Python ptt 爬蟲 B0843026 資工二 謝奇容

程式流程圖：

程式開始

進入ptt八卦版

是否是昨天的po文

存入陣列

是

判斷下篇貼文

爬取陣列內po文詳細內容

程式結束

存入CSV檔

判斷下篇貼文

此篇文蓋樓數是否為前10名

否

陣列內文章是否都判斷過了

是

現在時間減po文時間是否小於3天

否

是

存入陣列

否

是

否

使用套件：

1. selennium：原為網頁測試工具，由於是真正的瀏覽器在運作，所有網頁的操作都可由程式碼控制、執行。但也導致運行速度較慢，占用較高電腦效能。
2. datetime ：取得時間、計算時間、時間格式轉換

一張含有 文字 的圖片

自動產生的描述

程式解說及流程說明：

1. 設定瀏覽器及啟動瀏覽器
2. 進入ptt八卦版，按下同意按紐一張含有 文字 的圖片

   自動產生的描述
3. 設定時間單位一張含有 文字 的圖片

   自動產生的描述

設定時間格式為 月/日

設定昨日時間

設定前天時間

儲存昨日po文

1. 因為ptt最新的po文會在頁面的最下面，所以當此頁的第一則po文是前天時，有可能此頁還是有昨日的po文。判斷此頁第一則po文，若是前天、昨天或是今日的po文，則進入迴圈，判斷有沒有昨日的po文，否則跳出迴圈。因為有些貼文被刪除，會沒有標題及連結，所以在迴圈內，判斷貼文的日期是否為昨日，若是昨日，試試看能不能抓取到連結，不行的話跳過這篇貼文，可以的話就將連結存入陣列。此頁貼文都判斷完後，按下”上頁”的按鈕，抓取上頁的貼文及日期，判斷是否進迴圈。一張含有 文字 的圖片

   自動產生的描述

取得第一則po文日期

取得這頁所有po文

點擊上頁按鈕

設定上頁按鈕

將文章連結存入陣列

判斷文章是否被刪除

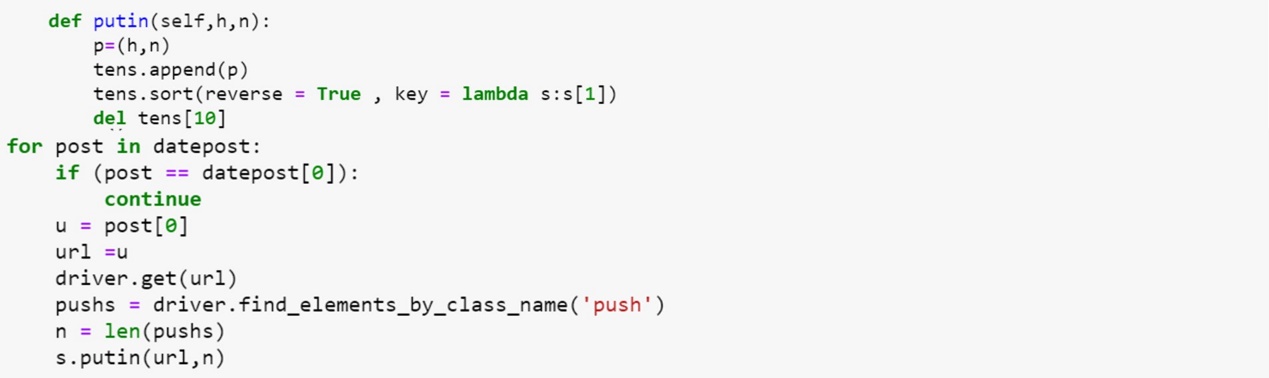
若是昨天的po文

取得第一則po文日期

取得這頁所有po文

1. 將建立的類別初始化，才能使用裡面的副程式

sort為類別的名字

1. 爬取昨日的每篇貼文，計算蓋樓數，使用副程式排序

陣列長度=push的數量=蓋樓的數量

陣列的第一個元素是程式開始時賦予的，為空字串

依蓋樓數，由大到小排列陣列內元素

陣列的第11個元素，蓋樓數為第11名，不在我們需要爬取的範圍，可以刪除

1. 利用副程式爬取10則po文的詳細內容，存入csv檔，再印出今日日期一張含有 文字 的圖片

   自動產生的描述

印出今天的日期

印出表格標題

將陣列內容印出來

將此陣列加入另一個陣列

將所有資訊存入一個陣列

抓取此篇貼文所有※欄位

抓取此篇貼文po文IP位置

用’—‘切割文字

抓取此篇貼文所有文字

抓取此篇貼文標題

抓取此篇貼文po文時間

抓取此篇貼文作者

抓取此篇貼文所需資訊

1. 關閉瀏覽器
2. 程式結束

程式執行結果：一張含有 桌 的圖片

自動產生的描述