Python & Data - Week 10

Web Scraping - Beautiful Soup

打包下載

步驟

1. 發現需要截取的 html tag 和 class / id 名
2. 使用 Beautiful Soup 獲取內容

使用 Beautiful Soup 獲取內容

安裝 Beautiful Soup

In [ ]:
```
pip install beautifulsoup4
```

```
DEPRECATION: Configuring installation scheme with distutils config files is
deprecated and will no longer work in the near future. If you are using a H
omebrew or Linuxbrew Python, please see discussion at https://github.com/Ho
mebrew/homebrew-core/issues/76621
Requirement already satisfied: beautifulsoup4 in /usr/local/lib/python3.9/s
ite-packages (4.10.0)
Requirement already satisfied: soupsieve>1.2 in /usr/local/lib/python3.9/si
te-packages (from beautifulsoup4) (2.3.1)
WARNING: You are using pip version 21.2.4; however, version 22.0.4 is avail
able.
You should consider upgrading via the '/usr/local/opt/python@3.9/bin/python
3.9 -m pip install --upgrade pip' command.
Note: you may need to restart the kernel to use updated packages.
```

基本用法

#1 使用 `BeautifulSoup(page.content, "html.parser")` 初始化功能，
page.content 是 requests 返回的原始內容 (bytes)
#2 使用 soup.find_all 找尋所有 `<a>` 的標籤
#3 印出每個 link 的內容
#4 使用 link['href'] 判斷 tag 的屬性

In [ ]:

```
import requests
from bs4 import BeautifulSoup

URL = "http://www.puiching.edu.mo/news"
page = requests.get(URL)

soup = BeautifulSoup(page.content, "html.parser") #1
links = soup.find_all("a") #2

for link in links:
    print(link.text) #3
    #if link['href'].startswith('/news'): #4
    #    print(link.text)
```

返回舊版網頁
中文
English

首頁
最新資訊
培正簡介
校徽及校歌
學校簡介
校務簡報
校曆資訊
校園地圖
基督教教育
團體組織
網址精選
聯絡我們
校園即時影像
入學及收費資訊
eClass
EVI
圖書館
中華文化館
全部
學校動態
學生獎項
其他資訊
媒體報導
「世界記憶學術中心」交流活動獲益良多
詳細
「職安健吉祥物」填色比賽表現出色
詳細
英國劍橋大學入學考試消息
https://shorturl.at/fgmEH
詳細
學界乒乓球單打賽獲五冠
詳細
全澳中學生衛星概念徵集比賽共兩隊獲獎
詳細
2
3
4
5
下一頁 >

共建群體免疫屏障
關注本校的 YouTube
關注本校的 Facebook
關注本校的 Instagram
特別資訊RSS

使用 tag 及 href 屬性不足以篩選出需要的內容

中文

學校動態

學生獎項

其他資訊

媒體報導

「世界記憶學術中心」交流活動獲益良多

詳細

「職安健吉祥物」填色比賽表現出色

詳細

英國劍橋大學入學考試消息

詳細

學界乒乓球單打賽獲五冠

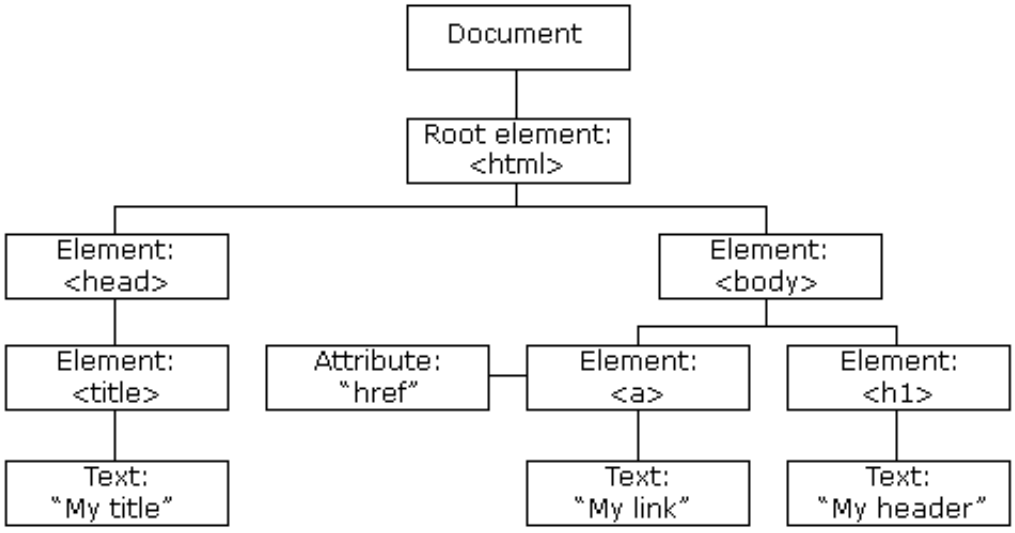
詳細

全澳中學生衛星概念徵集比賽共兩隊獲獎

詳細

Tag, Attribute...?

HTML DOM Tree



來源: HTML DOM

使用多重方法解析出需要的內容

步驟:

1. 先取得 div.news-item-container 的部分
2. 再針對每個 div 解析出 a, span.date, .news-content 等資料

In [ ]:

```
import requests
from bs4 import BeautifulSoup

URL = "http://www.puiching.edu.mo/news"
page = requests.get(URL)

soup = BeautifulSoup(page.content, "html.parser")
containers = soup.find_all(class_="news-item-container")

for news_container in containers:
    title_element = news_container.find("a") # 針對每個 news_container 找出 a
    print(title_element.text) # 印出 <a> 的文字內容
    print(title_element['href']) # 印出 a 的 href 屬性
```

「世界記憶學術中心」交流活動獲益良多
/news/view/3339
「職安健吉祥物」填色比賽表現出色
/news/view/3338
英國劍橋大學入學考試消息
/news/view/3337
學界乒乓球單打賽獲五冠
/news/view/3327
全澳中學生衛星概念徵集比賽共兩隊獲獎
/news/view/3334

下一步...

1. 找出日期
2. 繼續挖掘詳細頁內容
3. 繼續挖掘下一頁內容

Beautiful Soup Cheatsheet

1. 獲取 BS 實體

```
soup = BeautifulSoup(text, "html.parser")
```

> text 是 html 內容，可以通過讀取檔案或 requests.get 取得 "html.parser" 是其中一種 parser 類型

1. 找出所有指定 tag

```
elements = soup.find_all(tag)
```

> tag 是 HTML tag 的類型，例如：div, a, span 等 elements 是 iterable 類型，可以使用 for 取得所有內容

1. 指定某個 class 的 tag

```
elements = soup.find_all("div", class_="news-item-container")
element = soup.find("span", class_="date")
```

> find*all 找出所有匹配的內容，find 找出第一個匹配的內容 使用 class* 收窄結果的範圍到某一個 class 名稱

1. 指定某個 id 的元素

```
element = soup.find(id="page-nav")
```

> id 指定了某個 id="page-nav" 的 tag 理論上每個 id 在 html 是唯一的，所以用 find 找出第一個匹配的元素

1. 元素的文字內容

```
element.text
```

> 使用 .text 返回文字內容，如元素中有 html 碼會被去除

1. 元素的 Attribute 內容

```
element[attr]
```

> attr 是屬性的名稱，例如 `element['href']` 可找出 href 屬性的內容