Python 爬蟲

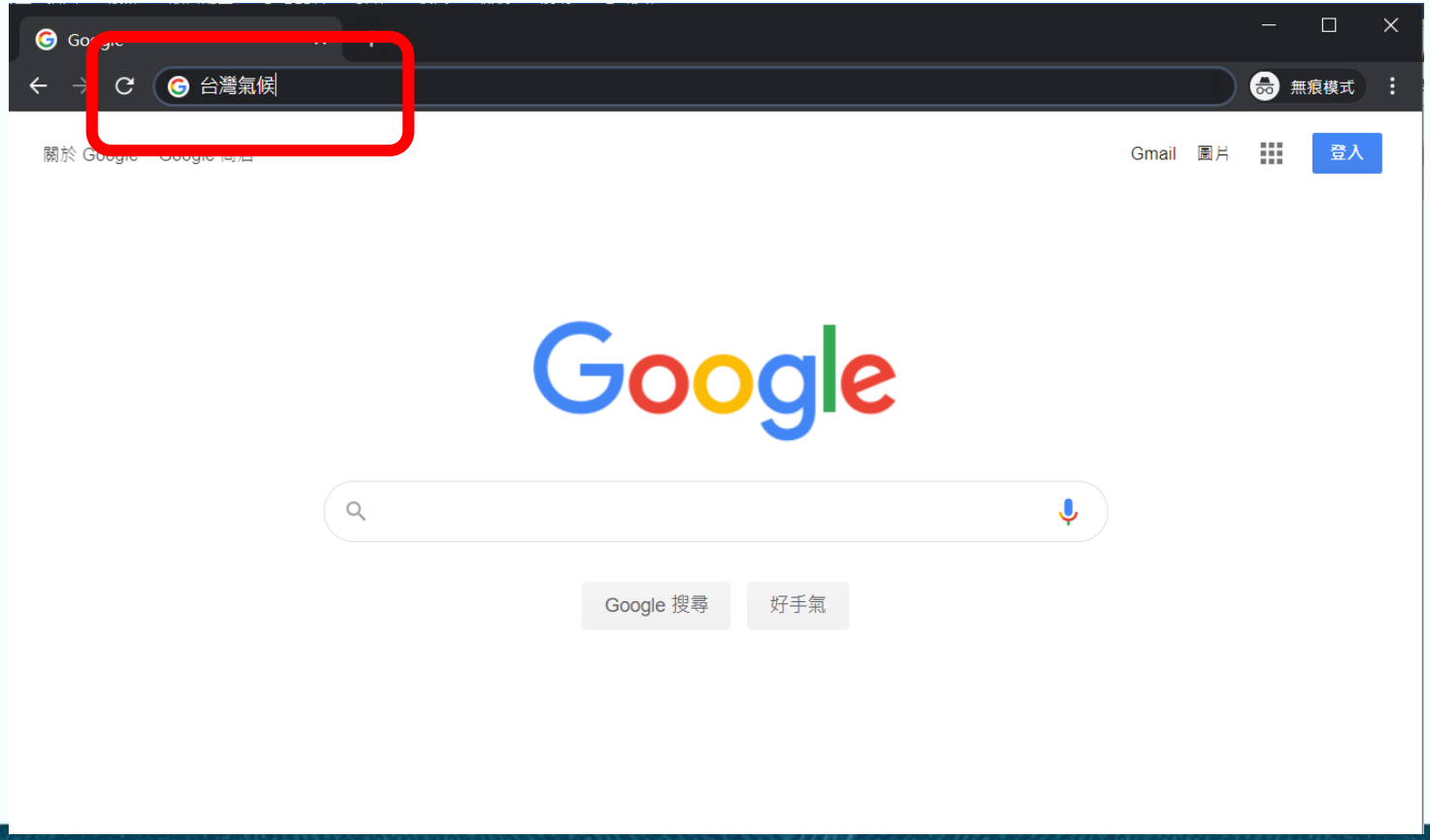
報告人 ： 王彤云

什麼是爬蟲？

- **網路爬蟲**（英語：web crawler），也叫**網路蜘蛛**（spider），是一種用來自動瀏覽全球資訊網的網路機器人
- 大數據時代，網路上有充足的資料
- 學習爬蟲，可以讓我們獲取更多的資料來源，並且這些資料來源可以按我們的目的進行採集，去掉很多無關資料

網路爬蟲必備基本的最基本知識“GET”請求

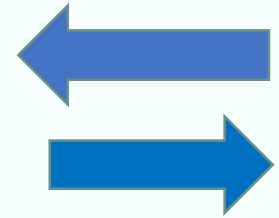
●在瀏覽器(如：Chrome)的網址列輸入文字，按下Enter鍵



一般使用者瀏覽網路的行為：

1. 瀏覽器發出Get請求！
https://www.youtube.com/?gl=TW&hl=zh-TW

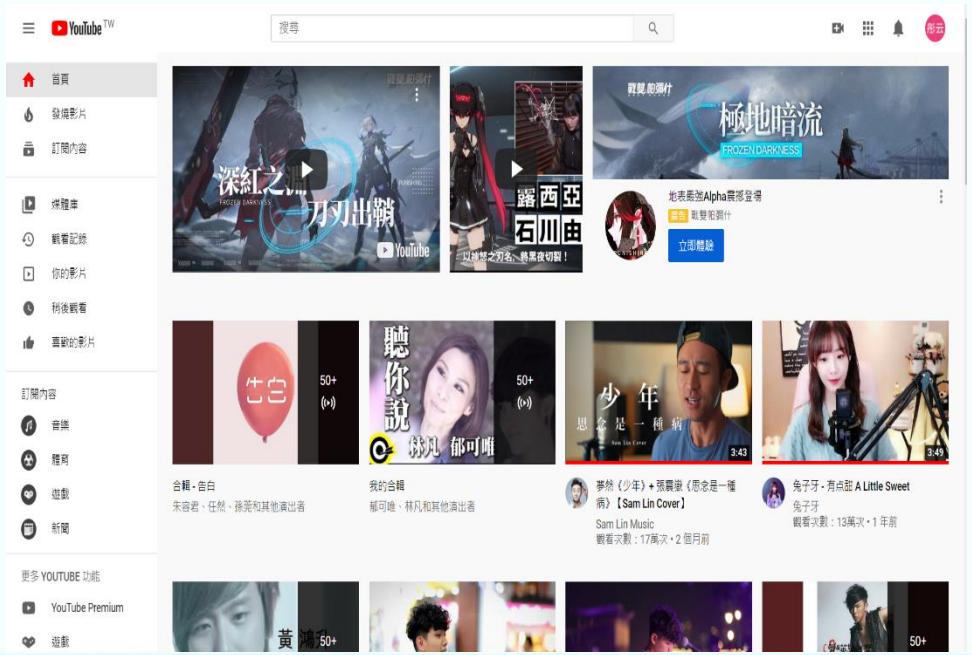
網頁伺服器



2. 回傳網頁資料

```
<!DOCTYPE html>
<html style="font-size: 10px;font-family: Roboto, Arial, sans-serif;" lang="zh-TW" dir="ltr" gl="TW">
▶<head>…</head>
…▼<body dir="ltr"> == $0
  ▶<ytd-app mini-guide-visible_>…</ytd-app>
  <script>if (window.ytcsi) {window.ytcsi.tick("gcc", null, '');}</script>
  <script>
      window['ytInitialGuideDataPresent'] = true;
    </script>
  <script>if (window.ytcsi) {window.ytcsi.tick("nc_pj", null, '');}</script>
  <script src="https://www.youtube.com/s/desktop/aa71f599/jsbin/www-i18n-constants-zh_TW.vflset/www-i18n-constants.js" type="text/javascript" name="www-i18n-constants/www-i18n-constants" class="js-httpswwwyoutubecomsdesktopaa71f599jsbinwwwi18nconstantszh_TWvflsetwwwi18nconstantsjs"></script>
  <script>if (window.ytcsi) {window.ytcsi.tick("rsbe_dpj", null, '');}
  </script>
  <script src="https://www.youtube.com/s/desktop/aa71f599/jsbin/desktop_polymer_in…d_html_polymer_flags.vflset/desktop_polymer_inlined_html_polymer_flags.js" type="text/javascript" name="desktop_polymer_inlined_html_polymer_flags/desktop_polymer_inlined_html_polymer_flags" class="js-
```

3. 呈現在網頁上

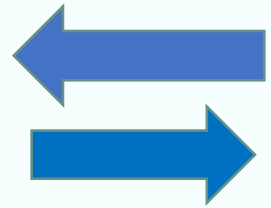


Python 網路爬蟲瀏覽網路的行為：

1. Python 發送Get請求!

requests.get('https://www.youtube.com/?gl=TW&hl=zh-TW')

網頁伺服器



2. 回傳網頁資料

3. 透過python 分析

```
<!DOCTYPE html>
<html style="font-size: 10px;font-family: Roboto, Arial, sans-serif;" lang="zh-TW" dir="ltr" gl="TW">
▶<head>…</head>
…▼<body dir="ltr"> == $0
  ▶<ytd-app mini-guide-visible_>…</ytd-app>
   <script>if (window.ytcsi) {window.ytcsi.tick("gcc", null, '');}</script>
   <script>
       window['ytInitialGuideDataPresent'] = true;
     </script>
   <script>if (window.ytcsi) {window.ytcsi.tick("nc_pj", null, '');}</script>
   <script src="https://www.youtube.com/s/desktop/aa71f599/jsbin/www-i18n-constants-zh_TW.vflset/www-i18n-constants.js" type="text/javascript" name="www-i18n-constants/www-i18n-constants" class="js-httpswwwyoutubecomsdesktopaa71f599jsbinwwwi18nconstantszh_TWvflsetwwwi18nconstantsjs"></script>
   <script>if (window.ytcsi) {window.ytcsi.tick("rsbe_dpj", null, '');}
   </script>
   <script src="https://www.youtube.com/s/desktop/aa71f599/jsbin/desktop_polymer_in…d_html_polymer_flags.vflset/desktop_polymer_inlined_html_polymer_flags.js" type="text/javascript" name="desktop_polymer_inlined_html_polymer_flags/desktop_polymer_inlined_html_polymer_flags" class="js-
```

import requests套件包

```
[21] import requests
```

GET請求抓取資料

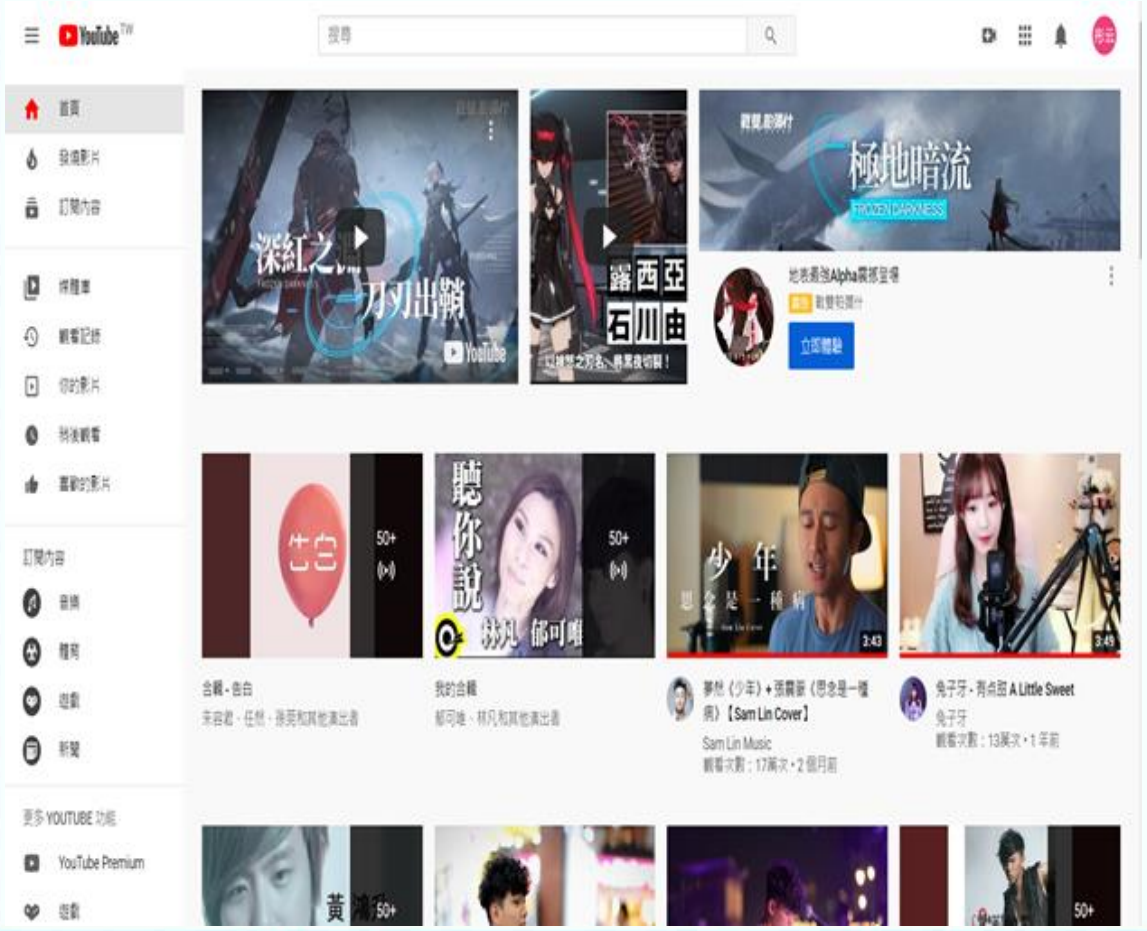
```
[22] page = requests.get('https://www.youtube.com/?gl=TW&hl=zh-TW')
```

列印出網頁程式碼

```
print(page.text)
```

人 VS 爬蟲所見

- 人們所見



- 爬蟲所見

```
<!DOCTYPE html>
<html style="font-size: 10px;font-family: Roboto, Arial, sans-serif;" lang="zh-TW" dir="ltr" gl="TW">
▶<head>…</head>
…▼<body dir="ltr"> == $0
  ▶<ytd-app mini-guide-visible_>…</ytd-app>
  <script>if (window.ytcsi) {window.ytcsi.tick("gcc", null, '');}</script>
  <script>
      window['ytInitialGuideDataPresent'] = true;
    </script>
  <script>if (window.ytcsi) {window.ytcsi.tick("nc_pj", null, '');}</script>
  <script src="https://www.youtube.com/s/desktop/aa71f599/jsbin/www-i18n-constants-zh_TW.vflset/www-i18n-constants.js" type="text/javascript" name="www-i18n-constants/www-i18n-constants" class="js-httpswwwyoutubecomsdesktopaa71f599jsbinwwwi18nconstantszh_TWvflsetwwwi18nconstantsjs"></script>
  <script>if (window.ytcsi) {window.ytcsi.tick("rsbe_dpj", null, '');}
  </script>
  <script src="https://www.youtube.com/s/desktop/aa71f599/jsbin/desktop_polymer_in…d_html_polymer_flags.vflset/desktop_polymer_inlined_html_polymer_flags.js" type="text/javascript" name="desktop_polymer_inlined_html_polymer_flags/desktop_polymer_inlined_html_polymer_flags" class="js-
```

結論

- Python 網路爬蟲只是模擬使用者操作瀏覽器的行為
- 透過Get 請求可以向網頁伺服器請求資料
- 收到的**資料是網頁程式碼(HTML語法)**

程式碼嘗試

```
# import requests 套件包
    import requests

# GET請求抓取資料
page =  requests.get( ‘ 網址‘)


# 列印出網頁程式碼
  print(page.text)
```

實作

爬取「ETtoday旅遊雲」網頁，
擷取台北旅遊景點的標題資訊

使用軟體？

在開發的過程中，常會需要搜尋HTML的節點，分享幾個常用的方法，包含:

1. BeautifulSoup安裝
2. 以HTML標籤及屬性搜尋節點
3. 以CSS屬性搜尋節點
4. 搜尋父節點
5. 搜尋前、後節點
6. 取得屬性值
7. 取得連結文字

一、BeautifulSoup安裝

```
pip  install  beautifulsoup4  #安裝Beautifulsoup套件(Package)
```

```
Requirement already satisfied: beautifulsoup4 in /usr/local/lib/python3.7/dist-packages (4.6.3)
```

```
[ ] pip  install  requests    #安裝Python的requests套件(Package)，將要爬取的網頁HTML程式碼取回來
```

```
Requirement already satisfied: requests in /usr/local/lib/python3.7/dist-packages (2.23.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist-packages (from requests) (2021.5.30)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-packages (from requests) (3.0.4)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (from requests) (2.10)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.7/dist-packages (from requests) (1.24.3)
```

```python
import  requests
from  bs4  import  Beautif

#引用requests套件(Package)
response  =  requests.get(

#引用BeautifulSoup類別(Cla
soup  =  BeautifulSoup(res

#輸出排版後的HTML內容
print(soup.prettify())
```

```
<div class="block_content sidebar-newest-news">
 <div class="inner">
  <!--part_list_1 文字列表 開始-->
  <div class="part_list_1">
   <h2>
    <a href="/article/2055091.htm" title="飛天掃帚＋無敵海景！基隆海濱公園">
     飛天掃帚＋無敵海景！基隆海濱公園
    </a>
   </h2>
   <h2>
    <a href="/article/2045346.htm" title="慕斯冰入口化開！高雄療癒「冰泡芙」">
     慕斯冰入口化開！高雄療癒「冰泡芙」
    </a>
   </h2>
   <h2>
    <a href="/article/2055601.htm" title="美如攝影棚！宜蘭無死角歐洲莊園民宿">
     美如攝影棚！宜蘭無死角歐洲莊園民宿
    </a>
   </h2>
   <h2>
    <a href="/article/2056265.htm" title="民俗專曝家鬼月13大忌　小心別被抓交替">
     民俗專曝家鬼月13大忌　小心別被抓交替
    </a>
   </h2>
   <h2>
    <a href="/article/2056247.htm" title="花近萬元住到發霉帳篷！她退房被收錢">
     花近萬元住到發霉帳篷！她退房被收錢
    </a>
   </h2>
```

二、以HTML標籤及屬性搜尋節點

```
[ ] #搜尋第一個符合條件的HTML節點，傳入要搜尋的標籤名稱
    result = soup.find("h3")

    print(result)
```

```
<h3 itemprop="headline">
<a href="https://travel.ettoday.net/article/2054494.htm" itemprop="url" title="免跑花東！竹子湖「2處金針花海」滿開　採滿回家煮才200元">免跑花東！竹子湖「2處金針花海」滿開　採滿回家煮才200元</a>
</h3>
```

```
#搜尋網頁中所有符合條件的HTML節點，傳入要搜尋的HTML標籤名稱。itemprop屬性值為headline的節點。利用limit關鍵字參數(Keyword  Argument)限制搜尋的節點數量
result  =  soup.find_all("h3",itemprop="headline",limit=3)
print(result)
```

```
[<h3 itemprop="headline">
<a href="https://travel.ettoday.net/article/2054494.htm" itemprop="url" title="免跑花東！竹子湖「2處金針花海」滿開　採滿回家煮才200元">免跑花東！竹子湖「2處金針花海」滿開　採滿回家煮才200元</a>
</h3>, <h3 itemprop="headline">
<a href="https://travel.ettoday.net/article/2053454.htm" itemprop="url" title="水手服氣泡飲、符號杯子蛋糕太Q！全台首間「美少女戰士咖啡廳」來了">水手服氣泡飲、符號杯子蛋糕太Q！全台首間「美少女戰士咖啡廳」來了</a>
</h3>, <h3 itemprop="headline">
<a href="https://travel.ettoday.net/article/2053408.htm" itemprop="url" title="最新必比登！34年「巷子龍家常菜」首入榜　必點全台最好吃宮保雞丁">最新必比登！34年「巷子龍家常菜」首入榜　必點全台最好吃宮保雞丁</a>
</h3>]
```

```
#搜尋了網頁中所有<h3>及<p>的HTML標籤內容。限定只搜尋兩個節點。
result  =  soup.find_all(["h3","p"],limit=2)
print(result)
```

```
[<h3 itemprop="headline">
<a href="https://travel.ettoday.net/article/2054494.htm" itemprop="url" title="免跑花東! 竹子湖「2處金針花海」滿開　採滿回家煮才200元">免跑花東! 竹子湖「2處金針花海」滿開　採滿回家煮才200元</a>
</h3>, <p class="summary" itemprop="description"></p>]
```

```
#當某一節點下只有單個子節點時，可以利用BeautifulSoup套件(Package)的select_one()方法(Method)，選取子節點
result = soup.find("h3",itemprop="headline")
print(result.select_one("a"))
```

```
<a href="https://travel.ettoday.net/article/2054494.htm" itemprop="url" title="免跑花東！竹子湖「2處金針花海」滿開　採滿回家煮才200元">免跑花東！竹子湖「2處金針花海」滿開　採滿回家煮才200元</a>
```

```
#如果某一節點下有多個子節點時，則使用select()方法(Method)，選取子節點。由於<div>標籤下有多個<a>標籤的子節點，所以可以利用select()方法(Method)，選取其下所有的<a>標籤，並且為串列(List)的資料型態
result  =  soup.find("div",itemprop="itemListElement")
print(result.select("a"))
```

```
[<a class="pic" href="https://travel.ettoday.net/article/2054494.htm" title="免跑花東! 竹子湖「2處金針花海」滿開　採滿回家煮才200元">
<img alt="免跑花東! 竹子湖「2處金針花海」滿開　採滿回家煮才200元" data-original="https://cdn2.ettoday.net/images/5811/c5811053.jpg" itemprop="image" onerror="this.src='//cdn2.ettoday.net/style/travel/images/fb_ettoday_trave
</a>, <a href="https://travel.ettoday.net/article/2054494.htm" itemprop="url" title="免跑花東! 竹子湖「2處金針花海」滿開　採滿回家煮才200元">免跑花東! 竹子湖「2處金針花海」滿開　採滿回家煮才200元</a>]
```

三、以CSS屬性搜尋節點

要依據HTML的css屬性來進行節點的搜尋，需使用 class_ 關鍵字參數(Keyword Argument)來進行css屬性值的指定

(以下為向下的搜尋節點方式)

✓ 0 秒

```
#搜尋第一個符合指定的HTML標籤及css屬性值的節點
titles  =  soup.find("p",class_  ="summary")
print(titles)
```

```
<p class="summary" itemprop="description"></p>
```

```
#搜尋網頁中符合指定的HTML標籤及css屬性值的所有節點
titles = soup.find_all("p",class_ = "summary",limit=3)
print(titles)
```

```
[<p class="summary" itemprop="description"></p>, <p class="summary" itemprop="description">把握暑假的尾巴，捷絲旅即日起推出「雙城買一送一」優惠，可任選台北西門、宜蘭礁溪、花蓮中正、及台南十鼓等熱門度假據點，專案可不須連住，</p>, <p class="summary" itemprop="description">力麗集團飯店宣布全台10間飯店從8/11起陸續推出24小時快閃特價，包括入住台北、高雄力麗哲園飯店三天兩夜只要1999元含早餐，等於每人每晚500有找；若想泡湯，宜蘭力麗威斯汀度假酒店豪華/</p>]
```

```
#單純只想要透過css屬性值來進行HTML節點的搜尋，則可以使用BeautifulSoup套件(Package)的select()方法(Method)
titles  =  soup.select(".summary",limit=3)
print(titles)
```

```
[<p class="summary" itemprop="description"></p>, <p class="summary" itemprop="description">把握暑假的尾巴，捷絲旅即日起推出「雙城買一送一」優惠，可任選台北西門、宜蘭礁溪、花蓮中正、及台南十鼓等熱門度假據點，專案可不須連住，</p>, <p class="summary" itemprop="description">力麗集團飯店宣布全台10間飯店從8/11起陸續推出24小時快閃特價，包括入住台北、高雄力麗哲園飯店三天兩夜只要1999元含早餐，等於每人每晚500有找；若想泡湯，宜蘭力麗威斯汀度假酒店豪華/</p>]
```

四、搜尋父節點

從某一個節點向上搜尋，則可以使用BeautifulSoup套件(Package)的find_parent()或find_parents()方法(Method)

```
[13] #搜尋<a>標籤且itemprop屬性值為url的節點
     result = soup.find("a",itemprop="url")

     #接著透過find_parents()方法(Method)，向上搜尋<h3>標籤的父節點
     parents = result.find_parents("h3")
     print(parents)
```

```
[<h3 itemprop="headline">
<a href="https://travel.ettoday.net/article/2054494.htm" itemprop="url" title="免跑花東！竹子湖「2處金針花海」滿開　採滿回家煮才200元">免跑花東！竹子湖「2處金針花海」滿開　採滿回家煮才200元</a>
</h3>]
```

五、搜尋前、後節點

[14] ✓ 0秒

```
result = soup.find("h3",itemprop="headline")

#同一層級的節點，想要搜尋前一個節點，可以使用BeautifulSoup套件(Package)的find_previous_siblings()方法
previous_node = result.find_previous_siblings("a")
print(previous_node)
```

```
[<a class="pic" href="https://travel.ettoday.net/article/2054494.htm" title="免跑花東！竹子湖「2處金針花海」滿開　採滿回家煮才200元">
<img alt="免跑花東！竹子湖「2處金針花海」滿開　採滿回家煮才200元" data-original="https://cdn2.ettoday.net/images/5811/c5811053.jpg" itemprop="image" onerror="this.src='//cdn2.ettoday.net/style/trav
</a>]
```

```
[15] result  =  soup.find("h3",itemprop="headline")

     #同一層級的節點，想要搜尋後一個節點，則使用find_next_siblings()方法(Method)
     next_node  =  result.find_next_siblings("p")
     print(next_node)
```

```
[<p class="summary" itemprop="description"></p>]
```

六、取得屬性值

```
#想要爬取「ETtoday的旅遊雲」台北景點首頁的標題連結。首先，利用find_all()方法搜尋網頁中所有<h3>標籤且itemprop屬性值為headline的節點
titles = soup.find_all("h3",itemprop="headline")

#透過for迴圈讀取串列(List)中的節點，由於<h3>標籤底下只有一個<a>標籤，所以可以利用BeautifulSoup套件的select_one()方法進行選取
for title in titles:
  print(title.select_one("a"))
```

```
<a href="https://travel.ettoday.net/article/2054494.htm" itemprop="url" title="免跑花東！竹子湖「2處金針花海」滿開　採滿回家煮才200元">免跑花東！竹子湖「2處金針花海」滿開　採滿回家煮才200元</a>
<a href="https://travel.ettoday.net/article/2053454.htm" itemprop="url" title="水手服氣泡飲、符號杯子蛋糕太Q！全台首間「美少女戰士咖啡廳」來了">水手服氣泡飲、符號杯子蛋糕太Q！全台首間「美少女戰士咖啡廳」來了</a>
<a href="https://travel.ettoday.net/article/2053408.htm" itemprop="url" title="最新必比登！34年「巷子龍家常菜」首入榜　必點全台最好吃宮保雞丁">最新必比登！34年「巷子龍家常菜」首入榜　必點全台最好吃宮保雞丁</a>
<a href="https://travel.ettoday.net/article/2052694.htm" itemprop="url" title="捷絲旅推「雙城買一送一」每人1700起　住台南新開糖廠主題旅店">捷絲旅推「雙城買一送一」每人1700起　住台南新開糖廠主題旅店</a>
<a href="https://travel.ettoday.net/article/2052227.htm" itemprop="url" title="把握暑假尾巴！力麗全台飯店快閃優惠　住一晚每人500元起含早餐">把握暑假尾巴！力麗全台飯店快閃優惠　住一晚每人500元起含早餐</a>
<a href="https://travel.ettoday.net/article/2049926.htm" itemprop="url" title="慶降級！限時9天　微風南山「世界最濃抹茶冰淇淋」90分鐘吃到飽">慶降級！限時9天　微風南山「世界最濃抹茶冰淇淋」90分鐘吃到飽</a>
<a href="https://travel.ettoday.net/article/2049732.htm" itemprop="url" title="每小時上百顆！今夏「最壯觀流星雨」下週登場　觀賞時間點出爐">每小時上百顆！今夏「最壯觀流星雨」下週登場　觀賞時間點出爐</a>
<a href="https://travel.ettoday.net/article/2049147.htm" itemprop="url" title="七夕這樣過！KKday推世民酒店住滿24小時　士林萬麗免費升等山景房">七夕這樣過！KKday推世民酒店住滿24小時　士林萬麗免費升等山景房</a>
<a href="https://travel.ettoday.net/article/2048876.htm" itemprop="url" title="全台4間絕美浴缸窗景住宿　泡澡獨享整片太平洋蔚藍海景">全台4間絕美浴缸窗景住宿　泡澡獨享整片太平洋蔚藍海景</a>
```

```
titles = soup.find_all("h3",itemprop="headline")
for title in titles:
    #利用get()方法(Method)取得href屬性值中的網
    print(title.select_one("a").get("href"))
```

```
https://travel.ettoday.net/article/2054494.htm
https://travel.ettoday.net/article/2053454.htm
https://travel.ettoday.net/article/2053408.htm
https://travel.ettoday.net/article/2052694.htm
https://travel.ettoday.net/article/2052227.htm
https://travel.ettoday.net/article/2049926.htm
https://travel.ettoday.net/article/2049732.htm
https://travel.ettoday.net/article/2049147.htm
https://travel.ettoday.net/article/2048876.htm
```

七、取得連結文字

```
titles  =soup.find_all("h3",itemprop="headline")
for  title  in  titles:
    #要取得<a>標籤的連結文字，可以利用BeautifulSoup套件(Package)的getText()方法(Method)
    print(title.select_one("a").getText())
```

免跑花東！竹子湖「2處金針花海」滿開　採滿回家煮才200元
水手服氣泡飲、符號杯子蛋糕太Q！全台首間「美少女戰士咖啡廳」來了
最新必比登！34年「巷子龍家常菜」首入榜　必點全台最好吃宮保雞丁
捷絲旅推「雙城買一送一」每人1700起　住台南新開糖廠主題旅店
把握暑假尾巴！力麗全台飯店快閃優惠　住一晚每人500元起含早餐
慶降級！限時9天　微風南山「世界最濃抹茶冰淇淋」90分鐘吃到飽
每小時上百顆！今夏「最壯觀流星雨」下週登場　觀賞時間點出爐
七夕這樣過！KKday推世民酒店住滿24小時　士林萬麗免費升等山景房
全台4間絕美浴缸窗景住宿　泡澡獨享整片太平洋蔚藍海景

謝謝聆聽