

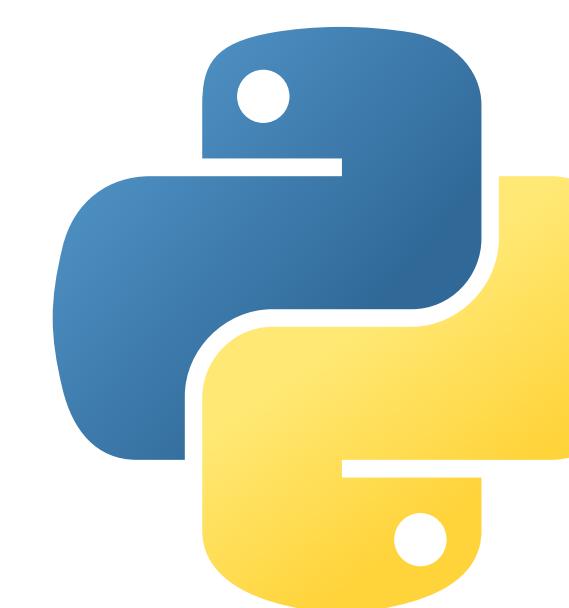
Day 09

HTTP 網頁架構 - 靜態網頁

靜態網頁的資料爬蟲策略



出題教練：楊鎮銘



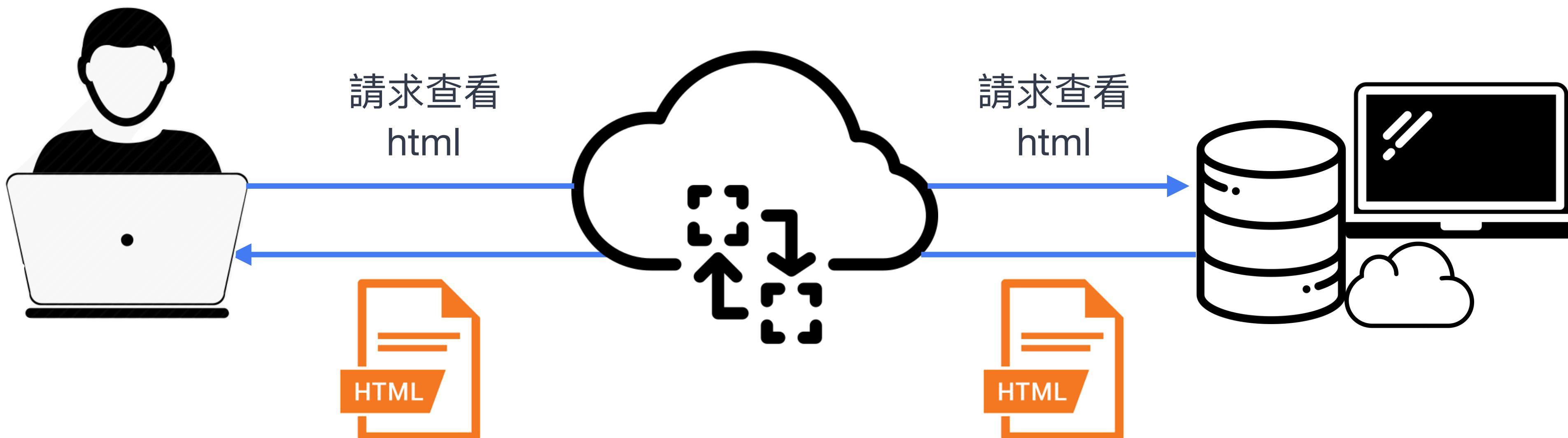
python

本日知識點目標

- 了解爬取圖片的流程
- 了解圖片延伸的相關問題
 - 副檔名的正確性
 - 如何透過程式判斷副檔名

圖片爬蟲流程 - 取得網頁

與一般爬蟲目標是文字的過程，圖片爬蟲其實只是要多送一次請求



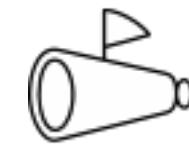
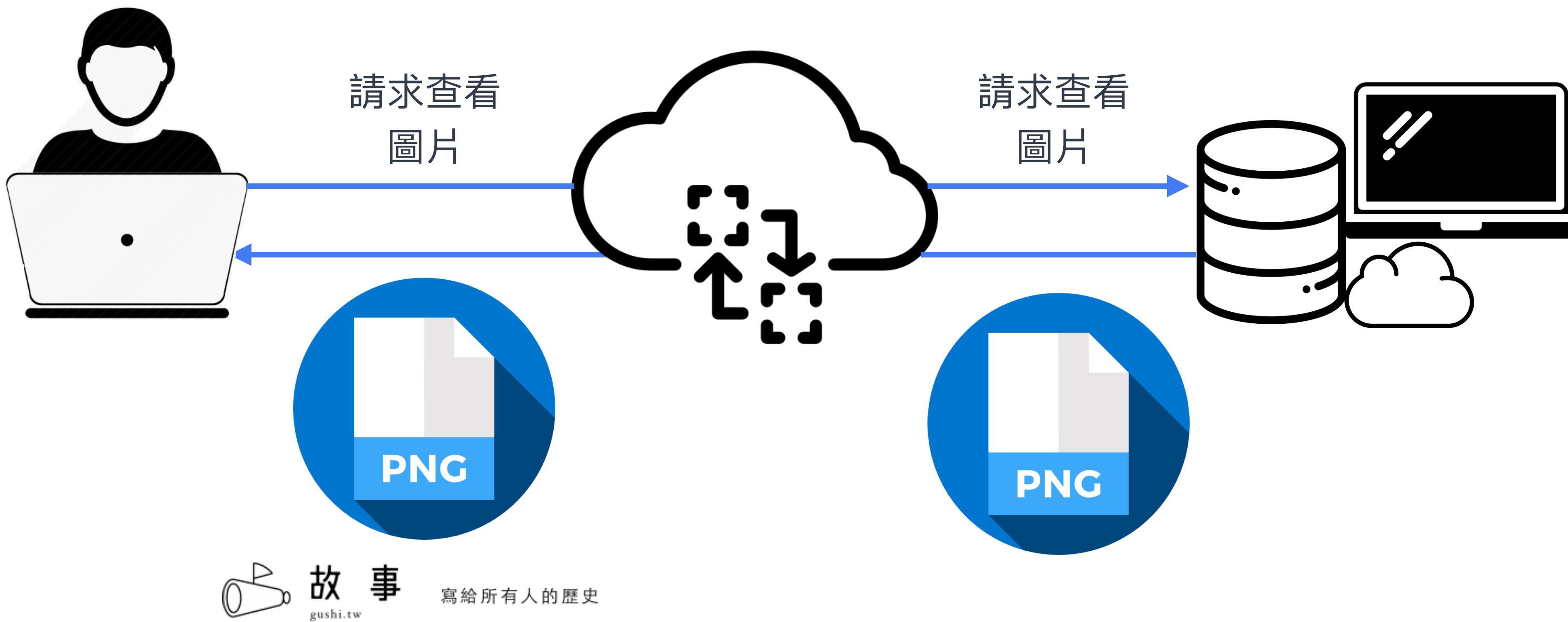
圖片爬蟲流程 - 從網頁檔中取得圖片網址

與一般爬蟲目標是文字的過程，圖片爬蟲其實只是要多送一次請求



圖片爬蟲流程 - 取得圖片檔

與一般爬蟲目標是文字的過程，圖片爬蟲其實只是要多送一次請求



圖片爬蟲流程 - 實作

簡單版本 - urllib 套件 [[官網](#)]

`urlretrieve(URL, FILENAME)`

沒有額外設定的話有時候會有例外發生，例如官網裏面有提到
假如目標檔案的 size 太小會有 `ContentTooShortError`

Note: 官網建議使用 requests 套件來取代

See also: The [Requests package](#) is recommended for a higher-level HTTP client interface.

圖片爬蟲流程 - 實作 (optional)

複雜版本 - requests 套件

```
def download_file(url):
    local_filename = url.split('/')[-1]
    with requests.get(url, stream=True) as r:
        r.raise_for_status()
        # receive 8192 bytes per chunk
        with open(local_filename, 'wb') as f:
            for chunk in
r.iter_content(chunk_size=8192):
                if chunk:
                    f.write(chunk)
```

設定為 stream 之後會跟網站建立通道

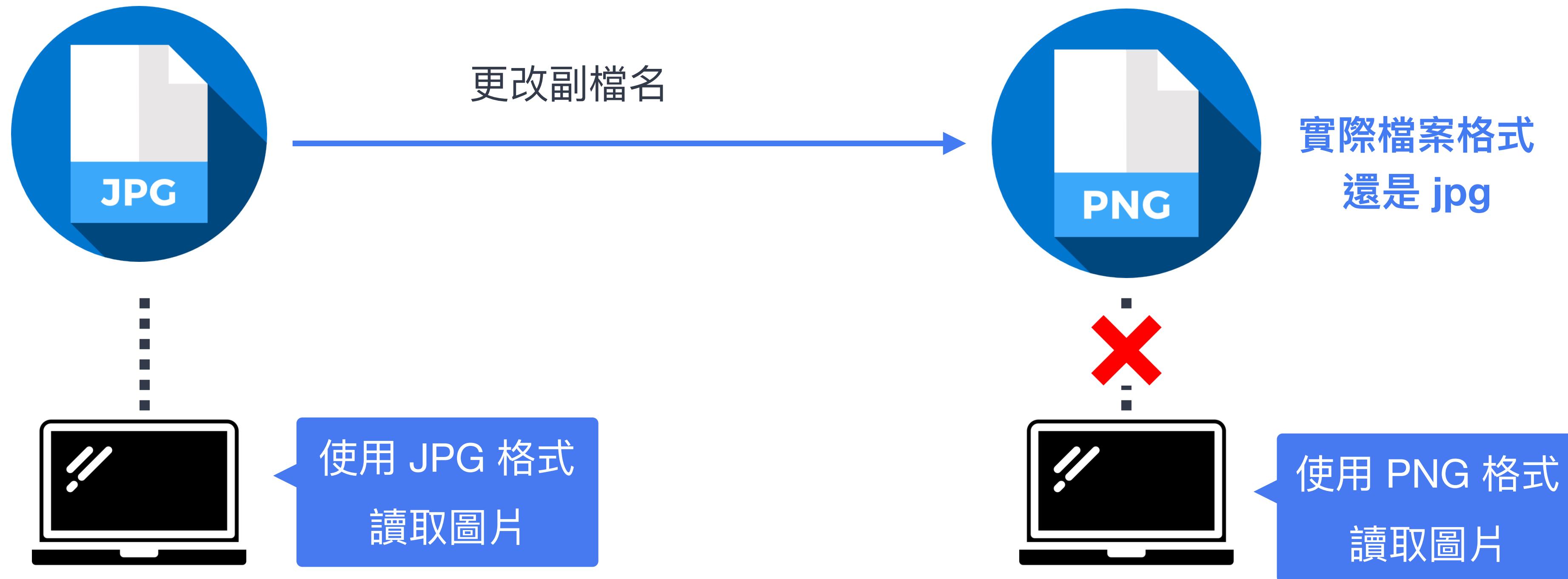
後面再透過 loop 把檔案拆成很多 chunk 下載

→主要下載圖片的程式碼

reference: [stackoverflow / download-large-file-in-python-with-requests](https://stackoverflow.com/questions/1669492/download-large-file-in-python-with-requests)

圖片副檔名會造成的問題

- 1.副檔名是讓電腦決定要用甚麼軟體開啟檔案的提示
- 2.更改副檔名不等於轉檔
- 3.副檔名錯誤就無法正確開啟檔案



網路上顯示的圖片格式不一定是正確的

除了原本就歸網站所有的圖片以外，其他通常會放在第三方服務，再藉由連結顯示在自己的網站



i.imgur.com/Cgb5oo1.jpg
是圖片

imgur.com/Cgb5oo1
是網站

ref: <M.1556291059.A.75A.html>

網路上顯示的圖片格式不一定是正確的

<http://i.imgur.com/Cgb5oo1.jpg>

<http://i.imgur.com/Cgb5oo1.png>

<http://i.imgur.com/Cgb5oo1.gif>

在這個第三方服務用不同副檔名
都可以檢視到同樣的圖片



下載圖片並以正確副檔名儲存

為了要用正確的副檔名存檔，我們必須下載下來之後先判斷圖片格式這邊可以藉由 `PIL.Image` 來判斷格式

```
from PIL import Image
resp = requests.get(image_url, stream=True)
image = Image.open(resp.raw)
print(image.format) # e.g. JPEG
# 假設我們重新組合圖片檔名與副檔名 logo.jpeg 之後
# 可以用 requests 的方式也可以用 PIL 儲存圖片
image.save('logo.jpeg')
```

重要知識點複習

- 爬取圖片只是要針對圖片網址多送一次請求
- 了解圖片延伸的相關問題
 - 副檔名對於在本地端讀取檔案會有影響
 - 建議透過 requests 套件下載
 - 可以透過 PIL.Image 檢查圖片格式



參考資料



- [floyernick/fleep-py](#)
 - 另外一個可以判斷圖片格式的套件，跟 PIL 一樣可以搭配 requests 使用，如果想知道更詳細的實作細節可以參考 [Medium](#) 文章

解題時間

LET'S CRACK IT

請跳出 PDF 至官網 Sample Code & 作業

開始解題

