國立清華大學
NATIONAL TSING HUA UNIVERSITY

AIMS - deep learning
網路爬蟲的介紹與實作

TA course 02

TAs: 蘇時頤, 廖柄淦

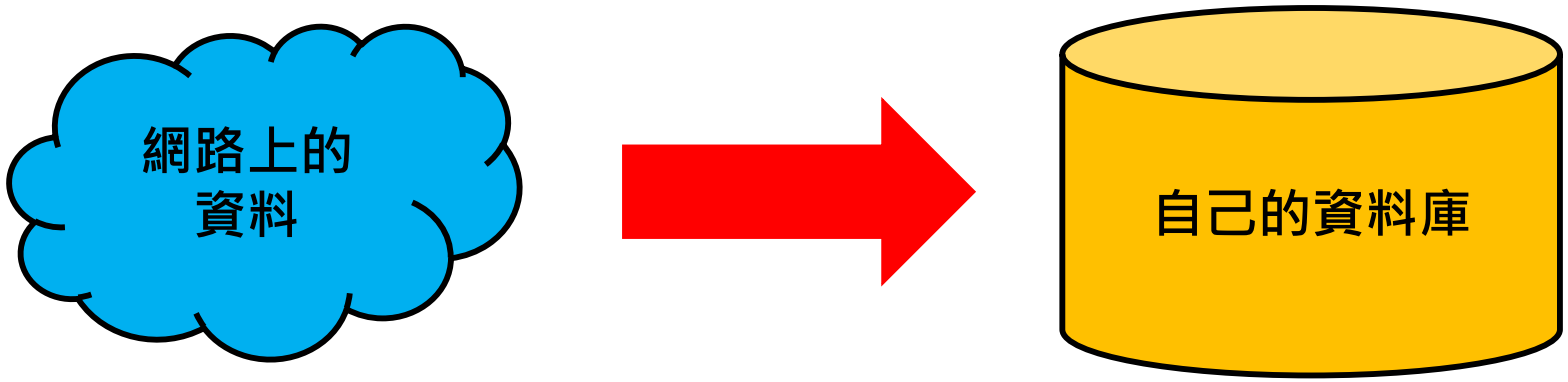
- Course website:
  https://github.com/matteosoo/aimsfellows_DL

Catalog

- Python 網路爬蟲
  - 什麼是網路爬蟲
  - 網路爬蟲的應用
  - 瀏覽器送出的請求
  - Method Get與Post
  - 爬蟲前的注意事項
  - 撰寫網路爬蟲的步驟
  - HTTP狀態碼
  - 標籤與CSS選擇器
  - 爬蟲實作-PTT表特版當日圖片下載

什麼是網路爬蟲

- 是一種利用http request抓取網路資料的技術

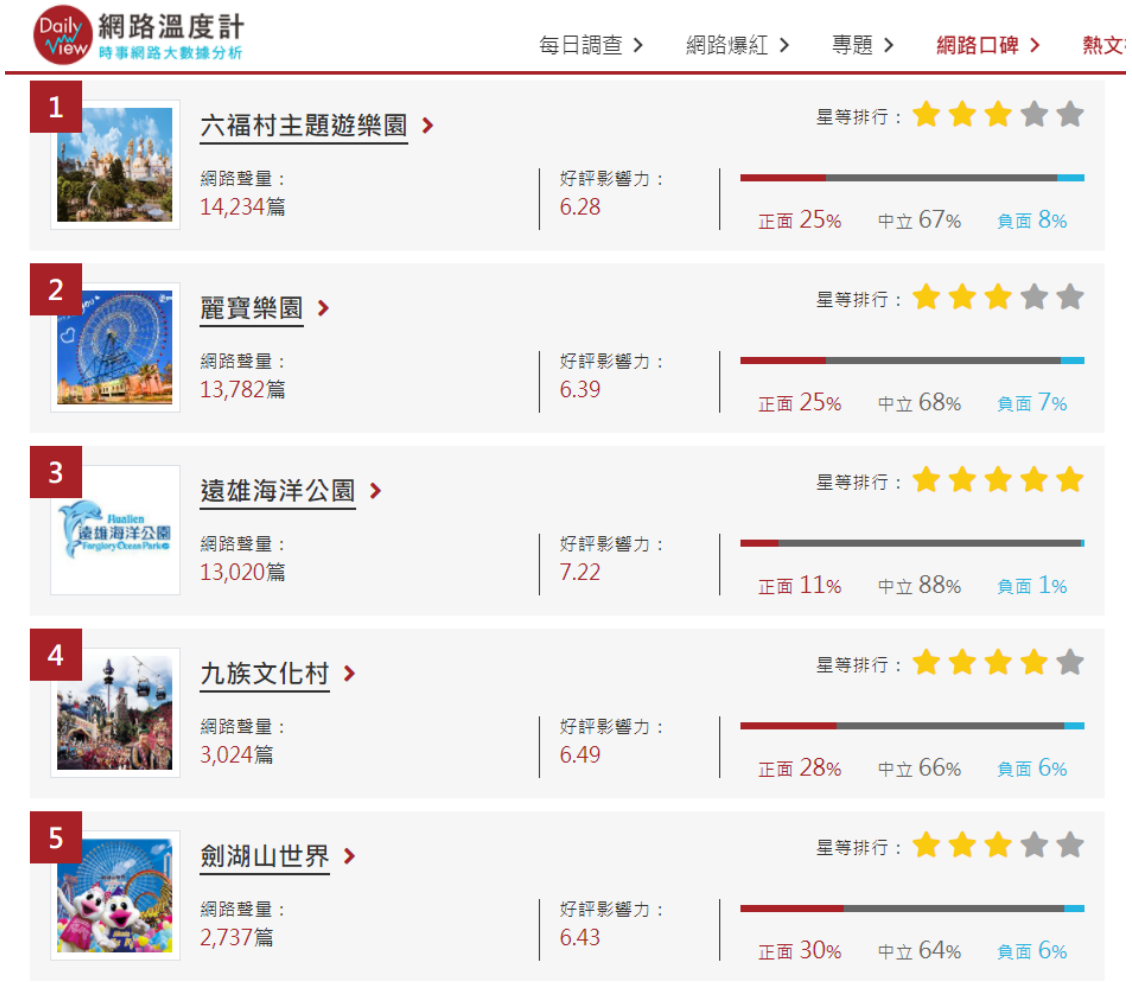


網路爬蟲的應用

- 網路輿情分析
  - 分析網路上消費者的意見
  - 了解目標客戶對於品牌、產品或是服務的想法

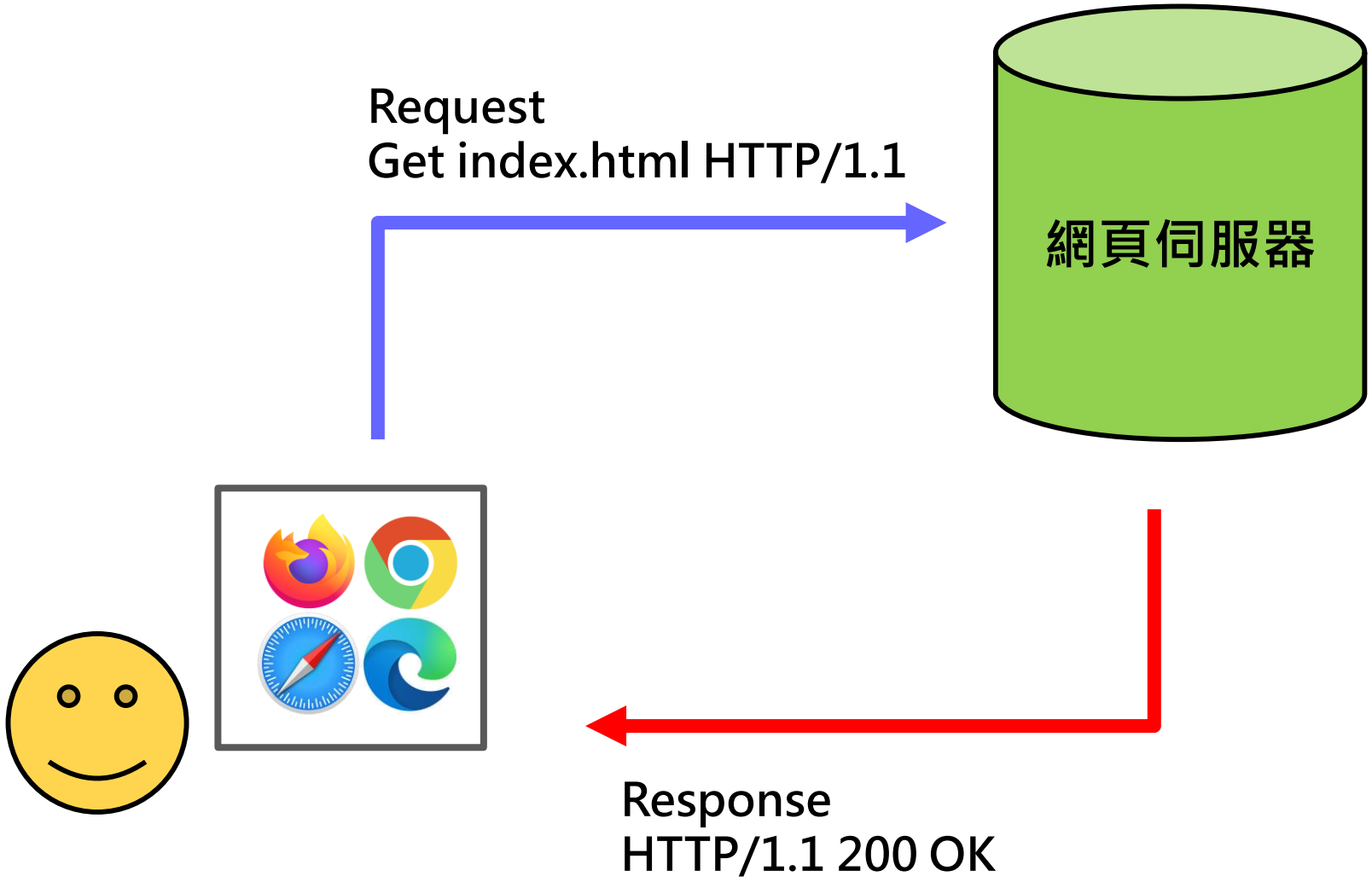
網路爬蟲的應用

- 網路輿情分析
  - 針對某一題進行群眾意見的分析



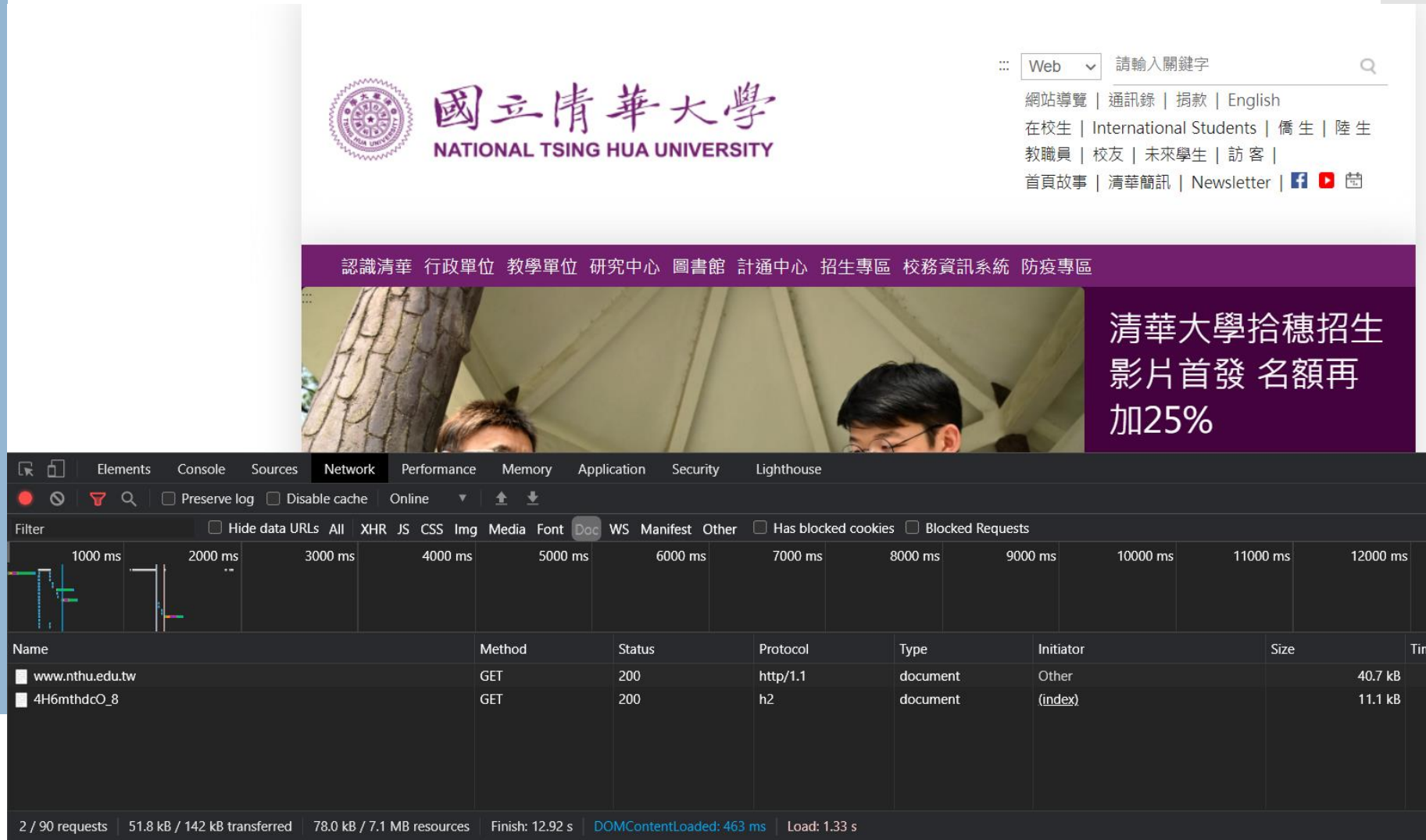
擷取自主題樂園 排行榜 | 2020主題樂園推薦 |旅遊 | 網路口碑 排行榜 | 旅遊 | 網路口碑 | 第1頁 | 網路聲量 | DailyView 網路溫度計

平常是如何取得網頁的內容



瀏覽器送出的請求

- 模擬瀏覽器送出請求的方法
- 但在這之前要先了解瀏覽器送了什麼請求



Method
Get 與 Post

- Get像是明信片
  - 網頁傳遞的參數可以透過網址發現

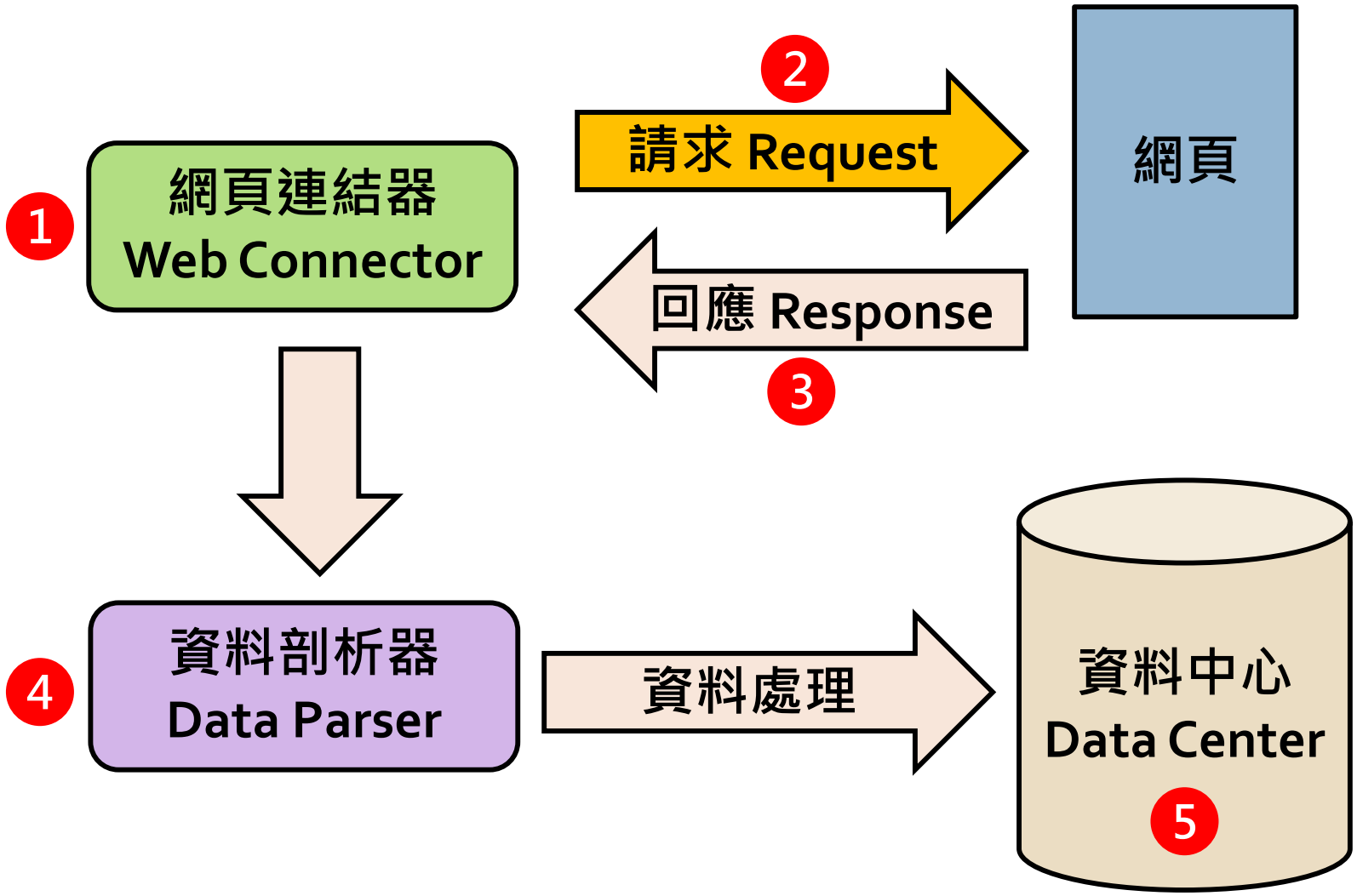
原始網址：
https://www.ptt.cc/bbs/C_Chat/index.html

經過查詢：
https://www.ptt.cc/bbs/C_Chat/search?q=%E7%B6%B4%E6%AD%8C

- Post像是信封
  - 網頁傳遞的參數看不到

- 本次實作範例要爬取的網站是PTT表特版
  - Method屬於Get
  - 只介紹Python抓取Method為Get的網站

使用Python撰寫網路爬蟲的步驟



為什麼要用Python實作網路爬蟲

- 有各種爬蟲框架
  - Scrapy
  - urllib2
  - BeautifulSoup
  - lxml
- 多線程、進程模型成熟穩定
  - 爬蟲是典型的多任務處理 → 需要等待
  - 多線程、進程可以優化效率
- 抓取下來的資料處理
  - 分詞處理
    - NLTK
    - Jieba
    - CKIP

在爬蟲前的注意事項

- 要檢視目標網站的規範「robots.txt」
- Robots協議
  - Robots Exclusion Protocol
  - 爬蟲協議、機器人協議
  - 網站通過此協議告訴搜尋引擎只有哪些頁面可以抓取
  - 這只是一種互相尊重的協議

使用Python實作網路爬蟲 - PTT表特版圖片下載

爬蟲的前置準備

```
[ ]  1 #匯入常用的套件
     2 import requests
     3 import os
     4 import re
     5 import time
     6 import urllib.request as ur
     7 from bs4 import BeautifulSoup
```

```
[ ]  1 #定義全域變數
     2 file_path = '/content/gdrive/My Drive/TA/1091/crawler/'
     3 ptt_url_head = 'https://www.ptt.cc'
```

取得網頁

定義爬蟲所需要用到的函式

```
[ ]
#取得網頁
def get_web_page(url):
  #取得網頁的回應
  response = requests.get(url=url, cookies={'over18':'1'})

  #如果回應的狀態碼是200以外的
  #代表網頁有問題
  if response.status_code != 200:
    print('Invalid url' + url)
    return None
  else:
    #如果回應的狀態碼是200代表網頁正常
    #回傳網頁的所有文字。
    return response.text
```

關於HTTP Status Code

- HTTP狀態碼
  - 用來表示網頁伺服器HTTP回應狀態的3位數代碼
  - 常見的代碼與其意義如下：
    - 200 OK
    - 403 Forbidden
    - 404 Not Found
    - 502 Bad Gateway

- 其他的狀態碼可以參考以下網站：

  HTTP狀態碼 - 維基百科，自由的百科全書

response.txt

```
<!DOCTYPE html>
<html>
  <head>
    <meta charset="utf-8">


<meta name="viewport" content="width=device-width, initial-scale=1">

<title>看板 Beauty 文章列表 - 批踢踢實業坊</title>

<link rel="stylesheet" type="text/css" href="//images.ptt.cc/bbs/v2.27/bbs-common.css">
<link rel="stylesheet" type="text/css" href="//images.ptt.cc/bbs/v2.27/bbs-base.css" media="screen">
<link rel="stylesheet" type="text/css" href="//images.ptt.cc/bbs/v2.27/bbs-custom.css">
<link rel="stylesheet" type="text/css" href="//images.ptt.cc/bbs/v2.27/pushstream.css" media="screen">
<link rel="stylesheet" type="text/css" href="//images.ptt.cc/bbs/v2.27/bbs-print.css" media="print">




  </head>
  <body>

<div id="topbar-container">
  <div id="topbar" class="bbs-content">
    <a id="logo" href="/bbs/">批踢踢實業坊</a>
    <span>&rsaquo;</span>
    <a class="board" href="/bbs/Beauty/index.html"><span class="board-label">看板 </span>Beauty</a>
    <a class="right small" href="/about.html">關於我們</a>
    <a class="right small" href="/contact.html">聯絡資訊</a>
  </div>
</div>

<div id="main-container">
  <div id="action-bar-container">
    <div class="action-bar">
      <div class="btn-group btn-group-dir">
        <a class="btn selected" href="/bbs/Beauty/index.html">看板</a>
```

要如何取得需要的資料

- 網頁是由標籤 (Tag) 所組成的階層式文件
- 主要由3個部分構成：
  - HTML (骨架)
  - CSS (階層樣式表，樣式)
  - JavaScript(負責與使用者互動的程式功能)
- 可以透過標籤與相關屬性去定位資料的位置

階層式文件

我是變色且置中的抬頭

我是段落一

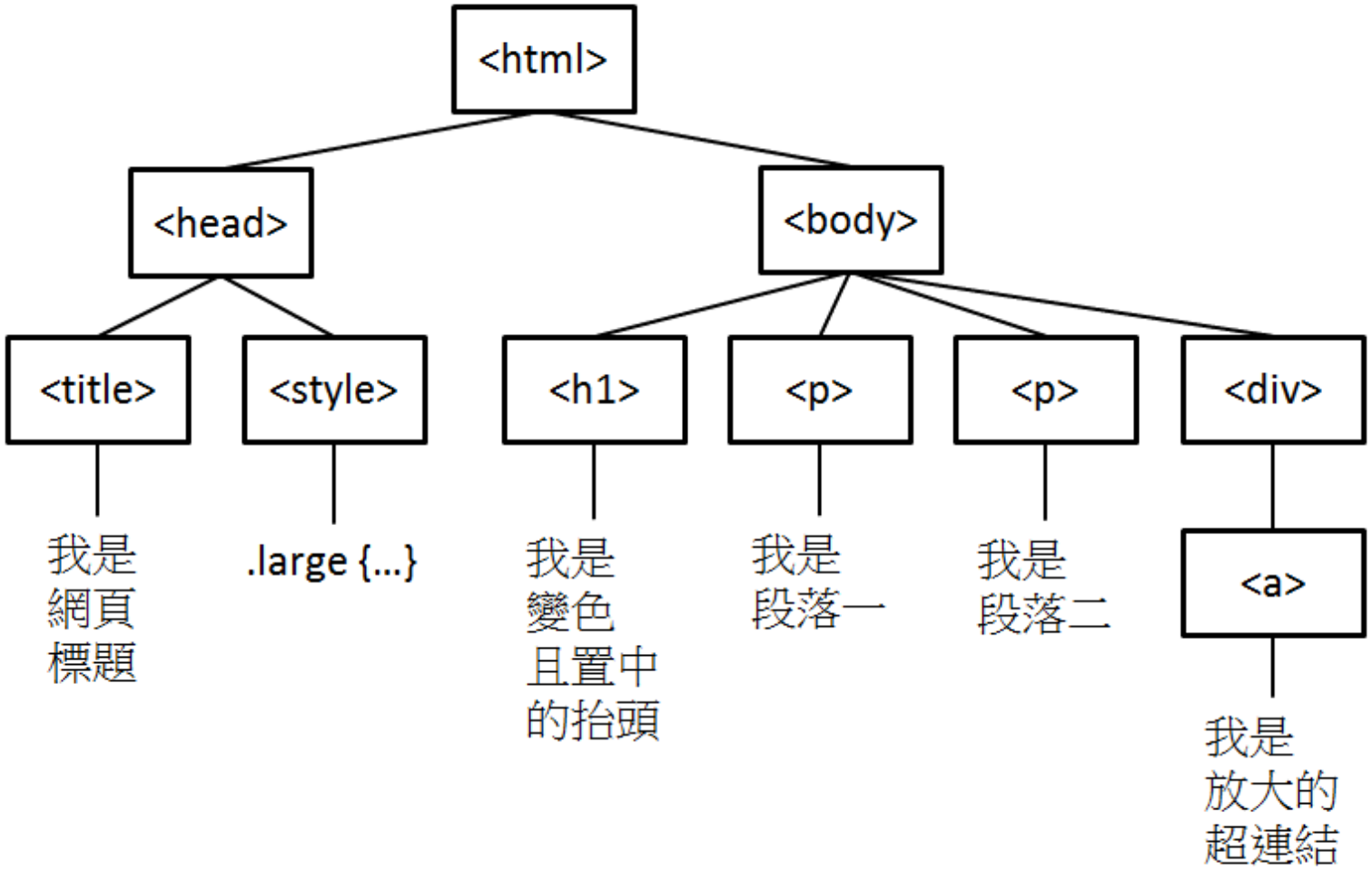
我是段落二

我是放大的超連結

test.html

```
1   <html>
2     <head>
3       <title>我是網頁標題</title>
4       <style>
5       .large {
6         color:blue;
7         text-align: center;
8       }
9       </style>
10    </head>
11    <body>
12      <h1 class="large">我是變色且置中的抬頭</h1>
13      <p id="p1">我是段落一</p>
14      <p id="p2" style="">我是段落二</p>
15      <div><a href='https://www.cycu.edu.tw/' style="font-size:200%;">我是放大的超連結</a></div>
16    </body>
17  </html>
```

網頁架構


<html>
<head>
<body>
<title>
<style>
<h1>
<p>
<p>

我是
網頁
標題
.large {...}
我是
變色
且置中
的抬頭
我是
段落一
我是
段落二
<a>
我是
放大的
超連結

CSS Selector

- CSS
  - 階層式樣式表，Cascading Style Sheets。
  - 用來位網頁添加樣式的電腦語言

- CSS選擇器
  - 可以根據class或是id，將符合的標籤套上樣式。
  - class以「.」作為開頭
  - id以「#」作為開頭

選擇器	範例	意思
.class	.large	選擇所有class=“.large”的標籤
#id	#name	選擇id=“name”的標籤

BeautifulSoup

- 使用BeautifulSoup進行解析
- 解析完畢後可以使用以下的函式去定位資料區塊：
  - find()
  - find_all()
  - select()

```
#定義最一開始的網頁超連結
ptt_beauty_url = 'https://www.ptt.cc/bbs/Beauty/index.html'
page = get_web_page(ptt_beauty_url)

#解析
soup = BeautifulSoup(page, 'html.parser')
#使用 find()尋找 body 標籤的文字
print(soup.find('body').text)
```

```
批踢踢實業坊
›
看板 Beauty
關於我們
聯絡資訊
```

取得想要的資料

```
[ ]
#取得網頁的內容
def get_articles(web_page, date):
  #解析網頁
  soup = BeautifulSoup(web_page, 'html.parser')

  #取得上一頁按鈕的超連結
  a_tags = soup.find_all('a', 'btn wide')
  previous_page_url = a_tags[1]['href']

  #用來儲存取得的文章資料
  articles = []

  divs = soup.find_all('div', 'r-ent')
  for d in divs:
    #符合條件而且超連結存在，代表文章沒有被刪除。
    if d.find('div', 'date').text == date and d.find('a'):

      #取得文章的超連結
      href = d.find('a')['href']

      #取得文章的網頁
      content = get_web_page(ptt_url_head + href)
      content_soup = BeautifulSoup(content, 'html.parser')
      content_divs = content_soup.find_all('div', id='main-container')

      #用來儲存文章標題
      title = ''

      #用來儲存要下載的圖片超連結清單
      image_url_list = []
```

取得想要的資料
-
接續上頁

```
31
32      #開始進行取得
33      for cd in content_divs:
34        #取得文章標題
35        content_spans = cd.find_all('span', 'article-meta-value')
36        #去除取得的文章標題的多餘空白
37        title = content_spans[2].text.strip()
38
39        #取得要下載的圖片超連結清單
40        content_a_hrefs = cd.find_all('a',{'href':re.compile('https:\/\/(imgur|i\.imgur)\.com\/.*')})
41        for cah in content_a_hrefs:
42          #cah的長度不是0代表有取得圖片的超連結
43          if len(cah) != 0:
44            #這裡必須處理網址最後面沒有.jpg的情況
45            if cah.text.find('.jpg') == -1:
46              image_url_list.append(cah.text+'.jpg')
47            else:
48              image_url_list.append(cah.text)
49
50      #儲存取得的文章標題與下載圖片的超連結清單
51      articles.append({'title':title, 'image_url_list':image_url_list})
52
53  #回傳文章資料與上一頁的超連結
54  return articles, previous_page_url
```

儲存圖片

```
[ ]
#儲存圖片
def save_images(folder_name, image_url_list):
  #將資料夾路徑與名稱串接起來
  folder_path = os.path.join(file_path,folder_name)

  #創建資料夾
  os.makedirs(folder_path)

  print('-----------------------------------------------')
  print('Folder name:', folder_name)

  #使用例外處理不可預知的例外發生
  try:
    for image_url in image_url_list:
      #取得圖片超連結的最後一段做為儲存圖片的檔名
      #例如kes83e.jpg
      file_name = image_url.split('/')[-1]

      #透過超連結下載圖片
      ur.urlretrieve(image_url, os.path.join(folder_path, file_name))
      print('{} saved...'.format(file_name))
  except Exception as e:
    #發生例外時印出例外
    print(e)
```

開始執行爬蟲

開始執行爬蟲

```
[ ]
#定義最一開始的網頁超連結
ptt_beauty_url = 'https://www.ptt.cc/bbs/Beauty/index.html'

#today是爬蟲的條件
#意思是只爬取指定日期的文章
today = time.strftime('%m', time.localtime()) + time.strftime('/%d', time.localtime())
print('-----------------------------------------')
print('Today is', today)

#開始進行爬蟲
print('-----------------------------------------')
print('Start')
while True:
  #呼叫get_web_page()函式取得網頁的所有文字
  page = get_web_page(ptt_beauty_url)

  #如果page不是None就進行爬蟲
  if page:
    #呼叫get_articles()函式取得文章與上一頁的超連結
    articles, previous_page_link = get_articles(page, today)

    #如果articles不是None就下載圖片
    if articles:
      for article_content in articles:
        #文章標題作為資料夾名稱
        title = article_content['title']

        #取得要下載的圖片超連結清單
        image_url_list = article_content['image_url_list']

        #呼叫save_images()函式下載圖片
        save_images(title, image_url_list)

      #更新網址進行下一輪的爬蟲
      ptt_beauty_url = ptt_url_head + previous_page_link
    else:
      #如果articles是None代表沒有符合條件的文章了
      #此時可以停止爬蟲
      break
print('-----------------------------------------')
print('End')
```

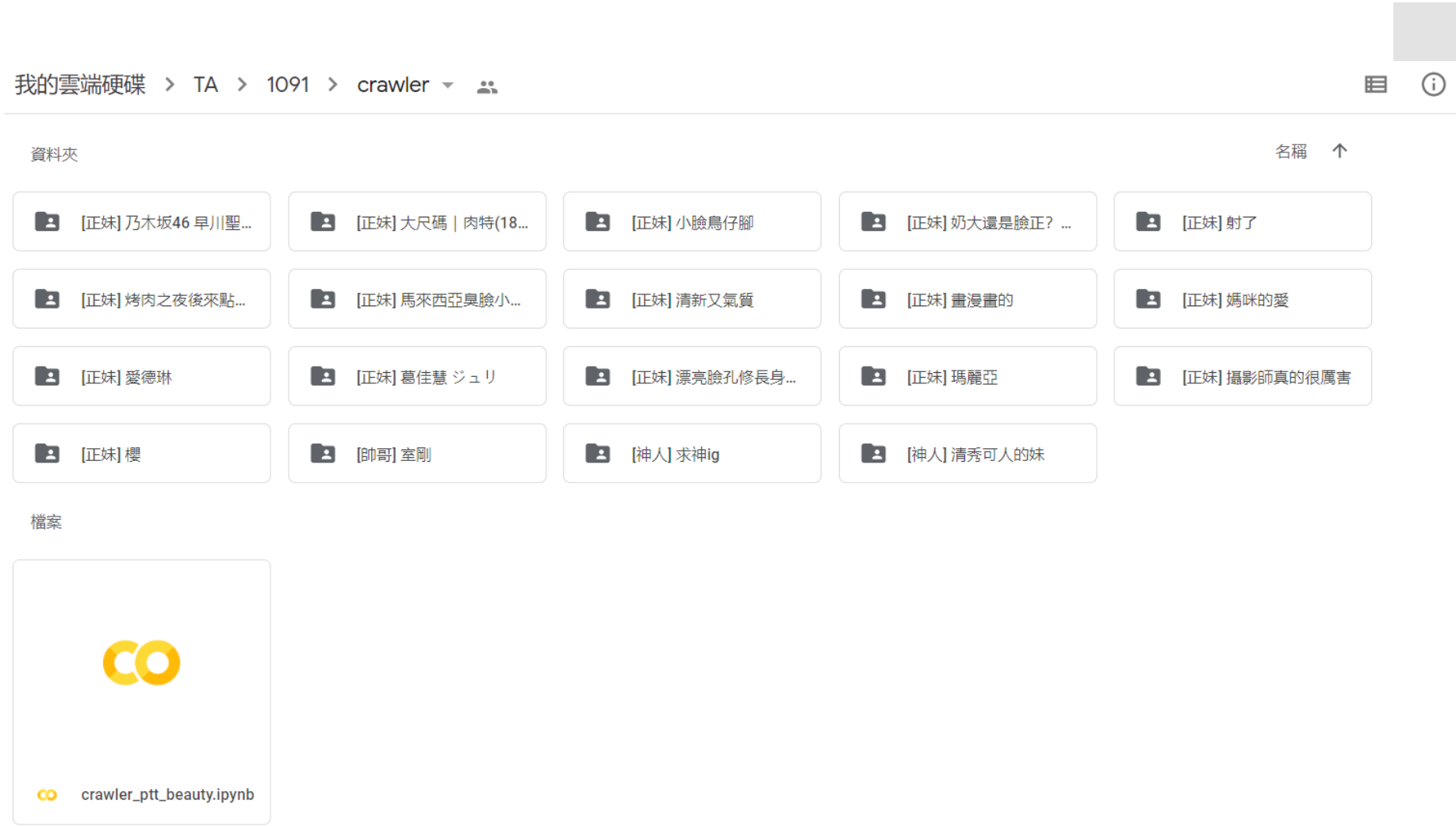
執行結果

```
-------------------------------------------
Today is 10/02
-------------------------------------------
Start
-------------------------------------------
Folder name: [正妹] 奶大還是臉正？抖幾？
WgFpFuu.jpg saved...
Ib7ImPt.jpg saved...
1yl49ib.jpg saved...
4z32Z7s.jpg saved...
-------------------------------------------
Folder name: [正妹] 攝影師真的很厲害
n0abslH.jpg saved...
45C9g6Y.jpg saved...
Sgupv6w.jpg saved...
HDyYZAB.jpg saved...
z8vktsk.jpg saved...
```

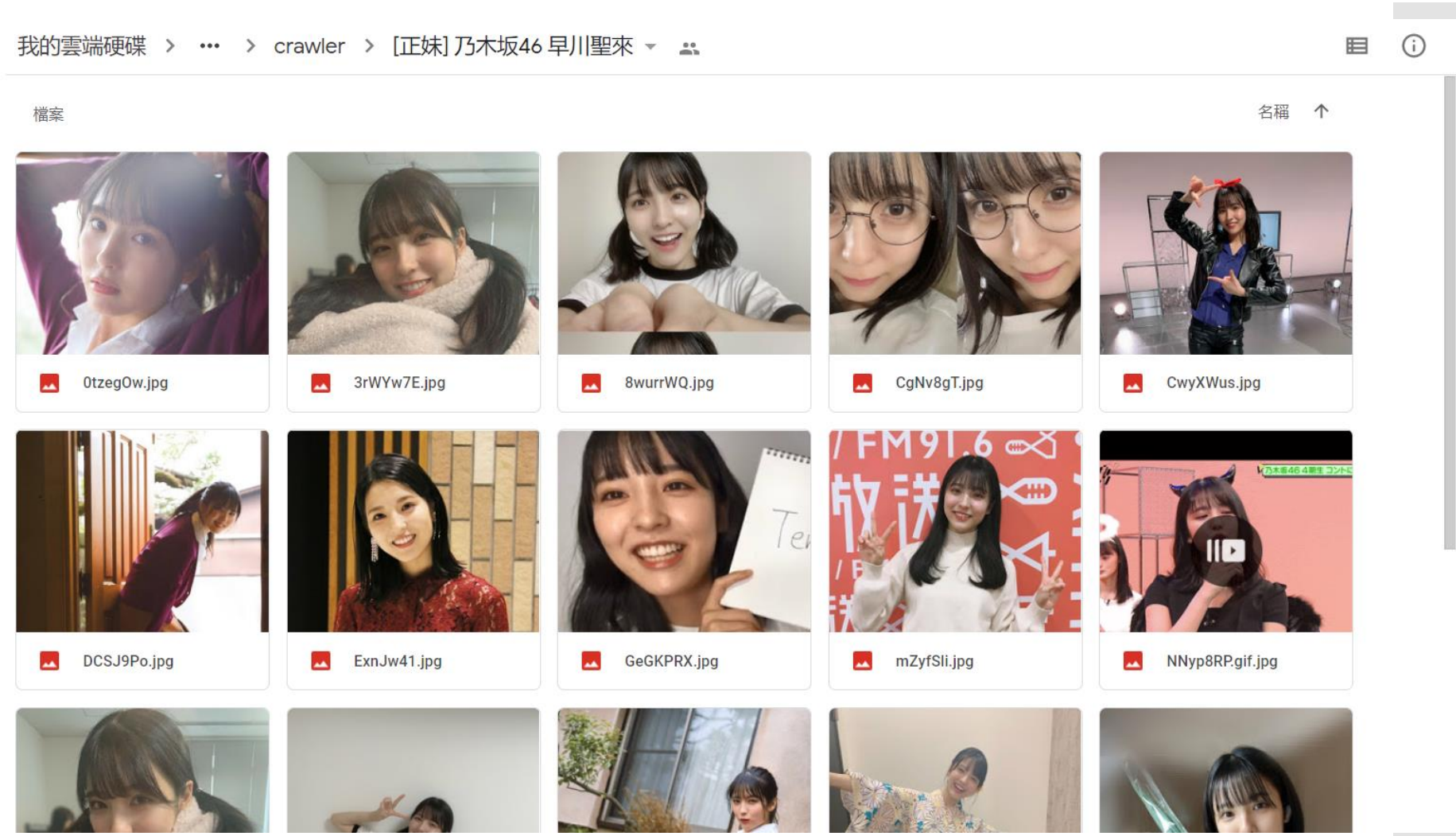
因為印出來的訊息太長了，所以中間的訊息就不貼出來給大家看。

```
oNunG1k.jpg saved...
Qcllb9t.jpg saved...
Qv9DHKO.jpg saved...
jw9D7uz.jpg saved...
-------------------------------------------
Folder name: [正妹] 畫漫畫的
O2C0PTv.jpg saved...
rgUT31T.jpg saved...
l5zqNGz.jpg saved...
TJh0ZHv.jpg saved...
59ajKeC.jpg saved...
Pxg7Y9r.jpg saved...
oTdyvLG.jpg saved...
lOX8ulz.jpg saved...
-------------------------------------------
End
```

執行結果


我的雲端硬碟 > TA > 1091 > crawler
資料夾
名稱
[正妹] 乃木坂46 早川聖...
[正妹] 大尺碼｜肉特(18...
[正妹] 小臉鳥仔腳
[正妹] 奶大還是臉正? ...
[正妹] 射了
[正妹] 烤肉之夜後來點...
[正妹] 馬來西亞臭臉小...
[正妹] 清新又氣質
[正妹] 畫漫畫的
[正妹] 媽咪的愛
[正妹] 愛德琳
[正妹] 葛佳慧 ジュリ
[正妹] 漂亮臉孔修長身...
[正妹] 瑪麗亞
[正妹] 攝影師真的很厲害
[正妹] 櫻
[帥哥] 室剛
[神人] 求神ig
[神人] 清秀可人的妹
檔案
crawler_ptt_beauty.ipynb

執行結果


我的雲端硬碟 > ••• > crawler > [正妹] 乃木坂46 早川聖來
檔案
名稱
0tzegOw.jpg
3rWYw7E.jpg
8wurrWQ.jpg
CgNv8gT.jpg
CwyXWus.jpg
DCSJ9Po.jpg
ExnJw41.jpg
GeGKPRX.jpg
mZyfSli.jpg
NNyp8RP.gif.jpg