**1-爬蟲基礎知識**

## D1：資料來源與檔案存取

1.（簡答題）檔案、API、爬蟲三種取得資料方式有什麼不同？

2.（實作）完成一個程式，需滿足下列需求：

「下載指定檔案到 Data 資料夾，存成檔名 Homework.txt」的檔案網址：**[https://www.w3.org/TR/PNG/iso\_8859-1.txt](https://www.w3.org/TR/PNG/iso_8859-1.txt" \t "_blank)** 或任一個 .txt 的檔案網址

檢查 Data 資料夾是否有 Homework.txt 檔名之檔案

將「Hello World」字串覆寫到 Homework.txt 檔案

檢查 Homework.txt 檔案字數是否符合 Hello World 字數

## D2：實作：Python 下載CSV檔案與解析

比較一下範例檔案中的「File I/O」與「CSV Reader」讀出來的內容有什麼差異

根據範例檔案的結果：

1. 取出班次一的每一個時間，印出來就好

2. 將班次一的每一個時間**用一個變數**保存

3. 將所有班次和其每一個時間**用一個變數**保存

(Hint： 2&3 要想一下用什麼的資料型態做整理比較適合)

## D3：實作：Python 下載XML檔案與解析

比較一下範例檔案中的「File I/O」與「xmltodict」讀出來的內容有什麼差異

根據範例檔案的結果：

1. 請問高雄市有多少地區有溫度資料？

2. 請取出**每一個地區**所記錄的**第一個時間點跟溫度**

3. 請取出**第一個地區**所記錄的**每一個時間點跟溫度**

## D4：HTTP Server-Client 架構說明與 利用 Python 存取 API

1. 比較一下範例檔案中的「r.text」與「json.loads(r.text)」讀出來的內容有什麼差異
2. 自行尋找一個合適的 API 接口做練習，並且查看其回傳內容：
   * [**https://cat-fact.herokuapp.com/facts**](https://cat-fact.herokuapp.com/facts) (來源：**[https://alexwohlbruck.github.io/cat-facts/](https://alexwohlbruck.github.io/cat-facts/" \t "_blank)**)
   * [**http://odata.wra.gov.tw/v4/RealtimeWaterLevel**](http://odata.wra.gov.tw/v4/RealtimeWaterLevel) (來源：**[https://data.gov.tw/dataset/25768](https://data.gov.tw/dataset/25768" \t "_blank)**)

## D5：實作 Day：API 資料串接 - 基本 API + JSON

請利用 API: [**https://www.dcard.tw/\_api/forums/pet/posts?popular=true**](https://www.dcard.tw/_api/forums/pet/posts?popular=true) 回答下列問題：

1. 這個 API 一次會回傳幾筆資料？每一筆資料包含哪些欄位？

2. 取出每一筆資料的「標題」、「貼文時間」、「留言人數」、「按讚人數」

3. 計算熱門/非熱門文章的「平均留言人數」與「平均按讚人數」

## D6：實作 Day：API 資料串接 - Headers

根據範例提供的 API 網址 ，完成以下問題：

1. 取出知乎問題發問時間

2. 取出第一筆與最後一筆回答的時間

**2-靜態網頁爬蟲技術**

## D7：HTTP 靜態網頁架構說明與淺談HTML / CSS / JavaScript

1. （簡答題）請問目前提到的這種基於 HTTP 一來一回的網頁傳輸方式，有哪些缺點及問題？

2. （簡答題）第 16 頁所使用的 JavaScipt 程式語言，跟我們一般使用的 Python 在執行上有什麼不同。

## D8：靜態網頁的資料爬蟲策略

利用 Request + BeatifulSoup 爬取下列兩個網站內容並解析：

1. Dcared 網址： [**https://www.dcard.tw/f**](https://www.dcard.tw/f)

2. 知乎： [**https://www.zhihu.com/explore**](https://www.zhihu.com/explore)

並且回答下面問題：

1. Request 取回之後該怎麼取出資料，資料型態是什麼？

2. 為什麼要使用 BeatifulSoup 處理？

3. 觀察一下知乎回來的資料好像有點怪怪的，該怎麼解決？

## D9：圖片下載

以正確的副檔名下載網頁中的圖片

範例網頁:[**https://www.ptt.cc/bbs/Beauty/M.1556291059.A.75A.html**](https://www.ptt.cc/bbs/Beauty/M.1556291059.A.75A.html)

Note：因為 PTT 會詢問「是否滿 18 歲」，這邊可以用 cookies 繞過

requests.get(URL, cookies={'over18': '1'})

## D10：PyQuery/grab

將之前用 requests + beatifulsoup 實作的方式，改寫成 grab + pyquery，並且比較有哪些地方不同。

## D11：Regular expression

本日作業：正規表達式練習

在網路爬蟲當中，正規表達式常常用來過濾以及搜尋特定的 pattern 字串。

今天我們來練習過濾 IP address，以及 URL。

## D12：實作 Day：ETTODAY 靜態爬蟲實作練習

根據範例 ，完成以下問題：

1. 取出今天所有的發文
2. 如果想要依照類別分類，怎麼儲存會比較好？
3. 哪一個類別的文章最多

## D13：實作 Day：PTT(批踢踢)網路爬蟲實作練習

根據範例 ，完成以下問題：

1. 印出最新文章的「作者」「標題」「時間」
2. 印出第一頁所有文章的「作者」「標題」「時間」
3. 試著爬爬看其他版的文章

## D14：實作 Day：Yahoo! 電影網頁爬蟲實作練習

YAHOO電影爬蟲練習

練習爬取電影放映資訊。必須逐步獲取電影的代號、放映地區、放映日期後，再送出查詢給伺服器。

## D15：實作 Day：金融資訊網頁爬蟲實作練習

台灣銀行牌告匯率爬蟲練習

作業part 1：請參考上面的範例，試著爬取所有貨幣對台幣的匯率

作業part 2：爬取每種貨幣對台幣的一個月歷史匯率資料，並畫出趨勢圖

## D16：實作 Day：Wiki的爬蟲實作練習

Wikipedia爬蟲練習

練習從Wikipedia中爬取文章。先定義一個搜尋的關鍵字，擷取該關鍵字詞的文章。

**3. 動態網頁爬蟲技術**

## D17：HTTP 動態網頁架構說明與非同步取得資料

回答以下問題：

1. 動態網頁跟靜態網頁的差別是什麼？原本靜態網頁的做法會產生什麼問題或是缺點？
2. 還記得我們在 ETTODAY 靜態爬蟲實作練習中，有請大家完成一個題目「取出今天所有的發文」，但仔細看其實並沒有真的把當天所有的新聞都抓回來， 試著回答看看原因是什麼？及該如何檢查？

## D18：瀏覽器開發者工具介紹

利用開發者工具去尋找「[**https://www.zhihu.com/explore**](https://www.zhihu.com/explore)」和「[**https://aicoin.cn**](https://aicoin.cn/)」網站的 HTTP 網址，以及他所帶的 Headers 分別是哪些，試著用 request 去發送有沒有帶 Headers 的差異。

## D19：動態網頁爬蟲 - 使用Selenium + BeautifulSoup 模擬瀏覽器執行

根據範例使用 selenuim 取出 高雄市小港區 2019 年 空氣品質 資料，並用一個適合的資料結構儲存

Hint: 將所有資料存在一個變數當中

## D20：動態網頁爬蟲 - 利用開發者工具，觀察模擬 API 存取

根據範例程式，使用「API Request」進一步取出 台北市士林區 table 的資料

## D21：實作 Day：ETTODAY 動態爬蟲實作練習

今日作業：

1. 取出今天所有的新聞
2. 取出今天下午三點到五點的新聞
3. 根據範例，取出三天前下午三點到五點的新聞

## D22：實作 Day：空氣污染網站爬蟲實作練習

根據範例 ，完成以下問題：

1. 取出 台北市士林區 2018/01 – 2018/08 的 SO2 資料
2. 取出 台北市士林區 2018/01 – 2018/08 的 SO2、CO 資料

## D23：實作 Day：東森新聞雲網站爬蟲實作練習

練習從東森新聞雲網站中，爬取新聞摘要及詳細內容。學習利用Selenium模擬人為操作，更新動態網頁後爬取新聞內容。

## D24：實作 Day：104人力銀行網站爬蟲實作練習

今日練習：從104人力銀行網站爬取求職公司資訊

**4. Scrapy 網站爬蟲框架**

## D25：多網頁爬蟲實作策略介紹

【今日無需提交作業，請完成今日範例練習與思考題，按下「我已完成」按鈕即可】

今日練習：

搭配之前的練習，將 [**PTT**](https://www.ptt.cc/ask/over18?from=%2Fbbs%2FGossiping%2Findex.html) 上其中一頁的列表文章都爬下來

* 我們練習將八卦版最新一頁的列表都爬下來
* 爬蟲過程忽略置底文章



請練習範例內容，並思考下列題目：

前面範例題是同一個網站多個網頁，請問生活中有其他多個網站多個網頁的應用範例嗎？

若思考題想要跟出題老師討論交流，歡迎您到共學社團上提問。

請點擊下方檢視範例參考Day025\_MultiPage\_Sample.ipynb，完成今日練習，按下「我已完成」按鈕即算完成今日知識點。

## D26：Scrapy 爬蟲流程 (1) - 建立流程 + 送出請求

學習透過命令列建立專案與爬蟲

請在 [**https://www.ptt.cc/bbs/index.html**](https://www.ptt.cc/bbs/index.html) 中設定一篇文章當作目標網頁，

並改寫 scrapy 中的 start\_urls

撰寫完整邏輯並執行爬蟲

Hint: PTT 爬蟲先前為了避開「是否滿 18」的頁面，透過 cookies 解決，這邊可以透過相同的方式繞過

scrapy.Request(..., cookies={'over18': '1'})

## D27：Scrapy 爬蟲流程 (2) - XPath + Item Pipeline

* 參考D026的作業，在 [**https://www.ptt.cc/bbs/index.html**](https://www.ptt.cc/bbs/index.html) 中設定一篇文章當作目標網頁
* 透過 Item 設定爬完的文章資訊資料格式
* 透過 Item Pipeline 設定爬文存成 JSON 的流程
  + open spider 時機點檢查儲存路徑
  + close spider 時機點存檔

## D28：Scrapy 爬蟲流程 (3) - API

參考本次投影片中的例子撰寫 [**main.py**](http://main.py/)

從外部給予目標網址跟存檔名稱

## D29：Scrapy 爬蟲流程 (4) - 多網頁爬蟲

這次假設我們的爬蟲可以根據外部選項 board 來決定要爬 PTT 的哪一個版

外部可以給予參數決定目標的版 (e.g. Gossiping, NBA, …)

爬蟲內部處理文章列表與內容

透過 Item Pipeline 將結果存成 JSON

**5. 進階爬蟲技術**

## D30：爬蟲可能會遇到的問題

（簡答題）試著舉出一到三個爬蟲可能會遇到的問題導致無法抓取的？

## D31：反爬：瀏覽器標頭與基本資訊

1. 尋找一個網站是有有進行 Header 檢查的？（Hint: 可以利用有加跟沒有加回傳結果差異做判斷）
2. 用上述網站說明該如何判斷 Header 中必須加上哪些資料才會正確

## D32：反爬：驗證碼處理

1. 讀讀看「test」資料夾下的圖檔！
2. 像 test/test2.png 這種複雜的圖片怎麼辦？

## D33：反爬：登入授權模擬

找一個需要登入的網站試試看，並說明思考流程（如果不知道要用哪個網站的話，可以試試看 [**https://github.com/new**](https://github.com/new) 網址，未登入時會被導向登入頁）

## D34：反爬：代理 IP

目前程式中的 proxy\_ips 是手動輸入的，請根據 [**https://free-proxy-list.net/**](https://free-proxy-list.net/) 寫一個可自動化抓取可用 Proxy 的 proxy\_ips。

## D35：加速：多線程爬蟲

找一個之前實作過的爬蟲改用多線程改寫，比較前後時間的差異。

## D36：加速：非同步爬蟲

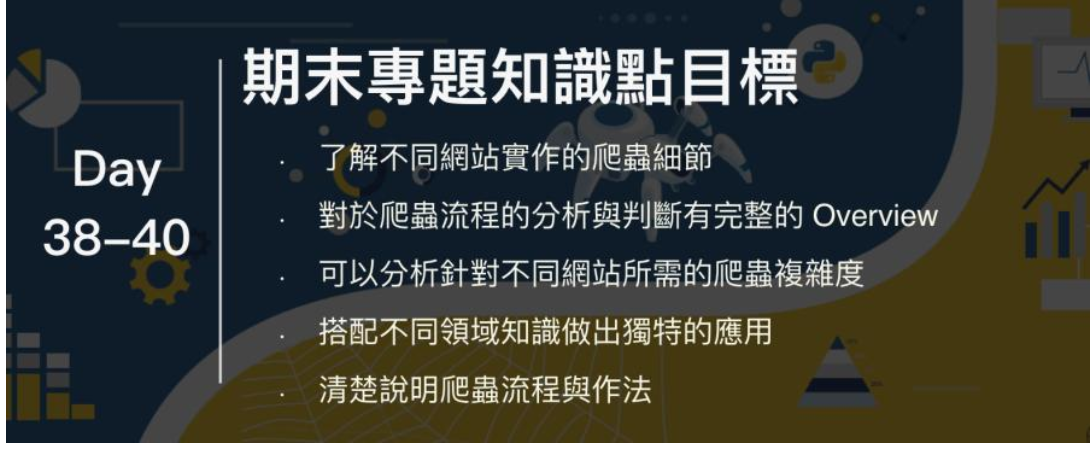
比較一下非同步爬蟲跟多線程爬蟲的差異是什麼？各自的優缺點為何？

## D37：自動化更新機制(排程)

列舉一個是需要用自動化更新的機制來達成的爬蟲應用。

**6. 期末實戰專題**

## D38 - D40：期末專題



### 期末專題完賽流程

Step 1. 學員選擇期末專題主題之一進行

Step 2. 將期末專題程式碼上傳至 github 後，至活動官網完成提交

Step 3. 請至 [**Python網路爬蟲實戰研習共學社團**](https://www.cupoy.com/club/ai_tw/0000016E62FB84E4000000026375706F795F72656C656173654B5741535354434C5542/content/post) 建立期末專案文章分享

Step 4. 請參考 [**Python期末專題\_Cupoy (格式參考)**](https://www.cupoy.com/clubnews/ai_tw/0000016E62FB84E4000000026375706F795F72656C656173654B5741535354434C5542/000001701E007A70000000026375706F795F70726572656C656173654B5741535354434C55424E455753) 文章作為專案文章建立基礎

Step 5. 由其他學員與出題教練進行按讚，主辦單位進行票數統計

Step 6. 票數最高的 10% 學員將獲得期末挑戰紀念品

Q.有關基礎目標與進階目標的差別?

只要達到基礎目標就可以發表專題文章，若學員想挑戰作到進階目標的話，對爬蟲的理解將會更全面，也會有加分。

# Python期末專題\_Cupoy (格式參考)

2020/02/07

### 一、專題摘要 ****(解釋實作與說明需要解決的問題，限300~500字。)****

1. 期末專題主題
2. 期末專題基本目標

### 二、實作方法介紹 ****(介紹使用的程式碼、模組，並附上實作過程與結果的截圖，需圖文並茂。)****

1. 使用的程式碼介紹
2. 使用的模組介紹

### 三、成果展示 (介紹成果的特點為何，並撰寫心得。)

### 四、結論 ****(總結本次專題的問題與結果)****

### 五、期末專題作者資訊 (請附上作者資訊)

1. 個人Github連結
2. 個人在百日馬拉松顯示名稱