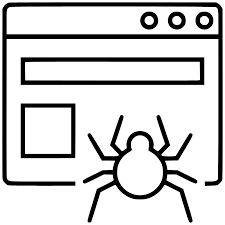
新聞搜尋、輕鬆學習

作者:陳亦揚





目錄

摘要

第一章  前言

   1-1  創作發想、動機與目的

   1-2  先前已發展的技術

第二章 作品特色與創意特質

2-1  獨創性(創新性)

2-2  實用性(功能性)

第三章 設計方法

3-1 使用設備

3-2 設計方法流程與說明

第四章 系統流程與架構

4-1 實驗描述

4-2 實驗成果

參考資料

摘要:

利用網路爬蟲抓取資訊新聞，將使用者感興趣的領域與關鍵字，學習資源等輸出在網頁上。

第一章  前言

   1-1  創作發想、動機與目的

我加入碩班開始接觸新事物與各種論文時，常常會擔心得太多:這個會不會太偏門?那個是不是跟別人重複了?那個東西會不會很難找到資料?這個做下去會不會根本沒有發展性?

於是我開始考慮與其在自己的思維裡打轉，不如去看看現今有甚麼資訊新聞，從中找出靈感，若該篇報導中還有附上關鍵字的連結或學習資源就更好了!

不管是找題目還是做研究，有時會陷入找不到靈感或者不知從何了解關鍵技術的困境，又或者鑽牛角尖錯過了近在眼前的最新資訊。

我希望用一個網頁，能提供各大資訊新聞網站的報導，也許是**資訊安全**(有甚麼電腦病毒出現了，微軟又出了什麼包)、**社交媒體**(youtube跟fb又知道你腦袋裡在想甚麼了，他們如何利用大數據知道你的一切)、**娛樂**(遊戲引擎的開發，遊戲公司的職缺開放)、**證照更新**(企業喜愛的熱門證照，考題更新)....等。並且提供這些報導的同時，在下方列出各種相關連結及國內外論文，希望能藉此方式提供更多靈感，不被侷限在某一領域。

1-2  先前已發展的技術

要使用網路爬蟲，有兩大套件必定會被提起，分別是Beatifulsoup與Selenium。

Beautifulsoup可以剖析HTML網頁，然後從剖析的網頁內容來擷取所需的目標資料。其實就是我們在使用瀏覽器時，按下F12跳出的網頁原始碼，Beautifulsoup可以把這個原始碼抓下來並且根據標籤或元件ID等..來抓出爬蟲設計者想要的資料。

Selenium是web應用程式的軟體測試框架，跨平台的自動瀏覽器，其目的是自動測試WEB應用程式。不過，Selenium還可以擷取動態網頁內容和HTML表單自動化。Selenium可以模擬人在網頁中的行為:點擊，拖曳，按下按鍵等等..。

若想利用python來快速創建一個網頁伺服器，Flask 是一個可建立 Web 應用程式的微架構。 這表示 Flask 會提供工具、程式庫和技術，用以建置 Web 應用程式。本次專題就是使用Flask在自己的電腦上建立一個伺服器(127.0.0.1)，將抓取到的資料直接透過Flask傳到網頁上，如此就不用再建立一個資料庫來儲存這些東西了。

第二章 作品特色與創意特質

2-1  獨創性(創新性)

網路爬蟲不是什麼新玩意兒，許多比價網都是利用網路爬蟲來蒐集各個購物網站的資料來將品項價格清單列給使用者看，甚至要說GOOGLE本身就是一個巨大的爬蟲，他透過我們想要找的關鍵字，在世界各地蒐集了與該關鍵字相關的資料來呈現。不過此次專題我是以”新聞搜尋”來”提供靈感與學習資源”的，對我來說以後當我對一個新事物感興趣時，可以直接用這個網頁來幫助自己找到最新的新聞資訊!

2-2  實用性(功能性)

目前使用上遇到比較大的問題是網頁載入比較久，因為每次載入都會跑一次爬蟲，再來就是美工方面非常的差勁、欠缺美感...

不過搜尋新聞這點倒是還挺不錯的，將來有空時我再新增更多網站的搜尋結果來充實網頁。

第三章 設計方法

3-1 使用設備

華碩X550V筆記型電腦、CPU :i7-6700HQ、記憶體:8G、顯示卡:NVIDIA GTX950M、硬碟:128G SSD+1TB HDD、作業系統:win10

開發環境:Visual Studio Code 。

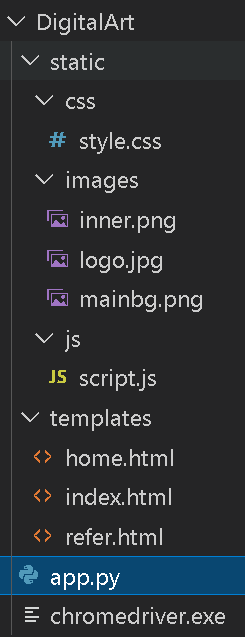
3-2 設計方法流程與說明

一開始的想法是單純地用爬蟲把資料抓下來到資料庫或是直接存在電腦中，然後令網頁到指定位置再把東西給拿出來呈現。但是當我用python寫完爬蟲後，我突然覺得既然都已經在這個專案中用python獲取了這些網頁資料，我能不能接著把這些資料丟到專案中的網頁裡?於是透過估狗搜尋關鍵字”python 網頁”這樣有點笨拙的方式找到了FLASK這個工具，摸索了一番後就變成不用透過資料庫和額外的伺服器，即可直接把資料丟到網頁中了。

第四章 系統流程與架構

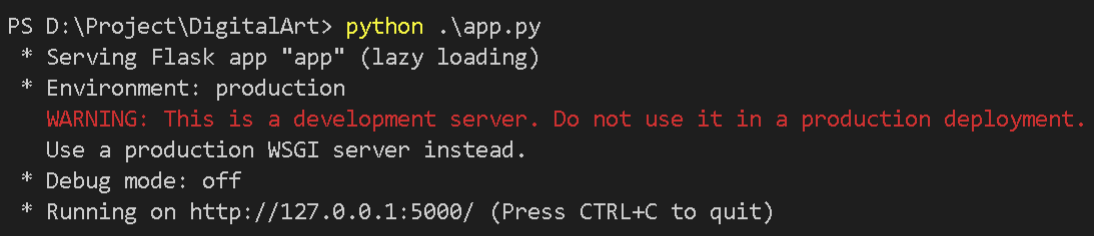
4-1 實驗描述

專案程式放置的相對路徑位置如圖:



* 根目錄中(:/)主程式(app.py)包含了FLASK的啟動本機伺服器與爬蟲程式(按理來說爬蟲應該是寫在別的地方再呼叫過來，但我在呼叫過程中有點問題所以就寫在一起了)。chromedriver.exe是selenium執行時會用到的，他會開啟瀏覽器並根據編寫的程序來抓取網頁。
* 會用到的網頁在(:/templates/)，三個網頁(home、index、refer)分別是主頁、搜尋後的跳轉頁、參考資料。
* 網頁會用到的圖片在(:/static/images/)
* 網頁會css在(:/static/css/)、JS在(:/static/js/)

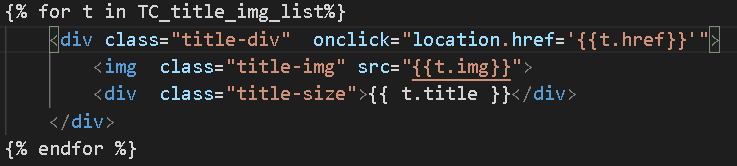
第一步是執行app.py，執行後會出現本機網址，可以ctrl+左鍵點擊該網址即可開啟首頁。



在首頁中可以點擊”隨便看看”，即會跳到下面的”最近的新聞”，這區的新聞每天都會不一樣。

在搜尋欄中輸入關鍵字並點擊搜尋後，網頁會以post的方式回傳欄內的文字到flask，這時爬蟲就會以該關鍵字進入新聞網站抓取資料並且把該站第一頁的搜尋結果列出來到網頁上(包含圖片與標題)。

由於我以前沒用過javascript，我在flask中第一次知道html還可以這樣寫迴圈。(圖片為index.html中第69行~74行)



第一行的TC\_title\_img\_list為爬蟲在TechCrunch中抓取下來的資料清單。

網頁上的將近十筆搜尋資料並不是寫十個div區塊，而是透過flask用這種迴圈將爬蟲的資料清單拿過來並尋訪該清單，如此即可建立一個包含了超連結、圖片、標題的區塊。

4-2 實驗成果

目前可以順利地列出新聞，且在下方增加了其他資源連結，若之後能加強美工與加快網頁載入速度就更好了。

參考資料:

網站:

https://www.youtube.com/watch?v=9X1graZtuPs

https://www.learncodewithmike.com/2020/10/django-web-scraping.html

https://codertw.com/%E7%A8%8B%E5%BC%8F%E8%AA%9E%E8%A8%80/454833/

https://segmentfault.com/q/1010000009970908

https://tw.alphacamp.co/blog/css-guide-box-model

https://lotus99.pixnet.net/blog/post/30096009

https://www.maxlist.xyz/2019/03/17/flask-get-post/

https://medium.com/seaniap/python-web-flask-get-post%E5%82%B3%E9%80%81%E8%B3%87%E6%96%99-2826aeeb0e28

https://www.itread01.com/content/1541434996.html

https://www.itread01.com/content/1497676809.html

https://ithelp.ithome.com.tw/articles/10223951

https://ithelp.ithome.com.tw/articles/10213503

https://selenium-python-zh.readthedocs.io/en/latest/locating-elements.html

https://medium.com/@bindaguo/%E7%B6%B2%E8%B7%AF%E7%88%AC%E8%9F%B2%E6%B7%BA%E8%AB%87-afcae0694f13

https://ithelp.ithome.com.tw/articles/10222132

https://www.ucamc.com/articles/317-vscode-windows-%E7%89%88%E6%9C%AC%E5%9F%BA%E6%9C%AC%E5%B8%B8%E7%94%A8%E5%BF%AB%E6%8D%B7%E9%8D%B5

https://blog.csdn.net/qq\_42991834/article/details/89480614

書籍:

Python從網路爬蟲到生活應用超實務: 人工智慧世代必備的資料擷取術, 作者:陳會安, 出版社:博碩文化 ,ISBN:9789864345397

影音:

https://www.youtube.com/watch?v=9X1graZtuPs

https://www.youtube.com/watch?v=RWviEK1Si68&list=PLDFBYdF-BxV1G4FBpG1EMyFtbsbZuJOvD