

網路程式設計網路爬蟲\_動態網頁解析

Instructor: 馬豪尚



### 下載網頁上的圖片資源

- > 基本原理
  - 向圖片資源的位置請求圖片內容
  - 開啟圖片資源,並存檔

#### > Example

- 圖片資源位置
  - https://imgur.com/gallery/uygEURT
- 請求圖片資源,以二進制的格式傳送圖片內容
  - > img = requests.get(url)
- 以二進制格式寫入檔案
  - > f = open('save.jpg', 'wb')
  - > f.write(img.content)





### 爬取圖片-靜態網頁

- › 使用chrome瀏覽器開發人員工具分析html
- › 使用Quick JavaScript Switcher判斷是否為動態嵌入網頁
- > 選擇取得資源的方式
  - Request
- > 定位圖片資源在網頁中的位置
  - Beautifulsoup



# 爬取圖片-靜態網頁#1 PTT\_Beauty

- > 爬取網頁情境
  - 要爬取的網頁內容為靜態網頁
  - 要將多張圖儲存下來, 一次爬一張圖
- ,爬取某一篇文章內的圖片
  - https://www.ptt.cc/bbs/Beauty/M.1638380033.A.7C7.html
- › 定位到html内的圖片位置



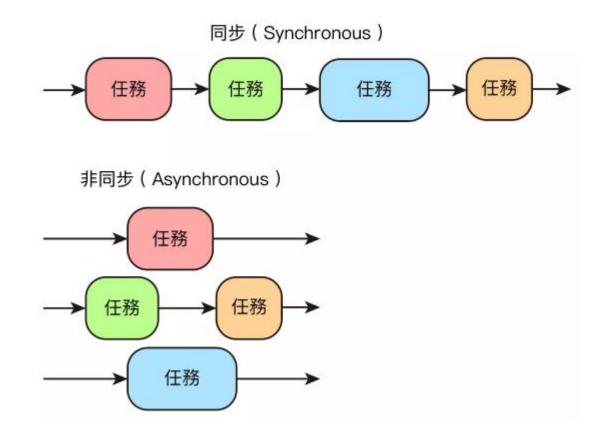
# 爬取圖片#1 PTT\_Beauty

- > 向圖片資源位置請求圖片資源內容
  - img = requests.get(url)
- , 儲存圖片檔案
  - -f = open('save.jpg', 'wb')
  - f.write(img.content)



# Python任務處理流程

> Python 在執行時,通常是採用同步的任務處理模式,一個 處理完成後才會接下去處理第二個





- › concurrent.futures標準函式提供了平行任務處理(非同步) 的功能,能夠同時處理多個任務
  - ThreadPoolExecutor 針對 Thread (執行緒)
  - ProcessPoolExecutor 針對 Process (程序)

英文	中文	說明
Thread	執行緒	程式執行任務的基本單位。
Process	程序	啟動應用程式時產生的執行實體,需要一定的 CPU 與記憶體資源,Process 由一到多個 Thread 組成,同一個Process 裡的 Thread 可以共用記憶體資源。



#### ThreadPoolExecutor

- > 會透過 Thread 的方式建立多個 Executors (執行器)
- > 執行並處理多個任務 (tasks)
- > ThreadPoolExecutor 有四個參數

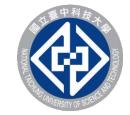
參數	說明
max_workers	Thread 的數量,預設 5 (CPU number * 5,每個 CPU 可以處理 5 個 Thread),數量越多,運行速度會越快,如果設定小於等於 0 會發生錯誤。
thread_name_prefix	Thread 的名稱,預設 "。
initializer	每個 Thread 啟動時調用的可調用對象,預設 None。
initargs	傳遞給初始化程序的參數,使用 tuple,預設 ()。



#### ThreadPoolExecutor

- > 創建一個執行 Thread 的啟動器
  - executor = ThreadPoolExecutor()
- › 使用 ThreadPoolExecutor 後,就能使用 Executors 的相關方法
  - executer.submit(fn, \*args)

方法	參數	說明
submit	fn, *args, **kwargs	執行某個函式。
map	func, *iterables	使用 map 的方式,使用某個函式執行可迭 代的內容。
shutdown	wait	完成執行後回傳信號,釋放正在使用的任何資源,wait 預設 True 會在所有對象完成後才回傳信號,wait 設定 False 則會在執行後立刻回傳。



#### ThreadPoolExecutor

- ,迴圈架構可以用map來執行
- Example
  with ThreadPoolExecutor() as executor:
  executor.submit(test, 2)
  executor.submit(test, 3)
  - executor.submit(test, 4)
  - →with ThreadPoolExecutor() as executor: executor.map(test, [2,3,4])



# 平行任務處理#2 threading

- › 載入threading 多執行緒處理模組
  - import threading
- › 建立 threading 的物件
  - thread = threading.Thread(target=function, args)
    - > target=function為指定執行的函式
    - › args為傳入函式的參數



# Threading

› 建立 threading 物件之後,就可以使用下列常用的方法

方法	說明	
start()	啟用執行緒。	
join()	等待執行緒,直到該執行緒完成才會進行後續動作。	
ident	取得該執行緒的標識符。	
native_id	取得該執行緒的 id。	
is_alive()	執行緒是否啟用,啟用 True,否則 False。	



### 爬取圖片#2 PTT\_Beauty

- > 爬取網頁情境
  - 要爬取的網頁內容為靜態網頁
  - 要將多張圖儲存下來,一次同時爬很多張圖
- > 使用concurrent.futures來批量下載圖片
  - 使用 map 的方式,使用某個函式執行可迭代的內容。
- > executor = ThreadPoolExecutor()
- › executor.map(下載的函式)



#### 練習1

- › 下載PTT\_Beauty最新的五篇文章內所有圖片
  - 使用selenium爬取最新的五篇文章
  - 使用selenium/Beautifulsoup定位圖片資源的元素位置
  - 取得圖片資源的網址
  - 用threading 多執行緒的方式下載圖片



### 爬取圖片#3 寶可夢圖鑑

- > 爬取網頁情境
  - 要爬取的內容是靜態的內容
  - 網址都是具有固定規律的
- › 使用chrome瀏覽器開發人員工具分析html
- > 選擇取得資源的方式
  - Request
- > 定位圖片資源在網頁中的位置
  - Beautifulsoup



### 爬取圖片#3 寶可夢圖鑑

- ,爬取內容子分頁的網址
  - https://tw.portal-pokemon.com/play/pokedex/0001
- > 定位圖片資源在網頁中的位置



> 搭配threading來批次下載



# 爬取圖片#4 imgur 圖片網站

- ,如果沒有可互動元素,網頁是採取往下滑就載入更多元素的方式 來設計
- › 可以用selenium套件中與javascript互動的功能來達成
  - driver.execute\_script(js)
- > Javascript控制網頁移動卷軸到某個位置
  - window.scrollTo(x, y)
- › Javascript取得瀏覽器的卷軸高度
  - document.body.scrollHeight
- > Example
  - js = "window.scrollTo(0, document.body.scrollHeight)"
  - driver.execute\_script(js)



# Python image downloader

- > 套件網址
  - https://github.com/webscraperio/image-downloader
  - 下載image\_downloader.py為自訂模組
- › google colab載入自訂模組
  - 上傳模組到工作資料夾下
  - import image\_downloader
- ,使用該套件
  - image\_downloader.download\_csv\_file\_images("img.csv")
    - > img.csv是一個csv檔裡面紀錄了要下載的圖片資源位置(url)
    - > csv檔內要讓套件下載的參考欄位名稱必須以'-src'結尾



#### 練習2

- > 下載imgur網站關鍵字為cat的圖片
  - 向下滑動10次獲取更多圖片
  - 爬取網頁內容並定位圖片資源的元素位置
  - 取得圖片資源的網址
  - 將每個網址位置輸出成一個csv檔, 欄位名稱為img-src
- › 使用image downloader下載該csv檔內的圖片