



第2章 爬取的資料來源：HTML、CSV和JSON (Catch data: HTML、CSV、JSON)

- 2-1 HTML與CSS基礎
- 2-2 資料標籤 – 文字和圖片標籤
- 2-3 群組標籤 – 清單、表格和結構標籤
- 2-4 網站巡覽 – 超連結標籤
- 2-5 互動介面 – 表單標籤
- 2-6 CSV與JSON





2-1 HTML與CSS基礎 (The basic concept of HTML and CSS)

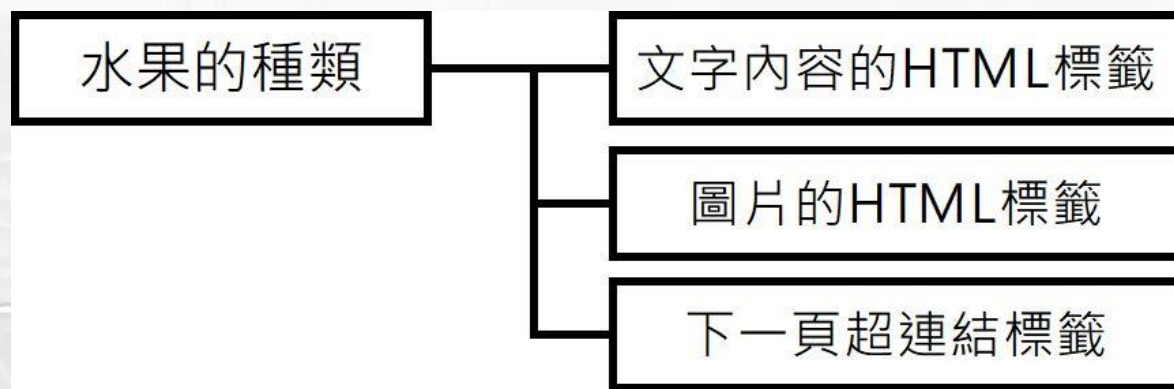
- 2-1-1 HTML標籤語法與結構
- 2-1-2 CSS的基礎





2-1 HTML與CSS基礎

- 網路爬蟲的資料來源是HTML5網頁，每一頁HTML網頁可以想像是網站水果園中的一棵水果樹，在樹上的水果是一個一個HTML標籤。本章內容可以讓讀者認識各種HTML標籤是如何建構出一棵水果樹，如下圖所示：





2-1-1 HTML標籤語法與結構

- 「HTML標示語言」（[HyperText Markup Language](#)）是文件內容的格式編排語言，在瀏覽器中顯示的網頁內容就是使用HTML語法所撰寫的標籤碼，這是Tim Berners-Lee在1991年建立，目前的最新版本是HTML5。
- 在HTML水果樹上的水果主要有三種：文字內容的HTML標籤、圖片和下一頁的超連結標籤，圖片是擷取標籤href屬性的URL網址（即圖片的URL網址），分頁資料需要擷取下一頁<a>標籤href屬性的URL網址。

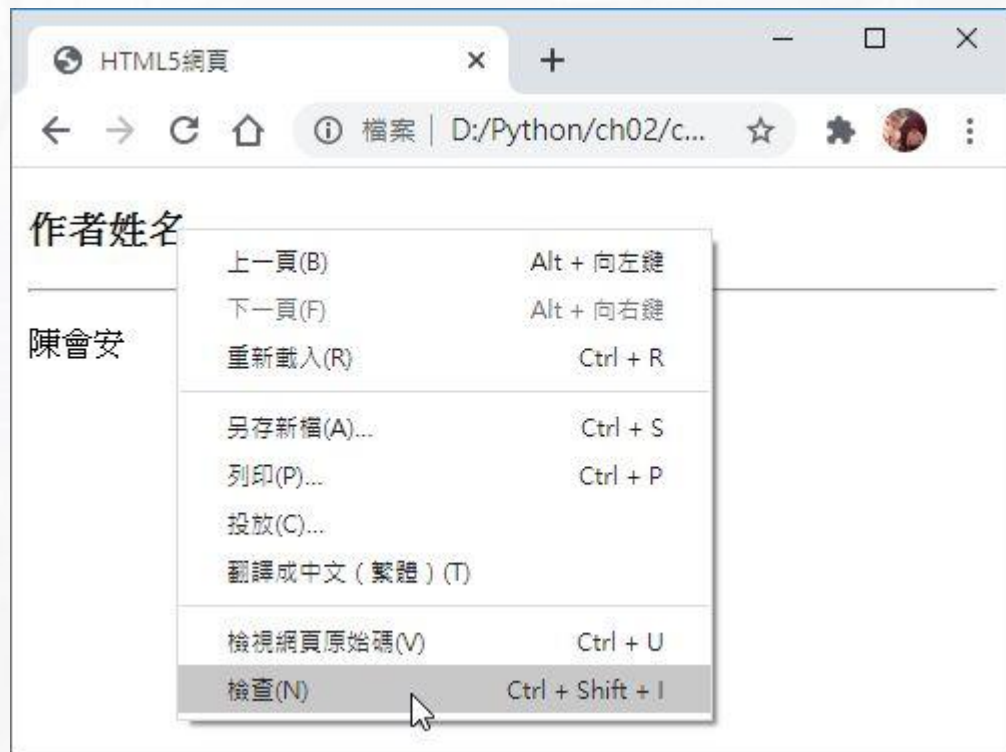




2-1-1 HTML標籤語法與結構

在瀏覽器檢查網頁內容的HTML標籤

- 請啟動Chrome瀏覽器開啟「file:///D:/Python/ch02/ch2-1-1.html」的HTML檔案，「D://Python/ch02」是書附範例檔的路徑，然後在【作者姓名】的文字內容上，執行【右】鍵快顯功能表的【檢查】命令（【檢視網頁原始碼】命令可以檢視整頁HTML標籤），如右圖所示：





2-1-1 HTML標籤語法與結構

在瀏覽器檢查網頁內容的HTML標籤

- 在Chrome開發人員工具可以看到此文字內容的HTML標籤<h3>（開發人員工具的詳細說明，請參閱第3-4節），在下方顯示標籤的階層結構html>body>h3#title（#title是指id屬性值title），如下圖所示：

Hashtag title means that attribute name (id) is title



hierarchical directory structure





2-1-1 HTML標籤語法與結構

HTML標籤語法

- HTML標籤語法是使用開始和結尾標籤所包圍的文字內容，其語法如下所示：

<Tag name, attribute name=attribute value>content </ Tag name>

<標籤名稱 屬性名稱=屬性值>文字內容</標籤名稱>

<h3 id="title">作者姓名</h3>





2-1-1 HTML標籤語法與結構

HTML網頁的標籤結構：ch2-1-1.html

- HTML標籤之中除了文字內容，還可以有其他子標籤，透過父子的巢狀標籤，可以建立出階層結構的HTML5網頁結構，如下所示：

```
<!DOCTYPE html>
```

```
<html>
```

```
<head>
```

```
<meta charset="utf-8">
```

```
<title>HTML5網頁</title>
```

```
</head>
```

```
<body>
```

```
<h3 id="title">作者姓名</h3>
```

```
<hr/>
```

```
<p class="author">陳會安</p>
```

```
</body>
```

```
</html>
```

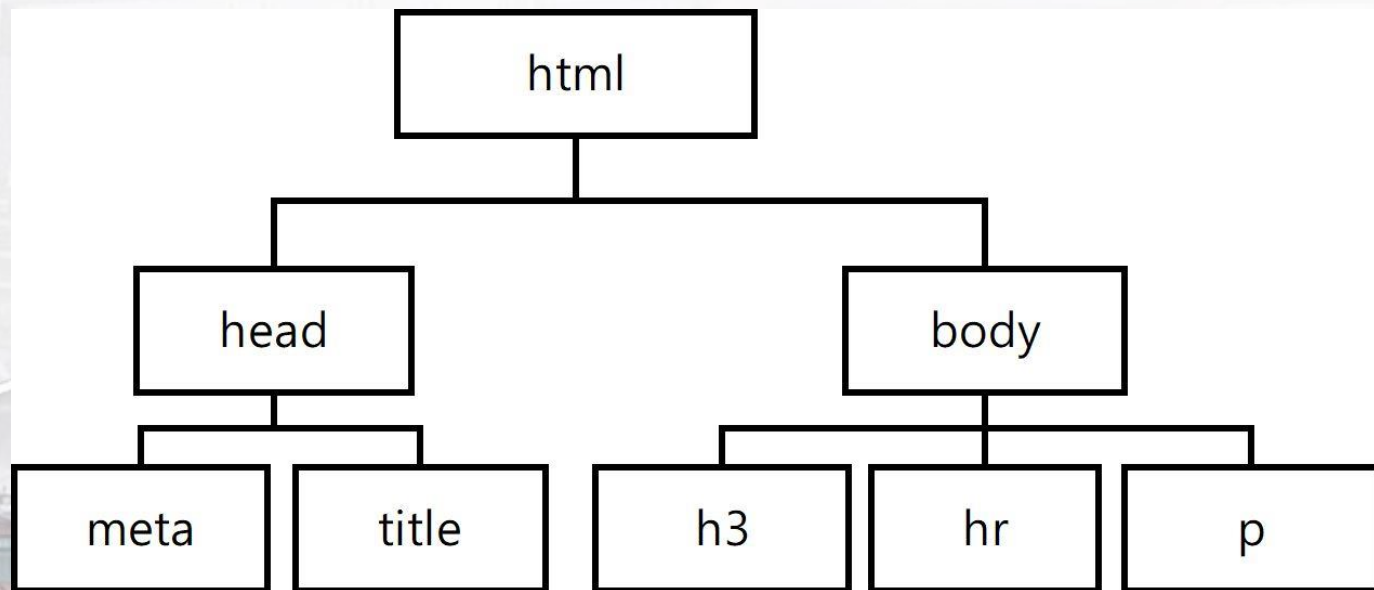




2-1-1 HTML標籤語法與結構

HTML網頁的標籤結構：ch2-1-1.html

- `<html>`標籤才是HTML網頁的根元素，一種容器元素，其內容是`<head>`和`<body>`兩種子標籤。我們也可以將網頁內容繪成樹狀結構，如同一棵倒立的水果樹，如下圖所示：





2-1-1 HTML標籤語法與結構

<head>子標籤

- 在<head>標籤的子標籤是描述HTML網頁本身，常用標籤的說明，如下表所示：

| 標籤 | 說明 |
|----------|--|
| <title> | 顯示瀏覽器視窗上方標題列或標籤頁的標題文字 |
| <meta> | 提供HTML網頁的metadata資料，例如：網頁描述、關鍵字、作者和最近修改日期等資訊 |
| <script> | 標籤內容是客戶端腳本程式碼，例如：JavaScript程式碼 |
| <style> | 在HTML網頁套用的CSS樣式碼 |
| <link> | 連接外部資源檔案，主要連接副檔名.css的CSS樣式表檔案 |



2-1-1 HTML標籤語法與結構

<body>子標籤

- <body>標籤才是瀏覽器看到的網頁內容，對於網頁爬蟲來說，<body>標籤的子標籤內容才是我們欲擷取的目標資料，如下所示：

<h3 id="title">作者姓名</h3>

<hr/>

<p class="author">陳會安</p>





2-1-2 CSS的基礎

CSS樣式

- 「CSS」（Cascading Style Sheets）層級式樣式表是一種樣式語言，用來描述標示語言的格式，可以重新定義HTML標籤的外觀。我們可以想像HTML標籤是位素顏的網紅，瀏覽器依據CSS替網紅化上妝後，就能成為網路上我們認識的網紅。
- HTML標籤<p>是一個段落，預設使用瀏覽器沒有色彩的預設字體與字型尺寸來顯示，我們可以使用CSS重新定義<p>標籤的樣式，如同替嘴唇（段落）化上小紅妝（HTML網頁：ch2-1-2.html），如下所示：

```
<style type="text/css">
p.author { font-size: 10pt;
           color: red; }
</style>
```




2-1-2 CSS的基礎

在HTML網頁套用CSS樣式

- HTML網頁的<p>標籤如果有class屬性值author，就符合CSS選擇器「p.author」的條件，瀏覽器就會在此標籤套用<style>標籤定義的CSS樣式，改為較小的紅色字來顯示，如下圖所示：

1.Please use css and print
:your name...

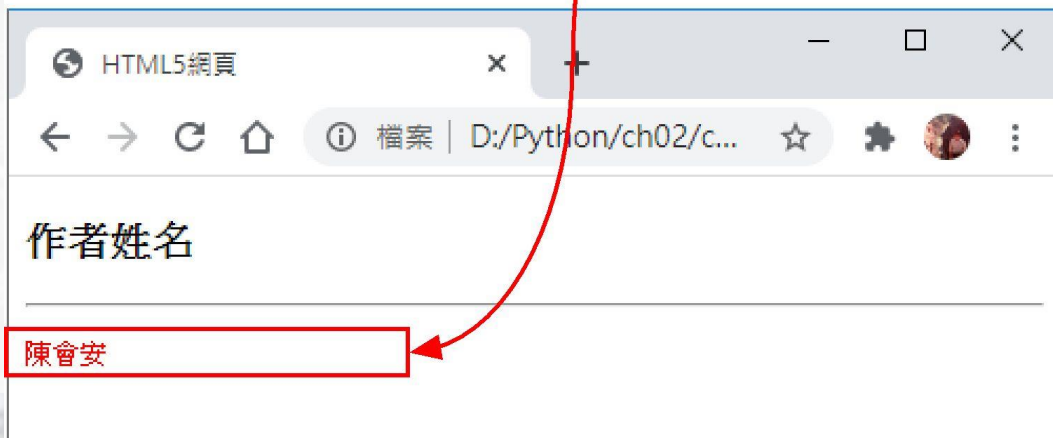
左小官的網頁

歡迎您的光臨

```
<html>
<head></head>
<body>
  <h3>作者姓名</h3>
  <hr/>
  <p class="author">陳會安</p>
</body>
</html>
```

CSS選擇器 `p.author`

```
<style type="text/css">
  p.author {font-size: 10pt;
            color: red;}
</style>
```





2-2 資料標籤 – 文字和圖片標籤

- 2-2-1 文字內容標籤
- 2-2-2 圖片標籤





2-2-1 文字內容標籤

標題文字標籤：ch2-2-1.html

- HTML網頁的標題文字是<hn>標籤，n是1~6，<h1>最重要，依序遞減至<h6>，提供6種不同尺寸變化的標題文字，如下所示：

<h1>HTML5網頁的標題文字</h1>

<h2>HTML5網頁的標題文字</h2>

<h3>HTML5網頁的標題文字</h3>

<h4>HTML5網頁的標題文字</h4>

<h5>HTML5網頁的標題文字</h5>

<h6>HTML5網頁的標題文字</h6>

Run

HTML5網頁的標題文字

HTML5網頁的標題文字

HTML5網頁的標題文字

HTML5網頁的標題文字

HTML5網頁的標題文字

HTML5網頁的標題文字





2-2-1 文字內容標籤

段落標籤：ch2-2-1a.html

- HTML網頁的文字內容不會換行（**Enter**鍵並沒有作用），我們可以使用<p>段落或
換行標籤來換行編排，
標籤沒有內容，<p>標籤的內容是段落文字，預設在前和後會增加邊界尺寸，如下所示：

<p>HTML網頁的文字內容是使用段落來編排</p>

Run (use
 and <p> separately)





2-2-1 文字內容標籤

容器標籤<div>和：ch2-2-1b.html

- HTML的<div>標籤可以在HTML網頁定義一個區塊來顯示文字內容，如下所示：

`<div>Python</div>`

- 上述<div>標籤會換行自成一個區塊。標籤也是容器標籤，不過這是單行元素，並不會換行建立獨立區塊，如下所示：

`<p>外國人很多都是淡藍色眼睛</p>`



跑一下
Run

Python

外國人很多都是淡藍色眼睛



2-2-1 文字內容標籤

跑一下

標示特定語意的文字內容標籤：**ch2-2-1c.html**

- **HTML**網頁的文字內容可能有些名詞或片語需要特別標示，我們只需將文字包含在下表標籤，就可以顯示不同的標示和語意效果，常用**HTML**標籤說明，如下表所示：

| 標籤 | 說明 |
|-----------------------|---|
| | 使用粗體字標示文字， HTML5 代表文體上的差異，例如：關鍵字和印刷上的粗體字等 |
| <i> | 使用斜體字標示文字， HTML5 代表另一種聲音或語調，通常是標示其他語言的技術名詞、片語和想法等 |
| | 顯示強調文字效果，在 HTML5 是強調發音上有細微改變句子的意義，例如：因發音改變而需強調的文字 |
| | HTML4 是更強的強調文字； HTML5 是重要文字 |
| <cite> | HTML4 是引言或參考其他來源； HTML5 是定義產品名稱，例如：一本書、一首歌、一部電影或畫作等 |
| <small> | HTML4 是顯示縮小文字； HTML5 是輔助說明或小型印刷文字，例如：網頁最下方的版權宣告等 |



2-2-2 圖片標籤

ch2-2-2.html

Run 跑一下

- HTML網頁可以使用標籤插入gif、jpg或png格式的圖檔，例如：顯示Penguins.jpg圖檔的標籤，如下所示：

```

```





2-3 群組標籤－清單、表格和結構標籤

- 2-3-1 清單標籤
- 2-3-2 表格標籤
- 2-3-3 結構標籤





2-3 群組標籤 – 清單、表格和結構標籤

- 群組標籤的目的是群組多個子標籤來建立階層結構，以水果樹來說，這些標籤是建立樹幹和樹枝，所以，群組標籤本身不是目標資料，其群組的文字或圖片子標籤才是目標資料，可以讓我們摘取連著樹枝的整串水果。
- **請注意！群組標籤位在最底層的項目、<td>和<th>儲存格標籤有可能是樹枝，也有可能本身就是目標資料的水果。**





2-3-1 清單標籤

項目編號：ch2-3-1.html

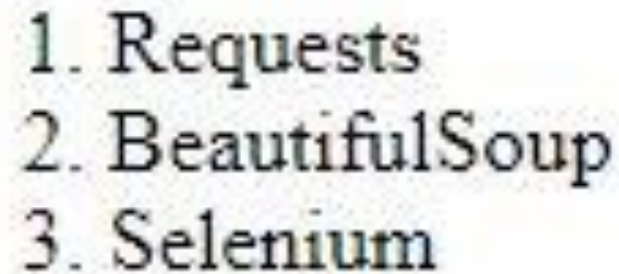
跑一下

- HTML清單提供數字順序的項目編號（Ordered List），如下所示：

Requests

BeautifulSoup

Selenium



```
1. Requests
2. BeautifulSoup
3. Selenium
```





2-3-1 清單標籤

項目符號：ch2-3-1a.html 跑一下

- HTML清單可以使用無編號的項目符號（Unordered List），即在項目前顯示小圓形、正方形等符號，父標籤是，子標籤也是標籤（目標資料是標籤或其子標籤的文字內容），如下所示：

Web API

AJAX

- Web API
- AJAX





2-3-1 清單標籤

定義清單：ch2-3-1b.html

跑一下

- HTML5定義清單（**Definition List**）是群組名稱和值成對的結合清單，例如：詞彙說明的每一個項目是定義和說明，如下所示：

<dl>

<dt>Requests</dt>

<dd>送出HTTP請求</dd>

<dt>BeautifulSoup</dt>

<dd>剖析HTML網頁</dd>

</dl>

Requests

送出HTTP請求

BeautifulSoup

剖析HTML網頁





2-3-2 表格標籤

- HTML表格是一組相關標籤的集合，我們需要同時使用多個標籤才能建立表格。HTML表格相關標籤的說明，如下表所示：

| 標籤 | 說明 |
|---------|---|
| <table> | 建立表格，其他表格相關標籤都位在此標籤之中 |
| <tr> | 定義表格的標題列（子標籤是<td>或<th>）或資料列（子標籤是<td>標籤） |
| <th> | 定義表格標題列的儲存格 |
| <td> | 定義表格資料列的儲存格 |





2-3-2 表格標籤

跑一下

```
<table border="1">
<tr>
  <th>客戶端</th>
  <th>伺服器端</th>
</tr>
<tr>
  <td>JavaScript</td>
  <td>ASP.NET</td>
</tr>
<tr>
  <td>AJAX</td>
  <td>PHP</td>
</tr>
</table>
```

| 客戶端 | 伺服器端 |
|------------|---------|
| JavaScript | ASP.NET |
| AJAX | PHP |



2-3-3 結構標籤

跑一下

- 在HTML 4是使用第2-2-1節的<div>標籤來建立版面配置（HTML網頁：ch2-3-3.html），如下所示：

```
<div>
<div>
  <h3>Python</h3>
  <p>程式語言</p>
</div>
<div>
  <h3>JavaScript</h3>
  <p>網頁語言</p>
</div>
</div>
```

Python

程式語言

JavaScript

網頁語言



Take a rest





2-4 網站巡覽 – 超連結標籤 跑一下

- HTML的<a>超連結標籤可以連接網站的其他網頁，或其他網站的網頁，超連結預設是使用藍色底線字；瀏覽過是顯示紫色底線字（HTML網頁：ch2-4.html），如下所示：

`清單標籤`

文字超連結

清單標籤





2-4 網站巡覽－超連結標籤

圖片和區塊超連結：ch2-4a.html 跑一下

- 超連結<a>標籤可以使用子標籤建立圖片超連結，當游標移至圖片上就會成為手形圖示，如下所示：

```
<a href="http://www.yahoo.com.tw">
```

```

```

```
</a>
```

- 在<a>標籤中還可以使用區塊元素，例如：<h3>標籤，如下所示：

```
<a href="http://www.hinet.net"><h3>中華電信  
HiNet</h3></a>
```

圖片超連結



區塊超連結

中華電信HiNet





2-4 網站巡覽－超連結標籤

網站選單的超連結清單：ch2-4b.html


跑一下

- 我們可以使用清單加上<a>標籤來建立網站選單，如下所示：

項目編號

項目符號

定義清單

- 
- 項目編號
 - 項目符號
 - 定義清單



2-5 互動介面 – 表單標籤

- 2-5-1 HTML表單標籤結構
- 2-5-2 文字內容欄位
- 2-5-3 選擇欄位





2-5-1 HTML表單標籤結構

- HTML網頁表單也是HTML標籤的集合，其根標籤是<form>，如下所示：

```
<form name="name" method="post | get" action="URL">  
    <input type=...>  
    <textarea> .... </textarea>  
    <select>  
        <option> .... </option>  
    </select>  
    <input type="submit" ...>  
</form>
```

"http://httpbin.org/post"





2-5-2 文字內容欄位

文字與密碼方塊欄位：ch2-5-2.html

跑一下

- 文字與密碼方塊可以傳遞使用者以鍵盤輸入的文字內容。例如：姓名、帳號和電話等資料；密碼欄位是將輸入資料在顯示時改用圓點或「*」星號取代，其使用上和文字方塊並沒有什麼不同，如下所示：

```
<input type="text" name="User" size="15"/>
```

```
<input type="password" name="Pass" size="15"/>
```

名稱: Joe

密碼:

註冊使用者





2-5-2 文字內容欄位

多行文字方塊欄位：ch2-5-2a.html

跑一下

- 多行文字方塊可以輸入多行或整篇文字內容，特別適合使用在地址、意見、描述或備註等文字資料的輸入，如下所示：

```
<textarea name="Address" rows="5" cols="50">  
</textarea>
```



The screenshot shows a web form with a label '地址:' (Address) and a '送出' (Submit) button. The text area is empty and has a small 'X' icon in the bottom right corner.



2-5-2 文字內容欄位

隱藏欄位：ch2-5-2b.html

跑一下

- 隱藏欄位不需要使用者輸入資料，這是不可見欄位，可以直接將**value**屬性值傳送到伺服器端。在**HTML**表單使用隱藏欄位的目的是可以傳送一些不需要輸入的參數值至伺服器，如下所示：

<input type="hidden" name="Type" value="Member"/>

- 上述標籤的**type**屬性值是**hidden**，**value**屬性是送出的值。





2-5-3 選擇欄位

核取方塊欄位：ch2-5-3.html

跑一下

- 核取方塊是一個開關，可以讓使用者選擇是否開啟功能或設定參數。**HTML**表單的核取方塊欄位是複選題，因為每一個都是可勾選或取消勾選的獨立開關，如下所示：

```
<input type="checkbox" name="GC"
```

```
checked="True"/>Chrome
```

```
<input type="checkbox" name="FF"/>Firefox
```

瀏覽器：☒ Chrome ☐ Firefox

送出





2-5-3 選擇欄位

選擇鈕欄位：ch2-5-3a.html

跑一下

- 選擇鈕是一組選項，每一個選項名稱旁有一個圓形選擇鈕，這是多選一的單選題，例如：性別是男或女，如下所示：

```
<input type="radio" name="Gender" value="male" checked="True"/>男
```

```
<input type="radio" name="Gender" value="female"/>女
```

性別: ☐ 男 ☒ 女

送出





2-5-3 選擇欄位

下拉式清單方塊欄位：ch2-5-3b.html

跑一下

- HTML的<select>標籤配合<option>標籤的選項可以建立下拉式清單方塊欄位，size屬性值1是下拉式清單方塊；大於1是清單方塊，如下所示：

```
<select name="Webs" size="4" multiple="True">  
  <option value="w1" selected="True">Yahoo!奇摩</option>  
  <option value="w2">中華電信Hinet</option>  
  <option value="w3">Google台灣</option>  
</select>
```





2-6 CSV與JSON

- 2-6-1 CSV
- 2-6-2 JSON





2-6-1 CSV

- **CSV（Comma-Separated Values）**檔案的內容是使用純文字方式表示的表格資料，這是一個文字檔案，其中的每一行是表格的一列，每一個欄位是使用「,」逗號來分隔。例如：現在有一個表格資料，我們準備將此表格轉換成**CSV**資料，如下表所示：

| 程式語言 | 開發者 | 上市年 | 副檔名 |
|--------|-------------------|------|-------|
| Python | Cuido van Rossum | 1991 | .py |
| Java | James Gosling | 1995 | .java |
| C++ | Bjarne Stroustrup | 1983 | .cpp |

程式語言,開發者,上市年,副檔名 Python,Cuido van Rossum,1991,.py Java,James Gosling,1995,.java C++,Bjarne Stroustrup,1983,.cpp



2-6-1 CSV

讀取CSV檔案：ch2-6-1.py Homework

- Python程式存取CSV檔案是使用csv模組，例如：讀取pl.csv檔案的內容（即前述表格資料），如下所示：

```
import csv
```

join()：連接字符串數組。將列表中的元素以指定的字符(分隔符)連接生成一個新的字符串

```
csvfile = "pl.csv"
```

```
with open(csvfile, 'r') as fp:
```

```
    reader = csv.reader(fp)
```

```
    for row in reader:
```

```
        print(','.join(row))
```

程式語言,開發者,上市年,副檔名

Python,Cuido van Rossum,1991,.py

Java,James Gosling,1995,.java

C++,Bjarne Stroustrup,1983,.cpp





- 語法： `'sep'.join(seq)`
- 參數說明 **sep**：分隔符。可以為空
- **seq**：要連接的元素序列、字符串、元組、字典
- 上面的語法即：以**sep**作為分隔符，將**seq**所有的元素合併成一個新的字符串
- 返回值：返回一個以分隔符**sep**連接各個元素後生成的字符串





2-6-1 CSV

寫入資料至CSV檔案：ch2-6-1a.py

跑一下

- 我們也可以將清單寫入CSV檔案，例如：將CSV清單寫入pl2.csv檔案，如下所示：

```
import csv
```

```
csvfile = "pl2.csv"
```

```
lst1 = [ ["Python", "Cuido van Rossum", 1991, ".py"],  
         ["Java", "James Gosling", 1995, ".java"],  
         ["C++", "Bjarne Stroustrup", 1983, ".cpp"] ]
```

```
with open(csvfile, 'w+', newline='') as fp:
```

```
    writer = csv.writer(fp)
```

```
    writer.writerow(["程式語言", "開發者", "上市年", "副檔名"])
```

```
    for row in lst1:
```

```
        writer.writerow(row)
```

Newline刪除多餘換行



2-6-2 JSON

認識JSON

- 「JSON」的全名為（JavaScript Object Notation），這是一種資料交換格式，JSON就是JavaScript物件的文字表示法，其內容只有文字（Text Only）。
- JSON是由Douglas Crockford創造的一種輕量化資料交換格式，因為比XML來的快速且簡單，JSON資料結構就是JavaScript物件文字表示法，不論是JavaScript語言或其他程式語言都可以輕易解讀，這是一種和語言無關的資料交換格式。





2-6-2 JSON

認識JSON

- **JSON**是一種可以自我描述和容易了解的資料交換格式，使用大括號定義成對的鍵和值（**Key-value Pairs**），相當於物件的屬性和值，類似**Python**語言的字典和串列，如下所示：

```
{  
    "key1": "value1",  
    "key2": "value2",  
    "key3": "value3",  
    ...  
}
```

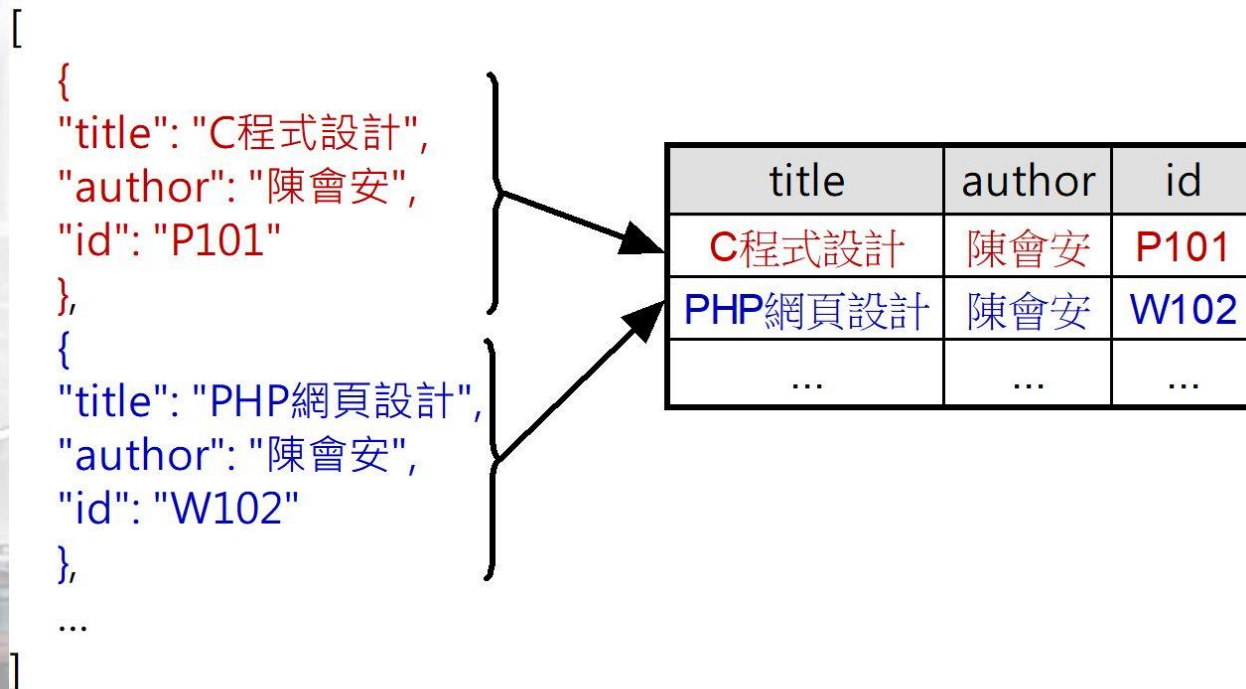




2-6-2 JSON

認識JSON

- **JSON**如果是物件陣列，每一個物件是一筆記錄，我們可以使用方括號「**[]**」來定義多筆記錄，如同是一個表格資料，如下圖所示：





2-6-2 JSON

JSON和Python字典的轉換：ch2-6-2.py 跑一下

- 在json模組的**dumps()**方法可以將Python字典轉換成JSON字串，**loads()**方法從JSON字串轉換成Python字典，如下所示：

```
import json
data = {
    "name": "Joe Chen",
    "grade": 95,
    "tel": "0933123456" }
json_str = json.dumps(data)
print(json_str)
data2 = json.loads(json_str)
print(data2)
```

`{"name": "Joe Chen", "grade": 95, "tel": "0933123456"}`
`{'name': 'Joe Chen', 'grade': 95, 'tel': '0933123456'}`



2-6-2 JSON

將JSON資料寫入檔案：ch2-6-2a.py

跑一下

- Python程式可以使用json模組的dump()方法將Python字典寫入JSON檔案，如下所示：

```
import json
data = {
    "name": "Joe Chen",
    "grade": 95,
    "tel": "0933123456" }
jsonfile = "Student.json"
with open(jsonfile, 'w') as fp:
    json.dump(data, fp)
```





2-6-2 JSON

讀取JSON檔案：ch2-6-2b.py

- Python是使用json模組的load()方法將JSON檔案內容讀取成Python字典，如下所示：

```
import json
```

```
jsonfile = "Student.json"
```

```
with open(jsonfile, 'r') as fp:
```

```
    data = json.load(fp)
```

```
json_str = json.dumps(data)
```

```
print(json_str)
```

1.請顯示

```
{'班級': '資三丙', '姓名': '小春', '學號': 's10813001', '電話': '0933123456'}  
{"班級": "資三丙", "姓名": "小春", "學號": "s10813001", "電話": "0933123456"}
```

