

網頁爬蟲 實戰演練

Web Crawler Practice



Python Web Crawler

Outline

- ◆ 蘋果日報今日焦點新聞
- ◆ 大樂透開獎
- ◆ 教育部網頁圖片下載



蘋果日報今日焦點新聞

- ◆ 蘋果日報今日焦點的網址為：

<https://tw.appledaily.com/hot/daily>

- ◆ 首先透過開發者工具觀察蘋果今日焦點的網頁，我們所需要的是新聞標題的數字、標題文字。

The screenshot shows the homepage of the Apple Daily website. At the top, there is a blue header bar with the Apple Daily logo and a navigation menu. Below the header, there is a list of ten news items, each consisting of a red number followed by a short headline. The headlines are:

- 01 | 《蘋果》踢爆 原能會怠惰 害3命 輻射檢驗師集體癌亡
- 02 | A4743 海陸悍將癌逝 5歲子抱父衣「爸爸別走」
- 03 | 入門 却不入门 BMW新320i M Sport獨家搶試
- 04 | 滿臉「爛痘」宅男 花兩週變天菜
- 05 | 搶70萬台筆電市場 頂級顯卡吸買氣 實測NVIDIA RTX顯卡筆...
- 06 | 人都死了原能會主委還在拗 「業者鐵定有問題... 但查無不...
- 07 | 內科心靈綠洲底加！漫漫點用茶香還上班族片刻寧靜
- 08 | 統一獅教頭蔭總 紿A-Lin完整的浪漫
- 09 | 蔡淑臻 老夫老妻吃路邊攤 黏李沛旭分食滷蛋
- 10 | 白色恐懼 港警逼停地鐵濫搜 掃蕩機場 埋伏碼頭

蘋果日報今日焦點新聞

◆ 我們會發現整個新聞區塊都是在`<ul class="all">`之下。

The screenshot shows the homepage of Apple Daily. At the top, there is a navigation bar with links to International, Finance,副刊 (Supplement), Sports, Real Estate, and Forum & Special Columns. Below the navigation bar is a list of 17 news items, each with a thumbnail image, title, and timestamp. To the right of the news list is a sidebar titled "即時新聞" (Latest News) which includes a video player and several news snippets. At the bottom right of the page is a "回到最上面" (Back to Top) button.

- 01 【降血糖食物】血糖超標易有糖尿病 營養學家推薦4大降血糖...
- 02 決裂 今天中執會攤牌 然英文：黨只要初選 就從沒贏過 賴清德...
- 03 【小心膽固醇】四大蛋類 膽固醇含量排行
- 04 賈靜雯長女入坑追《與惡》拋夫1打3享受爭寵
- 05 勝利鄭俊英淫片群組流出暗號「吃內嗑糖」爆沾毒
- 06 賴清德：初選輸的支持贏的 館長首播 神魔對談
- 07 郭彥甫首播失控 淚崩嘆笨想自殘
- 08 郭董吃格力傳言 專家打槍 中媒報導砸1800億買15% 鴻海不評論
- 09 8大話題殘酷票選 喬丹史上最強 完爆詹皇
- 10 光電千金砸10萬逼姪輸精管 嫁胎3月夫妻吃肥26kg
- 11 Selina桃花再等3年 唐綺陽勸退姊弟戀
- 12 一週破千萬觀看，劇情超展開！渡邊直美搞笑催逼「青春露」...
- 13 【櫃姐怎麼了】銷售奇招！櫃姐上班狂自拍 業績竟翻倍
- 14 《蘋果新聞網》今天起全面訂閱制 免費註冊一秒開禁
- 15 雪莉直播激凸 怒被視線強姦
- 16 許瑋甯瘦瘦5kg沒婚頭 等等演男友Cue開鏡
- 17 蔡資穎積分進補 預約冠軍戰

The screenshot shows the browser's developer tools with the DOM panel open. It highlights the `<ul class="all">` element, which contains all the news items. The CSS panel shows the following styles applied to the `.aht_board ul.all` selector:

```
html #section div div div article#maincontentvertebrae div.aht_board ul.all li div.aht_title
```

```
element.style { }
```

```
.aht_board ul.all { border-bottom: 2px solid #ffa800; }
```

```
ul { list-style: none; }
```

The right side of the developer tools shows the element's bounding box with dimensions of 650 x 1350 pixels, a padding of 2 pixels, and a margin of 0 pixels.

蘋果日報今日焦點新聞

◆ 其下的每個標籤內，需要的資訊都很有條理地放在相對應的div中，之後我們就可以在程式碼內利用這些屬性。

➤ 標題數字：`<div class='aht_title_num'>`



```
> <script>...</script>
> <script>...</script>
<!--roundabout slider-->
<div class="aht_board">
  ...
    <ul class="all">
      <li>
        <div class="aht_title_num atopred">
          <span>01</span> == $0
        </div>
        <div class="aht_title">
          <a target="_blank" href="https://tw.news.appledaily.com/headline/daily/20190410/38305308" title="【降血糖食物】血糖超標易有糖尿病 嘗養學家推薦4大降血糖食物">【降血糖食物】血糖超標易有糖尿病 嘗養學家推薦4大降血糖食物</a>
        </div>
      </li>
    ...
```

➤ 標題文字：`<div class='aht_title'>`



```
> <script>...</script>
<!--roundabout slider-->
<div class="aht_board">
  ...
    <ul class="all">
      <li>
        <div class="aht_title_num atopred">
          <span>01</span>
        </div>
        <div class="aht_title">
          <a target="_blank" href="https://tw.news.appledaily.com/headline/daily/20190410/38305308" title="【降血糖食物】血糖超標易有糖尿病 嘗養學家推薦4大降血糖食物">【降血糖食物】血糖超標易有糖尿病 嘗養學家推薦4大降血糖食物</a>
        </div>
      </li>
    ...
```

蘋果日報今日焦點新聞

```
1 import requests
2 from bs4 import BeautifulSoup
3
4 print('蘋果今日焦點')
5 print('-----')
6 dom = requests.get('https://tw.appledaily.com/hot/daily').text
7 soup = BeautifulSoup(dom, 'html.parser')
8 for news in soup.find('ul', 'all').find_all('li'):
9     print(news.find('span').text,news.find('div', 'aht_title').text,)
```

蘋果今日焦點

-
- 01 《蘋果》踢爆 原能會怠惰 害3命 輻射檢驗師集體癌亡
 - 02 A4743 海陸悍將癌逝 5歲子抱父衣「爸爸別走」
 - 03 入門 却不入门 BMW新320i M Sport獨家搶試
 - 04 滿臉「爛痘」宅男 花兩週變天菜
 - 05 搶70萬台筆電市場 頂級顯卡吸買氣 實測NVIDIA RTX顯卡筆電 3D算圖更勝MacBook Pro
 - 06 人都死了 原能會主委還在拗 「業者鐵定有問題... 但查無不法」
 - 07 內科心靈綠洲底加！漫漫點用茶香還上班族片刻寧靜
 - 08 蔡淑臻 老夫老妻吃路邊攤 黏李沛旭分食滷蛋
 - 09 統一獅教頭蔗總 紿A-Lin完整的浪漫
 - 10 白色恐懼 港警逼停地鐵濫搜 掃蕩機場 埋伏碼頭

大樂透開獎號碼

- ◆臺灣彩券的網址為：<http://www.taiwanlottery.com.tw>」，在首頁中會公告大樂透的中獎號碼。
- ◆本例子會輸出大樂透的黃球資料，接著輸出開獎的號碼順序，然後依照大小順序排序，最後印出特別號。

```
大樂透黃球資料：
[<div class="ball_tx ball_yellow">20 </div>, <div class="ball_tx ball_yellow">14 </div>, <div class="ball_tx ball_yellow">45 </div>,
<div class="ball_tx ball_yellow">29 </div>, <div class="ball_tx ball_yellow">27 </div>, <div class="ball_tx ball_yellow">05 </div>,
<div class="ball_tx ball_yellow">05 </div>, <div class="ball_tx ball_yellow">14 </div>, <div class="ball_tx ball_yellow">20 </div>,
<div class="ball_tx ball_yellow">27 </div>, <div class="ball_tx ball_yellow">29 </div>, <div class="ball_tx ball_yellow">45 </div>]
=====
開出順序：20    14    45    29    27    05
大小順序：05    14    20    27    29    45
特別號（紅球）：07
```

大樂透開獎號碼

- ◆檢視該網頁的原始碼，可以發現class名稱為「contents_box02」有4筆，分別對應到「威力彩區塊」、「38樂合彩區塊」、「大樂透區塊」與「38樂合彩區塊」，其中的「大樂透區塊」為第3個區塊。



```
<!doctype html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"  
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">  
<html xmlns="http://www.w3.org/1999/xhtml">  
  <head>...</head>  
  <body onload="showBanner();" data-feedly-mini="yes">  
    <form method="post" action="index_new.aspx" id="form1">  
      <div class="aspNetHidden">...</div>  
      <div class="aspNetHidden">...</div>  
      <div id="wrapper_overflow">  
        <iframe src="header.asp" width="1024" height="68" scrolling="no" frameborder="0">...</iframe>  
        <div id="left">...</div>  
        <div id="right_top">...</div>  
        <div id="rightdown">  
          <!--*****BINGO BINGO*****-->  
          <div class="contents_box01">...</div>  
          <div class="dotted01"></div>  
          <!--*****雙贏彩區塊*****-->  
          <div class="contents_box06">...</div>  
          <div class="dotted01"></div>  
          <!--*****威力彩區塊*****-->  
          <div class="contents_box02">...</div>  
          <div class="dotted02"></div>  
          <!--*****38樂合彩區塊*****-->  
          <div class="contents_box02">...</div>  
          <div class="dotted01"></div>  
          <!--*****大樂透區塊*****-->  
          ...  
          <div class="contents_box02" style="background-color: #000; color: #fff; padding: 5px; text-align: center; font-weight: bold;">...</div>  
        <div id="contents_logo_04"></div>  
        <div class="contents_mine_tx02">...</div>  
        <div class="contents_mine_tx04">...</div>  
        <div class="ball_tx ball_yellow">20</div>  
        <div class="ball_tx ball_yellow">14</div>  
        <div class="ball_tx ball_yellow">45</div>  
        <div class="ball_tx ball_yellow">29</div>  
        <div class="ball_tx ball_yellow">27</div>  
        <div class="ball_tx ball_yellow">05</div>  
        <div class="ball_tx ball_yellow">05</div>  
        <div class="ball_tx ball_yellow">14</div>  
        <div class="ball_tx ball_yellow">20</div>  
      </div>  
    </div>  
  </body>  
</html>
```

大樂透開獎號碼

1. 汇入requests、BeautifulSoup套件，並設定變數url。
2. 使用requests套件的get()函式取得網頁原始碼。
3. 在BeautifulSoup套件中使用html.parser解析原始碼，自訂bs為BeautifulSoup物件名稱。
4. 使用select()函式取出class名稱為「contents_box02」的串列：
 - data1[0].text為串列的第一項，也就是「威力彩區塊」字串。
 - data[1].text為串列的第二項，也就是「38樂合彩區塊」字串。
 - data1[2].text為串列的第三項，也就是「大樂透區塊」字串。

```
1 import requests
2 from bs4 import BeautifulSoup
3
4 url = 'http://www.taiwanlottery.com.tw/'
5 html = requests.get(url)
6 bs = BeautifulSoup(html.text, 'html.parser')
7
8 # 取出class名稱為contents_box02的串列
9 data1 = bs.select(".contents_box02")
```

大樂透開獎號碼

5. 在data1[2].text為串列的第三項，也就是「大樂透區塊」的字串中，使用find_all() 函式，找出所有的<div>標籤且class名稱為「ball_tx」的項目，也就是黃色彩球的部份，並將其資料指定給data2並輸出。(line8-10)
6. 使用for迴圈印出data2前6個與後6個text的內容。 (line14-15 · line17-18)
7. 使用find()函式，找出第一個<div>標籤且class名稱為「ball_red」項目，也就是紅色彩球的部份，並將其資料指定給red並輸出。(line20~21)

```
11 # 在第3個區塊中抓出黃球
12 data2 = data1[2].find_all('div', {'class': 'ball_tx'})
13 print('大樂透黃球資料：')
14 print(data2)
15 print('=====')
16
17 # 大樂透號碼
18 print("開出順序：", end="")
19 for n in range(0,6):
20     print(data2[n].text, end=" ")
21 print("\n大小順序：", end="")
22 for n in range(6,len(data2)):
23     print(data2[n].text, end=" ")
24
25 # 特別號
26 red=data1[2].find('div', {'class': 'ball_red'})
27 print("\n特別號（紅球）：%s" %(red.text))
```

```
大樂透黃球資料：
[<div class="ball_tx ball_yellow">45 </div>, <div class="ball_tx ball_yellow">44 </div>, <div class="ball_tx ball_yellow">10 </div>, <div class="ball_tx ball_yellow">36 </div>, <div class="ball_tx ball_yellow">46 </div>, <div class="ball_tx ball_yellow">10 </div>, <div class="ball_tx ball_yellow">36 </div>, <div class="ball_tx ball_yellow">37 </div>, <div class="ball_tx ball_yellow">44 </div>, <div class="ball_tx ball_yellow">45 </div>, <div class="ball_tx ball_yellow">46 </div>]
=====
開出順序：45 44 10 36 37 46
大小順序：10 36 37 44 45 46
特別號（紅球）：48
```

教育部圖片下載

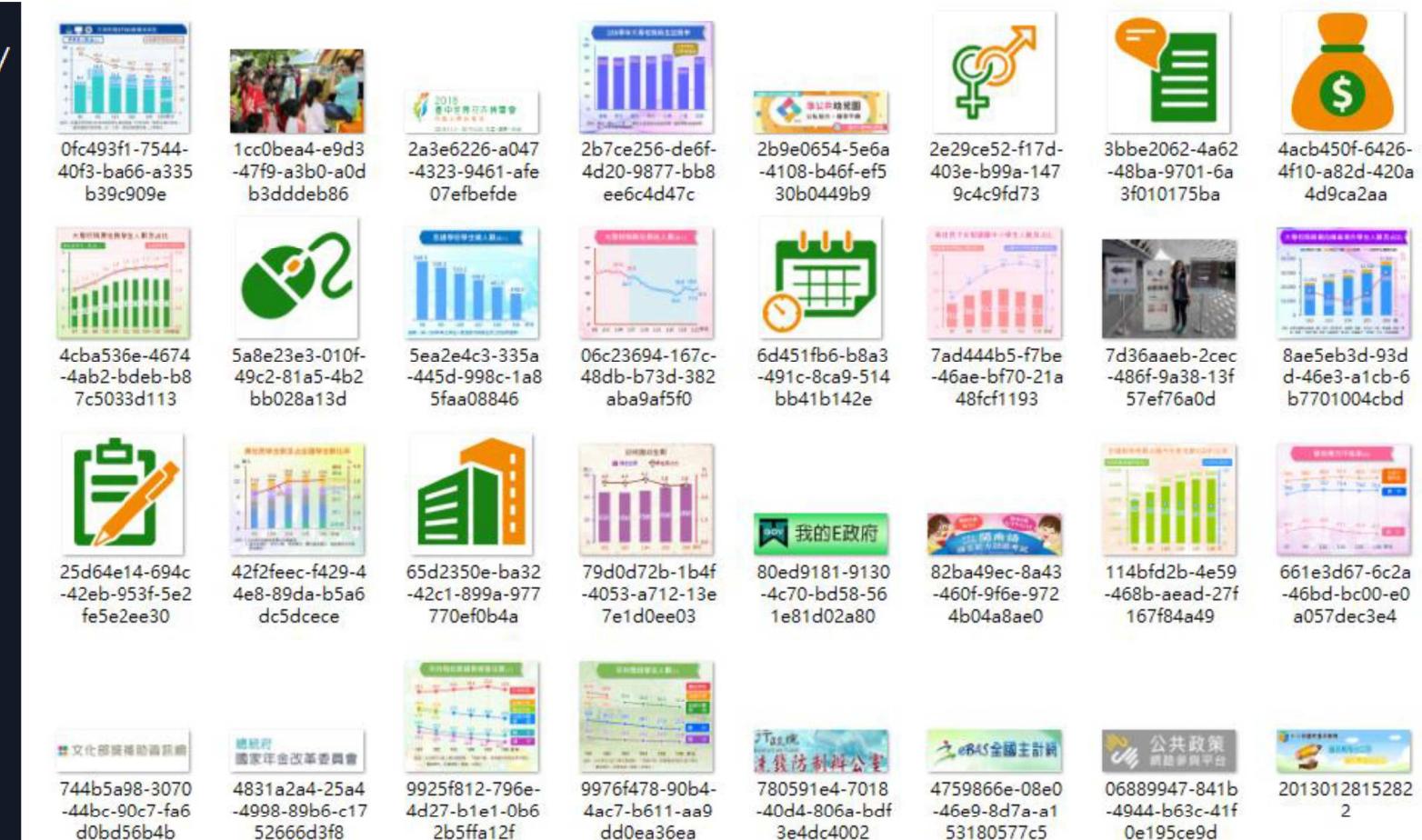
◆ 撰寫Python程式，自動擷取教育部網頁的圖片。教育部網站的網址為<https://www.edu.tw>。



教育部圖片下載

- ◆ 程式時會抓取網頁上所有的「img」標籤資料，進行圖檔的下載以及情況的描述：下載成功或無法下載。
- ◆ 程式自動在檔案工作目錄上建立「images」資料夾，並將網頁擷取下來的圖檔置於其中，下載結果如下。

```
=====
圖檔完整路徑： https://ws.moe.edu.tw/001/Upload/1/rempic/6203/688/
fc64ad98-3174-46b6-befb-0518935c1ff1.jpg
下載成功：fc64ad98-3174-46b6-befb-0518935c1ff1.jpg
=====
圖檔完整路徑： https://ws.moe.edu.tw/001/Upload/1/rempic/
6203/679/4759866e-08e0-46e9-8d7a-a153180577c5.jpg
下載成功：4759866e-08e0-46e9-8d7a-a153180577c5.jpg
=====
圖檔完整路徑： https://ws.moe.edu.tw/001/Upload/1/RelPic/
6203/66/20150909091005.png
下載成功：20150909091005.png
=====
圖檔完整路徑： https://ws.moe.edu.tw/001/Upload/1/rempic/
6203/79/06889947-841b-4944-b63c-41f0e195ce9d.jpg
下載成功：06889947-841b-4944-b63c-41f0e195ce9d.jpg
=====
圖檔完整路徑： https://ws.moe.edu.tw/001/Upload/1/rempic/
6203/110/80ed9181-9130-4c70-bd58-561e81d02a80.jpg
下載成功：80ed9181-9130-4c70-bd58-561e81d02a80.jpg
```



教育部圖片下載

1. 匯入各項所需的套件。
2. 設定教育部網站網址給變數url。
3. 設定存放圖片檔案的資料夾名稱，如果資料夾不存在，則在工作目錄下建立目錄「images」來儲存圖片。
4. 使用find_all()函式取得所有標籤的內容。
5. 使用for迴圈對每一個做處理，讀取src屬性值的內容，然後將src字串轉成串列attrs。

```
1 import requests,os,urllib
2 from bs4 import BeautifulSoup
3 url='https://www.edu.tw/'
4 html=requests.get(url)
5 html.encoding="utf-8"
6 bs=BeautifulSoup(html.text,'html.parser')
7 pics_dir="pics"
8 if not os.path.exists(pics_dir):
9     os.mkdir(pics_dir)          # 在工作目錄下建立目錄pics來儲存圖片
10 all_links=bs.find_all('img')    # 用串列取得所有標籤的內容
11 for link in all_links:
12     src=link.get('src')        # 讀取src屬性值的內容
13     attrs=[src]                # 將src字串轉成串列attrs
```

教育部圖片下載

6. 對串列 `attrs` 每一項進行處理，如果該項非空，且副檔名為「`.jpg`」或「`.png`」，則設定圖檔完整路徑給變數「`full_pth`」，並從網址的最右側取得圖檔的名稱給變數「`file_n`」。
7. 使用`urlopen()`函式取得圖檔資料，並以二進位寫入模式「`wb`」，將下載的檔案以同檔名寫到工作目錄下之資料夾內，完成後印出下載成功：」加上檔案名稱字串，最後進行檔案關閉。
8. 當檔案下載失敗時，進入此「`except`」區塊，印出「無法下載：」加上檔案名稱字串。

```
1 import requests,os,urllib
2 from bs4 import BeautifulSoup
3 url='https://www.edu.tw/'
4 html=requests.get(url)
5 html.encoding="utf-8"
6 bs=BeautifulSoup(html.text,'html.parser')
7 pics_dir="pics"
8 if not os.path.exists(pics_dir):
9     os.mkdir(pics_dir)          # 在工作目錄下建立目錄pics來儲存圖片
10 all_links=bs.find_all('img')    # 用串列取得所有<img>標籤的內容
11 for link in all_links:
12     src=link.get('src')        # 讀取src屬性值的內容
13     attrs=[src]                # 將src字串轉成串列attrs
14     for attr in attrs:
15         if attr!=None and('.jpg'in attr or '.png'in attr):# 讀取.jpg或.png檔
16             full_path = attr           # 設定圖檔完整路徑
17             file_n=full_path.split('/')[-1] # 從網址的最右側取得圖檔的名稱
18             print('=====')           # 儲存圖片程式區塊
19             print('圖檔完整路徑：',full_path)
20             try:
21                 image = urllib.request.urlopen(full_path)
22                 f = open(os.path.join(pics_dir,file_n), 'wb')
23                 f.write(image.read())
24                 print('下載成功：%s' %(file_n))
25                 f.close()
26             except:                  # 無法儲存圖片程式區塊
27                 print("無法下載：%s" %(file_n))
```



Q & A