**巨量資料分析-期中報告**

**學號：1411031011**

**姓名：潘芝儀**

1. 資料來源： <https://www.ptt.cc/bbs/Japan_Travel/index.html>

疫情趨緩，許多人都報復性出國，所以我尋找PTT，想了解大家對於旅遊的討論度。最後找到Japan­\_Travel看板，擷取目標為今年四月的[日期]、[標題]、[連結]、[推文數]、[作者]。

使我能以標題的吸引程度(主觀)以及推文數多寡(客觀)來當作參考，並且規劃我的下一次日本旅遊!!

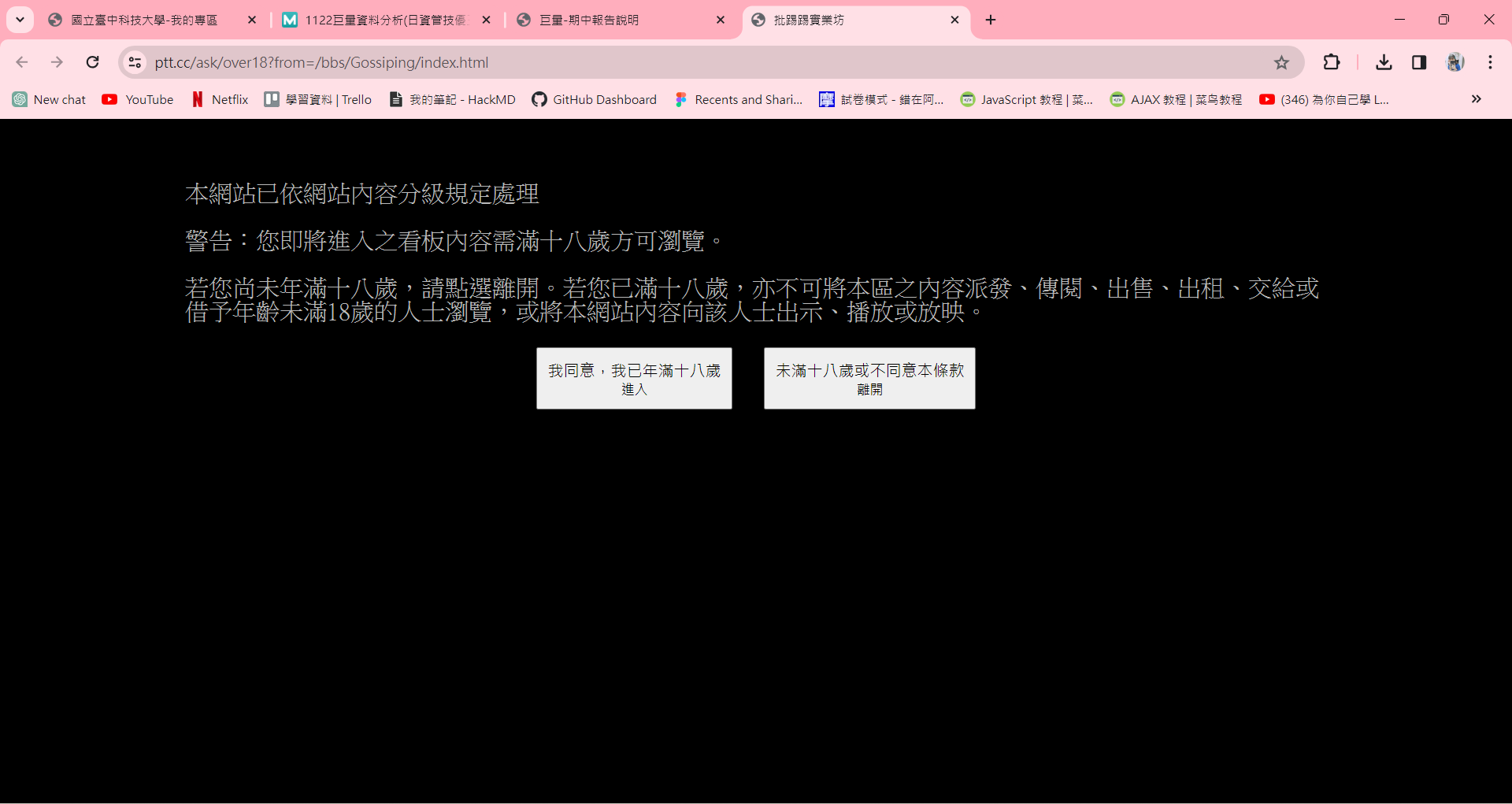
1. 爬蟲程式說明：

**這次使用到以下這四種套件：**

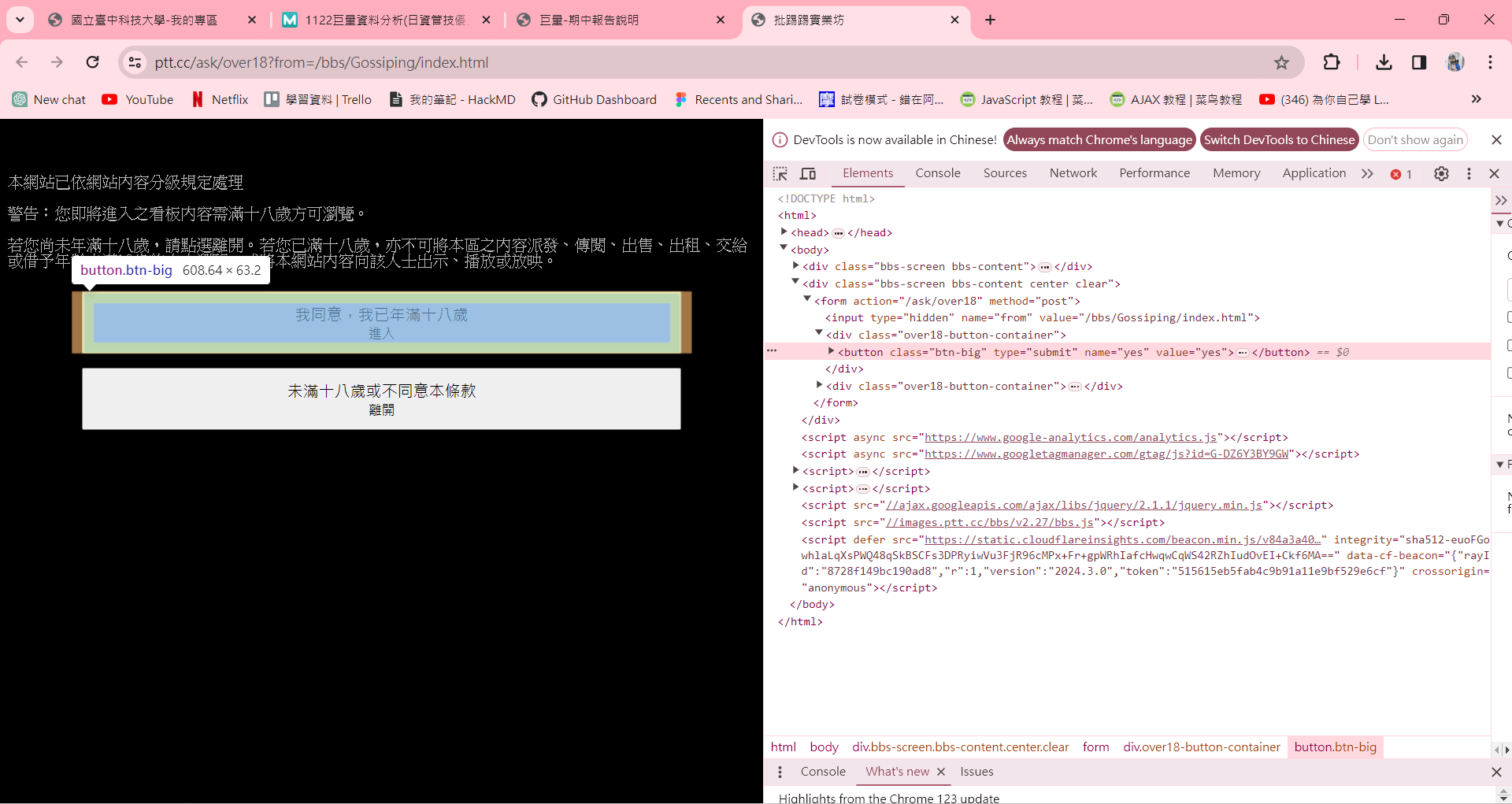


**主要的方法和流程：**

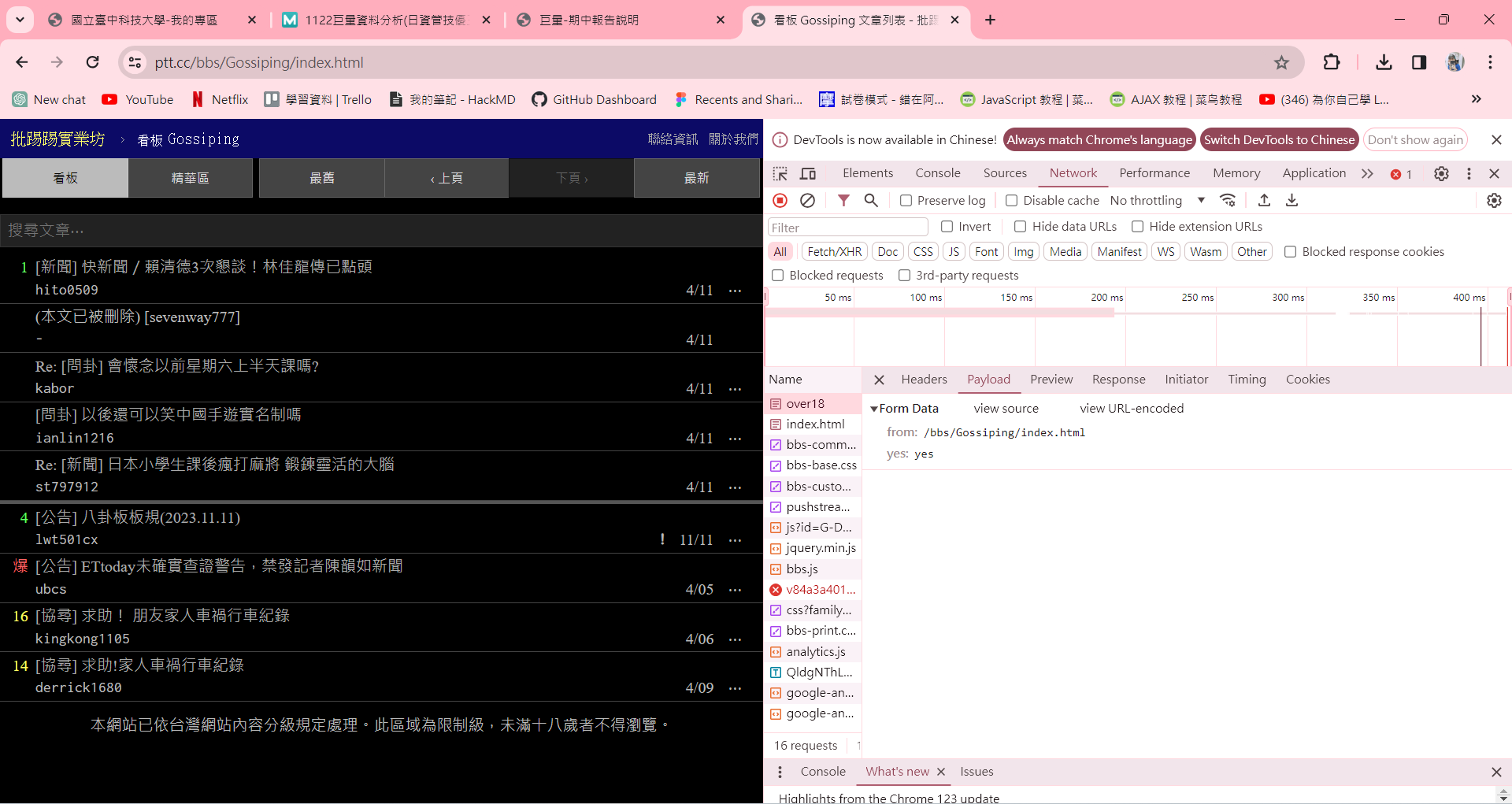
1. 這次所爬的PTT網頁有時會被轉址到出現需要表明的身分：



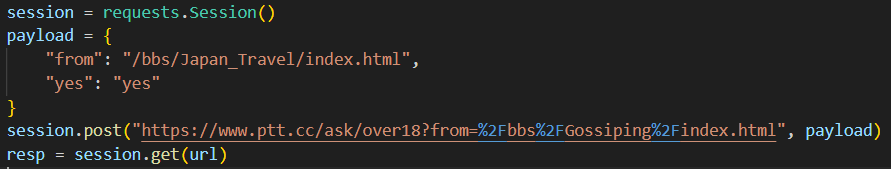
1. 按F12或右鍵「內容」來觀察其原始碼：



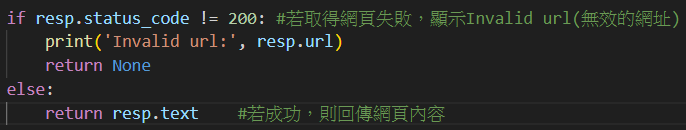
1. 按下同意後，觀察右側Network變化，點選「over18」可以看到表單使用POST傳遞資料，並在payload找到傳遞的資料：



1. 副程式，使用requests啟用session功能，使伺服器與瀏覽器達成溝通：



1. 副程式(續)，測試爬蟲是否有成功，如下圖所示：

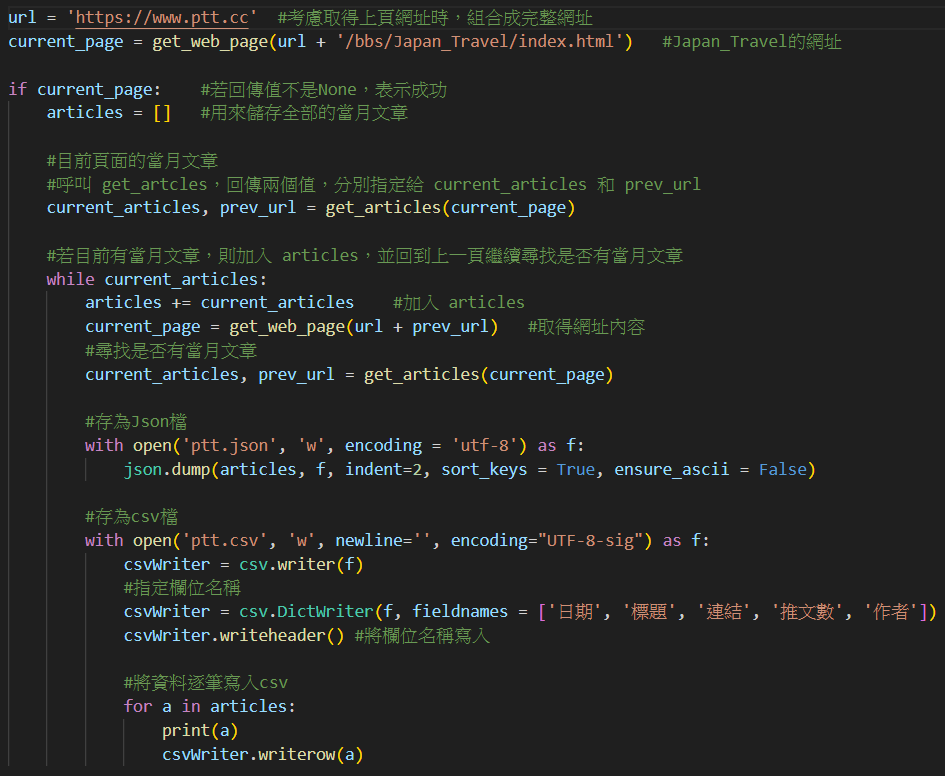


1. get\_web\_page副程式全貌：

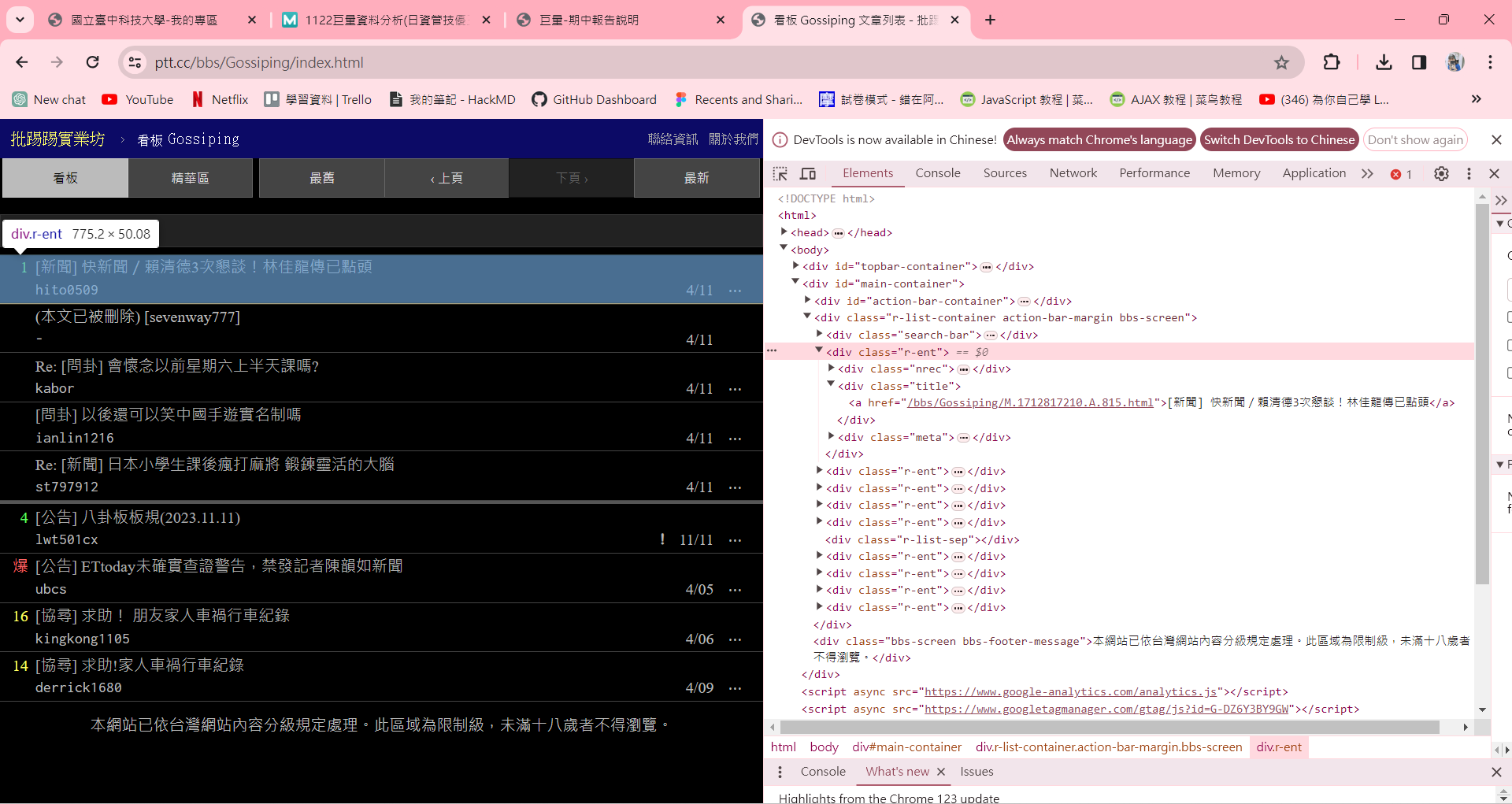


1. 延續上方副程式判斷回傳值，是否成功，如成功再來取得當月的文章：

(主程式)

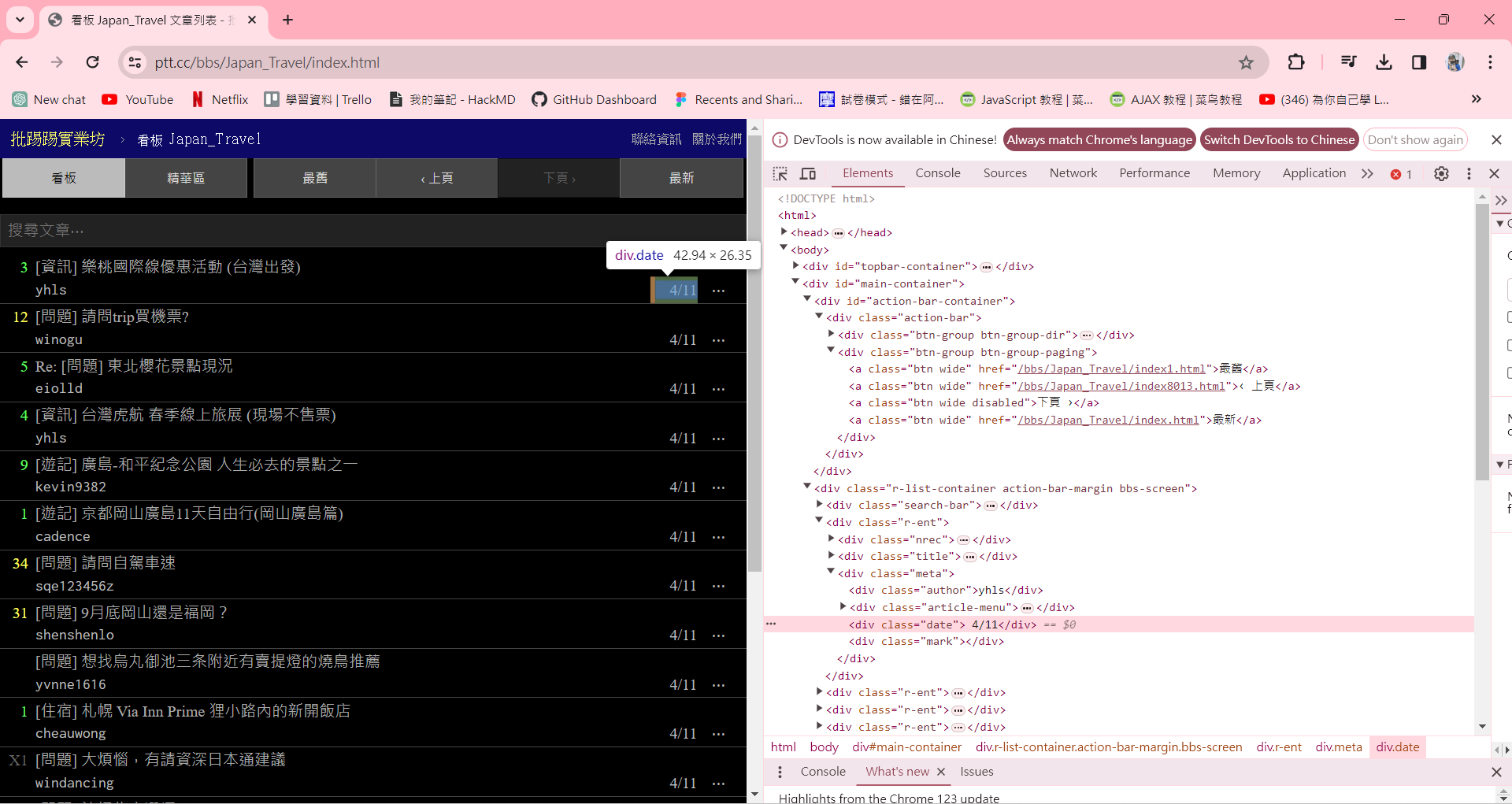


1. 再來解讀網頁標籤，如下圖所示：



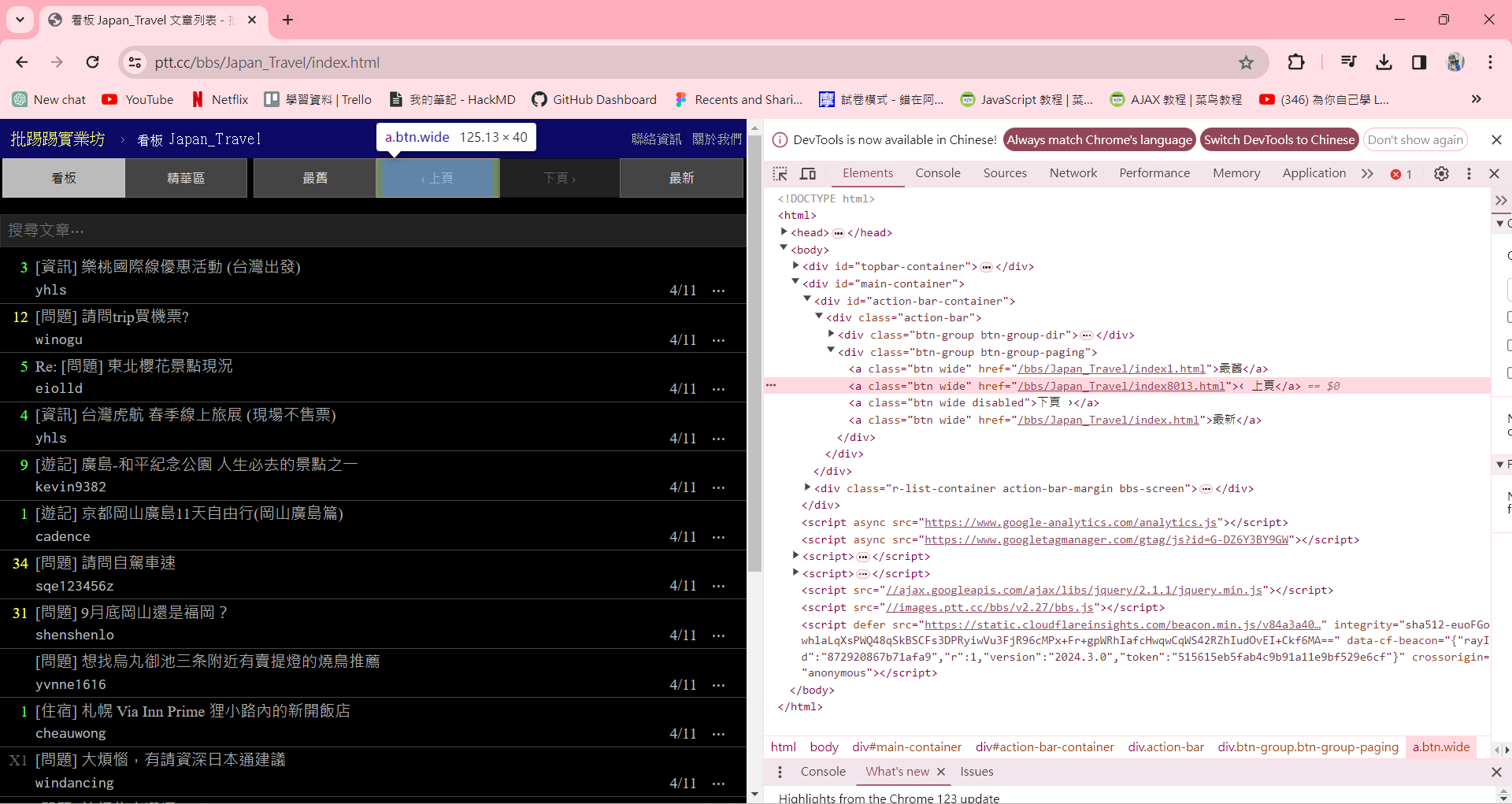
每筆文章在

class="r-ent"



每篇日期在

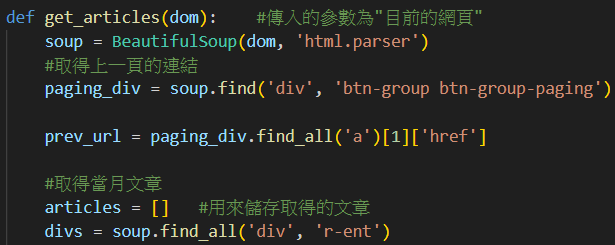
class="date"



爬4月份的文章，所以上頁還有文章要爬

1. get\_articles副程式

使用div中的class="btn-group btn-group-paging"取得上一頁的連結，創造articles用來儲存四月份的文章，並且每一筆文章在div中的class="r-ent"(可以看第7點的圖例)



1. 副程式(續)

給推文數(push\_count)在迴圈外多設一個初始值，不然會拋'UnboundLocalError'的錯誤。for迴圈中使用div中的class="date"取得文章日期，使用split(‘/’)從/分割，再來取得[0]的月份，並下方判斷指定4月份的文章。最後在div中的class="nrec"文章標籤取得[日期date]、[標題title]、[連結href]、[推文數push\_count]、[作者author]。



**資料儲存格式：使用Json檔以及csv檔來儲存**

