

Python與爬蟲基礎

國立臺北教育大學
數學暨資訊教育學系
蔡智孝


chtsai.ntue@mail.ntue.edu.tw

準備環境

- 安裝 [Visual Studio Code](#) 或 PyCharm


Download Visual Studio Code

Free and built on open source. Integrated Git, debugging and extensions.



↓ Windows
Windows 10, 11


User Installer	x64	Arm64
System Installer	x64	Arm64
.zip	x64	Arm64
CLI	x64	Arm64



↓ .deb
Debian, Ubuntu

↓ .rpm
Red Hat, Fedora, SUSE

.deb	x64	Arm32	Arm64
.rpm	x64	Arm32	Arm64
.tar.gz	x64	Arm32	Arm64
Snap	Snap Store		
CLI	x64	Arm32	Arm64

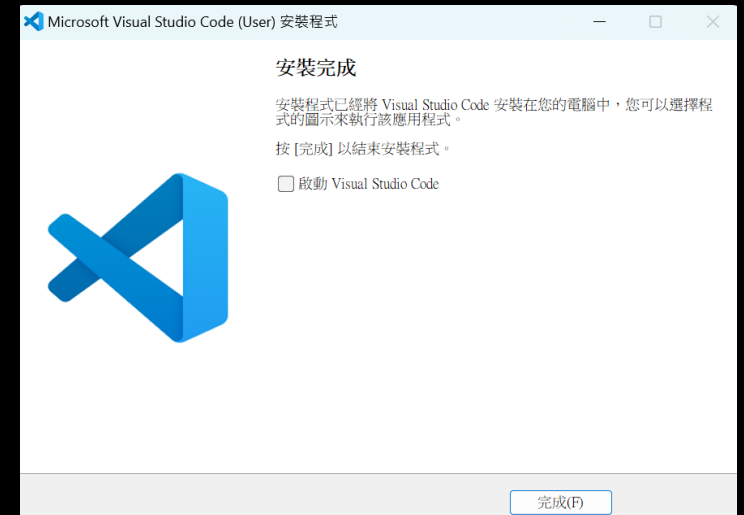


↓ Mac
macOS 10.15+

.zip	Intel chip	Apple silicon	Universal
CLI	Intel chip	Apple silicon	

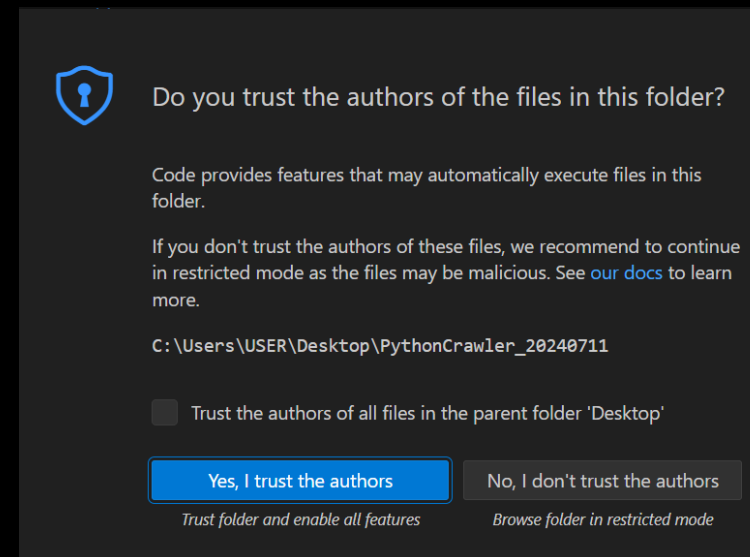
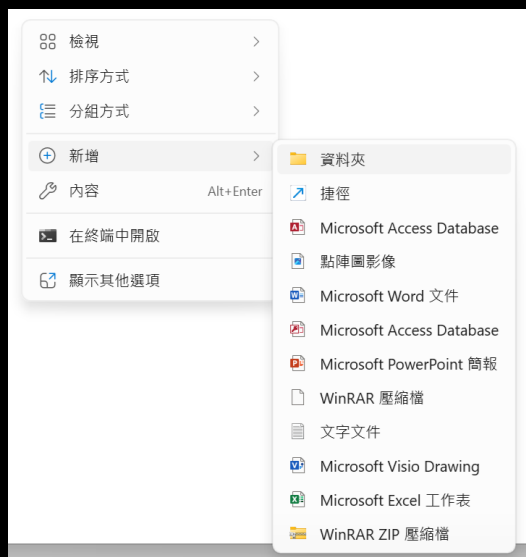
準備環境 Cont.

- 此處以安裝Windows版本為例
- 下載完成(VSCoDeUserSetup-x64-1.91.0)後，點擊兩次檔案進行安裝
- 安裝完成後先不啟動VSCode



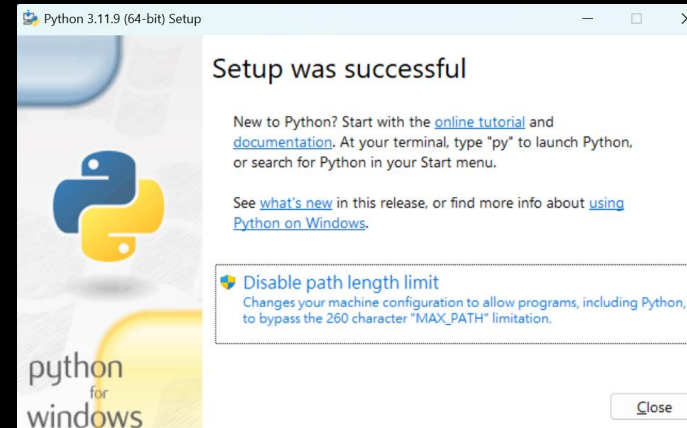
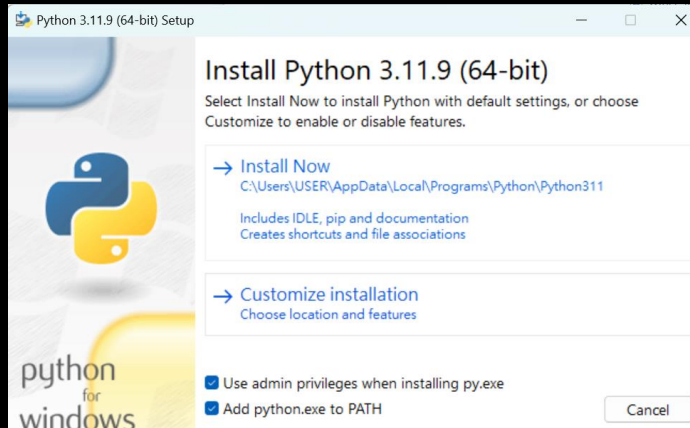
準備環境 Cont.

- 在想要儲存練習檔案的地方新增一個資料夾，例如桌面。
- 在新增好的資料夾上點擊滑鼠右鍵，選擇以VSCode開啟。
- 開啟後點選Yes, I trust the authors



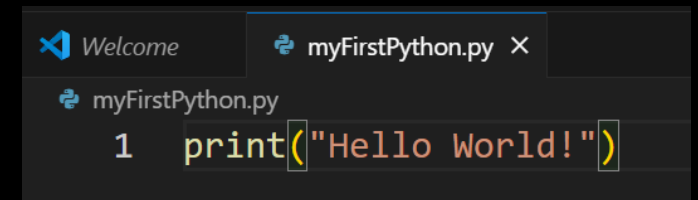
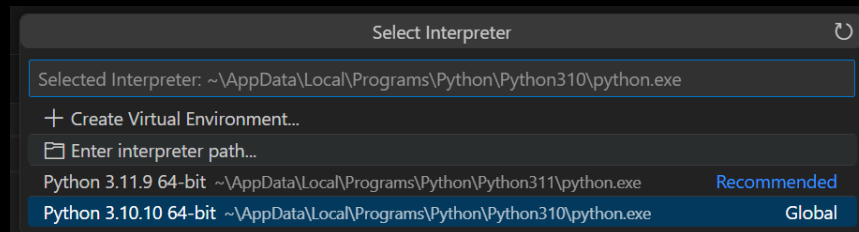
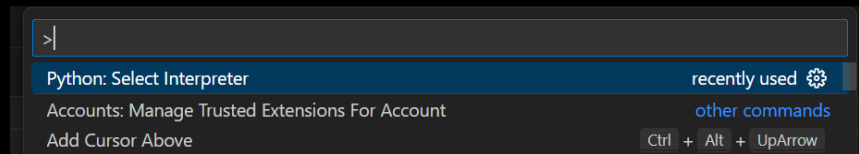
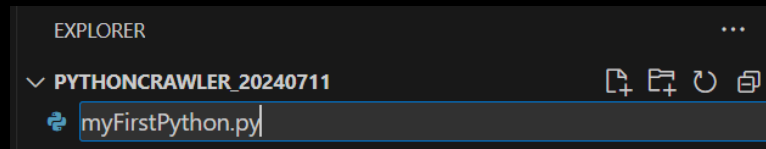
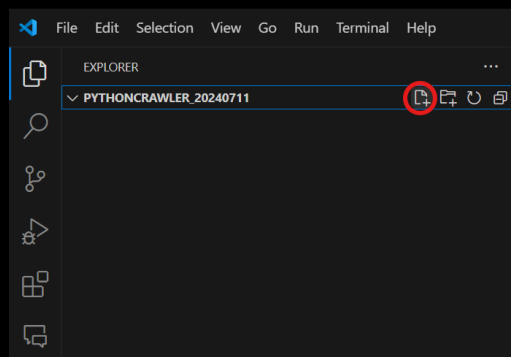
準備環境 Cont.

- 先安裝Python
- 建議不要安裝最新版，安裝3.11的版本即可
- 一樣點兩下進行安裝，記得勾選Add python.exe to PATH
- 然後選Install Now進行安裝，安裝完成後按下Close



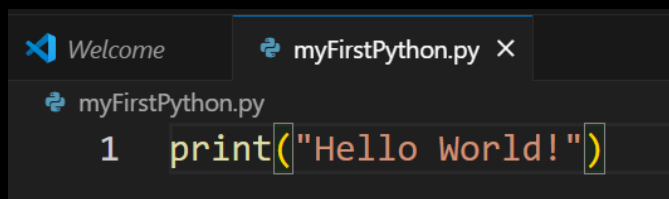
準備環境 Cont.

- 回到VSCode，左側的Explore會看到剛剛開啟的資料夾
- 按下新增檔案按鈕，新增一個.py檔
- 在程式編輯視窗按下Shift + Ctrl + p，輸入Python: Select Interpreter，選擇使用的Python直譯器版本



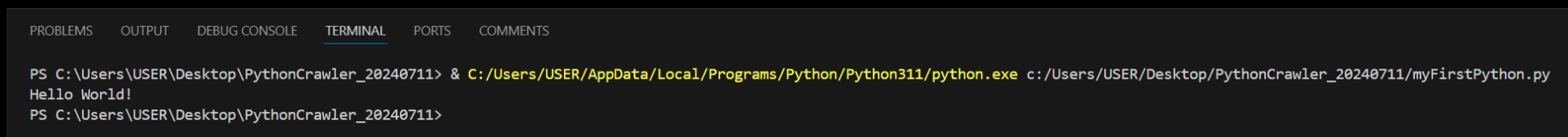
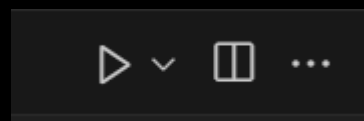
準備環境 Cont.

- 輸入程式並按下右上角的執行按鈕，確認可以正常執行



myFirstPython.py

```
1 print("Hello World!")
```



PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS COMMENTS

```
PS C:\Users\USER\Desktop\PythonCrawler_20240711> & C:/Users/USER/AppData/Local/Programs/Python/Python311/python.exe c:/Users/USER/Desktop/PythonCrawler_20240711/myFirstPython.py
Hello World!
PS C:\Users\USER\Desktop\PythonCrawler_20240711>
```

安裝套件

- 在下方的Terminal視窗輸入指令
 - `python --version`

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS  COMMENTS  
  
PS C:\Users\USER\Desktop\PythonCrawler_20240711> python --version  
Python 3.11.9  
PS C:\Users\USER\Desktop\PythonCrawler_20240711> █
```

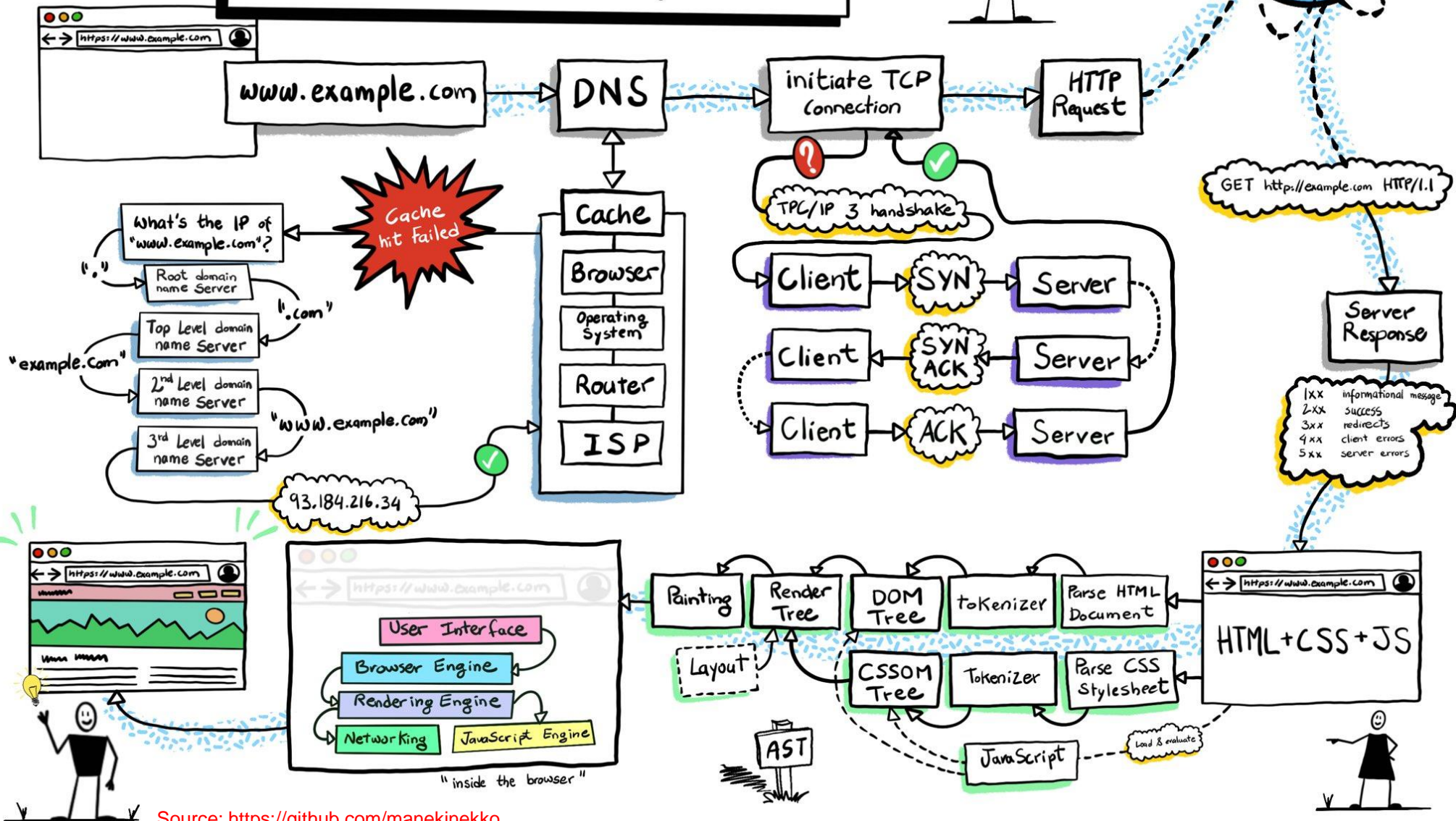
- 安裝套件指令
- `pip install -r requirements.txt`

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS  COMMENTS  
  
PS C:\Users\USER\Desktop\PythonCrawler_20240711> pip install -r .\requirements.txt █
```


A simple black and white cartoon of a stick figure with a large head and a wide smile. The figure is waving with its right hand. A speech bubble above its head contains the text "Hi!". To the right of the figure, there is a vertical watermark that reads "@manekinekko".

* a brief overview

The Internet is Complicated...



HTML: HyperText Markup Language

- 推薦網站 [w3schools](https://www.w3schools.com/)

```
test.html > html
1  <!DOCTYPE html>
2  <html>
3      <head>
4          <title>Page Title</title>
5      </head>
6      <body>
7          <h1>My First Heading</h1>
8          <p>My first paragraph.</p>
9      </body>
10 </html>
```

- [CSS Tutorial](#)
- [Python Tutorial](#)

什麼是爬蟲(web crawler)？

- 透過程式模擬網頁存取流程
 - 使用套件 `requests`
- 解析網頁超文件(HyperText)後，尋找需要的內容並擷取下來
 - 使用套件 `beautifulsoup4`
- 將擷取的內容進行整理，可以存成
 - JSON(以 `DICT` 字典的格式儲存)或 – 使用套件 `json`
 - `LIST` 格式(用來轉換為EXCEL資料) – 使用套件 `pandas`
- 根據需求進行分析等任務

範例1

範例2

範例3

範例4

範例1 - 基本操作

- 需要網頁網址

- 以ptt棒球版為例 - <https://www.ptt.cc/bbs/Baseball/index.html>



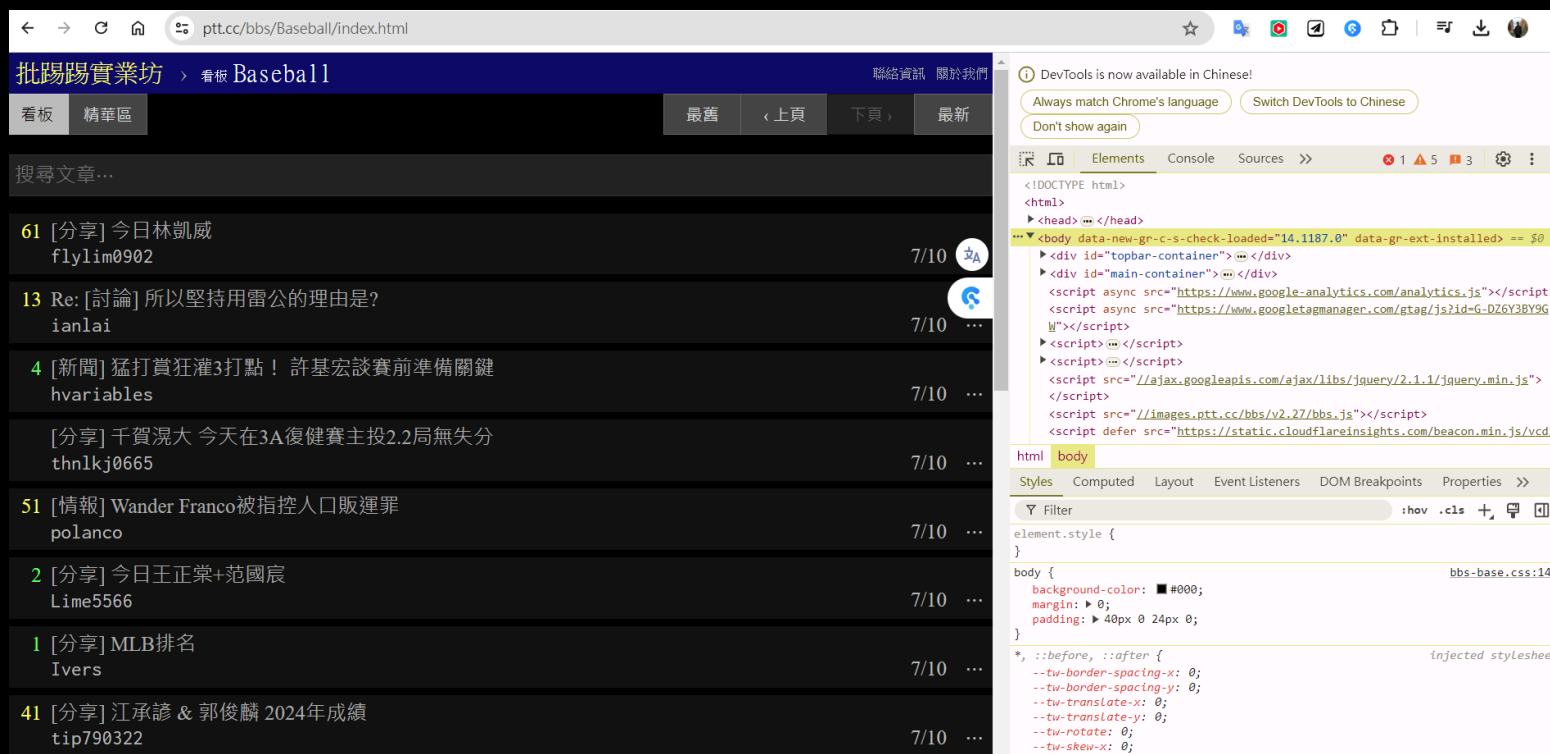
```
# 匯入 requests 套件
import requests

# 使用 requests.get 方法下載 PTT 棒球版首頁
url = 'https://www.ptt.cc/bbs/Baseball/index.html'
response = requests.get(url)
# 輸出網頁 HTML 原始碼
print(response.text)
```

ptt_crawler_01.py

範例1 Cont.

- 透過檢查開啟開發者工具 – 很重要
 - 在網頁空白處按下滑鼠右鍵，選擇最下面檢查
 - 或是按下F12功能鍵

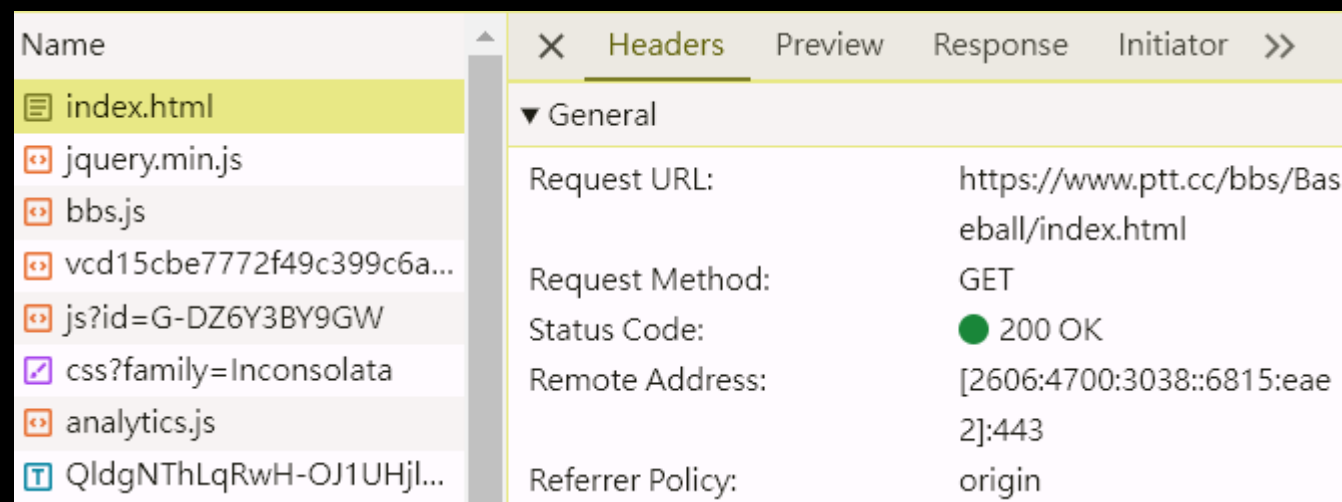
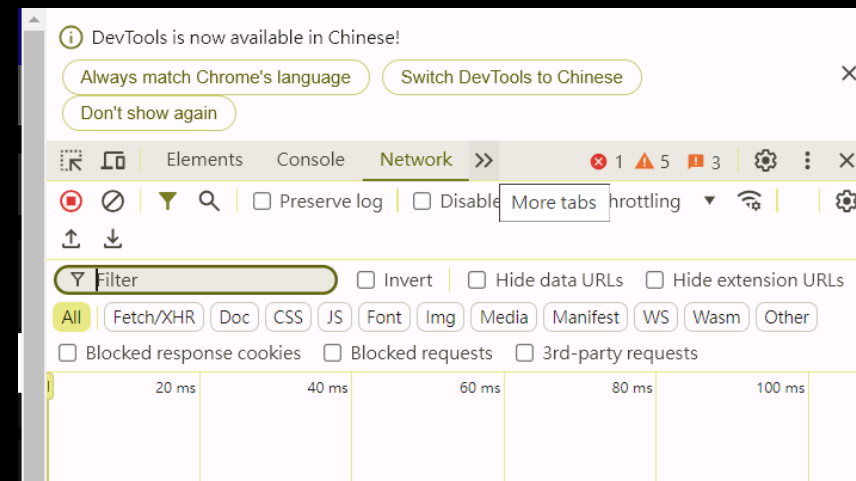


範例1 Cont.

- 為什麼抓取網頁超文件是使用`get()`方法？
- 請點選More tabs切換至Network
- 切換後按下F5更新網頁，觀察變化

Name	Status	Type	Initiator	Size	Time
index.html	200	docum...	Other	4.8 kB	585 ms
jquery.min.js	200	script	index.html:703	(memo...	0 ms
bbs.js	200	script	index.html:704	(memo...	0 ms
vcd15cbe7772f49c399c6a5b...	200	script	index.html:706	(memo...	0 ms
js?id=G-DZ6Y3BY9GW	200	script	index.html:679	(disk c...	13 ms
css?family=Inconsolata	200	stylesh...	bbs-base.css:1	(memo...	0 ms
analytics.js	200	script	index.html:691	(disk c...	7 ms
QldgNThLqRwH-OJ1UHjKE...	200	font	css	(memo...	0 ms
collect?v=1&_v=j101&a=19...	200	xhr	analytics.js:36	24 B	21 ms
collect?t=dc&aip=1&_r=3&...	200	xhr	analytics.js:36	22 B	81 ms
data:image/png;base...	200	png	content-script.js:	(memo...	0 ms

- 點選index.html選項



範例1 Cont.

- 讓程式更像是瀏覽器送出要求 – 加入User-Agent
- 一樣選取index.html，向下捲動找到Request Headers
 - 最下方即可看到User-Agent的資訊
- 修改程式內容加入headers資料

```
# 匯入 requests 套件
import requests

# 使用 requests.get 方法下載 PTT 棒球版首頁
url = 'https://www.ptt.cc/bbs/Baseball/index.html'

# 設定 User-Agent
headers = {
    'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/126.0.0.0 Safari/537.36'
}

# 帶入 headers 參數
response = requests.get(url, headers=headers)
# 輸出網頁 HTML 原始碼
print(response.text)
```

User-Agent:	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/126.0.0.0 Safari/537.36
-------------	---

ptt_crawler_02.py

範例2 – 解析超文件

- 超文件是複雜的HTML語言
- 透過beautifulsoup4套件可以快速的解析並找到所需要的內容

```
1 <!DOCTYPE html>
2 <html>
3   <head>
4     <meta charset="utf-8">
5
6
7     <meta name="viewport" content="width=device-width, initial-scale=1">
8
9     <title>看板 Baseball 文章列表 - 批踢踢實業坊</title>
10
11     <link rel="stylesheet" type="text/css" href="//images.ptt.cc/bbs/v2.27/bbs-common.css">
12     <link rel="stylesheet" type="text/css" href="//images.ptt.cc/bbs/v2.27/bbs-base.css" media="screen">
13     <link rel="stylesheet" type="text/css" href="//images.ptt.cc/bbs/v2.27/bbs-custom.css">
14     <link rel="stylesheet" type="text/css" href="//images.ptt.cc/bbs/v2.27/pushstream.css" media="screen">
15     <link rel="stylesheet" type="text/css" href="//images.ptt.cc/bbs/v2.27/bbs-print.css" media="print">
16
17
18
19
20   </head>
21   <body>
22
23     <div id="topbar-container">
24       <div id="topbar" class="bbs-content">
25         <a id="logo" href="/bbs/">批踢踢實業坊</a>
26         <span>&rsquo;</span>
27         <a class="board" href="/bbs/Baseball/index.html"><span class="board-label">看板 </span>Baseball</a>
28         <a class="right small" href="/about.html">關於我們</a>
29         <a class="right small" href="/contact.html">聯絡資訊</a>
30       </div>
31     </div>
```


範例2 – 解析超文件 Cont.

- 將伺服器回應的內容，直接傳給套件進行解析

```
# 匯入 requests 套件
import requests
# 匯入 BeautifulSoup 套件
from bs4 import BeautifulSoup

# 使用 requests.get 方法下載 PTT 棒球版首頁
url = 'https://www.ptt.cc/bbs/Baseball/index.html'

# 設定 User-Agent
headers = {
    'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Ge
}
# 帶入 headers 參數
response = requests.get(url, headers=headers)

# 使用 BeautifulSoup 解析 HTML 程式碼
soup = BeautifulSoup(response.text, 'html.parser')

# 輸出網頁 HTML 原始碼
print(soup.prettify())
```

ptt_crawler_03.py

範例2 – 解析超文件 Cont.

- 將解析後的內容存成.html檔案方便觀察

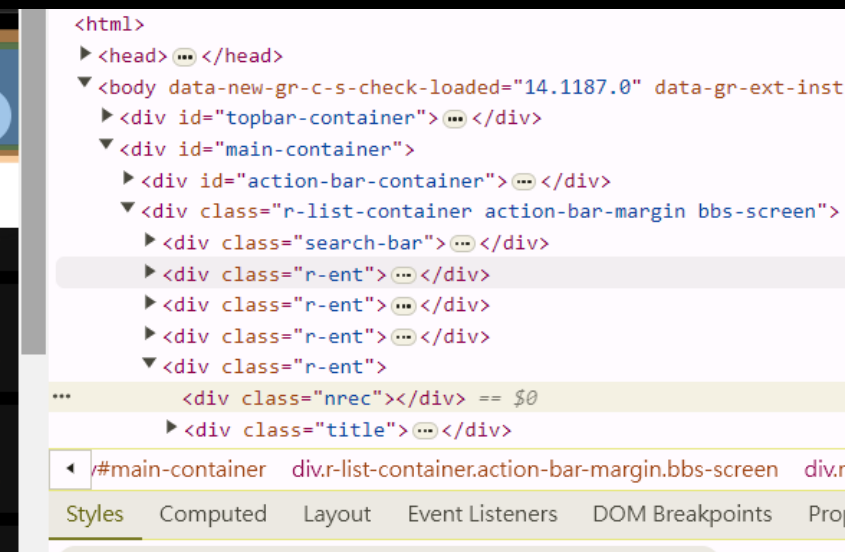
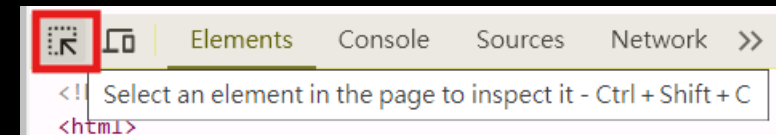
```
# 存成 HTML 檔案
with open('index2.html', 'w', encoding='utf-8') as f:
    f.write(soup.prettify())
```

- 考考大家，上面的程式碼要加在甚麼地方？比較兩種寫法差異

```
# 存成 HTML 檔案
with open('index1.html', 'w', encoding='utf-8') as f:
    f.write(response.text)
```

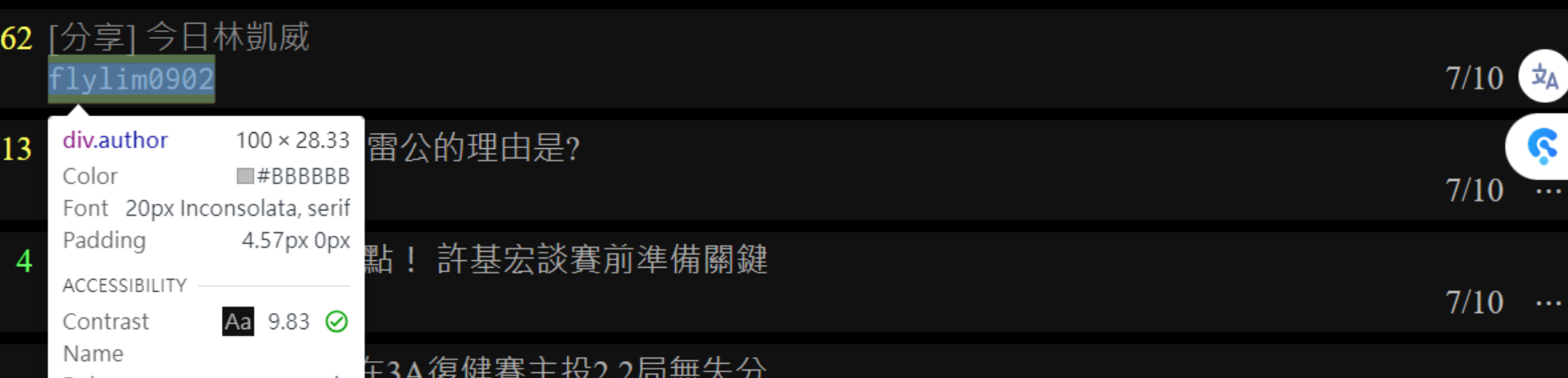
範例2 – 解析超文件 Cont.

- 爬取文章標題、作者、日期與人氣四個值
- 在開發者工具中，善用選取元素工具觀察程式
- 先看每一個發表貼文的架構
 - 每篇貼文都在 div 標籤 -> class 是 r-ent 下



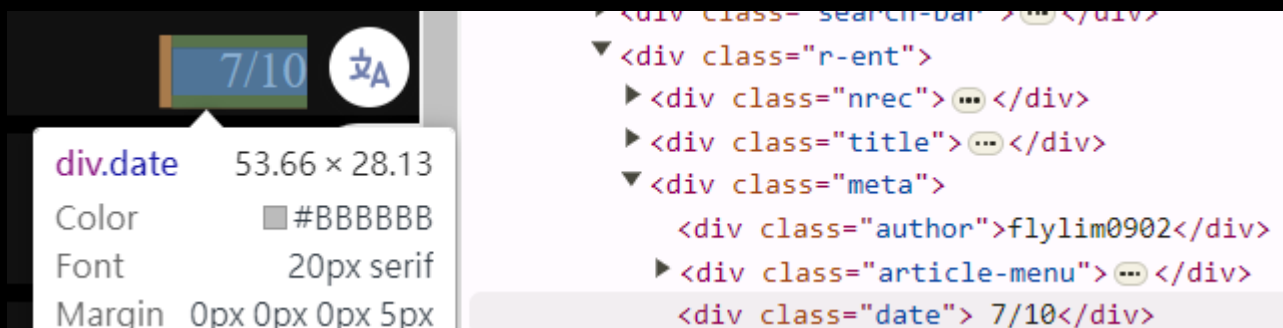
範例2 – 解析超文件 Cont.

- 作者文字在 div 標籤 -> class 是 author 下



```
<head>...</head>
<body data-new-gr-c-s-check-loaded="14.1187.0">
  <div id="topbar-container">...</div>
  <div id="main-container">
    <div id="action-bar-container">...</div>
    <div class="r-list-container action-bar-ma">
      <div class="search-bar">...</div>
      <div class="r-ent">
        <div class="nrec">...</div>
        <div class="title">...</div>
        <div class="meta">
          <div class="author">flylim0902</div>
          <div class="article-menu">...</div>
```

- 日期文字在 div 標籤 -> class 是 date 下

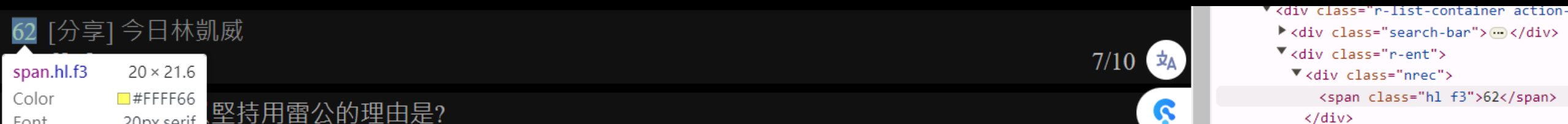


範例2 – 解析超文件 Cont.

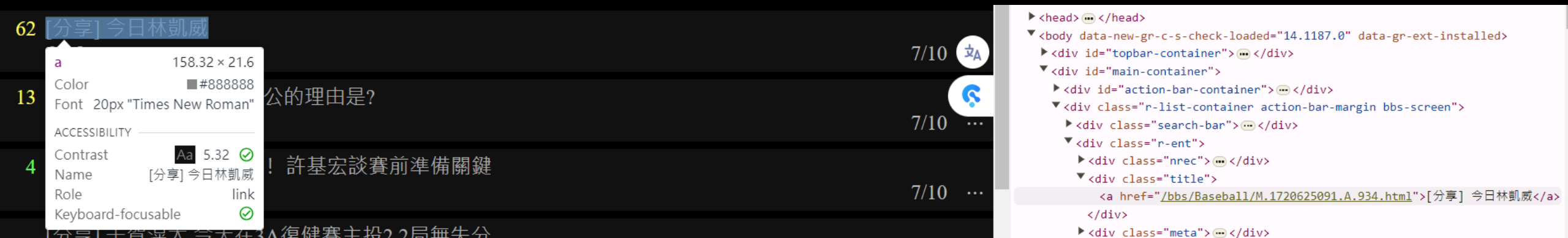
列表形式 - ptt_crawler_05.py

字典形式 - ptt_crawler_06.py

- 人氣文字在 div 標籤 -> class 是 nrec 下



- 標題文字在 div 標籤 -> class 是 title 下



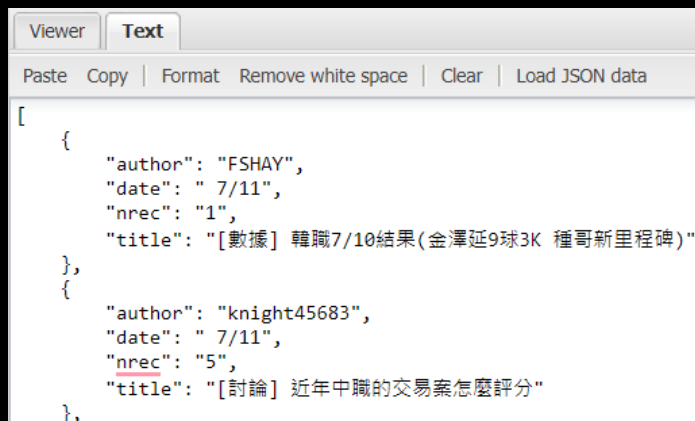
- 觀察所需資料的位置後，即可撰寫程式尋找結果

範例3 – 將結果存成JSON格式

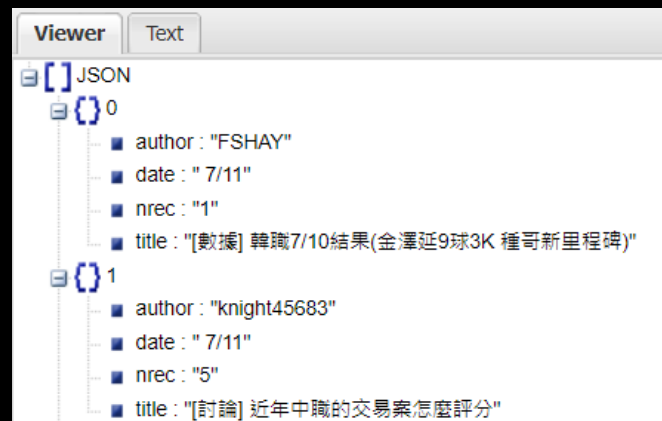
- 動手做做看，這段程式碼要加在哪裡？

```
# 存成 JSON 檔案
with open('index.json', 'w', encoding='utf-8') as f:
    json.dump(post_data, f, ensure_ascii=False, indent=4)
```

- 如果出來的JSON格式很醜不方便理解，可以將內容貼入下列網址 <https://jsonviewer.stack.hu/>



```
[
  {
    "author": "FSHAY",
    "date": " 7/11",
    "nrec": "1",
    "title": "[數據] 韓職7/10結果(金澤延9球3K 種哥新里程碑)"
  },
  {
    "author": "knight45683",
    "date": " 7/11",
    "nrec": "5",
    "title": "[討論] 近年中職的交易案怎麼評分"
  },
]
```



```
[ ] JSON
├── 0
│   ├── author: "FSHAY"
│   ├── date: " 7/11"
│   ├── nrec: "1"
│   └── title: "[數據] 韓職7/10結果(金澤延9球3K 種哥新里程碑)"
└── 1
    ├── author: "knight45683"
    ├── date: " 7/11"
    ├── nrec: "5"
    └── title: "[討論] 近年中職的交易案怎麼評分"
```

範例4 – 將結果存成EXCEL格式

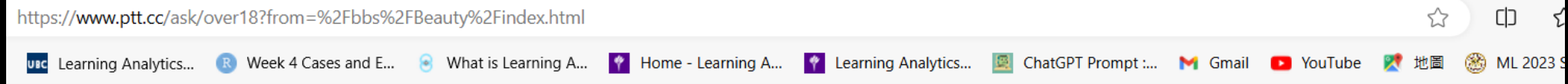
- 動手做做看，這段程式碼要加在哪裡？注意：要轉成pandas的DataFrame格式，資料要以List的格式儲存

```
# 存成 EXCEL 檔案
df = pd.DataFrame(post_data, columns=['作者', '日期', '人氣', '標題'])
df.to_excel('index.xlsx', index=False)
```

爬取單一文章內容

ptt_crawler_10.py

- 先至PTT選取有興趣的一篇文章
- <https://www.ptt.cc/bbs/Gossiping/M.1720632619.A.6B0.html>
- 若觀看討論版有年齡限制，可透過 cookies 繞過年齡檢查



本網站已依網站內容分級規定處理

警告：您即將進入之看板內容需滿十八歲方可瀏覽。

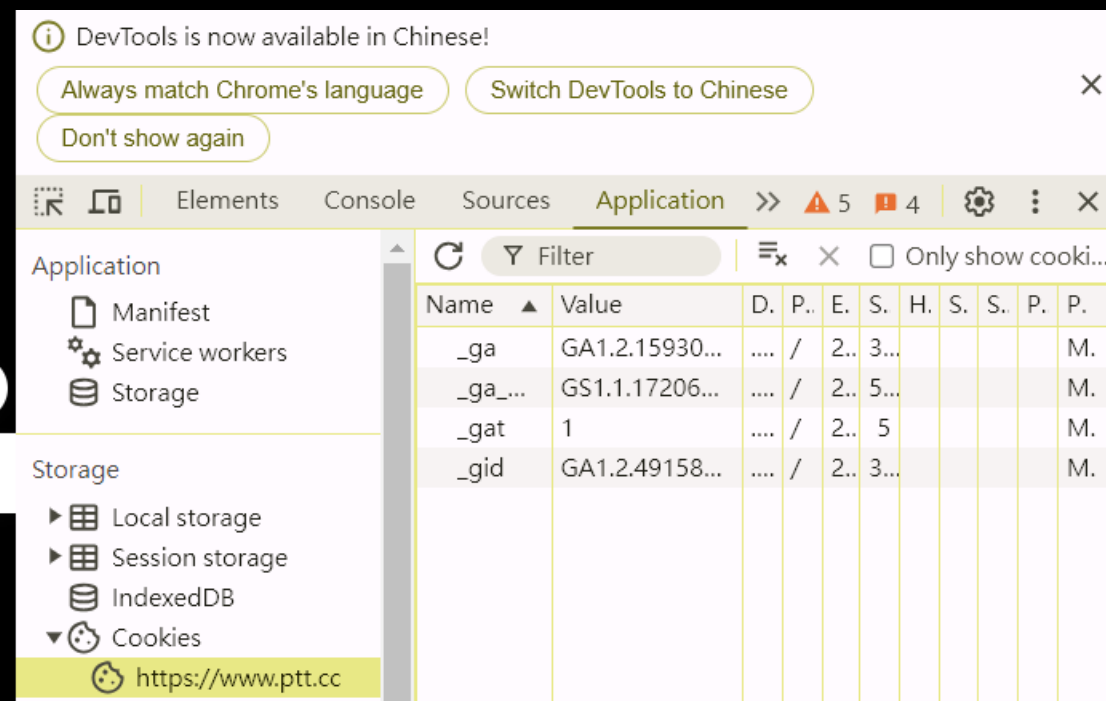
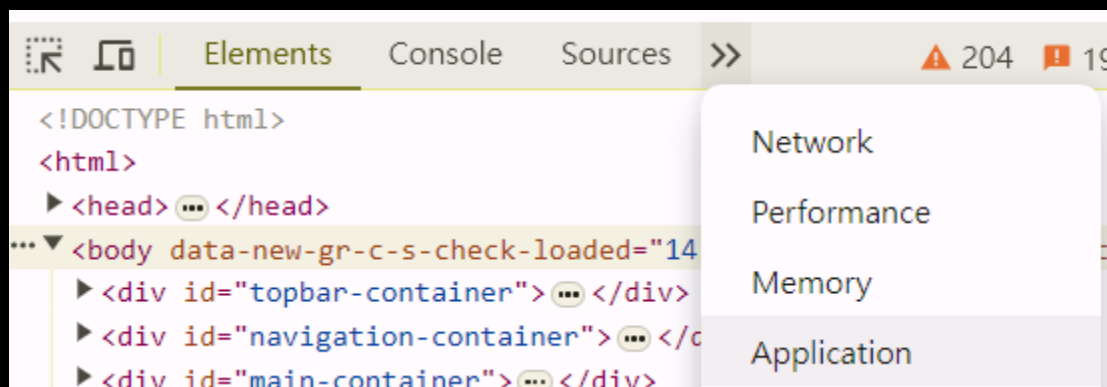
若您尚未年滿十八歲，請點選離開。若您已滿十八歲，亦不可將本區之內容派發、傳閱、出售、出租、交給或借予年齡未滿18歲的人士瀏覽，或將本網站內容向該人士出示、播放或放映。

我同意，我已年滿十八歲
進入

未滿十八歲或不同意本條款
離開

爬取單一文章內容 Cont.

- 開發工具中切換至Application
- 若觀看討論版有年齡限制，可透過 cookies 繞過年齡檢查



爬取留言

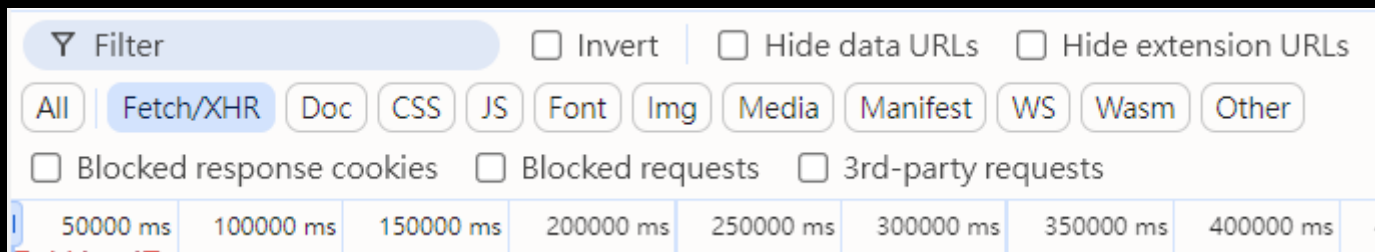
ptt_crawler_11.py

進階做法－給關鍵字，爬所有相關資料

ptt_crawler_12.py

爬取Ajax網頁

- 通常網頁可能為了保護資料，會先回應一個空白的頁面
- 確定使用者回應後，再把資料丟到使用者端呈現
 - 非同步的溝通與資料傳輸
- 所以就算獲得的status_code是200，但內容卻是空的
- 打開開發者工具，切換至Fetch/XHR，尋找可能的API內容



影音課程 999+ 直播 16 文章 999+ 組合 284

依時間 ▾

募資中課程 ▾

募資優惠



課程 無基礎也能完成質感畫：壓克力的肌理練習

by 許雲

募資倒數 20 天

17 %

NT\$ 1,390 ~~NT\$1,790~~

募資優惠

課程 TOPIK II 高分技巧：成為單字富翁的秘密

by SJ eCLASS

募資倒數 6 天



NT\$ 2,528 ~~NT\$3,200~~



DevTools is now available in Chinese!

Always match Chrome's language

Switch DevTools to Chinese

Don't show again

Elements Console Sources Network Performance >> 3 22 3

Filter Invert Hide data URLs Hide extension URLs

All Fetch/XHR Doc CSS JS Font Img Media Manifest WS Wasm Other Blocked response cookies

Blocked requests 3rd-party requests

50000 ms 100000 ms 150000 ms 200000 ms 250000 ms 300000 ms 350000 ms 400000 ms 450000 ms 500000 ms

Name Headers Payload Preview Response Initiator Timing

register-conversion?_c=1&cid=1237948...

register-trigger?partner_id=67811&uid...

register-trigger?partner_id=67811&uid...

register-trigger?partner_id=67811&uid...

salesPackages

search?anonymousId=29e62c34-3ebf-4...

search?anonymousId=29e62c34-3ebf-4...

search?anonymousId=29e62c34-3ebf-4...

search?category=COURSE&filter=INCUB...

search?category=COURSE&limit=8&pa...

search?filter=INCUBATING&limit=0&pa...

search?filter=INCUBATING&limit=0&pa...

search?limit=0&page=0

search?limit=0&page=0

spreadsheet?sheet=%2Fwording

status

status

status

```
{
  "_id": "603f1627b35e5faa941fde70",
  "name": "許雲",
  "username": "hsuyun0studio",
  "profileImageUrl": "https://images.hahow.com/...",
  "tags": [],
  "preOrderedPrice": 1390,
  "preOrderedPriceInMoneyPoint": 1390,
  "price": 1790,
  "priceInMoneyPoint": 1790,
  "successCriteria": {
    "numSoldTickets": 30
  },
  "numSoldTickets": 5,
  "reviewing": false,
  "isReject": false,
  "averageRating": 0,
  "numRating": 0,
  "bookmarkCount": 24,
  "campaign": {
    "types": []
  }
}
```


爬取Ajax網頁 Cont.

- 切換回Headers找出API
- https://api.hahow.in/api/products/search?category=COURSE&filter=INCUBATING&limit=24&page=0&sort=INCUBATE_TIME