紹 9](#_Toc143959878)

[三、成果展示 13](#_Toc143959879)

[1.PTT爬文結果 13](#_Toc143959880)

[2.匯入給ChatGPT總結 19](#_Toc143959881)

[四、結論 22](#_Toc143959882)

[1.文帖網址連結異常： 22](#_Toc143959883)

[2.文帖分類未用[…] 24](#_Toc143959884)

[3.文章未用「--」分隔主文和推文回應 25](#_Toc143959885)

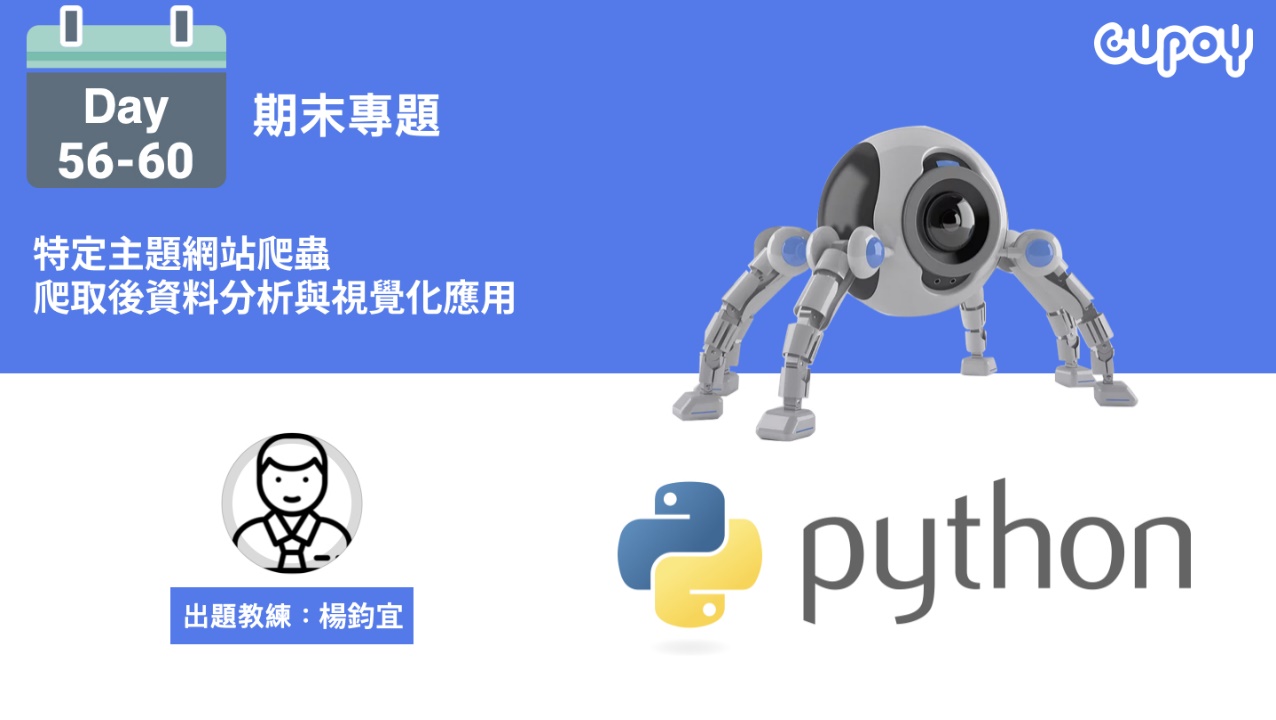
[4.ChatGPT API的應用和限制 25](#_Toc143959886)

[5.非同步方式爬文 27](#_Toc143959887)

[6文章作者、推文者統計 28](#_Toc143959888)

[五、期末專題作者資訊 31](#_Toc143959889)

[圖表目錄 31](#_Toc143959890)



# 一、專題摘要

## 1.期末專題主題-PTT Stock 版

本專題以PTT Stock股票版網頁每帖文章為網路爬虫爬取內容，整合之前所學的爬虫語法，搜集和整理PTT Stock版頁文章相關資訊，並將主帖主文內容傳給ChatGPT分析文章重點(summary )。

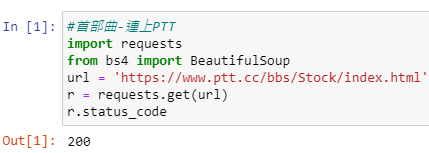
## 2.期末專題基本目標

PTT網站是以文帖為主要內容的靜態網頁，而且CSS格式固定，不需使用動態網頁爬虫技術。但網頁文章若沒處理會有很多空行，本專題首要主標是將網站整頁文帖爬文後清理文字內容再存檔，並可設定要爬文的帖頁數(程式碼也可指定整個版面所有文帖全部爬下存檔)，並使用非同步方式爬取主文，本專題範例以爬取約75篇文帖，比較非同步和一般方式效能差異，此外可統計分析貼文和推文者的發文次數，並結合ChatGPT API可取得主文摘要，方便查閱文章資訊重點。

# 二、實作方法介绍

## 1.使用的程式碼介紹

先測試基本語法，確定能正常執行，再將這些程式碼整合為函式。



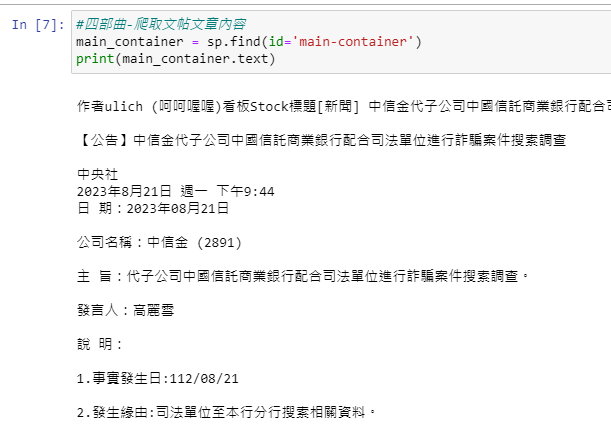
圖表 1 首部曲-連上PTT Stock



圖表 2 二部曲-爬取文帖標題和超連結



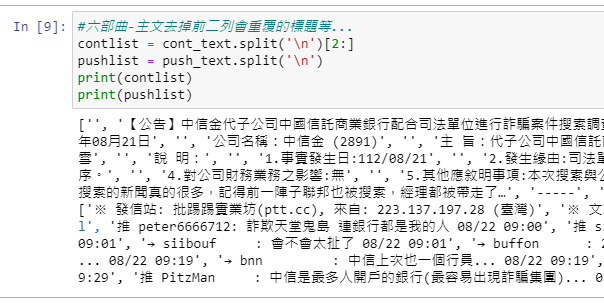
圖表 3 三部曲-測試爬取PTT文帖作者、看板、標題、日期



圖表 4 四部曲-爬取文帖文章內容



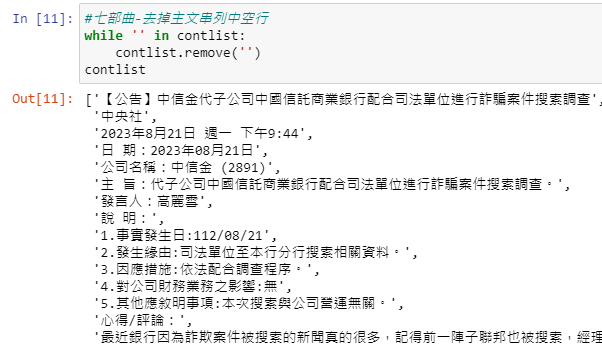
圖表 5 五部曲-以'\n--\n'分割主文和推文



圖表 6 六部曲-主文去掉前二列會重覆的標題等



圖表 7 主文轉串列後會有很多空行



圖表 8 七部曲-去掉主文串列中空行



圖表 9 文帖推文留言處理程式碼測試



圖表 10 留言者以字典統計方式

## 2.使用的模組和程式碼整合介紹

經過前面Python程式碼測試結果，我們先定義要爬PTT文章，要寫入檔案或串列內容的的程式碼。

這範例因PTT是標準靜態網頁，本範例僅用下列：

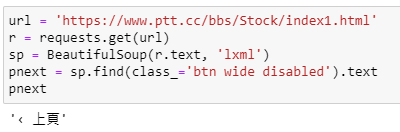
* import requests
* from bs4 import BeautifulSoup
* imprt time
* import openai

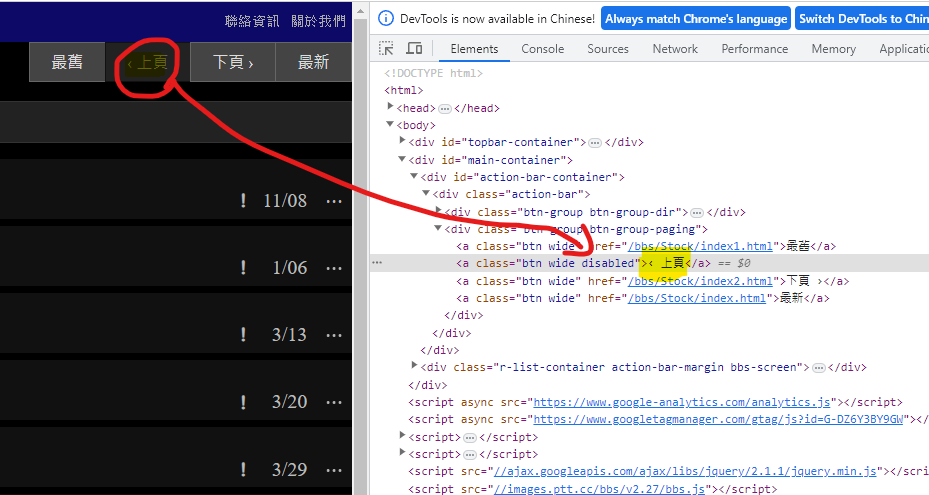


圖表 11 根據前面測試結果自訂爬PTT文章函式



圖表 12 用for迴圈設定爬PTT文帖的頁數方式爬文



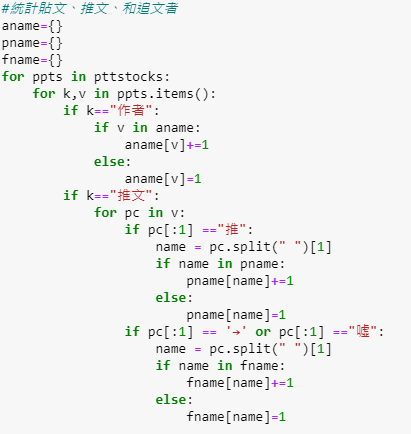


圖表 13 PTT最新和最舊會出現'btn wide disabled'類別

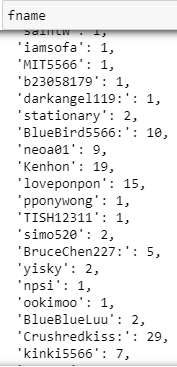
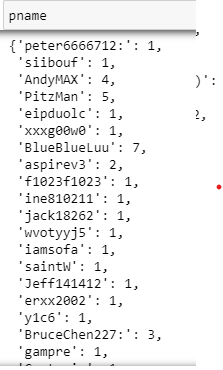
如上圖13 PTT文帖若爬到最新或最後一頁CSS中會出現'btn wide disabled'，若等於文字為'‹ 上頁'，表代沒有其他文帖，作為爬文停止點。我們也可用while 迴圈爬完所有文章(圖14)。



圖表 14 用while爬取PTT全部文的網頁程參考程式碼



圖表 15 用字典統計貼文、推文、和追文者程式碼

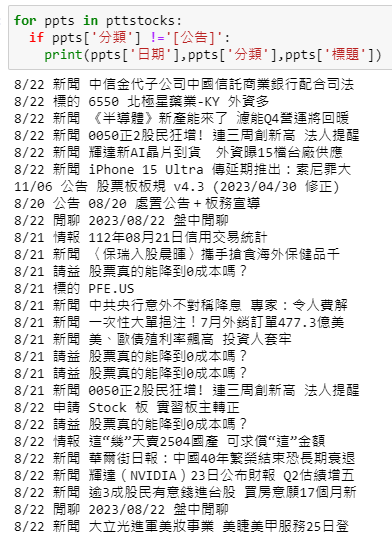


圖表 16 用字典統計貼文、推文、和追文者範例結果

# 三、成果展示

## 1.PTT爬文結果

以8/22爬取PTT Stock最近二頁文帖內容結果如下；



圖表 17爬取8/22 PTT Stock最近二頁文帖標題範例

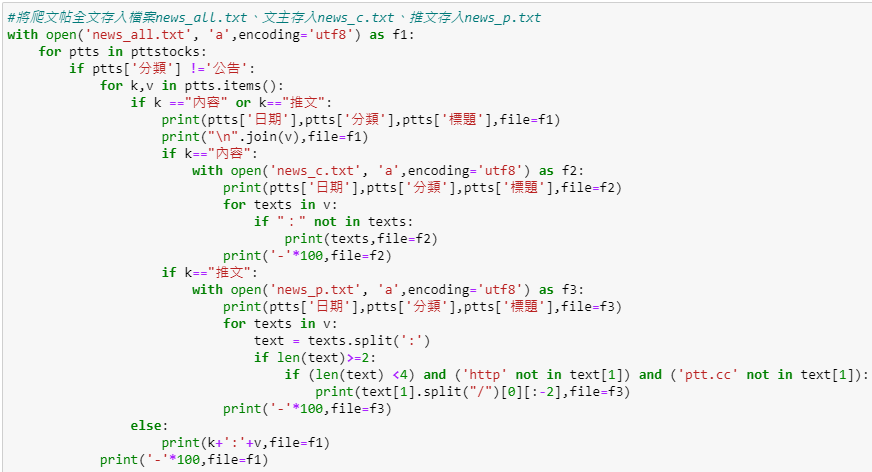


圖表 18輸出文帖全文的範例程式碼



圖表 19只輸出文帖主文和推文內容的範程式碼

最後整合前面程式碼，將爬文帖全文存入檔案news\_all.txt、文主存入news\_c.txt、推文存入news\_p.txt，作為傳入ChatGPT摘要分析用(圖20)。

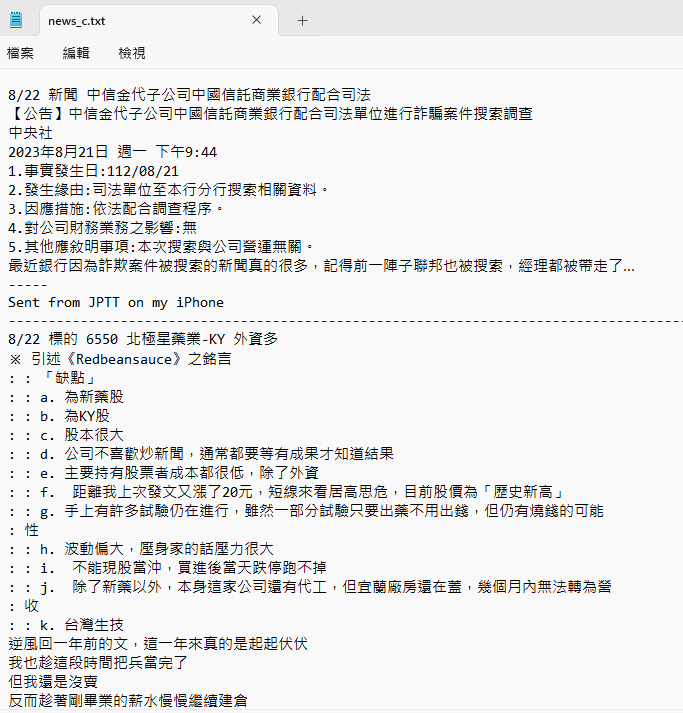


圖表 20 將爬文帖全文存入檔案範例程式碼

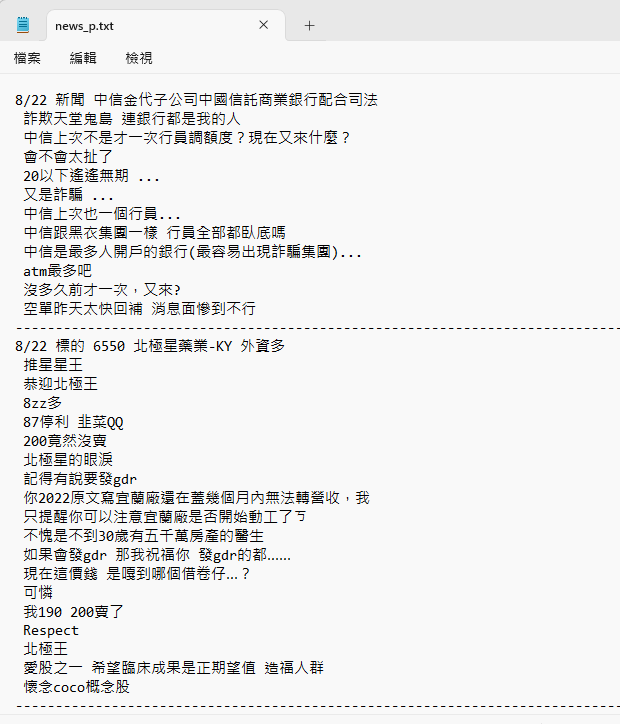
參考結果如下面三圖(圖21-23)；



圖表 21 news\_all.txt文帖全文範例



圖表 22 news\_c.txt文帖主文範例



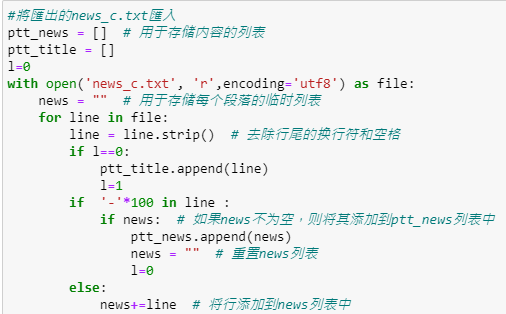
圖表 23 news\_p.txt文帖推文回應範例

## 2.匯入給ChatGPT總結

有了news\_c.txt我們可以將之匯入程式，再傳給Chat GPT去總結，再匯出成summary.txt，範例程式碼如圖24-27，結果如圖28。



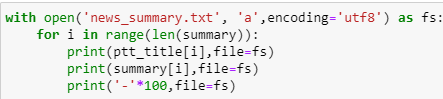
圖表 24定義和ChatGPT API溝通函式



圖表 25匯入news\_c.txt



圖表 26 用for迴圈傳給ChatGPT範例程式



圖表 27 將 ChatGPT摘要結果存檔範程式



圖表 28 ChatGPT摘要結果

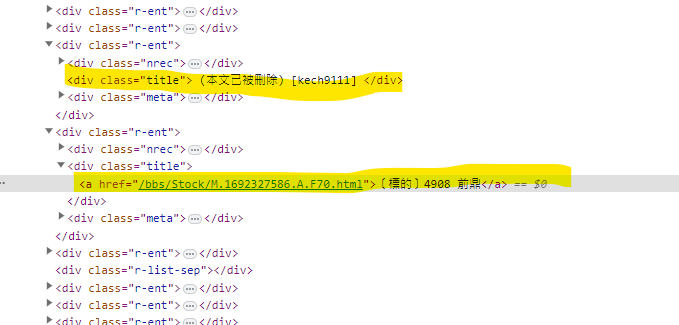
# 四、結論

我們在測試PTT網站上文帖爬文時，遇到一些特例，下面分享處理方式。

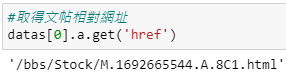
## 1.文帖網址連結異常：



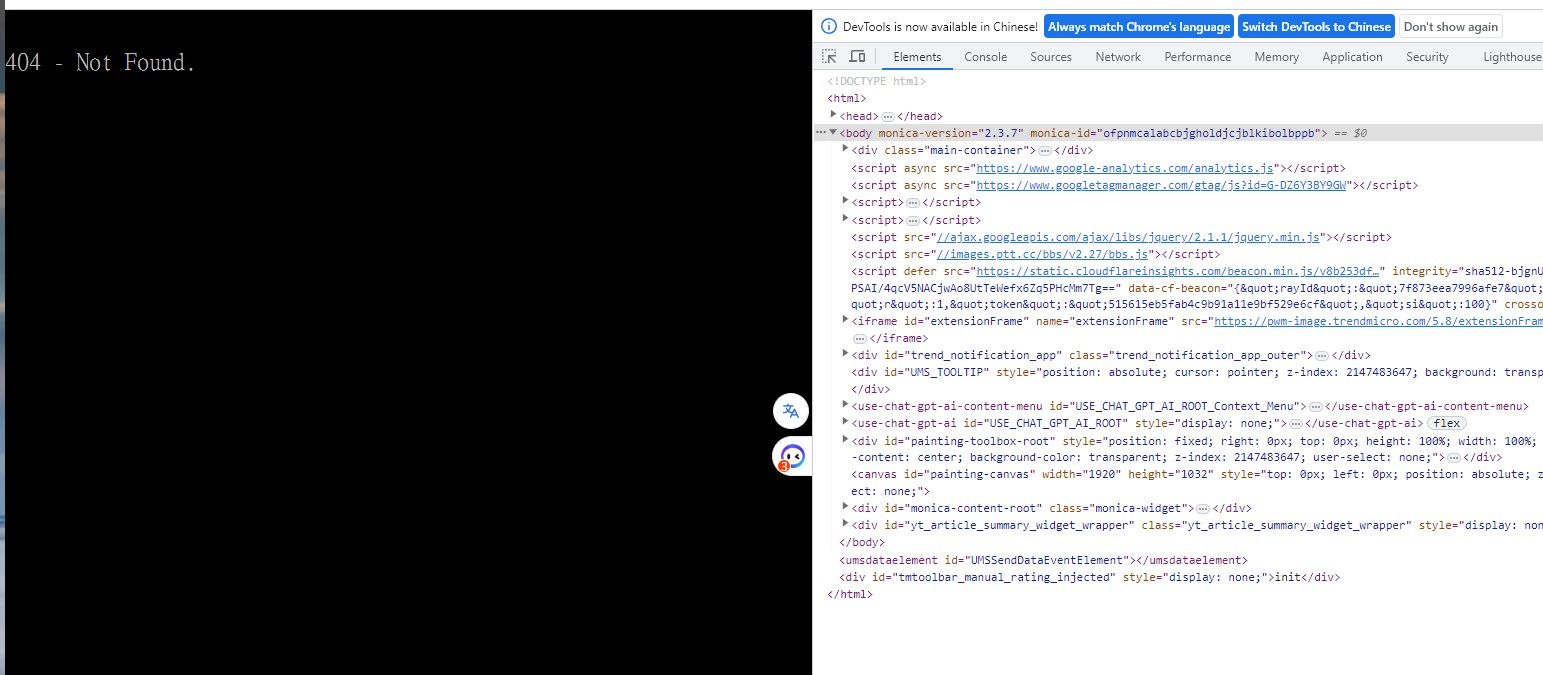
圖表 29 文帖連結異常但未標示已刪除特例

正常被刪的文是沒有超連結網址如下圖，就是沒有<a>TAG(圖30)。

圖表 30 PTT刪文和一般文章之差異-沒有<a>TAG



但有些文章超連結有問題(圖29)，版主尚未標註(本文己被刪除)，其網址是有問題的，程式執行會出現404 ERROR，如下圖31。



圖表 31 PTT 文帖出404現異常特例

跳過刪文只要判別「 if data.a: 」即可，就是找的到<a>TAG一定不是刪文，但測試時難得遇到特例，沒標註刪文但文帖對應的網址有問題，我們在函式中加了「 if rc.status\_code ==200: 」來避免這錯誤。

隔10幾分那個有問題的文帖也標記被刪如圖32：



圖表 32 異常文帖會被標示成-已被刪除

如果在我們在函式中程式碼

urlc ='https://www.ptt.cc' + data.a.get('href')

rc = requests.get(urlc)

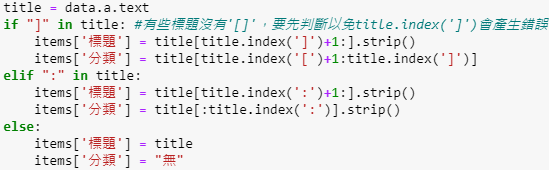
之後沒加上「 if rc.status\_code ==200: 」，接著執行下方的程式就會因找不到元素而掛了。

## 2.文帖分類未用[…]

PTT文帖幾乎99.9%標題一定會用中刮號來標註分類，但偶爾會遇到0.1%的例外，因我的程式是用”]“來找分類和標題，很難得遇到這0.1%的例外，如圖：



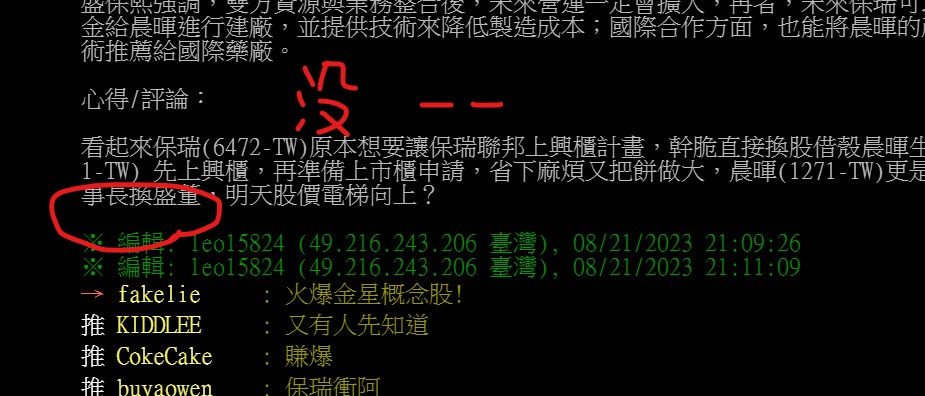
為了這0.1%的例外，在取得標題和分類的部份，程式碼增加了如果”]“不存在，甚至連”：”也沒有時的分類方式。



圖表 33解決 PTT文帖分類未使用中刮號的程式範例

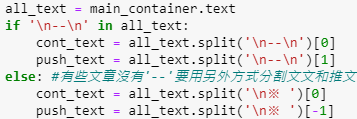
## 3.文章未用「--」分隔主文和推文回應

相同的PTT文帖幾乎99.99%文章一定會用「--」來分隔主文和回應，但偶爾會遇到0.01%的例外，如圖：



圖表 34 未使用「--」來分隔主文和回應特例

這種文帖可能是輯者誤刪的，因一般文主和回應推文間一定會有「--」，但這個部份程式是用「\n--\n」來取分和其他「-------」的分別，為了這0.01%的特列，程式修改如下：



圖表 35 解決未使用「--」來分隔主文和回應的程式碼

若「\n--\n」不存在時(多半是編輯時刪掉的，所以一定會有記錄)，以'\n※ '為分隔依據，因不確文章被編輯幾次，所以用…[**-1**](圖35)，代表取分隔最後部份。

## 4.ChatGPT API的應用和限制

因目前我還不是ChatGPT4付費會員，只能用ChatGPT3.5 API測試，因ChatGPT非常適合總結文章，對於爬取的PTT文章內容，多則千字以上，若可以簡略摘要，方便查閱，相信是十分實用的爬虫應用。

ChatGPT3.5的token限制，是以提問和回應總和字數計，文章內容超過2000千字以上，再加上回應文字容易產生錯誤如圖36：



圖表 36 ChatGPT3.5 token超過上限

這不是我們程式碼的問題，若升級到ChatGPT4，這問題會得到改善。GPT-4 API 的預設速率限制是token上限8K-8192(32K給特定用戶)和每分鐘 200 個請求。但我們用的ChatGPT3.5，每分只能有3個請求(平均每隔20秒一次)，而且token不能超過4000左右..@@!!..

若只是想測試ChatGPT3.5 API，目前想到的方式是主動限制文章內容不要超過2000字左右，程式碼如下；



圖表 37解決Chat GTP 3.5 token上限參考方式

如此ptt\_news就不會超過2000字，一下面程式碼就不會因token超過上限而終止。

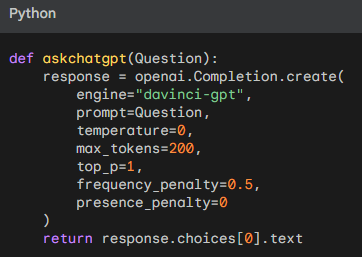


此外程式改用for迴圈如下圖，當ChatGPT3.5 API有異狀時，可以由len(summary)知道，要再從那段文章再開始總結。



圖表 38 解決ChatGPT3.5 API回應異常參考程式碼

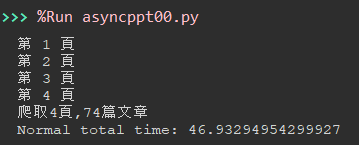
我也有問 Google Bard，如何連上GPT4 API，Bard給的回應如下(圖39)，理論上只要改engine=”dovinic-gpt”，回應的max\_tokens=可設512-1024大一些。因目前未付費，僅能提供程式碼參考。



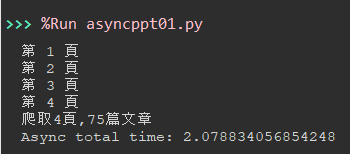
圖表 39 GPT4 API串接範例程式

## 5.非同步方式爬文

若我們取得PTT Stock股票版前面4 頁所有文帖的urls(約74-75頁)，再由迴圈方式爬完所有文章，若以一般方式約46秒(圖40)。但以非同步方式只需2秒左右。

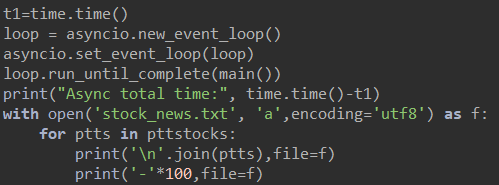


圖表 40 由迴圈方式爬完74篇文章約需47秒



圖表 41 非同步方式爬完75篇只需2秒左右

和前面程式碼相同方式，可將非同步爬取之文章存入檔案，再傳給ChatGPT總結(請參考圖37-38)。

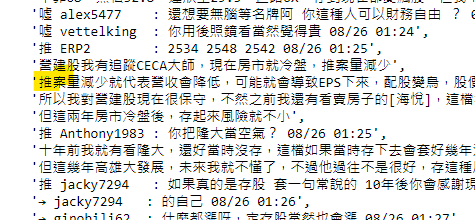


圖表 42 將非同步爬取之文章存入檔案範例程式

完整程式碼可由GitHub下載…^\_^…

## 6文章作者、推文者統計

圖15 為用字典統計貼文、推文、和追文者程式碼，測試發現一個0.001%特例會造成程式錯誤(圖43)。



圖表 43 推文中特例會造成程式執行錯誤

圖43因”推案量…”這行第一個字是”推”，但沒空格， pc.split(" ")[1]，

會造成out of index錯誤。因此增加判斷len(pc.split(" "))是否大於1。

if pc[:1] =="推":

if len(pc.split(" "))>1:

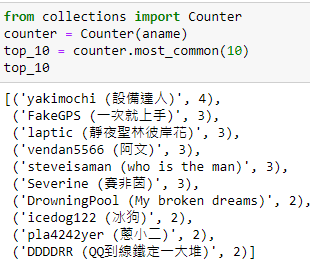
name = pc.split(" ")[1]

圖15修正後程式碼如圖44。

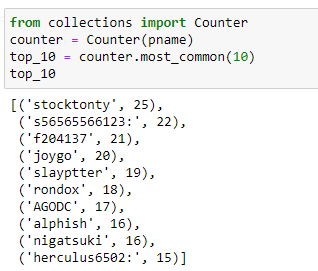


圖表 44 圖15修正後程式碼

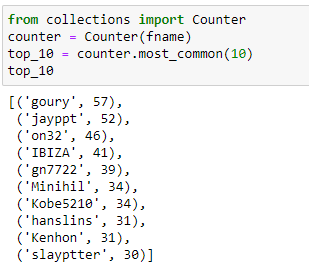
我們爬取8月26日前五頁約99篇文帖，前10名統計結果如圖45-47。



圖表 45 8/26約99篇文帖作者10名統計



圖表 46 8/26約99篇文帖推文者10名統計



圖表 47 8/26約99篇文帖追文者前10名統計

# 五、期末專題作者資訊

1.個人Github連結 <https://github.com/zionjac/PTT>

2.強哥



# 圖表目錄

[圖表 1 首部曲-連上PTT Stock 2](#_Toc143959810)

[圖表 2 二部曲-爬取文帖標題和超連結 2](#_Toc143959811)

[圖表 3 三部曲-測試爬取PTT文帖作者、看板、標題、日期 3](#_Toc143959812)

[圖表 4 四部曲-爬取文帖文章內容 4](#_Toc143959813)

[圖表 5 五部曲-以'\n--\n'分割主文和推文 4](#_Toc143959814)

[圖表 6 六部曲-主文去掉前二列會重覆的標題等 5](#_Toc143959815)

[圖表 7 主文轉串列後會有很多空行 5](#_Toc143959816)

[圖表 8 七部曲-去掉主文串列中空行 6](#_Toc143959817)

[圖表 9 文帖推文留言處理程式碼測試 6](#_Toc143959818)

[圖表 10 留言者以字典統計方式 7](#_Toc143959819)

[圖表 11 根據前面測試結果自訂爬PTT文章函式 8](#_Toc143959820)

[圖表 12 用for迴圈設定爬PTT文帖的頁數方式爬文 9](#_Toc143959821)

[圖表 13 PTT最新和最舊會出現'btn wide disabled'類別 9](#_Toc143959822)

[圖表 14 用while爬取PTT全部文的網頁程參考程式碼 10](#_Toc143959823)

[圖表 15 用字典統計貼文、推文、和追文者程式碼 10](#_Toc143959824)

[圖表 16 用字典統計貼文、推文、和追文者範例結果 11](#_Toc143959825)

[圖表 17爬取8/22 PTT Stock最近二頁文帖標題範例 12](#_Toc143959826)

[圖表 18輸出文帖全文的範例程式碼 13](#_Toc143959827)

[圖表 19只輸出文帖主文和推文內容的範程式碼 14](#_Toc143959828)

[圖表 20 將爬文帖全文存入檔案範例程式碼 15](#_Toc143959829)

[圖表 21 news\_all.txt文帖全文範例 16](#_Toc143959830)

[圖表 22 news\_c.txt文帖主文範例 17](#_Toc143959831)

[圖表 23 news\_p.txt文帖推文回應範例 18](#_Toc143959832)

[圖表 24定義和ChatGPT API溝通函式 19](#_Toc143959833)

[圖表 25匯入news\_c.txt 19](#_Toc143959834)

[圖表 26 用for迴圈傳給ChatGPT範例程式 19](#_Toc143959835)

[圖表 27 將 ChatGPT摘要結果存檔範程式 20](#_Toc143959836)

[圖表 28 ChatGPT摘要結果 20](#_Toc143959837)

[圖表 29 文帖連結異常但未標示已刪除特例 21](#_Toc143959838)

[圖表 30 PTT刪文和一般文章之差異-沒有<a>TAG 21](#_Toc143959839)

[圖表 31 PTT 文帖出404現異常特例 22](#_Toc143959840)

[圖表 32 異常文帖會被標示成-已被刪除 22](#_Toc143959841)

[圖表 33解決 PTT文帖分類未使用中刮號的程式範例 23](#_Toc143959842)

[圖表 34 未使用「--」來分隔主文和回應特例 24](#_Toc143959843)

[圖表 35 解決未使用「--」來分隔主文和回應的程式碼 24](#_Toc143959844)

[圖表 36 ChatGPT3.5 token超過上限 25](#_Toc143959845)

[圖表 37解決Chat GTP 3.5 token上限參考方式 25](#_Toc143959846)

[圖表 38 解決ChatGPT3.5 API回應異常參考程式碼 26](#_Toc143959847)

[圖表 39 GPT4 API串接範例程式 26](#_Toc143959848)

[圖表 40 由迴圈方式爬完74篇文章約需47秒 26](#_Toc143959849)

[圖表 41 非同步方式爬完75篇只需2秒左右 27](#_Toc143959850)

[圖表 42 將非同步爬取之文章存入檔案範例程式 27](#_Toc143959851)

[圖表 43 推文中特例會造成程式執行錯誤 27](#_Toc143959852)

[圖表 44 圖15修正後程式碼 28](#_Toc143959853)

[圖表 45 8/26約99篇文帖作者10名統計 28](#_Toc143959854)

[圖表 46 8/26約99篇文帖推文者10名統計 29](#_Toc143959855)

[圖表 47 8/26約99篇文帖追文者前10名統計 29](#_Toc143959856)