

# FB 爬蟲-貼文搜尋

## 動機：

家人因為工作需求常常要在一些 FB 社團上尋找別人張貼的資訊，然而以人力的方式瀏覽近百篇的文章既勞累又沒效率，所以我就想到了可以用網路爬蟲的方式來幫忙快速分析，並且趁著這次機會可以對網頁結構有基礎的認識。

## 構想：

我想要做一個應用程式，它可以讓我登入 FB 的帳號，接下來貼上想要爬取文章的網址，並且指定該頁面滾軸要往下拉幾次(加載多少的貼文量)後，把所有的文章顯示在介面上，還要附上作者的姓名以及發文時間，再透過關鍵字搜尋的方式篩選出符合的貼文。

## 程式規劃：

網路爬蟲的部分使用 python 來實作，利用 Selenium 模組來模擬使用者登入、瀏覽頁面等行為，還有 BeautifulSoup 模組來處理抓下來的 html 內容，最後將文章內容儲存到 txt 檔中。且為了讓程式在其他地方不需要安裝 python IDE 以及零零總總的模組也可以順利執行，所以我把.py 檔轉換成.exe 執行檔。而使用者介面以及文章分析篩選的部分則是用 c#來實作，嘗試了使用 Form Class 建立多個表單，讓應用程式內容更豐富。它會先執行上述的爬蟲執行檔，然後用 txt 檔作為媒介讀取抓下來的文章去分析。

## 遭遇的問題與解決方法：

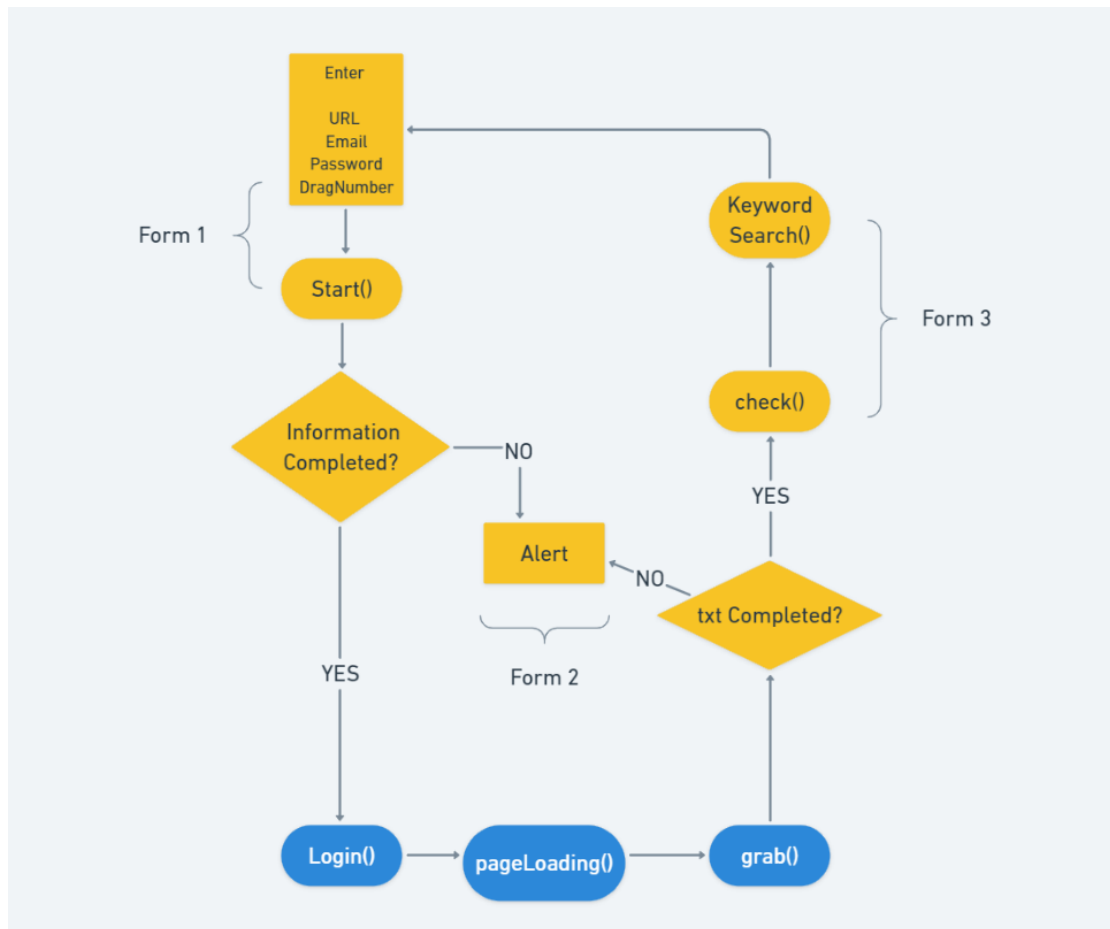
1. Selenium、Beautiful Soup 對於 html 中元素屬性 class 名稱格式不同：  
前者的 class 名稱中間空格要換成點，後者的 class 名稱則不用換。  
從網頁抓的 class 名稱不能直接用在 Selenium 上，雖然是很簡單的問題，但耗費了我很久的時間。  
EX:  
Beautiful Soup -> 'du4w351b k4urcfbm 19j0dhe7 sjgh65i0'  
Selenium -> 'du4w351b.k4urcfbm.19j0dhe7.sjgh65i0'
2. 因為網路速度的問題，所以不清楚頁面是否加載完成。參考網路的解決辦法是要加入等待機制，分為 explicit wait 和 implicit wait。前者是直接讓程式停止數秒，較為簡單，但是用多了會讓爬取資料的時間拖得很久。後者

是設定一個時限，若找到元素或是超過時限的話就不用繼續等待。然而它只能等待網頁元素，若我想要等待整個頁面加載則要另尋方法。我的方法是寫一個函式，它能記錄當前的網頁高度，若是網頁高度改變了才代表頁面加載完成，並且同樣地設定一個時限避免無止盡的等待。

3. 多數貼文因為字數問題都有” 查看更多” 的選項，遮擋了該貼文的大部分內容。解決辦法是找出” 查看更多” 專屬的屬性名稱，然後一個一個點開。然而我只找到’oajrlxb2 g5ia77ul qu0x051f esr5mh6w e9989ue4 r7d6kgcz rq0escxv nhd2j8a9 nc684nl6 p7hjln8o kvgmc6g5 cxmmr5t8 oygrvhab hcukyx3x jb3vyjys rz4wbd8a qt6c0cv9 a8nywdso ilao9s8h esuyzwwr flsip0of lzcic4wl oo9gr5id gpro0wi8 lrazzd5p’，這串 class 名稱經過我比對其他元素，應該是代表粗體、可點選的文字，如姓名、查看更多。因此還得多一到程序確認 element.text 是” 查看更多”。
4. 文字編碼問題：抓取到的文章能夠被 print() 出來，然後卻沒辦法使用 write() 的方式寫入到 txt 檔案中，最後發現可能是特殊字元在既有編碼中沒有被支援，因此在 open() 中多加一行 encoding=’utf8’ 的敘述就解決了，這個問題也是耗費了很多時間。
5. 無法刪除的問題：在程式中每次抓取文章都會建立 txt 檔來記錄，為了能夠重複使用，每次使用完畢後必須刪掉殘留下來的 txt 檔，然而常常發生無法刪除或是刪除不乾淨的狀況，最後我發現是 c# 的 FileStream 未關閉，這就如同不可以刪掉正在使用的程式，電腦會跳出提醒訊息一樣，只是換成用 c# 刪除時並沒有提醒我問題所在。最後記得使用 close() 就可以刪掉了。
6. McAfee 防毒軟體會把我 python 轉換的 exe 執行檔當成病毒!!!  
執行時可能必須關閉防毒軟體以及防火牆 不然會連 chrome 一起被刪掉  
萬分抱歉!!!!

**流程圖：（在下一張）**

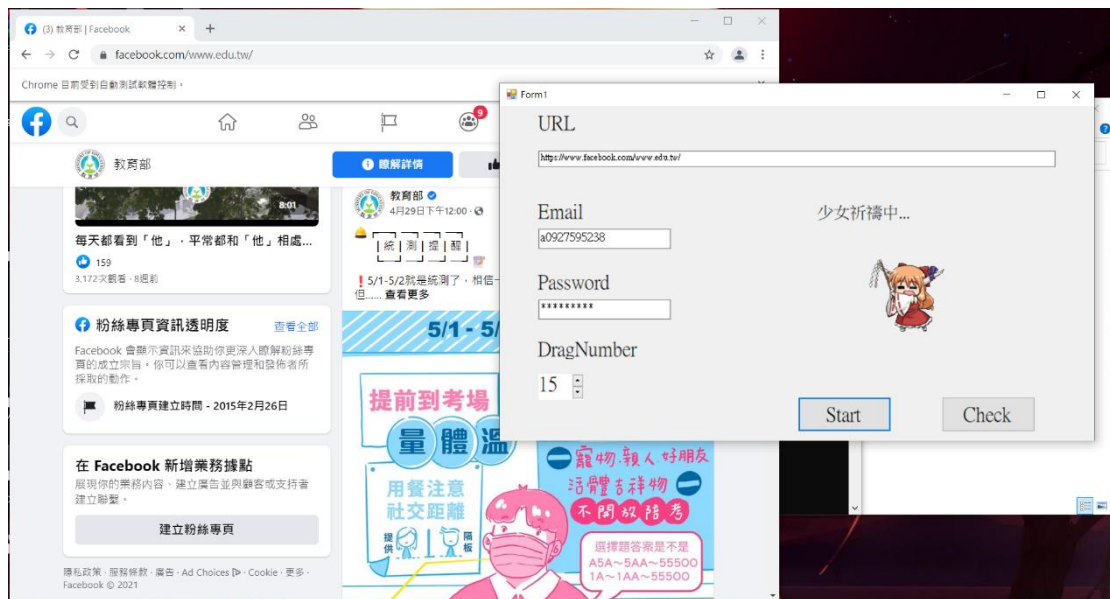
黃色為 c# 部分 藍色為 python 部分



程式測試執行結果：

The screenshot shows a window titled 'Form1'. It contains the following elements:

- A 'URL' label above a text input field.
- An 'Email' label above a text input field.
- A 'Password' label above a text input field.
- A 'DragNumber' label above a spinner control showing the value '0'.
- A text prompt '你想調查什麼呢?' (What do you want to investigate?).
- An image of a character (likely a mascot or avatar) next to the text prompt.
- Two buttons at the bottom: 'Start' and 'Check'.



Form1

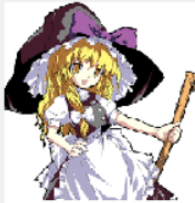
URL

Email

Password

DragNumber

捕獲完畢 DAZE



Start

Check

Form3

Keywords

Search

共有41篇

教育部

5月9日上

^ Mama購物網上線啦提供滿滿的購物選擇一

- 起支持媽媽追求自我實現和夢想吧(∩\_∩)
- #成就媽媽的成就在社會的
- 刻板印象中，媽媽總是「應該」將家中大小事全攬在身上，「需要」為了家庭與孩子犧牲奉獻。但實際上，照顧家庭和成就自我並不衝突，有伴侶和家人的支持及幫助，媽媽不需要受到傳統印象的束縛
- ✓ 從你我開始，作為伴侶或子

page 3



## 參考資料：

Selenium 的使用方法：

<https://medium.com/marketingdatascience/selenium%E6%95%99%E5%AD%B8-%E4%B8%80-%E5%A6%82%E4%BD%95%E4%BD%BF%E7%94%A8webdriver-send-keys-988816ce9bed>

<https://aitmr1234567890.medium.com/%E5%8B%95%E6%85%8B%E7%B6%B2%E9%A0%81%E7%88%AC%E8%9F%B2%E7%AC%AC%E4%BA%8C%E9%81%93%E9%8E%96-selenium%E6%95%99%E5%AD%B8-%E5%A6%82%E4%BD%95%E4%BD%BF%E7%94%A8find-element-s-%E5%8F%96%E5%BE%97%E7%B6%B2%E9%A0%81%E5%85%83%E7%B4%A0-%E9%99%84python-%E7%A8%8B%E5%BC%8F%E7%A2%BC-520fdaa983f9>

Beautiful Soup 的使用方法：

<https://blog.gtwang.org/programming/python-beautiful-soup-module-scrape-web-pages-tutorial/>

圖片來源：

[http://17woo.tgbusdata.cn/forum/month\\_1006/1006211942780b1920529f2815.gif](http://17woo.tgbusdata.cn/forum/month_1006/1006211942780b1920529f2815.gif)

<https://www.itsfun.com.tw/%E5%B0%84%E5%91%BD%E4%B8%B8%E6%96%87/wiki-9826246-7103126>

<http://dgta.hopto.org/wiki/%E6%9D%B1%E6%96%B9%E9%9D%9E%E6%83%B3%E5%A4%A9%E5%89%87:%E7%99%BB%E5%A0%B4%E8>

%A7%92%E8%89%B2:%E9%9C%A7%E9%9B%A8%20%E9%AD%94%E7%90%86%E6%B2%99/rev/18/show

<https://vp.uzkk.net/posts/%E5%9B%9B%E5%AD%A3%E6%98%A0%E5%A7%AC%C2%B7%E4%BA%9A%E7%8E%9B%E8%90%A8%E9%82%A3%E5%BA%A6.html>

等待功能：

<https://kkboxsq. wordpress. com/2017/06/16/the-difference-between-implicit-wait-and-explicit-wait/>