

# 國立屏東高級中學

## 資訊科報告

# 介紹網路爬蟲

學生： 2 年 14 班 29 號 謝易宸

指導老師：洪紹瑞

中華民國 110 年 12 月

# 目錄

壹、前言	3
一、動機	4
二、預設目標	5
貳、介紹與實作	6
一、網路爬蟲的介紹	7~8
二、使用 python 實作	9~10
參、問題討論與心得	11
一、問題討論	12
二、實作心得	13
肆、參考資料與網址	14

# 壹、前言

## 一、動機

在學校的電腦資訊概論課程中，老師有介紹與講解到一個程式—「**網路爬蟲**」，在此之前，這個程式名稱我只有聽過，卻不知其功能與用途，更不用說將其製作，但聽到老師上課的講解，我開始對這個「**網路爬蟲**」程式感到好奇和興趣，於是便想自己學習製作一個「**網路爬蟲**」的程式。



## 二、預設目標

先看別人的教學影片，將實作過程完全操作過一次，再改變成自己想搜尋的網站或網址。我打算用**網路爬蟲**將「**PTT 的動漫板文章列表**」資料給抓取下來，因為自己對**動漫、ACG 文化感興趣**，而且對網路爬蟲程式還是新手菜鳥的狀態，所以一開始選擇 PTT，是較簡單又容易操作的學習開始。

## 貳、介紹與實作

## 一、網路爬蟲的介紹

網路爬蟲是一種程式腳本，類似機器人，只要對其程式輸入一套規則、指令，網路爬蟲就會依照這套規則、指令，有規律性的自動性瀏覽網頁、抓取或蒐集資料。網路爬蟲可應用於票價比價、訂票、找工作資訊、查看網友留言、蒐集股市資料、下載及搜尋想要的圖片與資料。

使用網路爬蟲，必須輸入該網站的網址，而爬蟲程式就會依據指令，將那個網址的資料改抓取（依據不同的指令，會改變爬蟲程式抓取那些資料，例如找工作資料時，要求只抓取有 150 時薪的工作資料，而爬蟲程式就會自動將只有顯示 150 時薪的工作資料給抓取下來。）。

某些網站會阻止網路爬蟲來瀏覽與抓取資料，該網站的伺服器會偵測出哪些是

人類使用瀏覽器進來，而哪些是網路爬蟲，但網路爬蟲只要更改一些欄位資料，假裝自己是瀏覽器，能夠不被伺服器給偵測出。

網路爬蟲可以同時進行大量的連線，如果在短時間內大量連到某個網站，有一些小型的伺服器會承受不了而當機，也會造成過多的流量花費，所以在瀏覽小型伺服器時，可以將網路爬蟲的速度給放慢。而有些伺服器會為了避免出一些有關網路爬蟲造成的狀況，會設置一個 robots.txt 檔，裡面會有一些對網路爬蟲的規定。

如果想寫網路爬蟲程式，就必須先了解 HTML（一種用來組織架構並呈現網頁內容的程式語言。註 1），然後使用 PYTHON（註 2）或 NODEJS（註 3）之類的程式去撰寫。

（註 4 參考文獻）

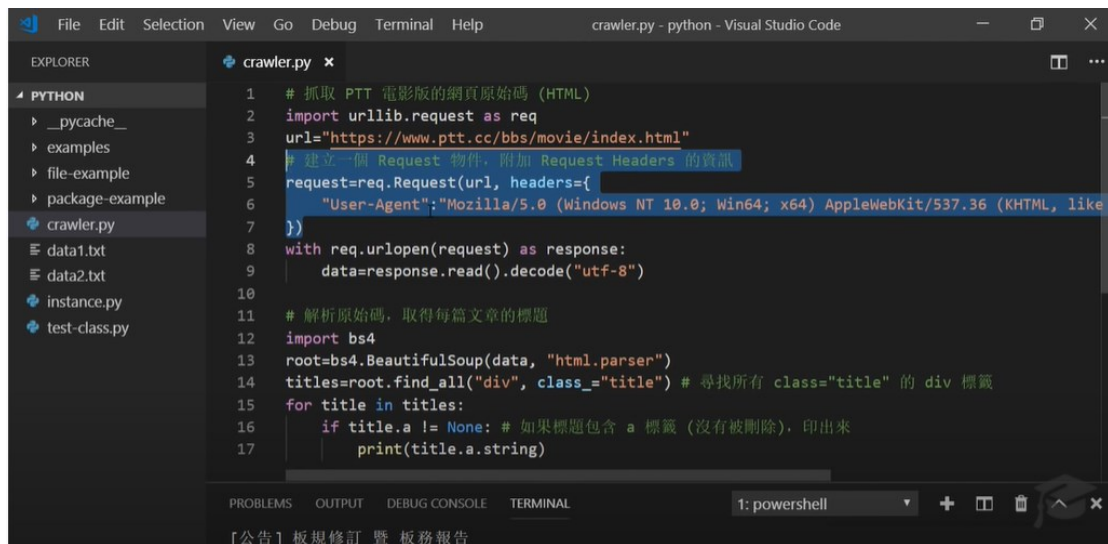


## 二、使用 python 實作

我這此選擇的教學影片是專門在教程式語言的 youtuber—彭彭的課程 的影片：

Python 網路爬蟲 Web Crawler 基本教學 By 彭彭

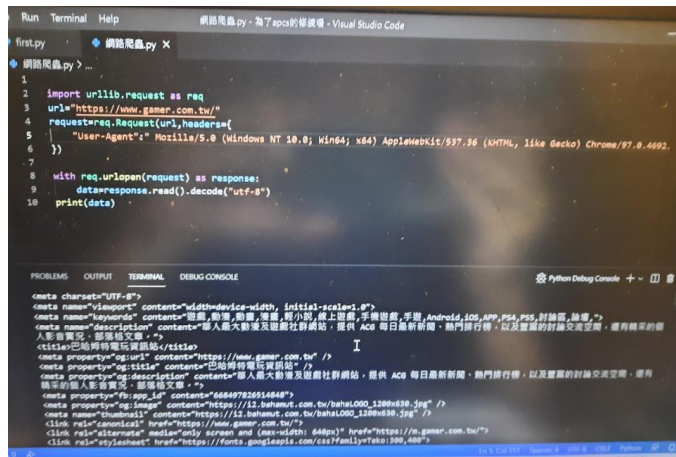
(註 5)



```
1 # 抓取 PTT 電影版的網頁原始碼 (HTML)
2 import urllib.request as req
3 url="https://www.ptt.cc/bbs/movie/index.html"
4 # 建立一個 Request 物件, 附加 Request Headers 的資訊
5 request=req.Request(url, headers={
6     "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like
7 })
8 with req.urlopen(request) as response:
9     data=response.read().decode("utf-8")
10
11 # 解析原始碼, 取得每篇文章的標題
12 import bs4
13 root=bs4.BeautifulSoup(data, "html.parser")
14 titles=root.find_all("div", class_="title") # 尋找所有 class="title" 的 div 標籤
15 for title in titles:
16     if title.a != None: # 如果標題包含 a 標籤 (沒有被刪除), 印出來
17         print(title.a.string)
```

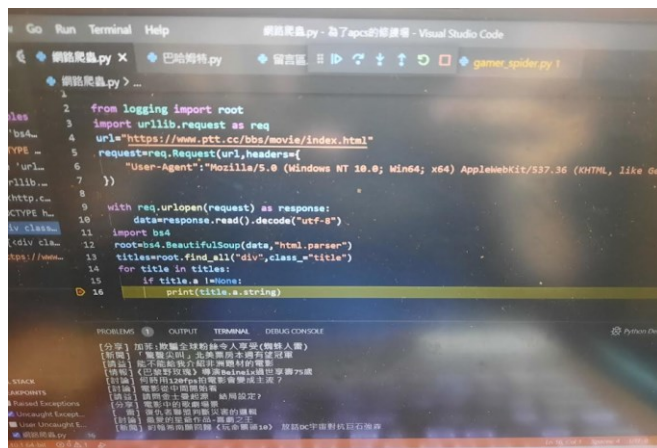
(圖 1 影片的範例)

我自己的實作圖片：



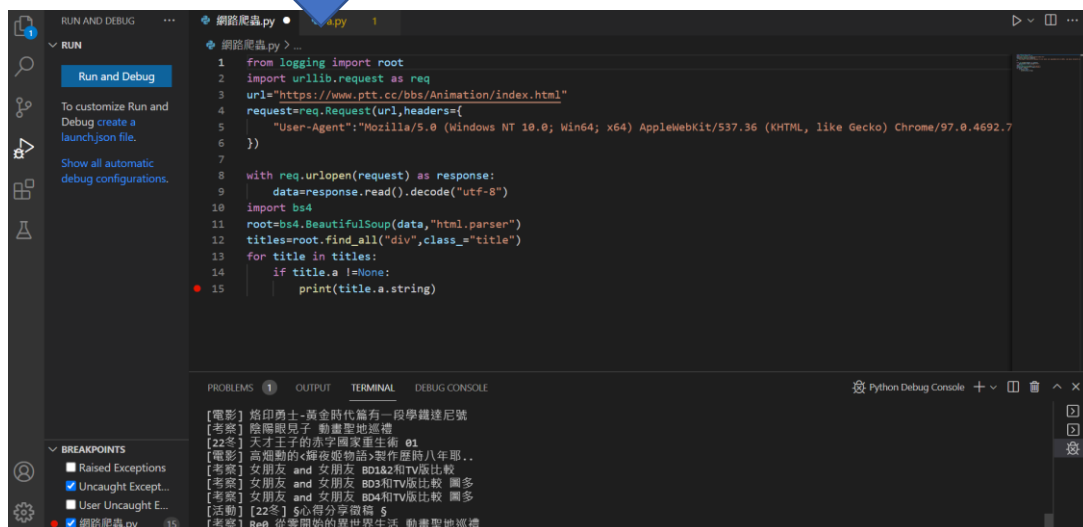
```
Run Terminal Help 網路爬蟲.py - 為了apex的學後者 - Visual Studio Code
網路爬蟲.py x
1
2 import urllib.request as req
3 url="https://www.gamer.com.tw/"
4 request=req.Request(url,headers={
5     "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/97.0.4692.7
6 })
7
8 with req.urlopen(request) as response:
9     data=response.read().decode("utf-8")
10 print(data)
```

(圖 2 刚开始製作的程式碼)



```
Go Run Terminal Help 網路爬蟲.py - 為了apex的學後者 - Visual Studio Code
網路爬蟲.py x 巴給姆特.py 網路爬蟲.py
1
2 from logging import root
3 import urllib.request as req
4 url="https://www.ptt.cc/bbs/movie/index.html"
5 request=req.Request(url,headers={
6     "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/97.0.4692.7
7 })
8
9 with req.urlopen(request) as response:
10     data=response.read().decode("utf-8")
11
12 import bs4
13 root=bs4.BeautifulSoup(data,"html.parser")
14 titles=root.find_all("div",class_="title")
15 for title in titles:
16     if title.a != None:
17         print(title.a.string)
```

(圖 3 完成影片範例的操作)



```
RUN AND DEBUG ... 網路爬蟲.py
1 from logging import root
2 import urllib.request as req
3 url="https://www.ptt.cc/bbs/Animation/index.html"
4 request=req.Request(url,headers={
5     "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/97.0.4692.7
6 })
7
8 with req.urlopen(request) as response:
9     data=response.read().decode("utf-8")
10
11 import bs4
12 root=bs4.BeautifulSoup(data,"html.parser")
13 titles=root.find_all("div",class_="title")
14 for title in titles:
15     if title.a != None:
16         print(title.a.string)
```

PROBLEMS OUTPUT TERMINAL DEBUG CONSOLE Python Debug Console

[電影] 槍印滿土-黃金時代竟有一段學鐵達尼號  
[遊戲] 陰陽眼易子 轟轟聖地巡禮  
[22冬] 天才王子的赤字國家重生術 01  
[電影] 高燃動的<輝夜姬物語>製作歷時八年耶..  
[考察] 女朋友 和 女朋友 BD1&2和TV版比較  
[考察] 女朋友 和 女朋友 BD3和TV版比較 圖多  
[考察] 女朋友 和 女朋友 BD4和TV版比較 圖多  
[運動] [22冬] 50心得分享徵稿 5  
[考察] Re0 從零開始的異世界生活 動畫聖地巡禮

(圖 4 自己所爬的 PTT 文章)

## 參、問題討論與心得

## 一、問題討論

問題 1：

請問這個網路爬蟲適用於每個網站嗎？

解答：

不行，每個網站都有不同的程式、開頭檔所組成，我所學習的這個程式只適用於 PTT 的文章列表。

問題 2：

請問在網路爬蟲程式的第一列所用的 request 是用於網路連線，那每個網路爬蟲程式都一定要有 request 嗎？

解答：

具我所查到的資料，以及所看過大部分的網路爬蟲程式，基本上都要有 request 這個程式塊件，因為沒有 request 就無法做到網路連線的功能，也就無法根據網址去找到該網站。

## 二、心得

雖然我是個剛接觸電腦程式不久，但聽到**網路爬蟲**這個程式，充滿好奇心的我，想著不管怎麼樣，也要去試試看、寫寫看，我知道這非常困難，但我義無反顧。因為是個程式新手，所以現在的我，面對這麼一大串的程式只能先照著學習一遍，我知道這根本是越級打怪，不過我就是想挑戰這個網路爬蟲。過程中，也挫折不少次，檢查過後還是無法跑出滿意的成果，不過我沒有放棄的念頭，仔細檢查與努力查找原因，讓自己的網路爬蟲程式可以執行。也是因為有學校的資訊課，讓我開始對網路爬蟲程式感興趣，進而開始練習與學習打程式。至此，我開始更努力往電腦程式學習鑽研。

# 肆、參考資料與網址

## 註 1

[HTML 基礎 - 學習該如何開發 Web | MDN \(mozilla.org\)](https://developer.mozilla.org/zh-CN/docs/Web/HTML)

## 註 2

[Python - 維基百科，自由的百科全書 \(wikipedia.org\)](https://zh.wikipedia.org/zh-tw/Python)

## 註 3

[Node.js - 維基百科，自由的百科全書 \(wikipedia.org\)](https://zh.wikipedia.org/zh-tw/Node.js)

## 註 4

[文組也看得懂的 - 網路爬蟲 - YouTube](#)

## 註 5

[Python 網路爬蟲 Web Crawler 基本教學 By 彭彭 - YouTube](#)

感謝觀閱