Web Crawler 網路爬蟲
以Python實作

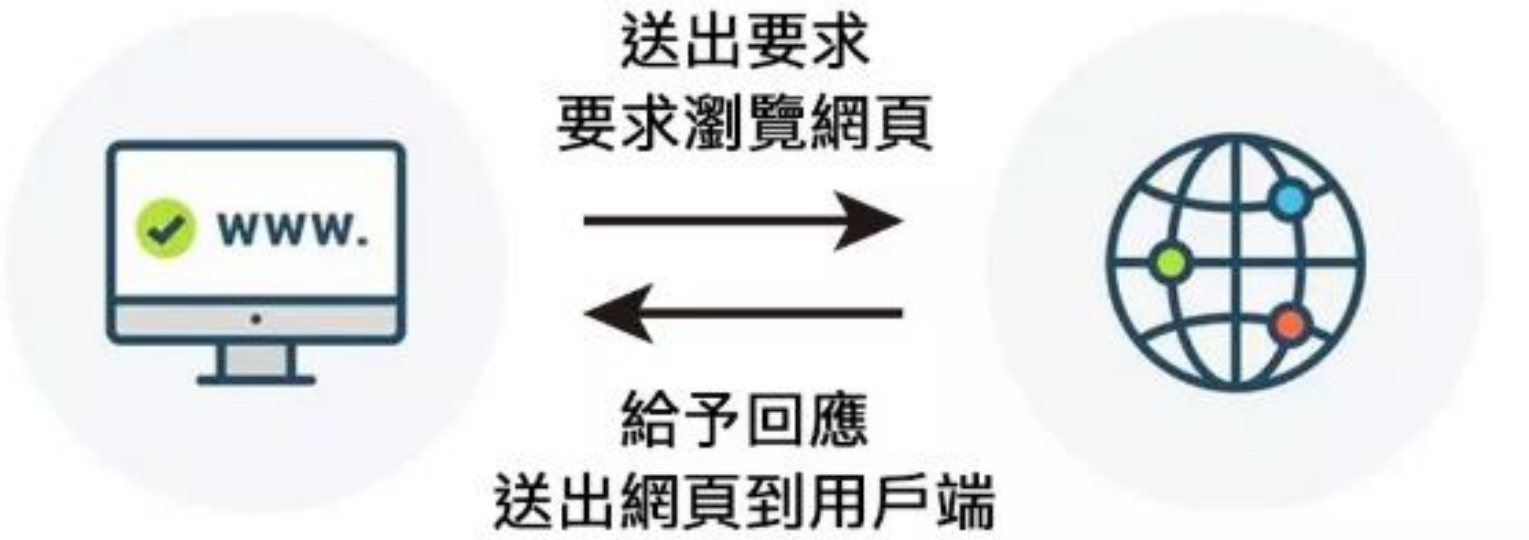
11/24/2024

Louisa Coffee

一、網路爬蟲概論

- 網路爬蟲 (spider 或 web crawler)，是一種可以「自動」瀏覽全球資訊網的網路機器人，許多的搜尋入口網站 (例如 Google)，都會透過網路爬蟲收集網路上的各種資訊，進一步分析後成為使用者搜尋的資料，許多開發者也會自行開發不同的爬蟲程式，進行大數據收集與分析的動作。

靜態網頁爬蟲



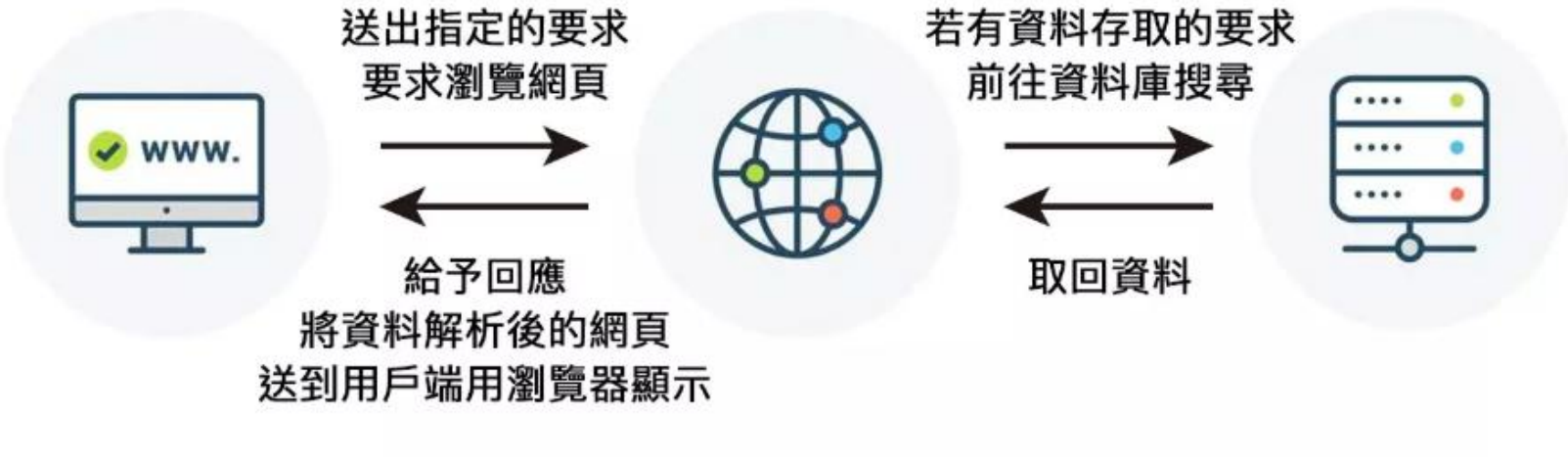
靜態網站是指**網站完成一個請求（request）與回應（response）後，用戶端即不再與伺服器有任何的交流**，所有的互動都只與瀏覽器的網頁互動，資訊不會傳遞到後端伺服器。

二、流程

1. 設定目標：即想要取得的資訊。
2. 觀察網頁：觀察網頁內容，如資料在HTML元素的位置等。
3. 解析內容：透過程式取得網頁，並取出所需的資訊。

(取得資料後要存在檔案或是資料庫就是看自己的需求如何再另行設計了。)

動態網頁爬蟲



動態網站是指**網站會依照使用者的行為不斷的與伺服器進行交流**，例如傳送了 apple 資訊給伺服器，資訊經過伺服器處理後，才會回應 apple 是甜的、紅的、脆的...等相關資訊，不少動態網站甚至需要進行「登入」的動作，像是 Facebook、Instagram...等。

通常動態網站爬蟲實作比較複雜，爬蟲必須要知道網站需要什麼「資訊」，提供了正確的資訊，才能取得所需要的資料（如同開啟保險箱一般，輸入了正確的密碼，才能開啟保險箱的內容）。

爬取並自動下載 PTT 正妹圖片

- 使用 Python 的 Requests 和 Beautiful Soup 函式庫，實作一個可以自動下載圖片的網路爬蟲，只要知道 PTT Beauty 板的網頁網址，就能將網頁內全部的正妹圖片，自動下載到電腦的資料夾中。

crawl_picture.py
crawl_picture2.py

Image in folder: down_img

爬取臺灣銀行牌告匯率

- 使用 Python 的 Requests 函式庫，實作一個爬取臺灣銀行營業時間的牌告匯率的網路爬蟲。

爬取統一發票號碼，自動對獎

- 這篇文章會使用 Python 的 Requests 和 Beautiful Soup 函式庫，實作一個爬取當期統一發票號碼，並進行自動對獎網路爬蟲。

Step 1:使用 Requests 抓取網頁內容

- 使用 Requests 函式庫的 get 的方法，抓取財政部稅務入口網裡統一發票網頁的內容，因為該網頁編碼為 utf-8，所以要加上 web.encoding='utf-8' 避免中文字出現亂碼。

lottery_2.py  統一發票網頁的內容

Step 2: 使用 Beautiful Soup 取出中獎號碼

- 使用 Beautiful Soup 函式庫的 select 的方法，從抓到的網頁內容裡，找到 class 為「container-fluid」的 div，將其內容輸出後就是中獎號碼，但需要注意的是，如果直接將內容輸出放入串列，會自動加上換行符號（因為原始資料裡有換行），此時可以透過串列 slice() 方法的操作，取出最後八碼即可。

lottery_2.py ⟶ 中獎號碼

Step3: 輸入號碼後自動對獎

- 使用 while 迴圈和 for 迴圈，就能做到不斷輸入號碼並自動對獎的功能。

lottery_3.py ➡ 輸入中獎號碼

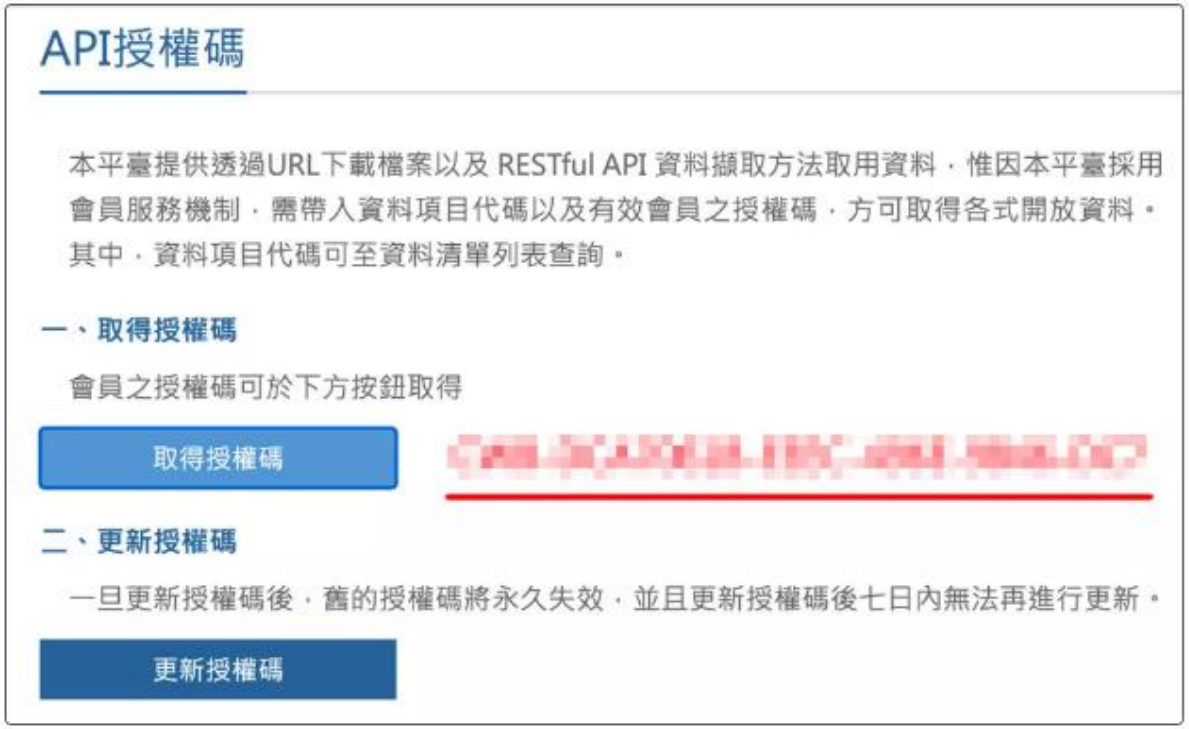
爬取天氣預報

- 從註冊氣象資料開放平臺開始，介紹如何取得天氣預報資料的 JSON 檔案，並使用 Python 的 Requests 函式庫，實作一個可以自動抓取氣象預報資料的網路爬蟲。

- 交通部中央氣象局為了便利民眾共享和應用政府資料，推出了「氣象資料開放平臺」，讓民眾可以在政府資源有限下，善用無限的創意，整合運用開放資料，提升政府資料品質及價值、優化政府服務品質。

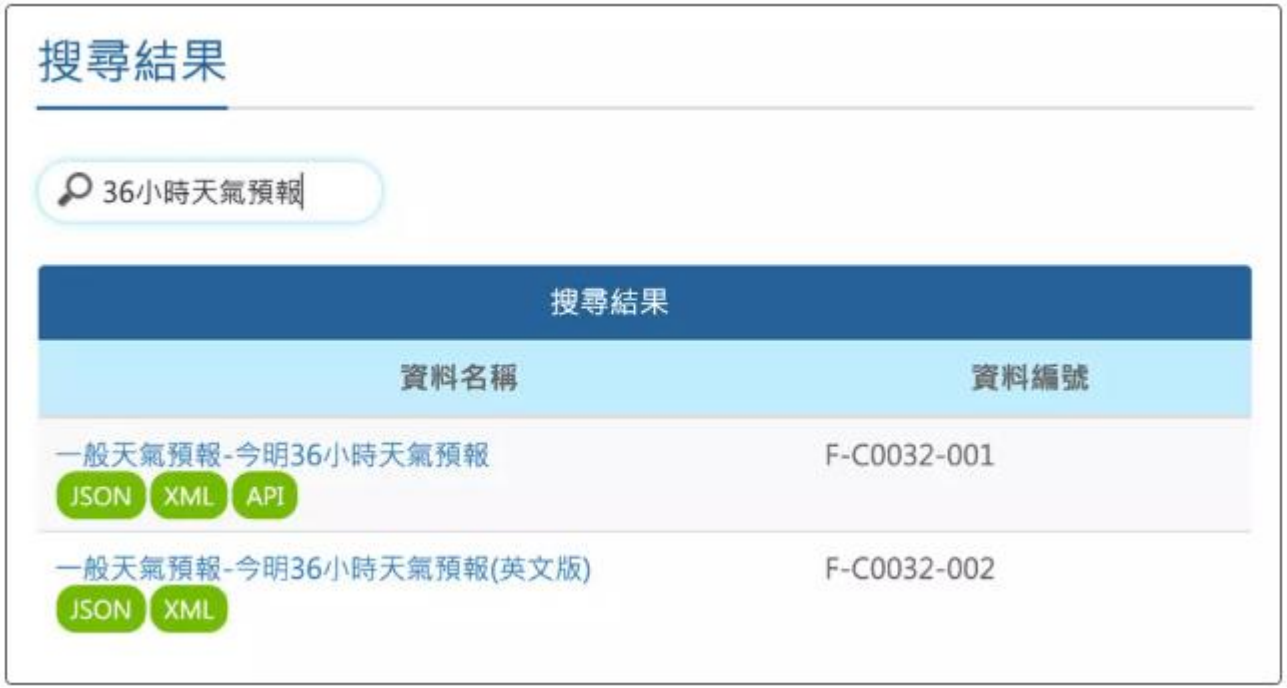
取得使用授權碼

- 要使用氣象資料開放平台的資料需要先「註冊」，點擊右上角的「註冊/登入」，點擊「氣象會員登入」，已有帳號的使用自己的帳號，沒有帳號可以點擊下方「加入會員」註冊帳號。
- 註冊成功後會成為「一般會員」，點擊「取得授權碼」按鈕，會出現個人的授權碼，如果授權碼被盜用或出現問題，可點擊「更新授權碼」重新產生。



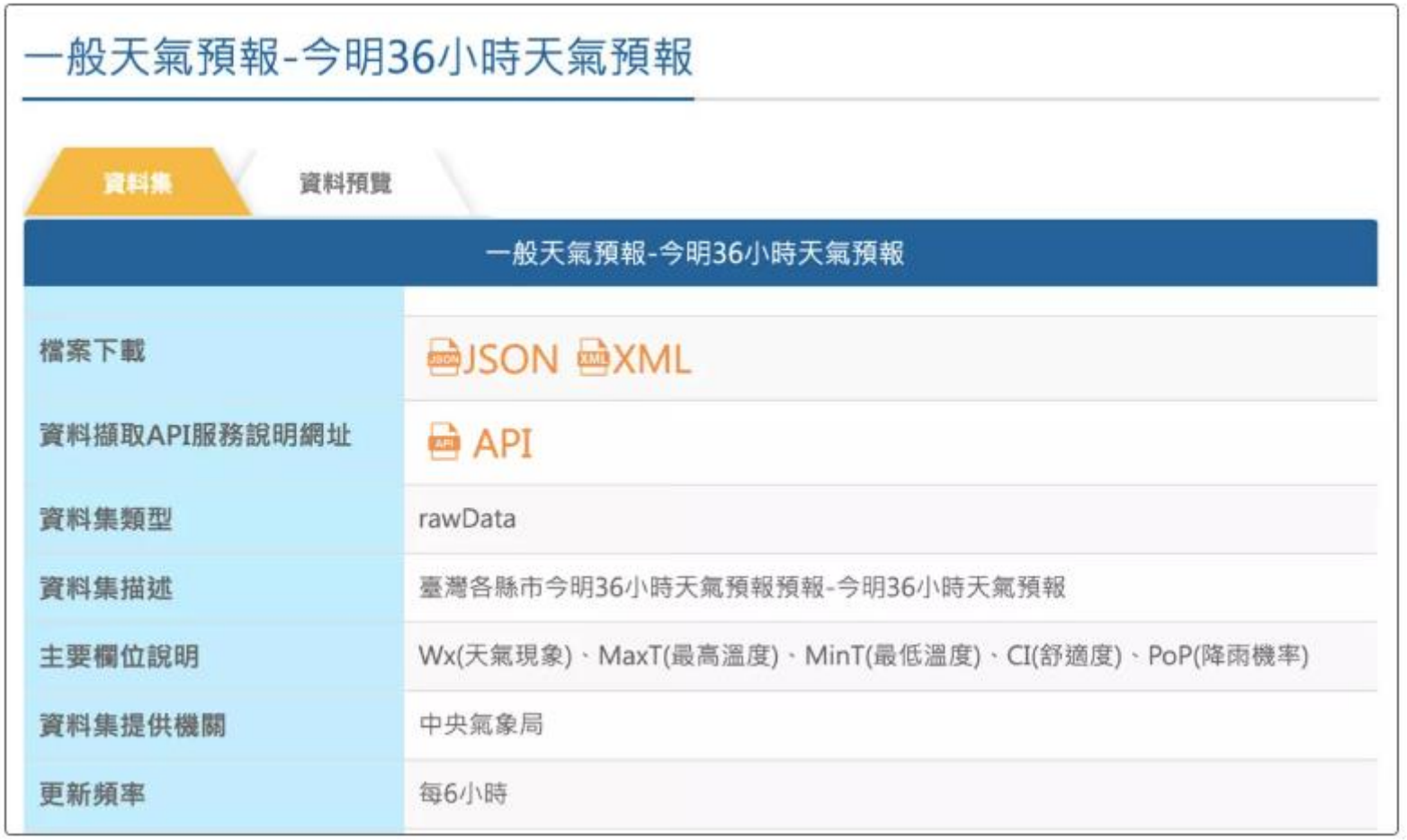
尋找氣象預報資料

- 點擊「資料主題」，選擇「預報」，搜尋「36小時天氣預報」，找到「一般天氣預報 - 今明 36 小時天氣預報」



開啟頁面後，點擊 JSON 或 XML 的連結可以取得台灣所有縣市的預報資料，點擊 API 連結，會開啟「中央氣象局開放資料平臺之資料擷取 API」的頁面，輸入個人的授權碼，就能透過 API 取出篩選後的資料。

尋找氣象預報資料



一般天氣預報-今明36小時天氣預報

資料集　資料預覽

一般天氣預報-今明36小時天氣預報	
檔案下載	JSON XML
資料擷取API服務說明網址	API
資料集類型	rawData
資料集描述	臺灣各縣市今明36小時天氣預報預報-今明36小時天氣預報
主要欄位說明	Wx(天氣現象)、MaxT(最高溫度)、MinT(最低溫度)、CI(舒適度)、PoP(降雨機率)
資料集提供機關	中央氣象局
更新頻率	每6小時

不論是用 JSON 檔案還是 API 的方式，開啟氣象預報資料後，可看見如下圖的 JSON 物件結構，weatherElement 裡的資料就是所需的天氣預報資料。


```
{
 "cwbopendata": {
  "@xmlns": "urn:cwb:gov:tw:cwbcommon:0.1",
  "identifier": "794f497b-743e-6365-d3cc-4a8427c97db0",
  "sender": "weather@cwb.gov.tw",
  "sent": "2022-12-14T11:00:03+08:00",
  "status": "Actual",
  "msgType": "Issue",
  "source": "MFC",
  "dataid": "C0032-001",
  "scope": "Public",
  "dataset": {
   "datasetInfo": {
    "datasetDescription": "三十六小時天氣預報",
    "issueTime": "2022-12-14T11:00:00+08:00",
    "update": "2022-12-14T11:00:03+08:00"
   },
   "location": [
    {
     "locationName": "臺北市",
     "weatherElement": [
      {
       "elementName": "Wx",
       "time": [
        {
         "startTime": "2022-12-14T12:00:00+08:00",
         "endTime": "2022-12-14T18:00:00+08:00",
         "parameter": {
          "parameterName": "陰有雨",
          "parameterValue": "14"
         }
        },
        {
         "startTime": "2022-12-14T18:00:00+08:00",
         "endTime": "2022-12-15T06:00:00+08:00",
```

weatherElement 裡有五種的預報因子，可透過程式單純取出所需的預報因子。

預報因子	說明
Wx	天氣現象
MaxT	最高溫度
MinT	最低溫度
CI	舒適度
PoP	降雨機率

Python 爬取天氣預報

```
(metaverse) c:\Python3\my_project\project_crawl>python weather_2.py
臺北市未來 8 小時陰天, 最高溫 19 度, 最低溫 19 度, 降雨機率 10 %
新北市未來 8 小時陰天, 最高溫 19 度, 最低溫 19 度, 降雨機率 10 %
桃園市未來 8 小時多雲, 最高溫 19 度, 最低溫 18 度, 降雨機率 10 %
臺中市未來 8 小時晴時多雲, 最高溫 20 度, 最低溫 19 度, 降雨機率 10 %
臺南市未來 8 小時多雲, 最高溫 22 度, 最低溫 21 度, 降雨機率 10 %
高雄市未來 8 小時陰天, 最高溫 24 度, 最低溫 22 度, 降雨機率 10 %
基隆市未來 8 小時陰天, 最高溫 20 度, 最低溫 20 度, 降雨機率 20 %
新竹縣未來 8 小時晴時多雲, 最高溫 20 度, 最低溫 19 度, 降雨機率 10 %
新竹市未來 8 小時晴時多雲, 最高溫 20 度, 最低溫 19 度, 降雨機率 10 %
苗栗縣未來 8 小時晴時多雲, 最高溫 19 度, 最低溫 18 度, 降雨機率 10 %
彰化縣未來 8 小時晴時多雲, 最高溫 20 度, 最低溫 19 度, 降雨機率 10 %
南投縣未來 8 小時多雲時晴, 最高溫 20 度, 最低溫 19 度, 降雨機率 10 %
雲林縣未來 8 小時多雲, 最高溫 20 度, 最低溫 20 度, 降雨機率 10 %
嘉義縣未來 8 小時多雲時晴, 最高溫 20 度, 最低溫 19 度, 降雨機率 10 %
嘉義市未來 8 小時多雲時晴, 最高溫 21 度, 最低溫 20 度, 降雨機率 10 %
屏東縣未來 8 小時陰時多雲, 最高溫 23 度, 最低溫 22 度, 降雨機率 10 %
宜蘭縣未來 8 小時陰短暫雨, 最高溫 18 度, 最低溫 18 度, 降雨機率 80 %
花蓮縣未來 8 小時陰時多雲, 最高溫 21 度, 最低溫 20 度, 降雨機率 20 %
臺東縣未來 8 小時陰時多雲, 最高溫 21 度, 最低溫 21 度, 降雨機率 10 %
澎湖縣未來 8 小時晴時多雲, 最高溫 22 度, 最低溫 21 度, 降雨機率 10 %
金門縣未來 8 小時晴時多雲, 最高溫 19 度, 最低溫 18 度, 降雨機率 10 %
連江縣未來 8 小時多雲時陰, 最高溫 17 度, 最低溫 17 度, 降雨機率 10 %
```

爬取 PTT 八卦版文章標題

- 使用 Python 的 Requests 和 Beautiful Soup 函式庫，實作一個網路爬蟲，利用傳送 cookie 的方式，突破未滿十八歲的按鈕檢查限制，取得 PTT 八卦版文章的標題，並更進一步使用 txt 儲存。

crawl_ptt.py

```
(metaverse) c:\Python3\my_project\project_crawl>python crawl_ptt.py
[新聞] 丟臉丟到國外！謝宜容案星媒關注　批「酷吏」重創台灣人權
https://www.ptt.cc//bbs/Gossiping/M.1732376731.A.52C.html

[問卦] 東京的天氣也太好了吧!!!
https://www.ptt.cc//bbs/Gossiping/M.1732376799.A.DDD.html

[問卦] 台灣違規繳罰款是日常吧
https://www.ptt.cc//bbs/Gossiping/M.1732376804.A.FA5.html

[問卦] 你們覺得BIGBANG是幾個人？
https://www.ptt.cc//bbs/Gossiping/M.1732376860.A.13A.html

[新聞] 菲律賓副總統語出驚人 稱已雇人刺殺總統
https://www.ptt.cc//bbs/Gossiping/M.1732376877.A.3BB.html

[公告] 八卦板板規(2024.07.21)
https://www.ptt.cc//bbs/Gossiping/M.1721519414.A.6A1.html

Fw: [公告] 請留意新註冊帳號使用信件詐騙
https://www.ptt.cc//bbs/Gossiping/M.1730554547.A.41C.html

[協尋] 11/8 22點多台北市汀州路三段行車紀錄
https://www.ptt.cc//bbs/Gossiping/M.1731514822.A.1F8.html

[協尋] 尋人啟事 朋友的爸爸有失智老人
https://www.ptt.cc//bbs/Gossiping/M.1731945199.A.90D.html
```