網路爬蟲與資料分析
動態網頁解析

Instructor: 馬豪尚

JavaScript

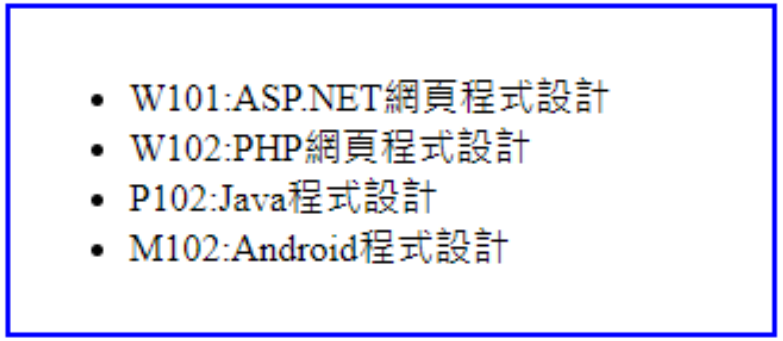
- JavaScript 是一種腳本，也能稱它為程式語言，可以讓你在網頁中實現出複雜的功能。
- JavaScript常用來完成以下任務
  - 嵌入動態文字於HTML頁面
  - 對瀏覽器事件作出回應
  - 讀寫HTML元素
  - 在資料被提交到伺服器之前驗證資料
  - 檢測訪客的瀏覽器資訊
  - 控制Cookie，包括建立和修改等

Quick JavaScript Switcher

› Chrome瀏覽器的擴充應用程式

Quick JavaScript Switcher

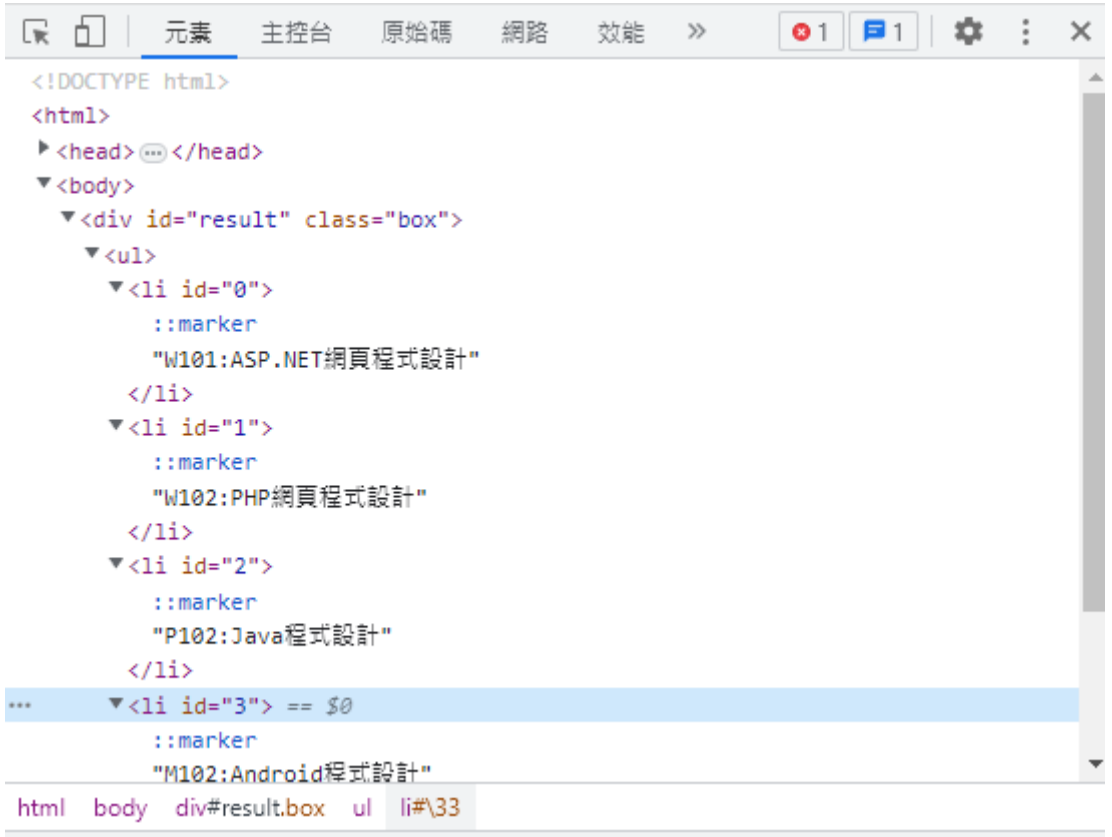
- 測試網址
  - https://fchart.github.io/books.html



Chrome瀏覽器開發人員工具

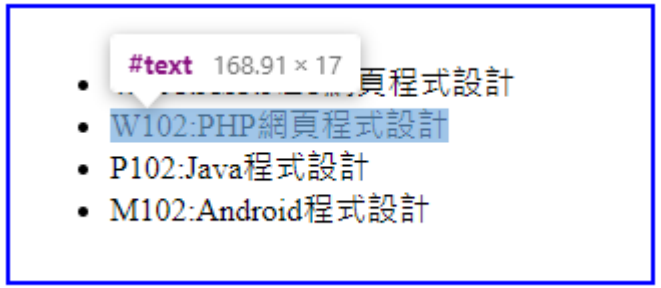
› 在瀏覽器的網頁頁面上點右鍵->檢查

- W101:ASP.NET網頁程式設計
- W102:PHP網頁程式設計
- P102:Java程式設計
- M102:Android程式設計



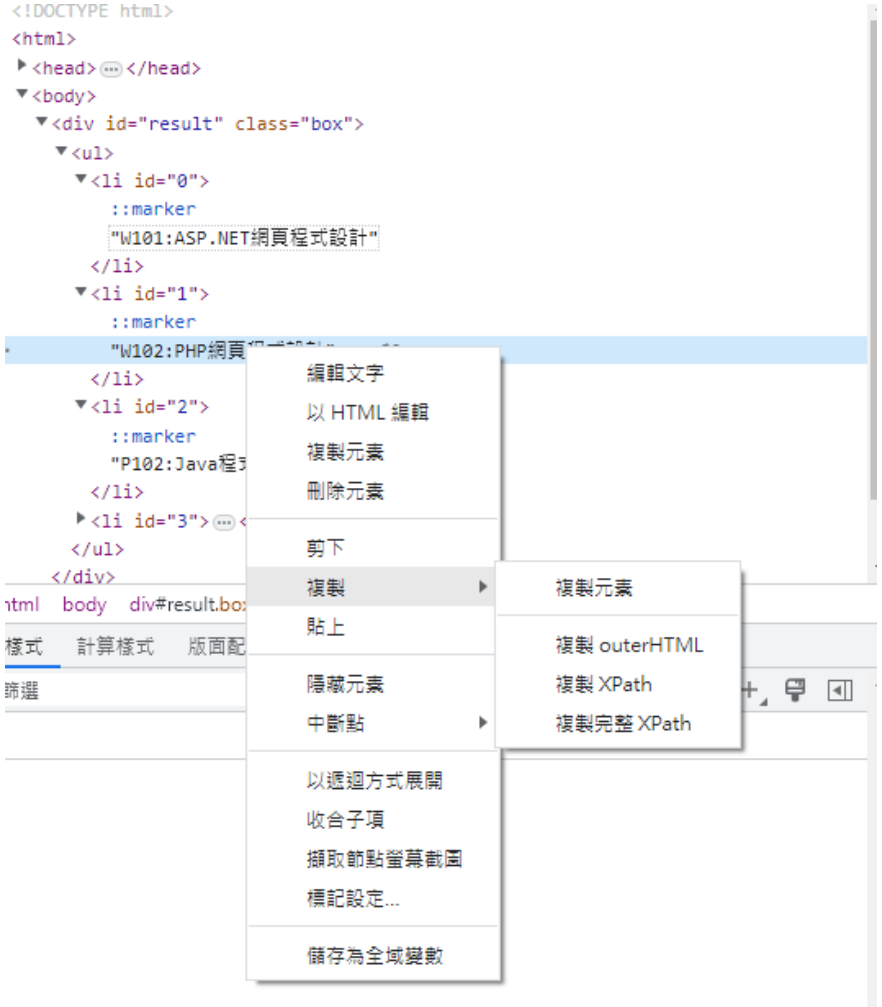
Chrome瀏覽器開發人員工具

› 可以檢視每一個html元素



取得選取元素的網頁定位資料

› 在該選取元速點選右鍵->複製



複製元素

<li id="1">W102:PHP網頁程式設計

爬蟲with JavaScript實務#1

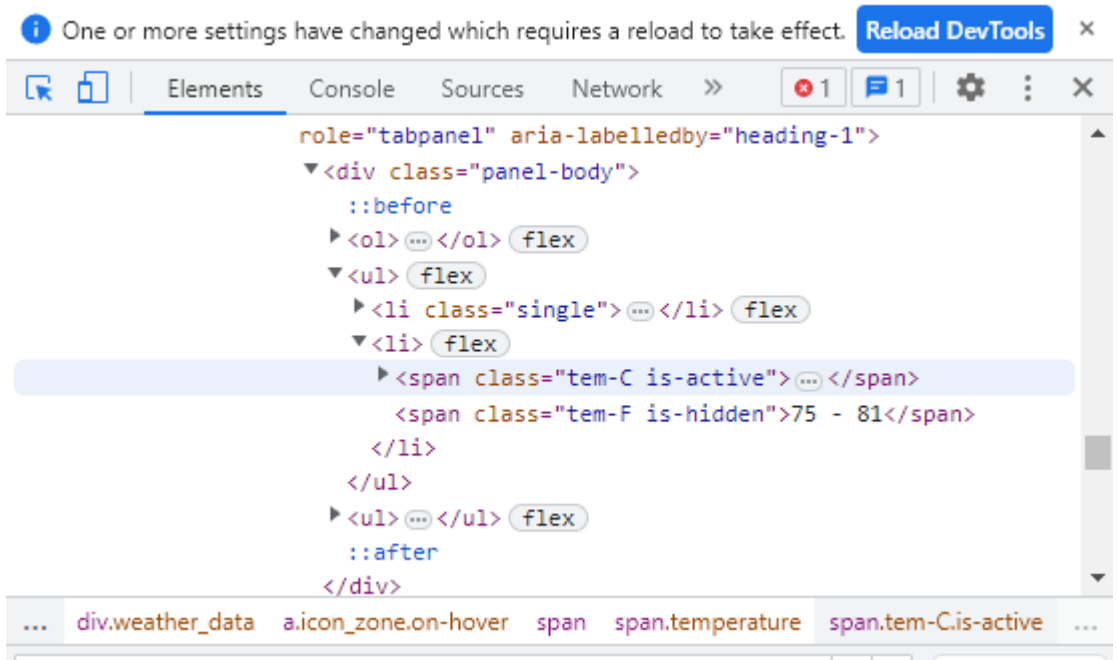
- 爬取氣象局天氣資訊
  - https://www.cwb.gov.tw/V8/C/W/County/County.html?CID=65

爬取網站

- 使用chrome瀏覽器開發人員工具分析html
- 情境: JavaScript會影響網站載入的內容
- 選擇獲得資源的方式
  - Selenium
- 爬取網站的過程
  - 定位到要互動的元素
  - 互動後等待網頁動態載入
  - 定位到要爬取的資料位置
  - 存下資料

Chrome瀏覽器開發人員工具

› 可以檢視每一個html元素



取得選取元素的網頁定位資料

› 在該選取元速點選右鍵->複製

複製元素

24 - 25

複製css selector

body > div.wrapper > main > div > div:nth-child(1) > div.d-xl-none.d-block > div.banner_wrap > ul > li:nth-child(1) > span.tem > span.tem-C.is-active

爬蟲with JavaScript實務#2

- 爬取momo購物網NBA球衣的商品資料
  - https://www.momoshop.com.tw/search/searchShop.jsp?keyword=nike NBA&searchType=1&curPage=1&_isFuzzy=0&showType=chessboardType
- 情境
  - 網站使用javascript來和使用者互動並載入內容
  - 並未隱藏其內容在frame裡
  - 要爬取的資料直接顯示在一個頁面內

爬取網站

- 使用chrome瀏覽器開發人員工具分析html
- JavaScript會影響momo購物網站內容
- 選擇獲得資源的方式
  - Selenium
- 解析網站的語法
  - Beautifulsoap
- 與網頁互動來獲取更多資料
  - Selenium

取得想要爬取元素的網頁定位資料

- 商品名稱
  - 複製元素的css selector
  - #BodyBase > div.bt_2_layout.searchbox.searchListArea > div.searchPrdListArea.bookList > div.listArea > ul > li:nth-child(1) > a > div.prdInfoWrap > div.prdNameTitle > h3

取得想要爬取元素的網頁定位資料

- 商品價錢
  - 複製元素的css selector
  - #BodyBase > div.bt_2_layout.searchbox.searchListArea.selectedtop > div.searchPrdListArea.bookList > div.listArea > ul > li:nth-child(1) > a > div.prdInfoWrap > p.money > span.price

取得下一頁的資料

- 找到下一頁按鈕的元素定位
  - 複製元素的css selector
  - body > div.web.header-fixed > div.bt_2_layout.searchbox.searchListArea.selectedtop > div:nth-child(5) > div > div.page-btn.page-next
  - .click()操作

爬蟲with JavaScript實務#3

- 爬取在網站內使用frame嵌入資料的內容- covid19網站
  - https://covid-19.nchc.org.tw/2023_city_confirmed.php?mycity=%E5%85%A8%E5%9C%8B
  - 因為使用frame嵌入所以無法在原始網頁中看到內容

全國 COVID-19 確診報表
[依個案研判日統計]

請點此按鈕,下載全部表單資料 (API) | 欄位 | CSV | Excel | Search:

個案研判日	縣市別	區域	新增確診人數	累計確診人數	七天移動平均新增確診人數
2023-09-07	全國	全區	2	10,241,523	0.43
2023-09-05	全國	全區	1	10,241,521	0.14
2023-08-07	全國	全區	1	10,241,520	0.14
2023-07-25	全國	全區	6	10,241,519	2.00
2023-07-24	全國	全區	4	10,241,513	1.14
2023-07-21	全國	全區	3	10,241,509	0.71
2023-07-20	全國	全區	1	10,241,506	0.29
2023-07-17	全國	全區	1	10,241,505	0.29
2023-07-13	全國	全區	1	10,241,504	0.29

個案研判日 | 縣市別 | 區域 | 新增確診人數 | 累計確診人數 | 七天移動平均新增確診

Showing 1 to 10 of 1,026 entries

```
▼<iframe src="dt02.php?encodeKey=NTc2Nzg0MzU4&dt_title=COVID-19 確診報表<b
[依個案研判日統計]_全國&dt…2&limitValue=全國&equalValue=5&limitColumn2=a03
tValue2=全區&equalValue2=5" height="530" width="100%" title="covidtable_
n_cdc6_1" frameborder="0">
  ▼#document (https://covid-19.nchc.org.tw/dt02.php?
  encodeKey=NTc2Nzg0MzU4&dt_title=COVID…
  equalValue=5&limitColumn2=a03&limitValue=%E5%85%A8%E5%8D%80&equalValu
    <!-- CSS -->
    ▼<html>
```

使用Selenium定位到frame裡面爬取資料

- driver.switch_to.frame()
  - 使用index來定位第幾個frame
    - driver.switch_to.frame(0)
  - 使用id/name來定位
    - driver.switch_to.frame(“id”)
  - 搭配driver.find的語法來定位
    - driver.switch_to.frame(driver.fine_element(By.CSS_SELECTOR ,”selector”))

練習

- 爬取內政部不動產交易實價查詢網站
  - https://lvr.land.moi.gov.tw/
- 使用Selenium模擬瀏覽器和網頁互動
  - 輸入縣市和鄉鎮市區(任意)
  - 爬取顯示出的每一筆資料(前100筆, 若查詢區域資料不足100則全部)
- 將資料存成csv檔