SCU LawTech

Class #05

LAWTECH
learn coding get outstanding

Content

- 作業回顧
- HTML / CSS 基礎
- XPATH 語言
- 爬蟲實作
- 課程作業

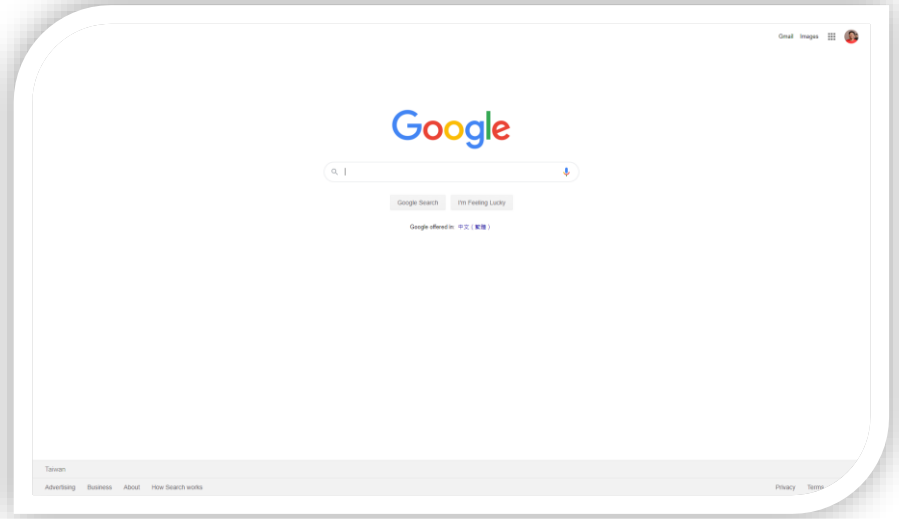
作業回顧

作業回顧

1 x 1 = 1	2 x 1 = 2	3 x 1 = 3	4 x 1 = 4
1 x 2 = 2	2 x 2 = 4	3 x 2 = 6	4 x 2 = 8
1 x 3 = 3	2 x 3 = 6	3 x 3 = 9	4 x 3 = 12
1 x 4 = 4	2 x 4 = 8	3 x 4 = 12	4 x 4 = 16
1 x 5 = 5	2 x 5 = 10	3 x 5 = 15	4 x 5 = 20
1 x 6 = 6	2 x 6 = 12	3 x 6 = 18	4 x 6 = 24
1 x 7 = 7	2 x 7 = 14	3 x 7 = 21	4 x 7 = 28
1 x 8 = 8	2 x 8 = 16	3 x 8 = 24	4 x 8 = 32
1 x 9 = 9	2 x 9 = 18	3 x 9 = 27	4 x 9 = 36

HTML
基礎

概念

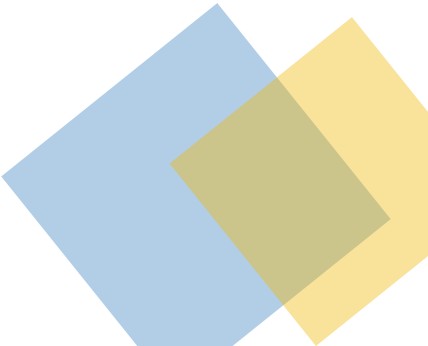


WEBSITE
= HTML + CSS + JS

什麼是 HTML？

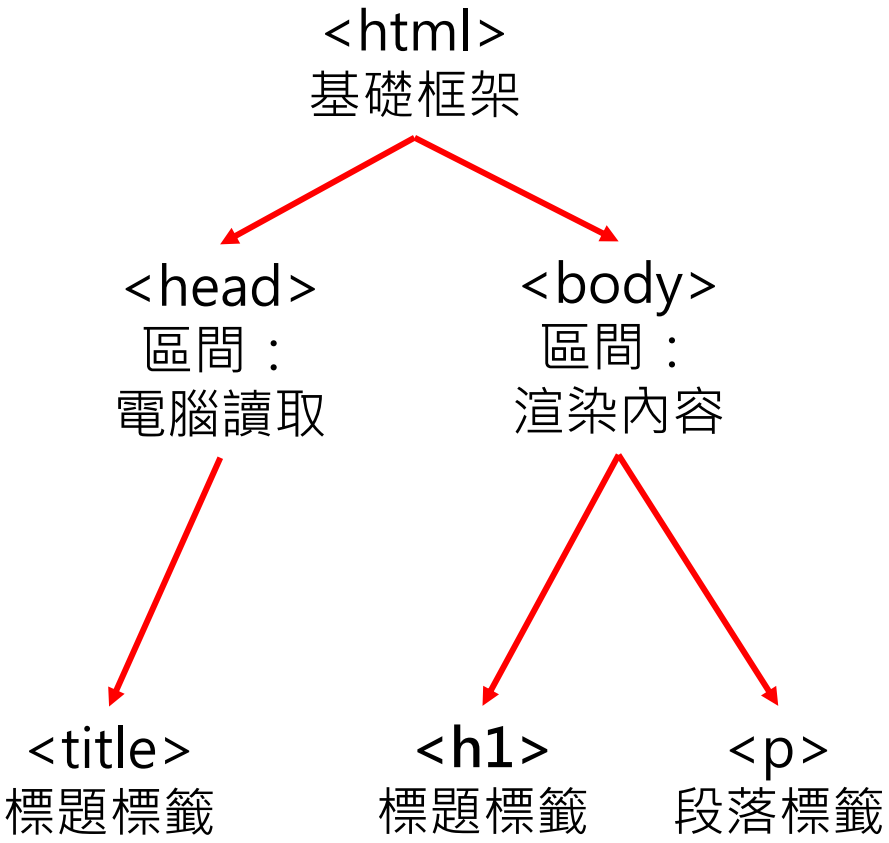
- HyperText Markup Language (超文件標示語言)
- 由一群元素 (Elements) 所組成的階層式文件
- 一個元素包含開始標籤、結束標籤、屬性以及內容

<標籤 屬性> 內容 </標籤>



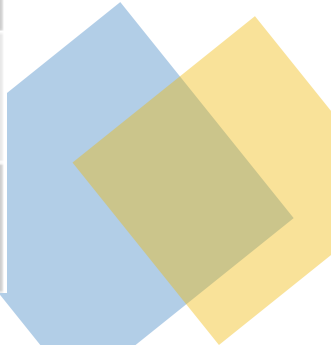
HTML 架構

```
<!DOCTYPE html>
<html>

    <head>
        <title>My First Webpage</title>
    </head>

    <body>
        <h1>
            My First Webpage
        </h1>
        <p>This is a paragraph...</p>
    </body>

</html>
```



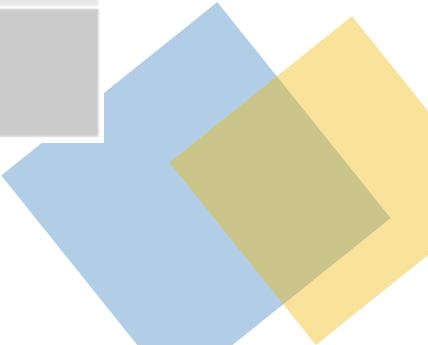
常用標籤（Tag）

標籤名稱	用途
<h1> - <h6>	標題
<p>	段落
<a>	超連結
<table>	表格
<tr>	表格內的 row
<td>	表格內的 cell
 	換行 (無結束標籤)



常用屬性（Attributes）

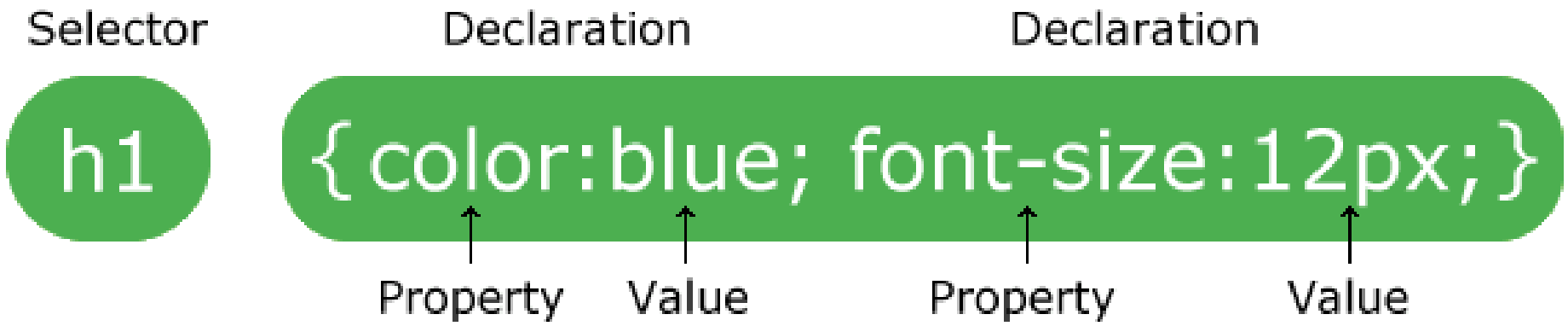
屬性名稱	意義
class	標籤的類別 (可重複)
id	標籤的 id (不可重複)
title	標籤的顯示資訊
style	標籤的樣式
data-*	自行定義新的屬性



CSS
基礎

什麼是 CSS？

- Cascading Style Sheets (串接樣式表)
- 一種用來替 HTML 增加 style 的語言，舉凡修改顏 色、字體大小、字體類型等等，皆由 CSS 完成

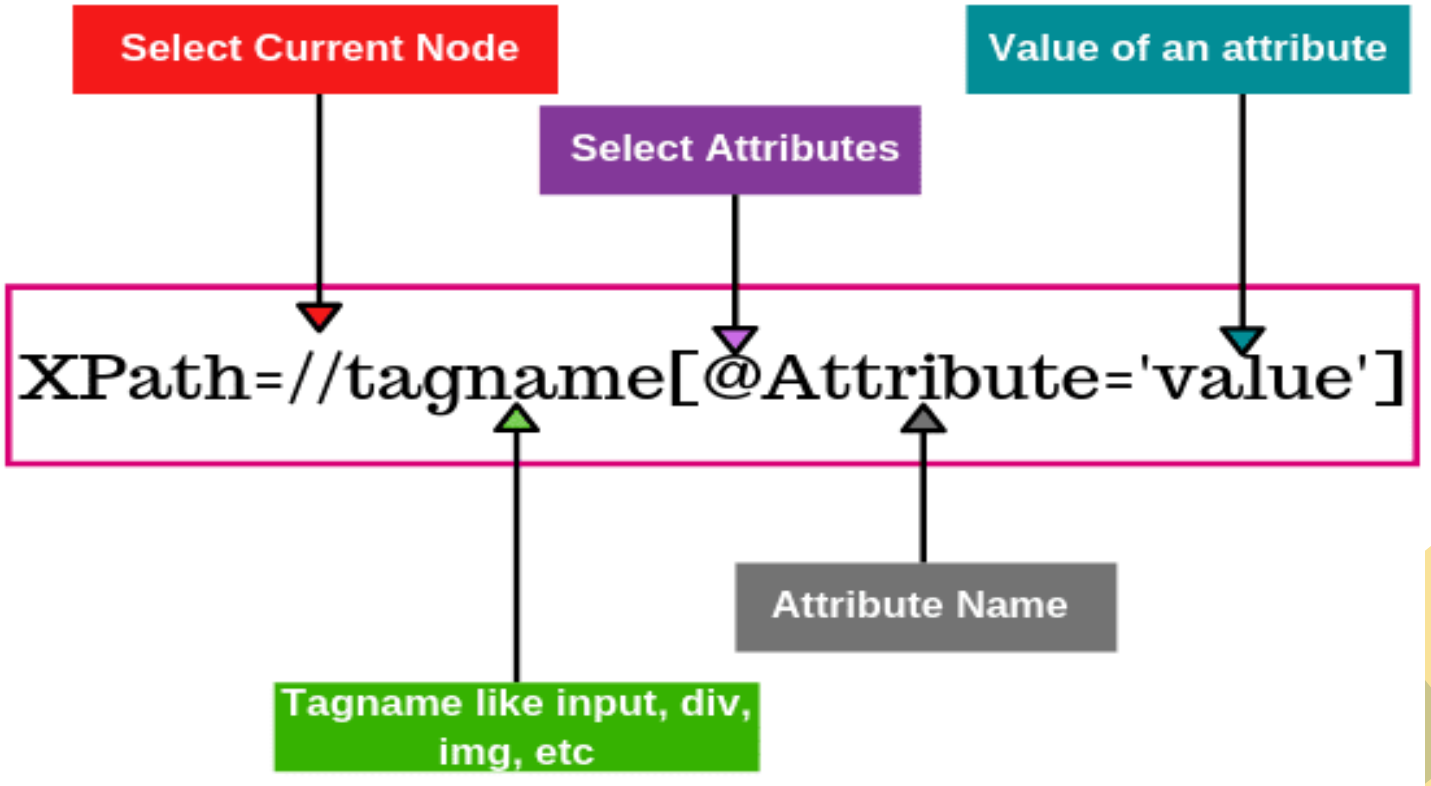


XPATH
語言

什麼是 XPATH？

- XML PATH Language（路徑語言）
- 一種在 XML 文檔中查找資訊的語言。

  XPath 最初設計是用來搜尋 XML 文檔的，但是它同樣適用於 HTML 文檔的搜索。



常用語法

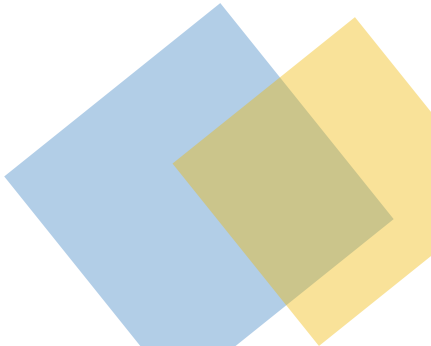
類別	運算式	描述
層級	/	從根結點選取(當前節點的下一級)
	//	從當前節點選取任意子孫節點(跨級)
	.	選取當前節點(如.//h3表示當前節點下跨級匹配h3標籤)
屬性	@	屬性訪問(如div[@class="xx"]或//a/@href)
函數	text()	獲取節點文本內容
	contains(A,B)	A是否包含B
	last()	標籤列表中的最後一個(如//a[last()])

小測驗

```
<!DOCTYPE html>
<html>

    <head>
        <title>My First Webpage</title>
    </head>

    <body>
        <h1>
            My First Webpage
        </h1>
        <p>This is a paragraph...</p>
    </body>

</html>
```

1.取 title 標籤的文字

2.取 p 標籤的文字



大測驗

https://law.moj.gov.tw/Index.aspx

1.截取【**刑事訴訟法**】的路徑

2.截取【**本月瀏覽人次的數字**】的路徑

大測驗

https://law.moj.gov.tw/Index.aspx

1.截取【**刑事訴訟法**】文字的路徑

//div[@class='hotlist']//li[3]/a/text()

2.截取【**本月瀏覽人次的數字**】的路徑

//div[@class='clearfix container']/div[@class='items'][2]//span/text()

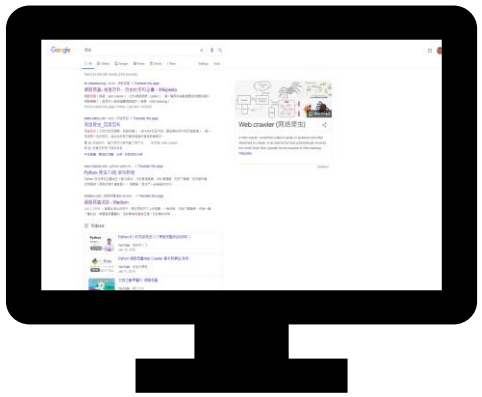
爬蟲實作

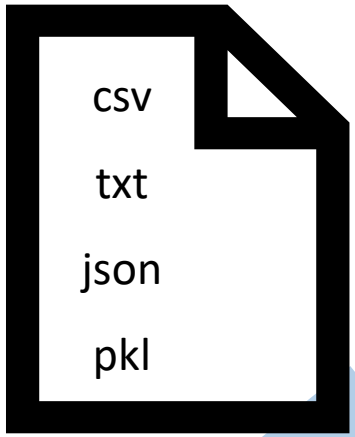
執行流程 – 用戶請求流程

登入網頁

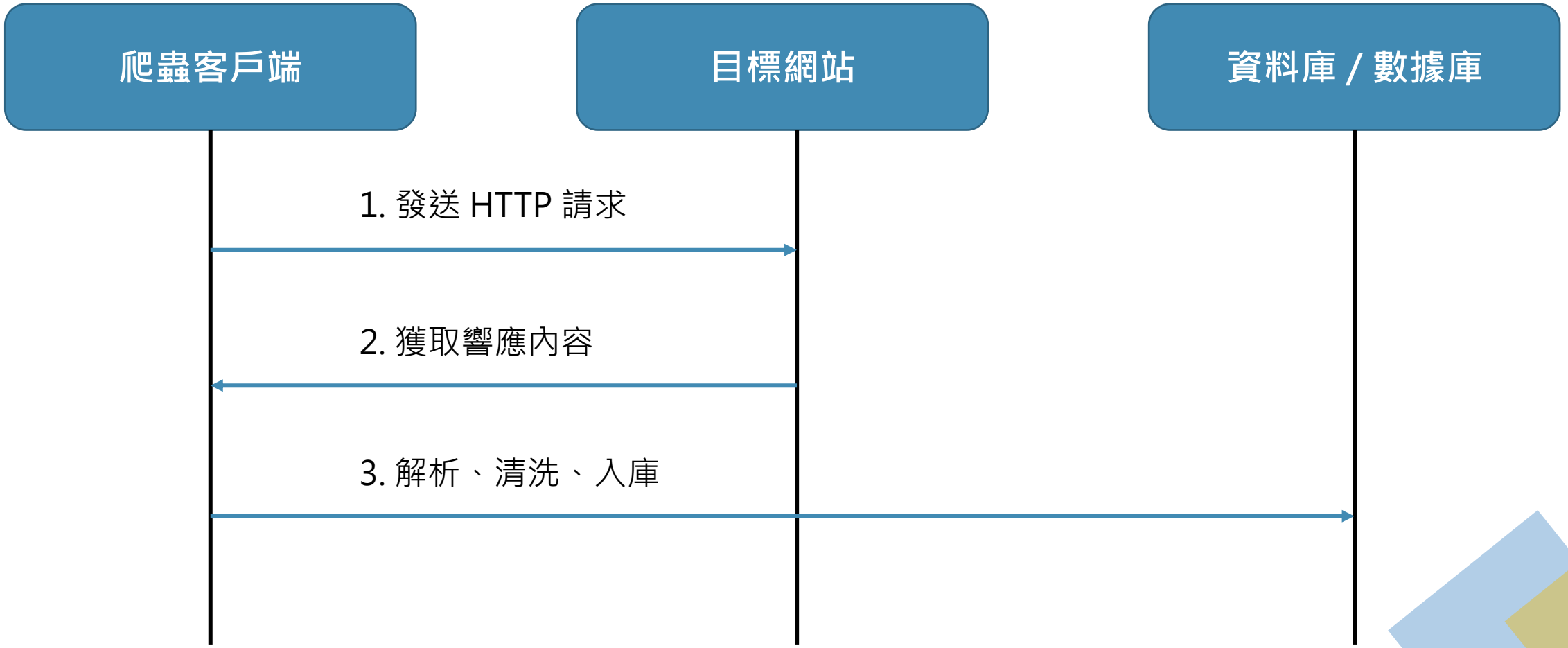
搜尋資料

下載資料





執行流程 – 程式請求流程



GET vs POST（封包）

```
GET /?id=010101 HTTP/1.1
Host: xxx.toright.com
User-Agent: Mozilla/5.0 (Windows; U; Windows NT 5.1; zh-TW; rv:1.9.2.13) Gecko
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: zh-tw,en-us;q=0.7,en;q=0.3
Accept-Encoding: gzip,deflate
Accept-Charset: UTF-8,*
Keep-Alive: 115
Connection: keep-alive
```

GET

```
POST / HTTP/1.1
Host: xxx.toright.com
User-Agent: Mozilla/5.0 (Windows; U; Windows NT 5.1; zh-TW; rv:1.9.2.13) Gecko
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: zh-tw,en-us;q=0.7,en;q=0.3
Accept-Encoding: gzip,deflate
Accept-Charset: UTF-8,*
Keep-Alive: 115
Connection: keep-alive

Content-Type: application/x-www-form-urlencoded
</code><code>Content-Length: 9
id=010101
```

POST

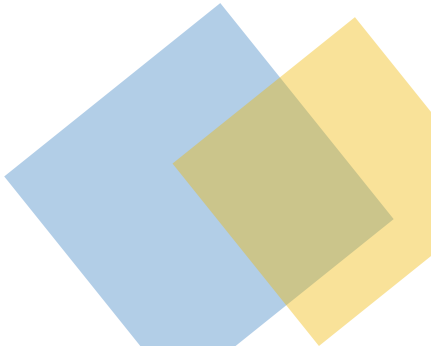
Python 套件安裝語法

- 方法 1： 在 Cell 中輸入 !pip install <套件名稱> ，然後執行

```
In [ ]: !pip install pandas
```

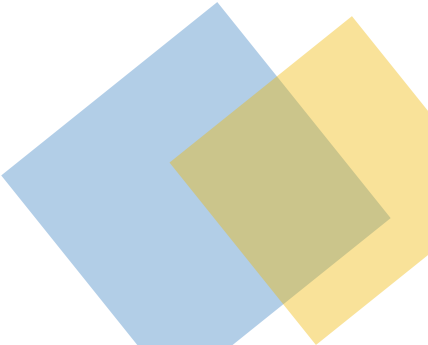
- 方法 2： 在終端機中輸入 pip install <套件名稱> ，然後按 ENTER 鍵

```
(base) C:\Users\sefx5>pip install pandas
```



本學年需下載的指定套件

安裝語法（套件）	用途
pip install pandas	資料處理、探勘
pip install numpy	向量處理
pip install requests	爬蟲
pip install selenium	爬蟲
pip install lxml	爬蟲定位工具
pip install beautifulsoup4	爬蟲定位工具
pip install plotly	繪圖



爬蟲：全國法規資料庫

https://law.moj.gov.tw/Index.aspx

今日任務：

在輸入指定關鍵字搜尋後，截取所有法規結果並將個別法規內之條文內容截取。

*** 在尋找目標網址時，通常其內容會出現在開發人員工具中的【XHR】及【HTML】項目中。

課程作業

作業

- 下載本學期所需套件（參考 1031_04.pdf ）
- 完成觀看 Class_04 影片，並將你對爬蟲流程的理解以流程圖繪出，其中可附加任何詳細資訊（工具、語言、注意事項等），包裹不在 PPT 範圍內之內容皆可接受。

*** 作業繳交格式範例： 06170171_陳偉傑_流程圖.jpg （僅接受 jpg / jpeg / png 格式）

THANKS

LAWTECH
learn coding get outstanding