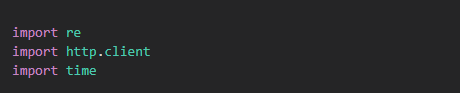
1. 前言



這次玩一個特別一點的好了 挑戰都用原生資料庫來寫爬蟲

好處是移植到大多環境應該也都正常直譯,不用另外安裝特定的套件

下次再來挑戰用C++刻爬蟲(?)

然後為啥要選表特版阿 裡面一堆18+照片耶

想必是單純以學術的角度,只是讓我們練習修改header

可見老師真的是用心良苦

一張含有 文字, 監視器, 螢幕擷取畫面, 室內 的圖片

自動產生的描述

1. 檔案相對位置

一張含有 文字 的圖片

自動產生的描述

這個專案的檔案其實沒有很多,由於我這次的環境是windows 10,

.dat都是我的cmd shell script

原則上使用者只要 點擊 init.dat

就會把 data.bat放入系統排程

而data.bat主要是直譯ptt.ipynb方法

後面會再提,等等先來查看ptt.ipynb,爬蟲的主體腳本

1. 引用函式庫&建立第一個連線

一張含有 文字 的圖片

自動產生的描述

這邊import re會用到正規表示法，作為清洗資料所用

import http.client 是為了要跟伺服器請求資料

Import time是為了記錄爬下來的時間以及判定昨天是哪一天

主機位置這設定為 [www.ptt.cc](http://www.ptt.cc) 也是這次作業唯一所求的網域 , 因此我後面也沒有再定義其他連線方式

有稍微修改一下header, 加上‘cookie:’ : ‘over18=1’ 原因是剛剛提到的,為了讓伺服器知道我已經長大了,\_\_cf\_bm=”…. , 其實我目前也還不知道原因,只是觀察各個瀏覽器都會送這個項目,我就先加上去,讓自己看起來沒這麼像是爬蟲(?)

然後conn.request就先放表特版首頁路徑,payload做任何沒有調整,payload通常是post的時候比較常用

來到表特首頁後

我在這邊截取的資訊只有上頁網址的編號而已

一張含有 文字 的圖片

自動產生的描述

由於我有觀察到表特版甚至是全部的ptt



Html檔案命名方式只是 index(流水號).html

因此可以很輕易更改數字請求而輕鬆回朔,真是貼心的網站

這邊採取得到倒數第二頁的方式+1 就知道最後一頁的路徑了

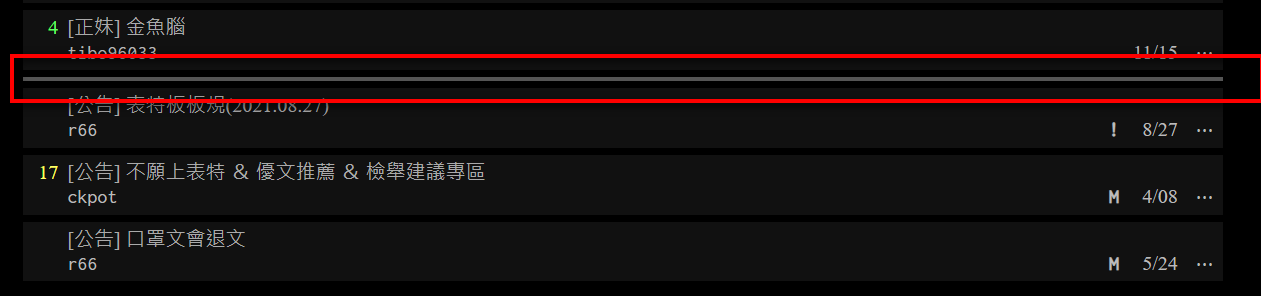
1. 瀏覽各個公告欄

先跳過一些中間得程式碼,來看看我是怎麼歷遍所求的公佈欄

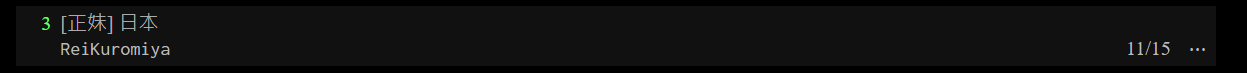
按照這次作業要求,只要昨天的資料,所以如果我讀取道的公布欄有前天甚至是更早之前的文章會停止搜索一張含有 文字 的圖片

自動產生的描述

稍微說一下 r-list-sep是下圖所說的分隔線的class,由於置頂公告的時間通常都是昨天以前的,會影響我上述的判斷(看到前天的文章作為停止條件),分隔線下面的都是置頂公告,所以我把後面分隔線後的資料都砍了



“regex” 的re策略是汲取是在公告欄的每個”文章”,而其中的是包含newline的任意字元



由於我必須要看過每個文章的時間,以及連結網址,而不是找尋到第一個而已

一張含有 文字, 螢幕擷取畫面, 螢幕, 銀色 的圖片

自動產生的描述

因此我把所有matches都存起來,再慢慢列舉

articleDate是每個文章的日期

articleDate的第一個group是月份,第二個是日期

利用DataCmp來判定是不是昨天 來決定要不查看該文章,並加入資料集

如果發現太前面的日期就不會繼續回朔到前一個公告欄了

如果是今天或是昨天的文章,迴圈會繼續

1. 文章搜索&時間判定

一張含有 文字 的圖片

自動產生的描述

這邊就透過參數提供的超連結路徑請求每一篇文章

依序採集了 推文數量,噓文數量,作者,標題,公告時間,第一個圖片

注意一下 編碼格式以及資料型態

其中的 \d+是n個任意數字(0<=n<無限大)

\s+是n個空白字串 .\* 是不包含換行的任意字串

time.ctime()記錄下我當下爬取的時間

還有要注意一下不是每個文章都有圖片(其實我也搞不表特版發文不附圖的邏輯)

所以簡單寫了一下沒有圖片的例外處理

時間判定方式 如下(自己看應該不難理解(?))

一張含有 文字, 牆, 監視器, 室內 的圖片

自動產生的描述

檔案輸出我這邊就省略不寫了

1. windows bash

我在init.dat的指令其實只有一行而已



SCHTASKS呼叫系統排程設定

/CREATE 是建立一個新的排程

/SC DAILY 是說我一天做一次

/TN <任務名稱> ”

/TR <任務路徑> : 其中%~dp0是bash內建的變數,表明所屬dat檔案的絕對路徑

%ch是使用者cmd的路徑,兩者不一定相同,小心不要混用用了

/ST 表示我想採用24小時制表明時間

以下是daily.bat

一張含有 文字 的圖片

自動產生的描述

由於我在cmd沒有設定jupyter環境變數

因此呼叫了anaconda的shell來跑這第二行指令

直譯 ptt.ipynb