

# 人工智慧

## Python網頁爬蟲 教學

### request, beautiful soup 套件

---

吳智鴻 教授

國立臺中教育大學 數位內容科技學系

2024/10/16

# 網頁爬蟲套件

---

(1) request套件 (抓取網頁資料用)

安裝方式

```
pip install requests
```

(2) BeautifulSoup (解析網頁用)

安裝方式

```
pip install bs4
```

# Request 套件與網頁解析

---

# Read\_ntcu.ipynb

## 讀取HTML資料，就是這麼簡單

---

```
import requests
url = 'http://www.ntcu.edu.tw/newweb/index.htm'
html = requests.get(url)
html.encoding="utf-8"
print(html.text)
```

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"
"http://www.w3.org/TR/html4/loose.dtd">
<html>
<head>
<meta http-equiv="Cache-Control" content="no-cache">
<meta http-equiv="Content-Type" content="text/html; charset=utf-8">
<meta name="google-site-verification" content="G5a3X9dDaYMMqBWiNKwvDixlSsP_Fz4utiCLhFysNTI" />
<META NAME="Keywords" CONTENT="國立臺中教育大學'國立台中教育大學'National Taichung University of Education'臺中教育大學'台中教育
大學'教育大學'">
<META name="description" content="國立臺中教育大學 National Taichung University of Education (NTCU) 包含校史、校園公告、行政教
學資訊及學生活動。" >
<script language="javascript" type="text/javascript">
</script>
<title>國立臺中教育大學-NTCU-National Taichung University of Education</title>
<style type="text/css">
<title>國立臺中教育大學-NTCU-National Taichung University of Education</title>
<style type="text/css">
<!--
body {
```

# Read\_dct.ipynb

## 試著抓取系上網頁看看

```
import requests
url = 'https://dct.ntcu.edu.tw/'
html = requests.get(url)
html.encoding="utf-8"
print(html.text)
```

```
<!doctype html>
<html>
<head>
<meta charset="UTF-8">
<title>國立臺中教育大學 - 數位內容科技學系</title>
<meta name="viewport" content="width=device-width, initial-scale=1">
<meta http-equiv="X-UA-Compatible" content="IE=edge">
<meta name="keywords" content="國立台中教育大學,數位內容科技學系">
<meta name="description" content="國立台中教育大學數位內容科技學系敘述">
<meta http-equiv="refresh" content="5;url=/news.php" />
<link rel="shortcut icon" href="img/favicon.ico" />
<link rel="shortcut icon" href="img/favicon.ico" /><link rel="stylesheet" href="css/flexslider.css">
<link rel="stylesheet" href="css/colorbox.css">
<link rel="stylesheet" href="css/main_style.css">
<style>
    .slides {
        position: fixed;
        top: 0;
        z-index: 1;
```



最新消息	
請輸入關鍵字	🔍
【新聞】財團法人資訊工業策進會-「NVIDIA」與「本會數位教育研...	2019.10.08
【新聞】為辦理「2019臺中國際動畫影展」活動，敬請鼓勵貴校師生踴...	2019.09.28
【新聞】「臺中國際動畫影展-2018得獎影片大賞」自9月6日至9月2...	2019.09.02
【新聞】【數位系】恭賀~由吳智鴻教授指導數位四甲莊汶權、以研究...	2019.07.11
【新聞】【數位系】108學年度大學部轉學考試時程公告	2019.07.01
【新聞】【數位系】恭賀~由吳育龍副教授指導數位二甲鄭羽涵、葉洵...	2019.06.26
【新聞】國家災害防救科技中心辦理「民生公共物聯網資料應用競賽」...	2019.06.11
【活動】【數位系】研究生學位考試申請	2019.05.22

# Read\_dct.ipynb 網頁檢查碼 (200)

```
import requests
url = 'https://dct.ntcu.edu.tw/'
html = requests.get(url)
html.encoding="utf-8"

# 檢查HTTP回應碼
print(html.status_code)
# print(html.text)
```

200

回應碼=200  
表示讀取成功



最新消息	
【新聞】財團法人資訊工業策進會函-「NVIDIA」與「本會數位教育研...	2019.10.08
【新聞】為辦理「2019臺中國際動畫影展」活動，敬請鼓勵貴校師生講...	2019.09.28
【新聞】「臺中國際動畫影展-2018得獎影片大突擊」自9月6日至9月2...	2019.09.02
【新聞】【數位系】恭賀~由吳智鴻教授指導數位四甲班沈權·以研究...	2019.07.11
【新聞】【數位系】108學年度大學部轉學考試時程公告	2019.07.01
【新聞】【數位系】恭賀~由吳育龍副教授指導數位二甲鄭羽涵·葉海...	2019.06.26
【新聞】國家災害防救科技中心辦理「民生公共物聯網資料應用競賽」...	2019.06.11
【活動】【數位系】研究生學位考試申請	2019.05.22

# Read\_dct.ipynb

## 檢查是否正確讀取網頁並印出

---

```
▶ import requests
url = 'https://dct.ntcu.edu.tw/'
html = requests.get(url)
html.encoding="utf-8"

# 檢查HTTP回應碼
if html.status_code == requests.codes.ok:
    print(html.text)
```

如果成功讀取，則  
列印抓取的網頁



# KeywordSearch2.ipynb

```
In [1]: import requests
url = 'https://dct.ntcu.edu.tw/'
html = requests.get(url)
html.encoding="utf-8"
#print(html.text)

# 把HTML原始碼以每一列分割成串列，並去除跳列字元
htmlList = html.text.splitlines()
#print(htmlList)

for row in htmlList:
    print(row)
```

把網頁串列htmlList  
每一列印出來

結果  
與上一頁程式結果  
比較看看

```
<!doctype html>
<html>
<head>
<meta charset="UTF-8">
<title>國立臺中教育大學 - 數位內容科技學系</title>
<meta name="viewport" content="width=device-width, initial-scale=1">
<meta http-equiv="X-UA-Compatible" content="IE=edge">
<meta name="keywords" content="國立台中教育大學,數位內容科技學系">
<meta name="description" content="國立台中教育大學數位內容科技學系敘述">
<meta http-equiv="refresh" content="5;url=/news.php" />
<link rel="shortcut icon" href="img/favicon.ico" />
<link rel="shortcut icon" href="../img/favicon.ico" /><link rel="stylesheet" href="css/flexslider.c
ss">
<link rel="stylesheet" href="css/colorbox.css">
<link rel="stylesheet" href="css/main_style.css">
<style>
    .slides {
        position: fixed;
        top: 0;
```

```
In [2]: # 統計個數
n=0
keyword = "數位"
for row in htmlList:
    if keyword in row: n+=1
print("找到 {} 次!".format(n))
```

改用變數來搜尋  
Keyword，更有彈性了

找到 4 次!

# Cursor

## Ctrl+I 詢問程式指令

```
1 # 导入所需的库
2 import requests # 用于发送HTTP请求
3 from bs4 import BeautifulSoup # 用于解析HTML
4
5 # 设置目标网页的URL
6 url = 'https://2023ntcu.ntcu.edu.tw/'
7 # 发送GET请求获取网页内容
8 response = requests.get(url)
9 # 设置响应的编码为UTF-8，确保正确处理中文
10 response.encoding = 'utf-8'
11
12 # 使用BeautifulSoup解析HTML内容
13 soup = BeautifulSoup(response.text, 'html.parser')
14 # 打印格式化后的HTML结构，方便查看网页结构
15 print(soup.prettify())
16
17 # 设置要搜索的关键词
18 keyword = "臺中"
19 # 初始化计数器
20 count = 0
21
22 # 遍历网页中的所有文本节点
23 for text in soup.stripped_strings:
24     # 检查当前文本是否包含关键词
25     if keyword in text:
26         # 如果包含关键词，计数器加1
27         count += 1
28         # 打印找到关键词的信息和当前计数
29         print(f"找到关键词 '{keyword}'，当前计数：{count}")
30
31 # 打印关键词出现的总次数
32 print(f"关键词 '{keyword}' 共出现 {count} 次")
33
```



The screenshot shows the Cursor AI interface with a question box at the top containing the text "stripped\_strings的用途". Below the question box, the AI provides a detailed answer in Chinese. The answer explains that `stripped_strings` is a useful method from the BeautifulSoup library for extracting text from HTML documents. It lists five key uses: 1. Text extraction, 2. Removing whitespace, 3. Generator objects, 4. Recursive traversal, and 5. Preserving text structure. At the bottom of the answer, it provides a code snippet for using `stripped_strings` in a loop.

**stripped\_strings的用途**

`stripped_strings` 是 BeautifulSoup 库中的一个非常有用的方法。让我为您详细解释它的用途：

1. 提取文本：  
`stripped_strings` 用于从 HTML 文档中提取所有的文本内容，不包括 HTML 标签。
2. 去除空白：  
它会自动去除文本开头和结尾的空白字符（如空格、换行符等），以及完全由空白字符组成的字符串。
3. 生成器对象：  
`stripped_strings` 返回一个生成器对象，这意味着它是内存高效的，特别是在处理大型文档时。
4. 递归遍历：  
它会递归地遍历整个文档树，提取所有层级的文本内容。
5. 保留文本结构：  
虽然它去除了多余的空白，但仍然保留了文本的基本结构和顺序。

在您的代码中，`stripped_strings` 的使用如下：

```
for text in soup.stripped_strings:
    if keyword in text:
```

# 自訂HTTP Headers 偽裝瀏覽器操作

---

# 自訂Http headers的作用

---

有些網站會擋爬蟲程式，所以需要加入headers設定，把程式偽裝成瀏覽器操作才不會被擋住。

# Read thsrc 讀取高鐵網站

```
import requests
url = 'https://irs.thsrc.com.tw/IMNINT/'
html = requests.get(url)
html.encoding="utf-8"
```

```
# 檢查HTTP回應碼
print(html.status_code)

#if html.status_code == requests.codes.ok:
#print(html.text)
```

程式會無法正確執行。  
如果不加自訂的headers設定，  
會被認為是爬蟲而被禁止執行



The screenshot shows the THSR website interface. The browser address bar displays 'irs.thsrc.com.tw/IMINT/'. The page header includes the THSR logo and navigation links for '中文 | English | 日本語' and '台灣高鐵網路訂票系統 24hrs'. A navigation bar contains links for '一般訂票', '雙人55', '信用卡優惠專區', '訂位查詢/付款', and 'TGo會員兌換'. Below this, a progress indicator shows the current step: '您所在的訂位步驟： 1. 查詢車次 > 2. 選擇車次 > 3. 取票資訊 > 4. 完成訂位'. A notice states: '票務服務系統(含網路訂票、T-EX App、便利商店等)將於11/4(四)凌晨00:30至05:00暫停服務。詳見網站最新消息。'. The main content area is titled '一般訂票' and contains a form with the following fields:

起訖站	起程站 <input type="text" value="請選擇..."/> 到達站 <input type="text" value="請選擇..."/>
車廂種類	<input checked="" type="radio"/> 標準車廂 <input type="radio"/> 商務車廂
座位喜好	<input checked="" type="radio"/> 無 <input type="radio"/> 靠窗優先 <input type="radio"/> 走道優先
訂位方式	<input checked="" type="radio"/> 依時間搜尋合適車次 <input type="radio"/> 直接輸入車次號碼
時間	去程 <input type="text" value="2021/11/02"/> <input type="text" value="約 請選擇..."/> 出發 <input type="checkbox"/> 訂購回程
票數	全票 <input type="text" value="1"/> 孩童票(6-11歲) <input type="text" value="0"/> 愛心票 <input type="text" value="0"/> 敬老票(65歲以上) <input type="text" value="0"/> 大學生優惠票 <input type="text" value="0"/>
查詢早鳥優惠	<input type="checkbox"/> 僅顯示尚有早鳥優惠之車次

At the bottom, a security notice reads: '為了確保交易安全，請輸入右圖中之驗證碼：' followed by a CAPTCHA image showing the number '4470' and a text input field. To the right of the input field are links for '重新產生' and '語音播放'. An orange button labeled '開始查詢' is located at the bottom right of the form area.

# Read\_thsrc

---

```
#讀取高鐵網頁
import requests
url = 'https://irs.thsrc.com.tw/IMNINT/'
headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)'
           ' AppleWebKit/537.36 (KHTML, like Gecko)'
           ' Chrome/70.0.3538.102 Safari/537.36'}

html = requests.get(url, headers=headers)
html.encoding="utf-8"
# 檢查HTTP回應碼
print(html.status_code)

#if html.status_code == requests.codes.ok:
# print(html.text)
```

偽裝成瀏覽器操作

# 高鐵時刻表查詢教學(github)

參考以下網址

<https://github.com/music1353/pyHowFun/blob/master/LESSON4%20%E7%88%AC%E8%9F%B2%E5%AF%A6%E6%88%B0%20-%20%E9%AB%98%E9%90%B5%E6%99%82%E5%88%BB.ipynb>

台灣高鐵 TAIWAN HIGH SPEED RAIL

中文 | English | 日本語 台灣高鐵網路訂票系統 24hrs

一般訂票 雙人55 信用卡優惠專區 訂位查詢/付款 TGo會員兌換

您所在的訂位步驟： 1. 查詢車次 > 2. 選擇車次 > 3. 取票資訊 > 4. 完成訂位

票務服務系統(含網路訂票、T-EX App、便利商店等)將於11/4(四)凌晨00:30至05:00暫停服務。詳見網站

### 一般訂票

起訖站	起程站 <input type="text" value="請選擇..."/> 到達站 <input type="text" value="請選擇..."/>
車廂種類	<input checked="" type="radio"/> 標準車廂 <input type="radio"/> 商務車廂
座位喜好	<input checked="" type="radio"/> 無 <input type="radio"/> 靠窗優先 <input type="radio"/> 走道優先
訂位方式	<input checked="" type="radio"/> 依時間搜尋合適車次 <input type="radio"/> 直接輸入車次號碼
時間	去程 <input type="text" value="2021/11/02"/> <input type="text" value="票價"/> 約 <input type="text" value="請選擇..."/> 出發 <input type="checkbox"/> 訂購回程
票數	全票 <input type="text" value="1"/> 孩童票(6-11歲) <input type="text" value="0"/> 愛心票 <input type="text" value="0"/> 敬老票(65歲以上) <input type="text" value="0"/> 大學生優惠票 <input type="text" value="0"/>
查詢早鳥優惠	<input type="checkbox"/> 僅顯示尚有早鳥優惠之車次

為了確保交易安全，請輸入右圖中之驗證碼：  [重新產生](#) | [語音播放](#)

開始查詢

# Beautiful Soup 套件

---

## Beautifulsoup 的屬性和方法

BeautifulSoup 常用的屬性和方法如下：(表中假設已建立 BeautifulSoup 型別物件 `sp`)

屬性或方法	說明
<code>title</code>	傳回網頁標題，例如： <code>sp.title</code> 。
<code>text</code>	傳回去除所有 HTML 標籤後的網頁文字內容。
<code>find()</code>	傳回第一個符合條件的 tag。例如： <code>sp.find("a")</code> 。
<code>find_all()</code>	傳回所有符合條件的 tag。例如： <code>sp.find_all("a")</code> 。
<code>select()</code>	傳回指定 CSS 選擇器如 <code>id</code> 或 <code>class</code> 的內容，例如：以 <code>id</code> 讀取 <code>sp.select("#id")</code> 、以 <code>class</code> 讀取 <code>sp.select(".classname")</code> 。

# 詳細的BS4指令教學可參考

**Python 使用 BeautifulSoup 抓取與解析  
網頁資料，開發網路爬蟲教學**

2018/02/01

10 則留言



<https://blog.gtwang.org/programming/python-beautiful-soup-module-scrape-web-pages-tutorial/>

**[Python爬蟲教學]7個Python使用  
BeautifulSoup開發網頁爬蟲的實用技巧**

2月 28, 2020



*Photo by Stanley Dai on Unsplash*

<https://www.learncodewithmike.com/2020/02/python-beautifulsoup-web-scraper.html>

# 安裝必要套件

---

```
!pip install beautifulsoup4
```

```
Requirement already satisfied: beautifulsoup4 in /usr/local/lib/python3.6/dist-packages (4.6.3)
```

# 課堂活動

---

# [課堂活動#1]解析基本網頁 BS4\_basic

```
from bs4 import BeautifulSoup

# 原始 HTML 程式碼
html_doc = """
<html><head><title>Hello World</title></head>
<body><h2>Test Header</h2>
<p>This is a test.</p>
<a id="link1" href="/my_link1">Link 1</a>
<a id="link2" href="/my_link2">Link 2</a>
<p>Hello, <b class="boldtext">Bold Text</b></p>
</body></html>
"""

# 以 BeautifulSoup 解析 HTML 程式碼
#url = 'https://dct.ntcu.edu.tw/news.php'
soup = BeautifulSoup(html_doc, 'html.parser')
print(soup.prettify())
```

```
<html>
<head>
  <title>
    Hello World
  </title>
</head>
<body>
  <h2>
    Test Header
  </h2>
  <p>
    This is a test.
  </p>
  <a href="/my_link1" id="link1">
    Link 1
  </a>
  <a href="/my_link2" id="link2">
    Link 2
  </a>
  <p>
    Hello,
    <b class="boldtext">
      Bold Text
    </b>
  </p>
</body>
</html>
```

# BS4基本指令

---

```
[4] # 網頁標題 HTML 標籤
    title_tag = soup.title
    print(title_tag)
```

```
<title>Hello World</title>
```

```
[5] # 所有的超連結
    a_tags = soup.find_all('a')
    for tag in a_tags:
        # 輸出超連結的文字
        print(tag.string)
```

```
Link 1
Link 2
```

```
[6] for tag in a_tags:
        # 輸出超連結網址
        print(tag.get('href'))
```

```
/my_link1
/my_link2
```

```
<html>
<head>
  <title>
    Hello World
  </title>
</head>
<body>
  <h2>
    Test Header
  </h2>
  <p>
    This is a test.
  </p>
  <a href="/my_link1" id="link1">
    Link 1
  </a>
  <a href="/my_link2" id="link2">
    Link 2
  </a>
  <p>
    Hello,
    <b class="boldtext">
      Bold Text
    </b>
  </p>
</body>
</html>
```

# [ 課堂活動#2] 解析系網 BS4\_NTCU 爬蟲並指定關鍵字搜尋次數

```
1 # BS4_NTC
2 # 導入所需的庫
3 import requests # 用於發送HTTP請求
4 from bs4 import BeautifulSoup # 用於解析HTML
5
6 # 設置目標網頁的URL
7 url = 'https://2023ntcu.ntcu.edu.tw/'
8 # 發送GET請求獲取網頁內容
9 headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.124 Safari/537.36'}
10 # 使用requests.get()方法發送請求並獲取響應
11 response = requests.get(url, headers=headers)
12 # 設置響應的編碼為UTF-8，確保正確處理中文
13 response.encoding = 'utf-8'
14
15 # 使用BeautifulSoup解析HTML內容
16 soup = BeautifulSoup(response.text, 'html.parser')
17 # 定義關鍵字
18 keyword = "臺中" # 您可以根據需要更改關鍵字
19
20 # 在網頁中搜尋關鍵字
21 keyword_count = soup.text.count(keyword)
22
23 # 印出關鍵字出現的次數
24 print(f"關鍵字 '{keyword}' 在網頁中出現了 {keyword_count} 次。")
```

網址

瀏覽器表頭

抓網頁&設定編碼為utf-8

抓下來的網頁解析

關鍵字搜尋

# [課堂活動#3]解析系網 BS4\_DCT

```
import requests
from bs4 import BeautifulSoup

# 下載 數位系網 首頁內容
url = 'https://dct.ntcu.edu.tw/news_content.php'
r = requests.get(url)

# 確認是否下載成功
if r.status_code == requests.codes.ok:
    # 以 BeautifulSoup 解析 HTML 程式碼
    soup = BeautifulSoup(r.text, 'html.parser')

    a_tags = soup.find_all('a')
    for tag in a_tags:
        # 輸出超連結的文字
        print('網址: ' + str(tag.string) + ' -> ' + tag.get('href'))
        #print(tag.get('href'))
```

```
網址: 碩士班 -> article_list.php?aid=3&bid=5
網址: 碩士在職專班 -> article_list.php?aid=3&bid=6
網址: 課程資訊 -> #
網址: 規劃介紹 -> article_list.php?aid=3&bid=7

網址: 修業規定 -> article_list.php?aid=4&bid=8
網址: 大學部 -> article_list.php?aid=4&bid=9
網址: 課程地圖 -> article_list.php?aid=4&bid=10&cid=12
網址: 課程架構 -> article_list.php?aid=4&bid=10&cid=13
網址: 課程大綱 -> article_list.php?aid=4&bid=10&cid=14
網址: 群組課程 -> article_list.php?aid=4&bid=10&cid=15
網址: 碩士班 -> article_list.php?aid=4&bid=10&cid=12

網址: 課程地圖 -> article_list.php?aid=4&bid=11&cid=16
網址: 課程架構 -> article_list.php?aid=4&bid=11&cid=17
網址: 課程大綱 -> article_list.php?aid=4&bid=11&cid=18
網址: 碩士在職專班 -> article_list.php?aid=4&bid=11&cid=16
```

# Yahoo

---

# Yahoo 首頁新聞爬蟲程式#1

```
import requests
from bs4 import BeautifulSoup

# 下載 Yahoo 首頁內容
r = requests.get('https://tw.yahoo.com/')

# 確認是否下載成功
if r.status_code == requests.codes.ok:
    # 以 BeautifulSoup 解析 HTML 程式碼
    soup = BeautifulSoup(r.text, 'html.parser')

    # 以 CSS 的 class 抓出各類頭條新聞
    stories = soup.find_all('a', class_='story-title')
    for s in stories:
        # 新聞標題
        print("標題：" + s.text)
        # 新聞網址
        print("網址：" + s.get('href'))
```



The screenshot shows the Yahoo! Taiwan homepage. At the top, there is a navigation bar with links for Home, Mail, News, Stock Market, Weather, Sports, Yahoo TV, Entertainment, Shopping Center, City, Auction, and More. Below this is the Yahoo! logo and a search bar containing the text "張小燕女兒貝怡儂". A search button is visible on the right. Below the search bar, there are several categories: 焦點 (Focus), 運動 (Sports), 娛樂 (Entertainment), 新奇 (New & Odd), 生活 (Life), and 影音 (Audio & Video). The main content area features a large article with a red circle highlighting a bright object in a dark sky, titled "驚見14點不明物 30秒震撼畫面曝". To the right of this article are several smaller news snippets with titles like "女師戀16歲離職 13年後...", "不滿準媳嫂 她po這句8...", and "如何區分台陸客? 日人1...". At the bottom, there are more news items, including "狗狗肚裡藏... 醫挖出117條嚇壞" and "機場1台瘋狂暴走 結局川普也讚".

# Yahoo 頭條新聞程式

## 課堂練習

## 嘗試 Cursor AI 功能 寫出頭條新聞程式

read\_dct.py  
教我怎樣抓出頭條新聞內容

好的，我来教您如何抓取雅虎新闻的头条新闻内容。根据您的代码，我们可以看到您已经成功获取了一些新闻标题。让我们进一步改进代码，以获取更多头条新闻的详细信息。

以下是修改后的代码，专注于抓取头条新闻内容：

read\_dct.py 3/4 ✓

这段代码做了以下改进：

- 我们使用 `find_all('div', class_='Cf')` 来找到所有可能包含头条新闻的 `div` 元素。
- 对于每个找到的头条新闻，我们尝试提取以下信息：
  - 标题：使用 `find('h3', class_='Mb(5px)')`
  - 链接：在找到的元素中查找 `<a>` 标签
  - 来源：查找带有特定类的 `<span>` 标签
  - 摘要：查找 `<p>` 标签
- 我们使用条件语句来检查每个元素是否存在，以避免在元素不存在时出现错误。

Completed Reapply all Save all

+ read\_dct.py x

Edit, refactor, or add code (/ for commands)

Esc to close claude-3.5-sonnet Submit

```
1 # 导入所需的库
2 import requests # 用于发送HTTP请求
3 from bs4 import BeautifulSoup # 用于解析HTML
4
5 # 设置目标网页的URL
6 url = 'https://tw.news.yahoo.com/'
7 # 发送GET请求获取网页内容
8 headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) App
9 response = requests.get(url, headers=headers)
10 # 设置响应的编码为UTF-8，确保正确处理中文
11 response.encoding = 'utf-8'
12
13 # 使用BeautifulSoup解析HTML内容
14 soup = BeautifulSoup(response.text, 'html.parser')
15
16 # 找到所有的新闻文章
17 articles = soup.find_all('li', class_='js-stream-content')
18
19 # 遍历并打印每个新闻文章的信息
20 for article in articles:
21     # 尝试获取新闻标题
22     title = article.find('h3')
23     if title:
24         print("标题:", title.text.strip())
25
26     # 尝试获取新闻链接
27     link = article.find('a')
28     if link and 'href' in link.attrs:
29         full_link = 'https://tw.news.yahoo.com' + link['href']
30         print("链接:", full_link)
31
32     # 尝试获取新闻来源和时间
33     source_time = article.find('span', class_='C(#959595)')
34     if source_time:
35         print("来源和时间:", source_time.text.strip())
36
37     # 尝试获取新闻摘要
38     summary = article.find('p')
39     if summary:
40         print("摘要:", summary.text.strip())
41
42     print("---") # 分隔线，用于区分不同的新闻
43
44 # 如果没有找到任何新闻，打印提示信息
45 if not articles:
46     print("未找到任何新闻文章，可能需要调整选择器或检查网页结构。")
```

# Yahoo 首頁新聞爬蟲程式#2

```
In [2]: # 讀取所有 <b>的標籤內容
print(soup.find('b'))

<b class="Pos-a T-0 End-0">23°</b>
```

```
In [3]: # 讀取所有<a>的標籤內容
print(soup.find_all('a'))
```

```

<b class="Pos-a T-0 End-0">23°</b>
</a>, <a aria-haspopup="true" class="yucs-trigger D-1b MouseOver NoTextDecoration W-100" data-plugin="tooltip" data-tooltip
="{target-child": ".Icon", "click": {"toggle": {"position": "b", "minWidth": "170px", "id": "UHLogin", "parent": "tabbar-signin"}}, "mo
useenter": {"show": {"mouseover": true, "position": "b", "minWidth": "170px", "id": "UHLogin"}}, "mouseleave": {"hide": {"id": "UHLogi
n"}}}' data-trigger="mouseenter mouseleave click" data-ylk="subsec:login;" href="https://login.yahoo.com/config/login?.src=tw
wfp&lang=zh-Hant-TW&intl=tw&done=https://tw.yahoo.com/" id="uh-profile" role="button">
<i class="t-ico Icon Lh-1 C-darkGrey Zoom-1 Va-t portrait-light">|</i>
<b class="t-txt MouseOver-TextDecoration Va-b Lh-12 D-1b">登入</b>
</a>, <a class="D-b NoTextDecoration Va-t Trigger Btn-y W-100" href="https://login.yahoo.com/config/login?.src=twfp&lan
g=zh-Hant-TW&intl=tw&done=https://tw.yahoo.com/">
<i class="Va-m Icon Fz-xs login-btn-yahoo">|</i>
<b class="Va-m Fz-xs Px-4">Yahoo奇摩帳號登入</b>
</a>, <a class="sign-up Ta-start D-i" href="https://tw.edit.yahoo.com/registration?.intl=tw&src=fpctx&done=http
s://tw.yahoo.com/">請註冊</a>, <a aria-haspopup="true" class="yucs-trigger D-1b MouseOver NoTextDecoration" data-plugin="tool
tip" data-tooltip="{target-child": ".", "click": {"toggle": {"position": "b", "minWidth": "170px", "id": "", "parent": "tabbar-mai
l"}}, "mouseenter": {"show": {"mouseover": true, "position": "b", "minWidth": "170px", "id": ""}}, "mouseleave": {"hide": {"id": ""}}}' da
ta-trigger="mouseenter mouseleave click" data-ylk="subsec:mail;" href="https://mail.yahoo.com/?intl=tw&lang=zh-Hant-T
W" role="button">
```



# 執行結果 (抓出Yahoo首頁頭條內容)

標題：NBA主席不道歉 發聲明提台北  
網址：<https://tw.news.yahoo.com/nba-%E4%B8%8D%E6%9C%83%E5%B0%8D%E8%8E%AB%E9%9B%B7%E7%99%BC%E8%A8%0E%91%93%E6%AD%89-%E5%B8%AD%E4%BD%9B%E5%86%8D%E7%99%BC%E8%81%B2%E6%98%8E%E6%8F%90%E5%8F%B0%E5%8C%97-121221013.html>

標題：毀台大連儂牆 陸男遭驅逐出境  
網址：<https://tw.news.yahoo.com/%E6%92%95%E6%AF%80%E9%80%A3%E5%84%82%E7%89%86-%E9%99%B8%E7%94%B7%E6%83%90%E5%88%97%E4%B8%8D%E5%8F%97%E6%AD%A1%E8%BF%8E%E4%BA%BA%E7%89%A9-123011767.html>

標題：「安倍賀電」爆烏龍 蘇揆臉黑  
網址：<https://tw.news.yahoo.com/%E5%AE%89%E5%80%8D%E8%B3%80%E9%9B%BB%E7%A5%9D%E4%B8%AD%E8%8F%AF%E6%B0%91%E5%9C%8B%E5%9C%8B%E6%85%B6-%E7%A6%8F%E5%B2%A1%E5%A4%96%E9%A4%A8%E6%90%9E%E7%83%8F%E9%BE%8D%E9%81%AD%E6%97%A5%E6%9C%AC%E6%89%93%E8%87%89-050256236.html>

標題：小英出手「國民黨恐賠光家產」  
網址：<https://tw.news.yahoo.com/%E9%9F%93%E6%B0%91%E8%AA%BF%E4%B8%8B%E8%B7%8C-%E8%94%A1%E7%8B%82%E6%8C%96%E5%B7%A5%E5%95%86%E7%95%8C-%E5%9C%8B%E6%B0%91%E9%BB%A82020%E6%81%90%E8%B3%A0%E5%85%89%E5%AE%B6%E7%94%A2-045200322.html>

標題：NBA無法割捨的中國「大餅」  
網址：<https://tw.news.yahoo.com/-yahoo%E8%AB%96%E5%A3%87%E5%9A%B4%E9%9C%87%E7%94%9F-nba%E7%84%A1%E6%B3%95%E5%89%B2%E6%8D%A8%E7%9A%84%E4%B8%AD%E5%9C%8B%E5%A4%A7%E9%A4%85-094601382.html>

標題：被天敵橫掃 雙城悲憤放假去了  
網址：<https://tw.news.yahoo.com/mlb-%E6%B4%8B%E5%9F%BA%E4%B8%89%E9%80%A3%E5%8B%9D%E6%A9%AB%E6%8E%83%E9%9B%99%E5%9F%8E-%E7%8E%87%E5%85%88%E6%99%89%E7%B4%9A%E7%BE%8E%E8%81%AF%E5%86%A0%E8%BB%8D%E8%B3%BD-092631706.html>

標題：大復活？富爾茲「吃掉」玫瑰  
網址：<https://tw.news.yahoo.com/nba-%E7%B2%BE%E5%BD%A9%E6%9A%B4%E6%89%A3-%E7%BD%95%E7%97%85%E7%8B%80%E5%85%83%E5%A4%A7%E5%BE%A9%E6%B4%BB-081045887.html>

標題：小胖拚總冠軍：每場都當最終戰  
網址：<https://tw.sports.yahoo.com/news/%E4%B8%AD%E8%81%B7-%E5%82%99%E6%88%B0%E7%B8%BD%E5%86%A0%E8%BB%8D%E6%88%B0-%E6%9E%97%E6%B3%93%E8%82%B2-%E6%AF%8F%E5%A0%B4%E9%83%BD%E7%95%B6%E6%9C%80%E7%B5%82%E6%88%B0-113309286.html>

標題：內線傷光光 勇士緊急簽人補強  
網址：<https://tw.sports.yahoo.com/news/%E5%8B%87%E5%A3%AB%E7%B7%8A%E6%80%A5%E7%B0%BD%E4%B8%8Bpinne-n>

The screenshot shows the Yahoo! homepage with a search bar at the top containing the text "金鐘 左左右右". Below the search bar, there are navigation links for "新聞", "股市", "理財", "運動", "名人娛樂", "電影", "Yahoo TV", "遊戲", "電競", "動漫", and "日本代購". The main content area features a large banner for the movie "Spider-Man: Far From Home" with the text "更多漫威系列 一次看個夠! (HD)". Below the banner, there are several news snippets, including "NBA主席不道歉 發聲明...", "毀台大連儂牆 陸男遭驅逐...", "安倍賀電 爆烏龍 蘇揆臉黑", "小英出手「國民黨恐賠光家產」", "NBA無法割捨的中國「大餅」", "4G爆低月租\$1.1瘋潮來...", and "愛國藝人被洗版? 趙少康: 先把...". The right side of the page has a "YAHOO! TV" section with a video player showing a news broadcast.

# [課堂活動#4]解析新聞網站，以中時為例

## BS4\_chinatimes

試著以中時新聞網為例，抓取其新聞

The screenshot shows the chinatimes website interface. At the top, there is a '速報' (Breaking News) banner with the text '11:58 美豬協商共識 先辦公聽會'. Below this is a large featured article with a video player showing a man pointing upwards, with the headline '昔掀安心亞裙被譙 男星隱退當上班族'. To the right of the featured article is a '熱門新聞' (Hot News) section, which is highlighted with a red box. This section contains a list of news items with their respective categories. Below the featured article is a '焦點新聞' (Focus News) section with a grid of smaller news items. On the right side of the page, there are sections for '中天新聞直播' (CTV News Live) and '生活市集' (Life Market).

**速報** 11:58 美豬協商共識 先辦公聽會

**熱門新聞**

- 罷韓後高雄變冷清？他「1張空拍圖」曝現況 網大驚 - 政治
- 韓國瑜喊「風雨飄搖」藏玄機？高雄議長罕吐真言 - 政治
- iPhone 12系列不送充電器及耳機 充電線也改換一新 - 科技
- 林子祥73歲近況曝光 生日同框愛妻葉蒨文「竟像父女」 - 娛樂
- 馬英九台北市長政績如何？資料一攤開酸民全傻了 - 政治
- 得知韓國瑜確定被罷免 李四川立刻做了這件事 - 政治
- 歌仔戲天王李如麟「前夫家暴女兒不認她」淚嘆：想要家 - 娛樂

**焦點新聞**

- 范瑋琪貼「指責別人最容易」又被酸 網問為何不慶祝國慶 12:11 | 娛樂
- 中天關定、扁都看不下去？ 政院：尊重NCC裁決 12:09 | 政治
- 賣門號給詐團有罪 賣淫集團卻無罪 12:06 | 社會
- 因蔡英文「論文門」案遭控妨害名譽 彭文正請檢官查：求處死刑 12:11 | 政治
- 林右昌基隆捷運漂亮起手勢 台北市長排序出變化 12:12 | 政治
- Gogoro傳內鬼竊電池技術 2工程師投靠新東家遭聲押 12:08 | 社會

**中天新聞直播**

2:00 午間新聞

台北灣、江南大宅 贊助

江南大宅坐鎮交通樞紐  
淡水新市鎮漸漸成為輕移民的首選之一  
[了解詳情](#)

**生活市集**

SMART  
128

# 解析網頁

---

```
607 </div>
608 <div class="col-md-12 col-lg-12 col-xl-4">
609 <section class="hot-news">
610 <header>
611 <h4 class="section-title">熱門新聞</h4>
612 </header>
613 <ul class="vertical-list">
614 <li>
615 <h4 class="title"><a href="https://www.chinatimes.com/realtimenews/20201013005450-260407?ctrack=pc_main_hot_p01">罷韓後高雄變冷清？他「1張空拍圖」曝現況 網大驚 - 政治</a></h4>
616 </li>
617 <li>
618 <h4 class="title"><a href="https://www.chinatimes.com/realtimenews/20201014000915-260412?ctrack=pc_main_hot_p02">iPhone 12系列不送充電器及耳機 充電線也改換一新 - 科技</a></h4>
619 </li>
620 <li>
621 <h4 class="title"><a href="https://www.chinatimes.com/realtimenews/20201014000737-260407?ctrack=pc_main_hot_p03">韓國瑜喊「風雨飄搖」藏玄機？高雄議長罕吐真言 - 政治</a></h4>
622 </li>
623 <li>
624 <h4 class="title"><a href="https://www.chinatimes.com/realtimenews/20201013003240-260404?ctrack=pc_main_hot_p04">林子祥73歲近況曝光 生日同框愛妻葉蒨文「竟像父女」 - 娛樂</a></h4>
625 </li>
626 <li>
627 <h4 class="title"><a href="https://www.chinatimes.com/realtimenews/20201014001824-260407?ctrack=pc_main_hot_p05">馬英九台北市長政績如何？資料一攤開酸民全傻了 - 政治</a></h4>
628 </li>
629 <li>
630 <h4 class="title"><a href="https://www.chinatimes.com/realtimenews/20201014001381-260407?ctrack=pc_main_hot_p06">得知韓國瑜確定被罷免 李四川立刻做了這件事 - 政治</a></h4>
631 </li>
632 <li>
633 <h4 class="title"><a href="https://www.chinatimes.com/realtimenews/20201014001574-260407?ctrack=pc_main_hot_p07">藍委驚爆川普「使出這一招」 台灣就慘了 - 政治</a></h4>
634 </li>
635 </ul>
636 </section>
637 </div>
638 </div>
```

# [課堂活動#5]解析新聞網站，以中時為例

## BS4\_chinatimes

```
# Chinatimes頭條新聞
import requests
from bs4 import BeautifulSoup

# 下載 Chinatimes 首頁內容
url = 'https://www.chinatimes.com/'
r = requests.get(url)

# 確認是否下載成功
if r.status_code == requests.codes.ok:
    # 以 BeautifulSoup 解析 HTML 程式碼
    soup = BeautifulSoup(r.text, 'html.parser')

    # 以 CSS 的 class 抓出各類頭條新聞

    titles = soup.find_all('h2')
    for title in titles:
        print(title.select_one('a').text)
        print(title.select_one('img').src)

# 讀取所有 <b>的標籤內容
#print(soup.find('b'))

#讀取所有<a>的標籤內容
#print(soup.find_all('a'))
```

提示：

用soup.find\_all () 找到所需要的區塊  
用select\_one()指令，取出標題與超連結  
參考網頁BS4指令教學，找出適合的指令

# 執行結果

---

## 成功抓出新聞了

罷韓後高雄變冷清？他「1張空拍圖」曝現況 網大驚 - 政治

[https://www.chinatimes.com/realtimenews/20201013005450-260407?ctrack=mo\\_main\\_hot\\_p01](https://www.chinatimes.com/realtimenews/20201013005450-260407?ctrack=mo_main_hot_p01)

韓國瑜喊「風雨飄搖」藏玄機？高雄議長罕吐真言 - 政治

[https://www.chinatimes.com/realtimenews/20201014000737-260407?ctrack=mo\\_main\\_hot\\_p02](https://www.chinatimes.com/realtimenews/20201014000737-260407?ctrack=mo_main_hot_p02)

林子祥73歲近況曝光 生日同框愛妻葉蒨文「竟像父女」 - 娛樂

[https://www.chinatimes.com/realtimenews/20201013003240-260404?ctrack=mo\\_main\\_hot\\_p03](https://www.chinatimes.com/realtimenews/20201013003240-260404?ctrack=mo_main_hot_p03)

得知韓國瑜確定被罷免 李四川立刻做了這件事 - 政治

[https://www.chinatimes.com/realtimenews/20201014001381-260407?ctrack=mo\\_main\\_hot\\_p04](https://www.chinatimes.com/realtimenews/20201014001381-260407?ctrack=mo_main_hot_p04)

iPhone 12系列不送充電器及耳機 充電線也改換一新 - 科技

[https://www.chinatimes.com/realtimenews/20201014000915-260412?ctrack=mo\\_main\\_hot\\_p05](https://www.chinatimes.com/realtimenews/20201014000915-260412?ctrack=mo_main_hot_p05)

165cm辣模中空薄紗綁事業線 混血美顏引熱議 - 星光 - 時尚

[https://www.chinatimes.com/fashion/20201010000037-263903?ctrack=mo\\_main\\_hot\\_p06](https://www.chinatimes.com/fashion/20201010000037-263903?ctrack=mo_main_hot_p06)

韓國瑜留下爛攤子？黃捷要高雄對「這55位孩子」負責 - 政治

[https://www.chinatimes.com/realtimenews/20201014001466-260407?ctrack=mo\\_main\\_hot\\_p07](https://www.chinatimes.com/realtimenews/20201014001466-260407?ctrack=mo_main_hot_p07)

高雄左營機車撞傷行人棄車逃逸 騎士到案說明

[https://www.chinatimes.com/realtimenews/20201014002827-260402?ctrack=mo\\_main\\_rtime\\_p01](https://www.chinatimes.com/realtimenews/20201014002827-260402?ctrack=mo_main_rtime_p01)

《電子零件》定穎Q3營收回溫 Q4視歐美疫情而定

[https://www.chinatimes.com/realtimenews/20201014002848-260410?ctrack=mo\\_main\\_rtime\\_p02](https://www.chinatimes.com/realtimenews/20201014002848-260410?ctrack=mo_main_rtime_p02)

《國際金融》義大利前任財長 接任裕信銀行CEO

[https://www.chinatimes.com/realtimenews/20201014002847-260410?ctrack=mo\\_main\\_rtime\\_p03](https://www.chinatimes.com/realtimenews/20201014002847-260410?ctrack=mo_main_rtime_p03)

《體育》適合這10款了！北京冬奥会官員買保險 試

# Exercise#1

---

1. 利用指令列印出 Yahoo 首頁原始碼
2. 抓出其他內容練習

# 下載指定網站的圖檔(url\_images.ipnb) 抓取網站的所有圖片，並存入 images 目錄中

```
In [*]: import requests,os
from bs4 import BeautifulSoup
from urllib.request import urlopen

#指定要抓的網址 http://www.ntcu.edu.tw/newweb/index.htm
url = 'http://www.toopen.com/img/87.aspx'

html = requests.get(url)
html.encoding="utf-8"

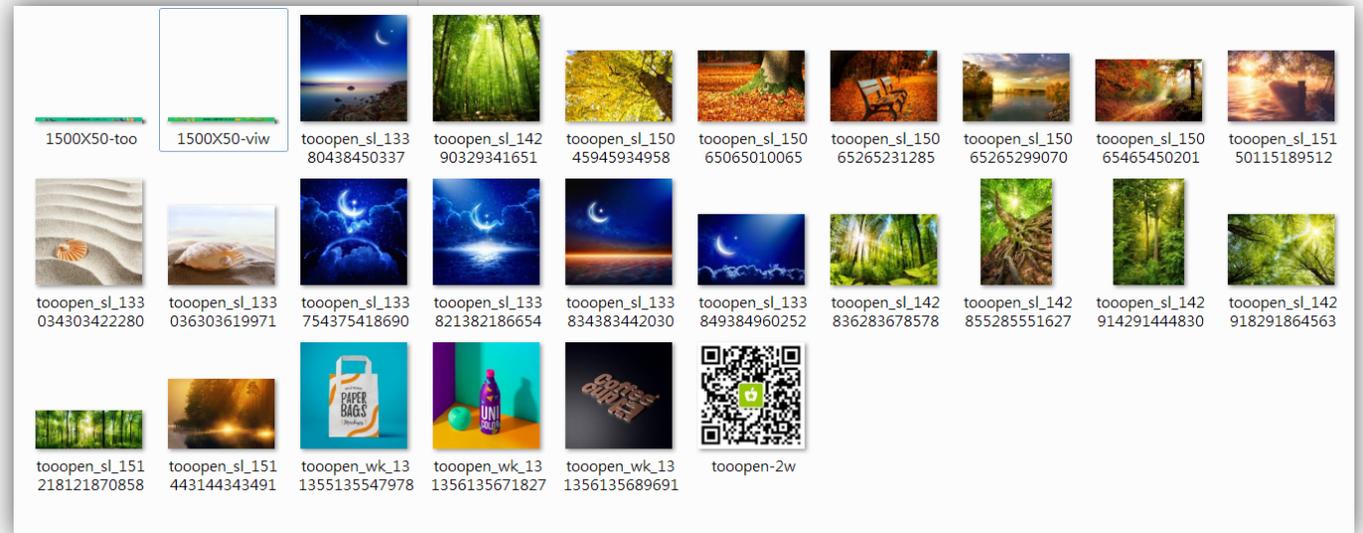
sp = BeautifulSoup(html.text, 'html.parser')

# 建立 images 目錄儲存圖片

# ex. images_dir="images/" images為指定的目錄
images_dir="images"

# 如果目錄不存在，則建立目錄
if not os.path.exists(images_dir):
    os.mkdir(images_dir)

# 取得所有 <a> 和 <img> 標籤
all_links=sp.find_all(['a','img'])
for link in all_links:
    # 讀取 src 和 href 屬性內容
    src=link.get('src')
    href = link.get('href')
    attrs=[src,src]
    for attr in attrs:
        # 讀取 .jpg 和 .png 檔
        if attr != None and ('.jpg' in attr or '.png' in attr):
            # 設定圖檔完整路徑
            full_path = attr
            filename = full_path.split('/')[-1] # 取得圖檔名
            ext = filename.split('.')[-1] #取得副檔名
            filename = filename.split('.')[-2] #取得主檔名
            if 'jpg' in ext: filename = filename + '.jpg'
            else: filename = filename + '.png'
            print(attr)
            # 儲存圖片
            try:
                image = urlopen(full_path)
                f = open(os.path.join(images_dir,filename),'wb')
                f.write(image.read())
                f.close()
            except:
                print("{} 無法讀取!".format(filename))
```



# 正規表示式與網頁擷取

---

# 使用正規表示式擷取網頁內容

## 常見的正規表示式

正規表示式	功能說明
.	代表一個除了換列字元 (\n) 以外的所有字元
^	代表輸入列的開始
\$	代表輸入列的結束
*	代表前一個項目可以出現 0 次或無限多次
+	代表前一個項目可以出現 1 次或無限多次
?	代表前一個項目可以出現 0 次或 1 次
[abc]	代表一個符合 a 或 b 或 c 的任何字元
[a-z]	代表一個符合 a、b、c ~z 的任何字元

正規表示式	功能說明
\	代表後面的字元以一般字元處理
{m}	代表前一個項目必須正好出現 m 數
{m,}	代表前一個項目出現次數最少 m 次，最多無限次。
{m,n}	代表前一個項目出現次數最少 m 次，最多 n 次。
\d	一個數字字元，相當於 [0123456789] 或 [0-9]。
^	反運算，例如：[^a-d] 代表除了 a、b、c、d 外的所有字元。
\D	一個非數字字元，相當於 [^0-9]。
\n	換列字元
\r	回列首字元 (carriage return)
\t	tab 定位字元
\s	空白字元，相當於 [\r\t\n\f]。
\S	非空白字元，相當於 [^\r\t\n\f]。
\w	一個數字、字母或底線字元，相當於 [0-9a-zA-Z_]。
\W	一個非數字、字母或底線字元，相當於 [^\w]，即 [^0-9a-zA-Z_]。

# 正規表示式的範例

語法	正規表示式	範例
整數	<code>[0-9]+</code>	33025
有小數點的實數	<code>[0-9]+\.[0-9]+</code>	75.93
英文詞彙	<code>[A-Za-z]+</code>	Python
變數名稱	<code>[A-Za-z_][A-Za-z0-9_]*</code>	<code>_pointer</code>
Email	<code>[a-zA-Z0-9_]+@[a-zA-Z0-9\._]+</code>	guest@kimo.com.tw
URL	<code>http://[a-zA-Z0-9\./_]+</code>	http://e-happy.com.tw/

# 建立正規表示式物件

語法範例：

---

```
import re
pat = re.compile('[a-z]+')
```

正規表示式物件提供下列的方法：

方法	說明
<code>match(string)</code>	傳回指定的字串中符合正規表示式的字串，直到不符合字元為止，並把結果存入 <code>match</code> 物件 ( <code>object</code> ) 之中；若無符合字元，傳回 <code>None</code> 。
<code>search(string)</code>	傳回指定的字串中第一組符合正規表示式的字串，並把結果存入 <code>match</code> 物件 ( <code>object</code> ) 之中；若無符合字元，傳回 <code>None</code> 。
<code>findall()</code>	傳回指定的字串中所有符合正規表示式的字串，並傳回一個串列。

# Google 查詢

---

# 觀察一下原始碼 (注意紅色框框)

```
<div>
  <div class="ZINbbc xpd 09g5cc uUPGi">
    <div class="kCrYT">
      <a href="/url?q=https://www.chinatimes.com/realtimenews/20181229002055-260405&sa=U&ved=2ahUKEwi4tJmD-Iz1AhWPw4sBHbxuDecQFjAEegQICRAB&usg=AOvVaw0GXvAFMBJVHRXT-T_3JtoR">
        <div class="BNeawe vvjwJb AP7Wnd">
          寒流、冷氣團傻傻分不清？一張圖秒懂六大冷空氣- 生活- 中時電子報
        </div>
        <div class="BNeawe UPmit AP7Wnd">
          https://www.chinatimes.com › 生活
        </div>
      </a>
    </div>
    <div class="x54gtf">
    </div>
    <div class="kCrYT">
      <div>
        <div class="BNeawe s3v9rd AP7Wnd">
          <div>
          <div>

```

# 程式碼

```
import requests
from bs4 import BeautifulSoup

# Google 搜尋 URL
google_url = 'https://www.google.com.tw/search'

# 查詢參數
my_params = {'q': '寒流'}

# 下載 Google 搜尋結果
r = requests.get(google_url, params = my_params)

# 確認是否下載成功
if r.status_code == requests.codes.ok:
    # 以 BeautifulSoup 解析 HTML 原始碼
    soup = BeautifulSoup(r.text, 'html.parser')

    # 觀察 HTML 原始碼
    print(soup.prettify())

    # 以 CSS 的選擇器來抓取 Google 的搜尋結果
    items = soup.select('div.kCrYT > a[href^="/url"]')
    for i in items:
        # 標題
        print("標題：" + i.text)
        # 網址
        print("網址：" + i.get('href'))
```

你想要查詢的關鍵字

自訂之CSS選擇器  
須依據不同網頁設計

google.com.tw/search?q=寒流

Prof. Chih-Hung... Google 地圖 Google Journal 2 Google 日曆 Meeting - Google...

寒流

全部 新聞 圖片 影片 地圖 更多 設定 工具

約有 7,830,000 項結果 (搜尋時間：0.33 秒)

**寒流、冷氣團怎麼分？3分鐘讓你變冷天氣大師#低溫(140563 ...**  
<https://www.cool3c.com> > 科學新知 ▾  
2019年1月21日 - 民眾時常可以從氣象預報或新聞報導中看到「寒流」、「冷氣團」等名詞，但是你知道其中的差別在哪裡嗎？當你看到「最強寒流來襲」或「最強冷氣團...

**寒流、冷氣團傻傻分不清？一張圖秒懂六大冷空氣- 生活- 中時 ...**  
<https://www.chinatimes.com> > 生活 ▾  
2018年12月29日 - 臉書《台灣颱風論壇 | 天氣特急》在臉書分享一張圖，用飲料的少冰、微冰、去冰、完全去冰等來分類，要讓民眾秒懂什麼是「寒流」、什麼是「冷氣團」。

**天氣即時預報- 入冬首波寒流！週四(9日)起，全台寒冷，低溫下探 ...**  
<https://www.facebook.com> > weather.taiwan > posts ▾  
2017年2月6日 - 但從目前預測資料看來，週四(9日)開始，今年冬天第一波寒流就要南下，尤其當晚，850hPa之0°C線就涵蓋北部，顯示這一波冷空氣是「快、強、冷」，...

**最強寒流來襲| 三立新聞網**  
<https://www.setn.com> > project ▾  
入冬最強寒流來襲，氣象局發布低溫特報，全台各地陸續刷新最新低溫紀錄！

# 透過print(soup.prettify())列印 檢視原始碼

用prettify()函數將soup這個物件美化，soup就是的html\_doc解析的結果，最後print將它印出來：

```
# Google 搜尋 URL
google_url = 'https://www.google.com.tw/search'

# 查詢參數
my_params = {'q': '寒流'}

# 下載 Google 搜尋結果
r = requests.get(google_url, params = my_params)

# 確認是否下載成功
if r.status_code == requests.codes.ok:
    # 以 BeautifulSoup 解析 HTML 原始碼
    soup = BeautifulSoup(r.text, 'html.parser')

# 觀察 HTML 原始碼
print(soup.prettify())

# 以 CSS 的選擇器來抓取 Google 的搜尋結果
items = soup.select('div.kCrYT > a[href^="/url"]')
for i in items:
    # 標題
    print("標題：" + i.text)
    # 網址
    print("網址：" + i.get('href'))
```

解析後的網頁原始檔

```
<!DOCTYPE doctype html>
<html lang="zh-TW">
  <head>
    <meta charset="utf-8"/>
    <meta content="/images/branding/googleg/1x/googleg_standard_color_128dp.png" itemprop="image"/>
    <title>
      寒流 - Google 搜尋
    </title>
    <script nonce="Dm3rNLn00rq6luQDQP3sAw==">
      (function(){var a=window.performance;window.start=(new Date).getTime();a:{var b=window;if(a){var c=a.timing;if(c){var d=c.navigationStart,e=c.responseStart;if(e>d&&e<=window.start){window.start=e;b.wsrt=e-d;break a}}a.now&&(b.wsrt=Math.floor(a.now()))}window.google=window.google||{};google.aft=function(f){f.setAttribute("data-impl",+new Date)};}).call(this);(function(){window.jsarwt=function(){return!1}}).call(this);(function(){var c=[],e=0;window.ping=function(b){-1==b.indexOf("&zx")&&(b+="%zx="+new Date().getTime());var a=new Image,d=e++;c[d]=a;a.onerror=a.onload=a.onabort=function(){delete c[d]};a.src=b};}).call(this);
    </script>
    <style>
      body{margin:0 auto;max-width:736px;padding:0 8px;a{color:#1967D2;text-decoration:none;tap-highlight-color:rgba(0,0,0,.1)}a:visited{color:#4B11A8}a:hover{text-decoration:underline}img{border:0}html{font-family:Roboto,HelveticaNeue,Arial,sans-serif;font-size:14px;line-height:20px;text-size-adjust:100%;color:#3C4043;word-wrap:break-word;background-color:#fff;hBswnc/hac
```

# 並設計一個萃取資料的規則 使用一個配合網頁設計規則的CSS檢視器

語法:

```
items=soup.select('div.kCrYT > a[href^="/url"]')
```

表示會先找出div名稱為 kCrYT的class，然後再抓a href超連結出來

```
<div>
  <div class="ZINbbc xpd 09g5cc uUPGi">
    <div class="kCrYT">
      <a href="/url?q=https://www.chinatimes.com/realtimenews/20181229002055-260405&sa=U&ved=2ahUKEwi4tJmD-Iz1AhWPw4SBHbxuDecQFjAEegQICRAB&usg=AOvVaw0GXvAfMBJVHRXT-T 3JtoR">
        <div class="BNeawe vvjwJb AP7Wnd">
          寒流、冷氣團傻傻分不清？一張圖秒懂六大冷空氣- 生活- 中時電子報
        </div>
        <div class="BNeawe UPmit AP7Wnd">
          https://www.chinatimes.com > 生活
        </div>
      </a>
    </div>
  <div class="x54gtf">
  </div>
  <div class="kCrYT">
    <div>
      <div class="BNeawe s3v9rd AP7Wnd">
        <div>
        </div>
      </div>
    </div>
  </div>
```

按下CTRL+U檢視原始碼

# 最後結果

標題：寒流、冷氣團怎麼分？3分鐘讓你變冷天氣大師#低溫(140563) - 癮科技 ...<https://www.cool3c.com> > 科學新知

網址：[/url?q=https://www.cool3c.com/article/140563&sa=U&ved=2ahUKEwjD6ciy\\_Iz1AhU2y4sBHZVEAd4QFjADegQICAB&usg=A0vVaw3b2D0ivah5Rzn5vAfq5MVV](https://www.cool3c.com/article/140563&sa=U&ved=2ahUKEwjD6ciy_Iz1AhU2y4sBHZVEAd4QFjADegQICAB&usg=A0vVaw3b2D0ivah5Rzn5vAfq5MVV)

標題：寒流、冷氣團傻傻分不清？一張圖秒懂六大冷空氣- 生活- 中時電子報<https://www.chinatimes.com> > 生活

網址：[/url?q=https://www.chinatimes.com/realtimenews/20181229002055-260405&sa=U&ved=2ahUKEwjD6ciy\\_Iz1AhU2y4sBHZVEAd4QFjAEegQICRAB&usg=A0vVaw2CC2MnGzJkZOVc9M2mTk0k](https://www.chinatimes.com/realtimenews/20181229002055-260405&sa=U&ved=2ahUKEwjD6ciy_Iz1AhU2y4sBHZVEAd4QFjAEegQICRAB&usg=A0vVaw2CC2MnGzJkZOVc9M2mTk0k)

標題：天氣即時預報- 入冬首波寒流！週四(9日)起，全台寒冷，低溫下探10°C ...<https://www.facebook.com> > weather.taiwan > posts

網址：[/url?q=https://www.facebook.com/weather.taiwan/posts/1276103999112022/&sa=U&ved=2ahUKEwjD6ciy\\_Iz1AhU2y4sBHZVEAd4QFjAFegQICBAB&usg=A0vVaw0-luqtSJPVu4fKltjxQwzu](https://www.facebook.com/weather.taiwan/posts/1276103999112022/&sa=U&ved=2ahUKEwjD6ciy_Iz1AhU2y4sBHZVEAd4QFjAFegQICBAB&usg=A0vVaw0-luqtSJPVu4fKltjxQwzu)

標題：最強寒流來襲| 三立新聞網<https://www.setn.com> > project

網址：[/url?q=https://www.setn.com/project.aspx%3FProjectID%3D1060&sa=U&ved=2ahUKEwjD6ciy\\_Iz1AhU2y4sBHZVEAd4QFjAGegQIBhAB&usg=A0vVaw3ZpksJ07g6yCZiZKrf1RlH](https://www.setn.com/project.aspx%3FProjectID%3D1060&sa=U&ved=2ahUKEwjD6ciy_Iz1AhU2y4sBHZVEAd4QFjAGegQIBhAB&usg=A0vVaw3ZpksJ07g6yCZiZKrf1RlH)

標題：2016年1月北半球寒流- 維基百科，自由的百科全书<https://zh.wikipedia.org> > zh-tw > 2016年1月北半球寒流

網址：[/url?q=https://zh.wikipedia.org/zh-tw/2016%25E5%25B9%25B41%25E6%259C%2588%25E5%258C%2597%25E5%258D%258A%25E7%2590%2583%25E5%25AF%2592%25E6%25B5%2581&sa=U&ved=2ahUKEwjD6ciy\\_Iz1AhU2y4sBHZVEAd4QFjAHegQIBxAB&usg=A0vVaw0UaKp6VItWgbvBP3weg6x](https://zh.wikipedia.org/zh-tw/2016%25E5%25B9%25B41%25E6%259C%2588%25E5%258C%2597%25E5%258D%258A%25E7%2590%2583%25E5%25AF%2592%25E6%25B5%2581&sa=U&ved=2ahUKEwjD6ciy_Iz1AhU2y4sBHZVEAd4QFjAHegQIBxAB&usg=A0vVaw0UaKp6VItWgbvBP3weg6x)

標題：氣象局一張圖來了！未來一周降溫但寒流不會來| 生活新聞| 生活| 聯合 ...<https://udn.com> > 生活

網址：[/url?q=https://udn.com/news/story/7266/3514053&sa=U&ved=2ahUKEwjD6ciy\\_Iz1AhU2y4sBHZVEAd4QFjAIegQIBBAB&usg=A0vVaw1DB28Uhajkbn21LrwwXyRB](https://udn.com/news/story/7266/3514053&sa=U&ved=2ahUKEwjD6ciy_Iz1AhU2y4sBHZVEAd4QFjAIegQIBBAB&usg=A0vVaw1DB28Uhajkbn21LrwwXyRB)

標題：溫度破紀錄？氣象局：今年全季沒寒流！是1947年來最暖冬季| 生活| 新 ...<https://newtalk.tw> > news > view

網址：[/url?q=https://newtalk.tw/news/view/2019-03-05/215262&sa=U&ved=2ahUKEwjD6ciy\\_Iz1AhU2y4sBHZVEAd4QFjAJegQIAhAB&usg=A0vVaw2U](https://newtalk.tw/news/view/2019-03-05/215262&sa=U&ved=2ahUKEwjD6ciy_Iz1AhU2y4sBHZVEAd4QFjAJegQIAhAB&usg=A0vVaw2U)

# Exercise#2

---

設定一個關鍵字（ex.流感疫苗），  
並抓取Google查詢的結果與連結

# 網頁分析與擷取技術

---

台彩中獎號碼擷取

# 威力彩 程式 抓取最新一期的中獎號碼

```
import requests  
from bs4 import BeautifulSoup
```

```
url = 'http://www.taiwanlottery.com.tw/'  
html = requests.get(url)  
sp = BeautifulSoup(html.text, 'html.parser')  
#print(sp.prettify())  
data1 = sp.select("#rightdown")  
#print(data1)
```

先把網頁印出來分析

```
data2 = data1[0].find('div', {'class': 'ball_tx ball_green'})  
#print(data2)  
data3 = data2.find_all('div', {'class': 'ball_tx ball_green'})  
print(data3)
```

指定印出特定class區塊

```
[<div class="ball_tx ball_green">29 </div>, <div class="ball_tx ball_green">27 </div>, <div class="ball_tx ball_green">15 </div>, <div class="ball_tx ball_green">05 </div>, <div class="ball_tx ball_green">31 </div>, <div class="ball_tx ball_green">36 </div>, <div class="ball_tx ball_green">05 </div>, <div class="ball_tx ball_green">15 </div>, <div class="ball_tx ball_green">27 </div>, <div class="ball_tx ball_green">29 </div>, <div class="ball_tx ball_green">31 </div>, <div class="ball_tx ball_green">36 </div>]
```

```
# 威力彩號碼  
print("開出順序:", end="")  
for n in range(0,6):  
    print(data3[n].text, end=" ")
```

```
print("\n大小順序:", end="")  
for n in range(6, len(data3)):  
    print(data3[n].text, end=" ")
```

```
## 第二區  
red = data2.find('div', {'class': 'ball_tx ball_green'})  
print("\n第二區: {}".format(red.text))
```

```
開出順序: 29 27 15 05 31 36  
大小順序: 05 15 27 29 31 36  
第二區: 04
```

預期結果

## 威力彩

### 本期中獎號碼

遊戲	威力彩
期別	第 <b>108000080</b> 期
開獎日期	民國 <b>108</b> 年 <b>10</b> 月 <b>7</b> 日
兌獎期限	至 <b>109</b> 年 <b>1</b> 月 <b>7</b> 日 (註3)
本期中獎號碼	第1區 依大小順序排列: <b>05 15 27 29 31 36</b> 第2區: <b>04</b> 第1區 依開出順序排列: <b>29 27 15 05 31 36</b>

先分析看看中獎號碼  
會在哪一個區段中

# 網頁解析

```
In [1]: import requests
        from bs4 import BeautifulSoup

        url = 'http://www.taiwanlottery.com.tw/'
        html = requests.get(url)
        sp = BeautifulSoup(html.text, 'html.parser')
        #print(sp.prettify())
        data1 = sp.select("#rightdown")
        print(data1)

        data2 = data1[0].find('div', {'class': '██████████'})
        #print(data2)

        data3 = data2.find_all('div', {'class': '██████████'})
        #print(data3)
```

```
[<div id="rightdown">
<!--*****BINGO BINGO*****-->
<div class="contents_box01">
<div id="contents_logo_01"> </div><div class="contents_mine_tx01"><span class="font_black15">108/10/16 第108058525期 </span>
<span class="font_red14"><a href="/Lotto/BingoBingo/history.aspx">開獎結果</a> | <a href="/Lotto/BingoBingo/drawing.aspx">各
期獎號查詢</a><a href="/lotto/BingoBingo/OEHLStatistic.htm"><div id="contents_logo_01-2"></div></a></span></div><div class="c
ontents_mine_tx04">開出獎號</div><div class="ball_box01"><div class="ball_tx ball_yellow">02 </div><div class="ball_tx ball_y
ellow">04 </div><div class="ball_tx ball_yellow">10 </div><div class="ball_tx ball_yellow">13 </div><div class="ball_tx ball
_yellow">15 </div><div class="ball_tx ball_yellow">20 </div><div class="ball_tx ball_yellow">25 </div><div class="ball_tx ba
ll_yellow">34 </div><div class="ball_tx ball_yellow">36 </div><div class="ball_tx ball_yellow">42 </div><div class="ball_tx
ball_yellow">47 </div><div class="ball_tx ball_yellow">55 </div><div class="ball_tx ball_yellow">58 </div><div class="ball_t
x ball_yellow">67 </div><div class="ball_tx ball_yellow">69 </div><div class="ball_tx ball_yellow">70 </div><div class="ball
_tx ball_yellow">71 </div><div class="ball_tx ball_yellow">72 </div><div class="ball_tx ball_yellow">78 </div><div class="ba
ll_tx ball_yellow">79 </div></div><div class="contents_mine_tx08"> 超級<br/> 獎號<br/><div class="ball_red">58</div></div><
div class="contents_mine_tx08"> 猜<br/> 大小<br/><div class="ball_blue_BB1">-</div></div><div class="contents_mine_tx0
8"> 猜<br/> 單 雙<br/><div class="ball_blue_BB2">小雙</div></div>
</div>
<div class="dotted01"></div>
<!--*****雙贏彩區塊*****-->
<div class="contents_box06">
```

Try it!

試著分析網頁，並找出其規則  
下一頁有解答，先不要看喔

---

# 威力彩 程式 抓取最新一期的中獎號碼

```
import requests
from bs4 import BeautifulSoup

url = 'http://www.taiwanlottery.com.tw/'
html = requests.get(url)
sp = BeautifulSoup(html.text, 'html.parser')
#print(sp.prettify())
data1 = sp.select("#rightdown")
#print(data1)

data2 = data1[0].find('div', {'class': 'ball_tx ball_green'})
#print(data2)

data3 = data2.find_all('div', {'class': 'ball_tx ball_green'})
print(data3)
```

```
[<div class="ball_tx ball_green">29 </div>, <div class="ball_tx ball_green">27 </div>, <div class="ball_tx ball_green">15 </div>, <div class="ball_tx ball_green">05 </div>, <div class="ball_tx ball_green">31 </div>, <div class="ball_tx ball_green">36 </div>, <div class="ball_tx ball_green">05 </div>, <div class="ball_tx ball_green">15 </div>, <div class="ball_tx ball_green">27 </div>, <div class="ball_tx ball_green">29 </div>, <div class="ball_tx ball_green">31 </div>, <div class="ball_tx ball_green">36 </div>]
```

```
# 威力彩號碼
print("開出順序:", end=" ")
for n in range(0,6):
    print(data3[n].text, end=" ")

print("\n大小順序:", end=" ")
for n in range(6, len(data3)):
    print(data3[n].text, end=" ")

## 第二區
red = data2.find('div', {'class': 'ball_red'})
print("\n第二區: {}".format(red.text))
```

```
開出順序: 29 27 15 05 31 36
大小順序: 05 15 27 29 31 36
第二區: 04
```

## 威力彩

### 本期中獎號碼

遊戲	威力彩
期別	第 <b>108000080</b> 期
開獎日期	民國 <b>108</b> 年 <b>10</b> 月 <b>7</b> 日
兌獎期限	至 <b>109</b> 年 <b>1</b> 月 <b>7</b> 日 (註3)
本期中獎號碼	第1區 依大小順序排列: <b>05 15 27 29 31 36</b> 第2區: <b>04</b> 第1區 依開出順序排列: <b>29 27 15 05 31 36</b>

---

思考一下這個能創造  
甚麼商業模式 & APP?

# Ex.

設計一個APP，能夠每天顯示各大新聞網站的頭條新聞。

統計新聞的熱門關鍵字



# 參考來源

---

文淵閣工作室，Python初學特訓班，碁峰。

網頁資料

<https://blog.gtwang.org/programming/python-beautiful-soup-module-scrape-web-pages-tutorial/>