Python的requests套件

requests.get()：

取得網頁資料

Python的Beautiful Soup套件

BeautifulSoup(res.text,'html.parser')：

將網頁的HTML程式碼取回來後，接著引用BeautifulSoup類別(Class)，傳入取回的HTML結構字串，並且指定HTML的解析型態來建立其物件。

soup.select('something')：

從soup裡面取得某些我們要的標題。

Python的re套件

re.sub(pattern,repl,string,count = 0,flags = 0)：

pattern:表示規則運算式中的模式字串

repl:被替換的字串（既可以是字串，也可以是函數）

string:要被處理的，要被替換的字串

count:匹配的次數

flag:編譯標誌。編譯標誌讓你可以修改正則表達式的一些運行方式

re.compile：

用來將正則表達式轉換爲一個「pattern object」，我們可以稱之爲「模式對象」。將正則表達式轉換爲模式對象的作用就是可以將其保存下來，已備後續之用。

findall(pattern, string, flags=0)：

可以找出string中所有匹配pattern的項。

re.search：

掃描整個字符串並返回第一個成功的匹配。

something.group(1)：

第一括號內的東西，且符合我們的要求。

Python的urlretrieve套件

urlretrieve(url, filename, reporthook=None, data=None)：下載檔案用

url:要下載的網址

filename:檔案名稱

Python的os套件

os.path.isdir(folderpath)：

檢查目前目錄下有沒有這個資料夾。

os.mkdir(foldername)：

建立一個資料夾。

os.path.join()：

將多個路徑組合後返回。

Python的sys套件

sys.argv[1]：

用來讀取從外部輸入的資料。



第1~6行：匯入我們所需要的一些函式庫

第8行：一個名為download\_images的function，可以傳入文章的title

第9行：把文章的title一個一個取出來做動作

第10、11行：把title做處理，刪去一些創資料夾時不能出現的符號以及去除頭尾的空白

第13、14行：檢查當前路徑有沒有以title為名稱的資料夾，如果沒有就建一個

第15~18行：進去我們要的網頁，給定網頁cookie，因為ptt表特版有18歲的限制

第19行：從裡面取得我們要的圖片名稱，用findall()，要找的規格在第49行有規範

第20~23行：從第20行取得的圖片庫裡面，把每張圖片依序下載，下載名稱用re.search()取得

第25行：名為crawler的finction，可以傳入要做幾次

第27~28行：檢查當前目錄下有沒有download這個資料夾，如果沒有就建立

第29行：設定初始的網頁位址

第30行：執行N次的迴圈

第32~35行：進去我們要的網頁，給定網頁cookie，因為ptt表特版有18歲的限制

第37行：透過BeautifulSoup取得網頁的內容建立一個物件

第39行：從物件裡面取得我們要的文章

第41行：取得上一頁的網址

第43行：從剛剛拿到的資料中取出我們要的部分

第45行：把剛剛拿出的部分設定成下一個要執行的網址

第47行：執行download\_images Function

第49行：設定給findall用的參數

第51行：從cmd執行時，獲取參數用，最主要讓使用者輸入要download多少頁的圖片