

下午場Outline

- Matplotlib
- Plotting with Pandas
- Seaborn

- HTML & CSS (基本概念)
- 網頁爬蟲

講師：方聖瑋(Willy Fang)

Email: w70024@gmail.com

國立中央大學-資管所管理組

網頁與爬蟲

HTML 與 CSS 初探

講師：方聖瑋(Willy Fang)

Email: w70024@gmail.com

國立中央大學-資管所管理組

01

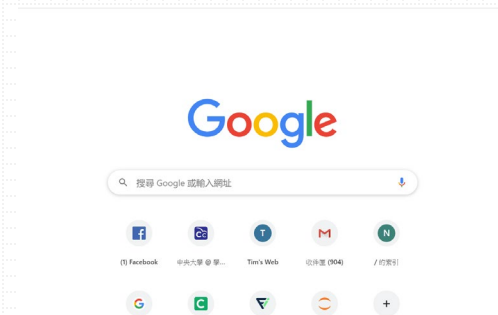
網頁的組成與架構 - HTML





1. 何謂HTML

- HTML 的全名為 “Hyper Text MarkUp Language” ，即「超文件標記語言」，是編寫網頁的一種標籤式的描述語言。透過不同的標籤製造出不同的東西。
- 瀏覽器(Chrome/IE Edge/Firefox) 解讀後呈現各種標籤的功能。

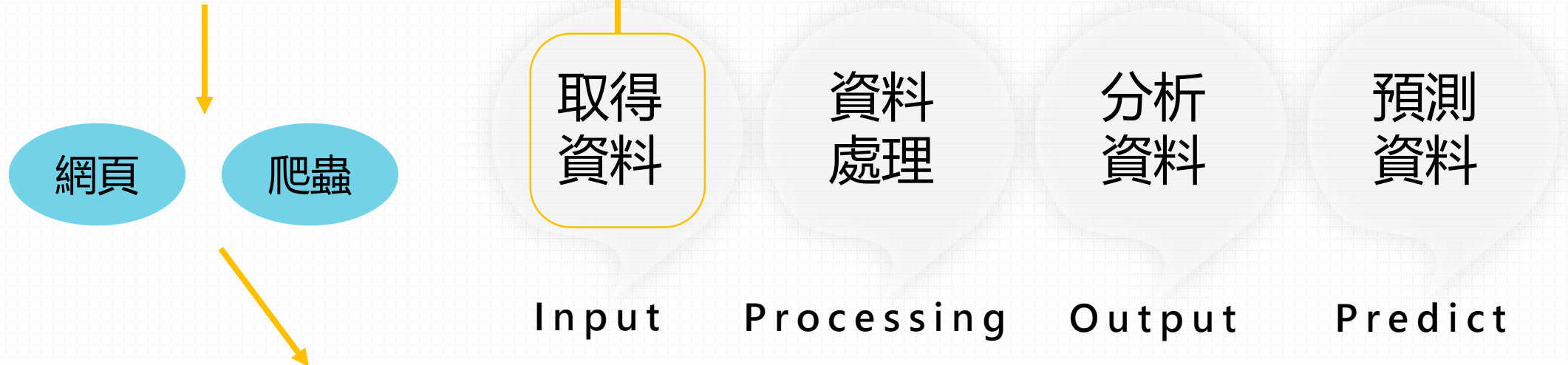




1.1 爬蟲與HTML

1. 網路上現有的資料集
2. 從企業資料庫取得
3. 從網頁取得原始資料

這是我們熟悉的資料分析步驟



因此我們必須要有網頁HTML的概念！



小考

1.2 HTML架構

HTML版本：文件類型 (doctype)

```
<!DOCTYPE html>
```

`<html>` 用來存放所有網頁元素、標籤的內容

```
<head>
```

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

```
<title>我的第一個網頁</title>
```

```
</head>
```

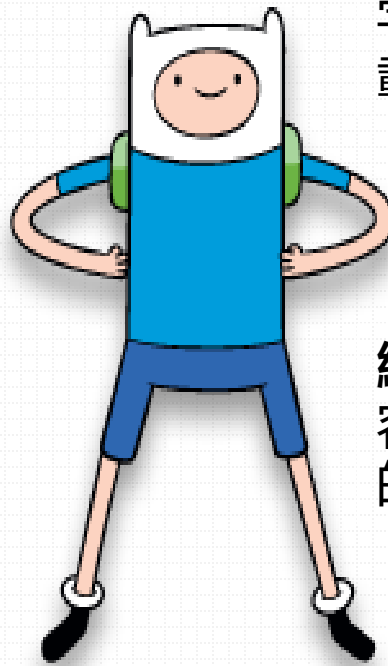
```
<body>
```

```
Hello World !
```

```
</body>
```

```
</html>
```

網頁基本資料與設定。 Ex: 文字編碼、標題、預設設備大小、載入css、js檔案等等



網頁的內容物。 將想呈現的內容透過排版與css、js與各式的標籤呈現出來。



1.3 標籤架構

開始標記

`<html>`包住整份文件`</html>`

結束標記

* `<head></head>` # 網頁的資訊、引用檔案，**不顯示於瀏覽器中**

① `<title></title>` # 網頁的標題

② `<meta charset="UTF-8">` # 文件編碼為UTF-8格式

* `<body></body>` # 網頁的主內容，**顯示於瀏覽器中**



1.3.1 標籤屬性

<meta charset="UTF-8">

標籤名稱

屬性名稱

屬性值

- * 表明了你要提供標籤內容甚麼類型的額外資訊
- * 記得內容要加上雙引號!



1.3.2 標籤屬性

屬性名稱 開始標籤

``點我進入Yahoo首頁``

標籤名稱 屬性值 結束標籤

給予初學者練習網頁設計的線上即時編輯器！

<https://codepen.io/willy-fang/pen/pogPKZG>



1.3.3 常用標籤-標題

`<h1>我是標題</h1>`

```
<body>

<h1>這是標題</h1>
<h2>這是標題</h2>
<h3>這是標題</h3>
<h4>這是標題</h4>
<h5>這是標題</h5>
<h6>這是標題</h6>

</body>
```



這是標題

這是標題

這是標題

這是標題

這是標題

這是標題

- *HTML的標題有六「階」

- *主標、次標...依序下去

- *每個瀏覽器顯示的文字大小都會有一點差距



1.3.4 常用標籤-段落

<p>我是段落</p>

```
<body>  
  <p>我是段落</p>  
  <p>我是段落</p>  
  <p>我是段落</p>  
</body>
```



我是段落

我是段落

我是段落

跳到下一個段落，中間空一行
(如同在Word中按"Enter")



1.3.5 常用標籤-換行

不需結束標籤!

```
<body>  
  <p>我是段落</p>  
  <p>我是段落</p><br>  
  <p>我是段落</p>  
</body>
```



我是段落

我是段落

我是段落

直接跳到下一行
(如同在Word中按"Shift"+"Enter")



1.3.6 常用標籤 – 不加

什麼都不加

```
<body>
```

我是段落

我是段落

我是段落

```
</body>
```



我是段落 我是段落 我是段落

都在同一行什麼都不做
(僅僅只是加個空格)



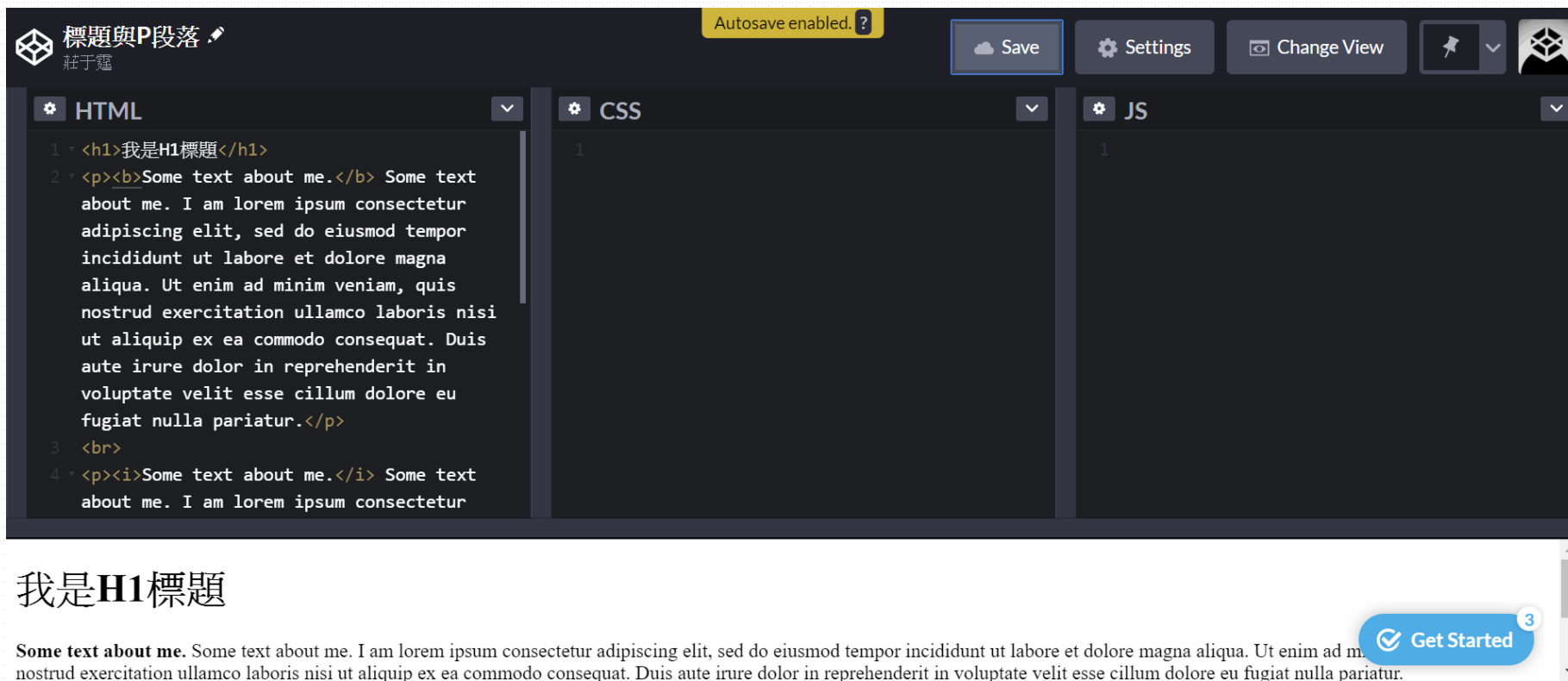
1.3.7 其他常用文字標籤

說明	tags	原始內容	輸出結果
Bold	<code></code>	<code>這是加粗的文字</code>	這是加粗的文字
Italic	<code><i></code>	<code><i>這是斜體的文字</i></code>	這是斜體的文字
Superscript	<code><sup></code>	這是 <code><sub></code> 上標 <code></sub></code> 的文字	這是上標的文字
Subscript	<code><sub></code>	這是 <code><sup></code> 下標 <code></sup></code> 的文字	這是下標的文字
Big	<code><big></code>	<code><big>這是放大的文字</big></code>	這是放大的文字
Small	<code><small></code>	<code><small>這是縮小的文字</small></code>	這是縮小的文字
Horizontal rules	<code><hr></code>	水平分隔線 <code><hr></code> 水平分隔線	水平分隔線 _____ 水平分隔線



1.3.8 自己試試看 - Codepen

給予初學者練習網頁設計的線上即時編輯器！



<https://reurl.cc/V60Gly>



1.3.9 常用標籤 - 編號與項目清單

` `

製作編號清單的元素

` `

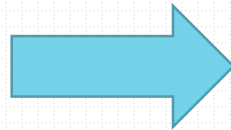
清單中的每個項目

```
<!DOCTYPE html>
<html>
<body>
```

```
<ol>
  <li>Coffee</li>
  <li>Tea</li>
  <li>Milk</li>
</ol>
```

```
<ol start="50">
  <li>Coffee</li>
  <li>Tea</li>
  <li>Milk</li>
</ol>
```

```
</body>
</html>
```



1. Coffee
2. Tea
3. Milk

50. Coffee
51. Tea
52. Milk



1.3.9 常用標籤 - 編號與項目清單

` `

製作**項目清單**的元素

` `

清單中的每個項目

```
<!DOCTYPE html>
<html>
<body>

<h4>An Unordered List:</h4>
<ul>
  <li>Coffee</li>
  <li>Tea</li>
  <li>Milk</li>
</ul>

</body>
</html>
```



An Unordered List:

- Coffee
- Tea
- Milk



1.3.10 常用標籤 - 連結

```
<a href="" target="_blank">我是連結</a>
```

檔案、圖片或網頁路徑

連結開啟方式

製作超連結(Ex.針對文字)

- * 連結到其他網站href的屬性值就是該網站的完整網址。
- * 連到同一個網站的不同網頁(網站中所有網頁都存在同一個資料夾裡), 那href屬性的值就是檔名即可。
- * 在新視窗開啟連結<a>標記中使用target屬性" _blank"
- * href="#id", 連到網頁特定部分(特定ID)(屬於css的部分)



1.3.11 常用標籤-圖片



圖片失效的替代文字

```
<img src = 'https://aaa/img/image.jpg' alt = '圖片簡述' title = '圖片標題'>
```

圖片位址/來源

替代文字

圖片標題

當網站圖片顯示不出來時，所顯示的文字

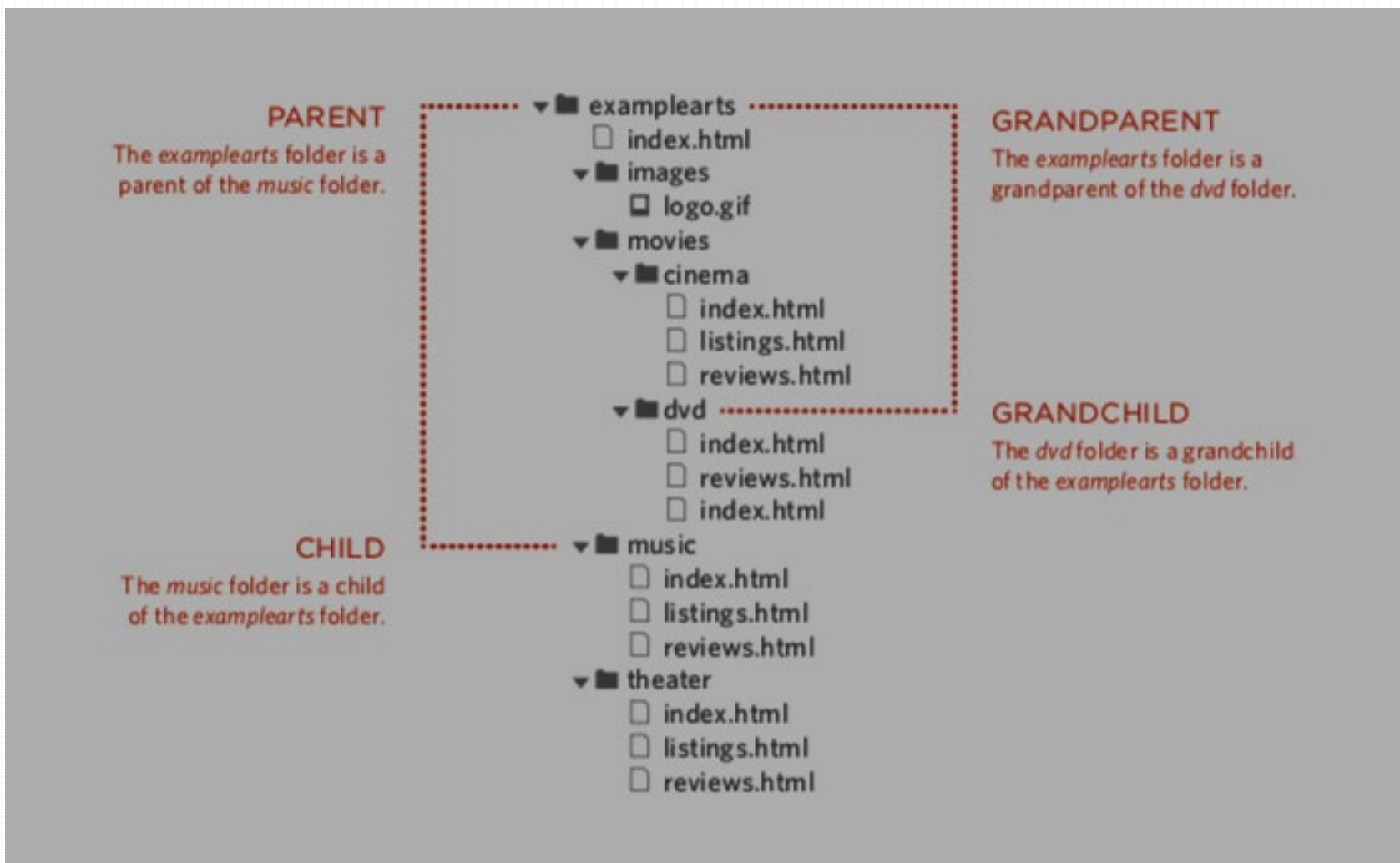
```
style = "width:300px; height:300px"
```

更改圖片寬跟高



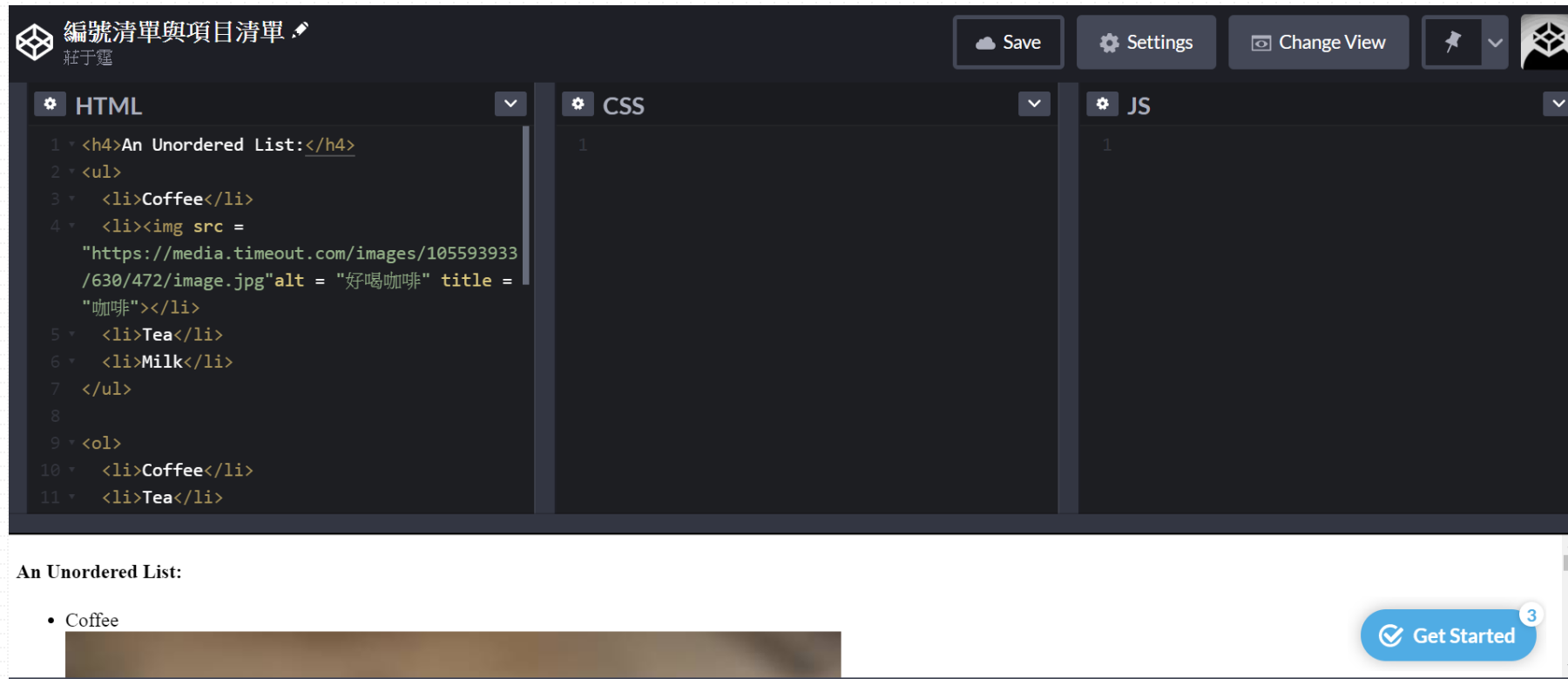


1.3.12 網站架構圖





1.3.13 自己試試看 - Codepen



<https://reurl.cc/exMv8W>



1.3.14 常用標籤 - table

`<table></table>`

表格

`<tr></tr>`

列

`<th></th>`

標頭欄

`<td></td>`

欄

colspan屬性 - 定義可橫跨的欄數

rowspan屬性 - 定義可橫跨的列數

```
<table>
  <tr>
    <th></th>
    <th>9am</th>
    <th>10am</th>
    <th>11am</th>
    <th>12am</th>
  </tr>
  <tr>
    <th>Monday</th>
    <td colspan="2">Geography</td>
    <td>Math</td>
    <td>Art</td>
  </tr>
  <tr>
    <th>Tuesday</th>
    <td colspan="3">Gym</td>
    <td>Home Ec</td>
  </tr>
</table>
```



	9am	10am	11am	12am
Monday	Geography		Math	Art
Tuesday	Gym			Home Ec



`<thead></thead>`

此段為表格的標題

`<tbody></tbody>`

此段為表格的內容

`<tfoot></tfoot>`

此段為表尾

Date	Income	Expenditure	表頭
1st January	250	36	表身
2nd January	285	48	表身
31st January	129	64	表身
	7824	1241	表尾

```
<table>
  <thead>
    <tr>
      <th>Date</th>
      <th>Income</th>
      <th>Expenditure</th>
    </tr>
  </thead>
  <tbody>
    <tr>
      <th>1st January</th>
      <td>250</td>
      <td>36</td>
    </tr>
    <tr>
      <th>2nd January</th>
      <td>285</td>
      <td>48</td>
    </tr>
    <!-- additional rows as above -->
    <tr>
      <th>31st January</th>
      <td>129</td>
      <td>64</td>
    </tr>
  </tbody>
  <tfoot>
    <tr>
      <td></td>
      <td>7824</td>
      <td>1241</td>
    </tr>
  </tfoot>
</table>
```

表頭

表身

表尾



1.3.15 版本標籤與註解

HTML版本(Doctype)

因HTML的版本不同，標籤語法可能有少許的差異，需在文件第一行宣告

```
HTML5 HTML
<!DOCTYPE html>

HTML 4
<!DOCTYPE html PUBLIC
"-//W3C//DTD HTML 4.01 Transitional//EN"
"http://www.w3.org/TR/html4/loose.dtd">
```

註解


<!--我是這段程式的註解-->

在開發程式時，寫下註解幫助未來維護此程式的人快速讀懂此程式的功能

```
1 <!--我是這段程式的註解-->
2 <p>我是段落</p>
3 <p>我也是段落</p>
4 <table border='1'>
5   <tr>
6     <th>標題一</th>
7     <th>標題二</th>
8     <th>標題三</th>
9   </tr>
10  </table>
```




1.3.16 自己試試看 - Codepen

 講義
莊子霆

SaveSettingsChange View

HTML

1 * <!--我是這段程式的註解-->
2 * <p>我是段落</p>
3 * <p>我也是段落</p>
4 * <table border='1'>
5 * <tr>
6 * <th>標題一</th>
7 * <th>標題二</th>
8 * <th>標題三</th>
9 * </tr>
10 * <tr>
11 * <th>內容1</th>
12 * <th>內容2</th>
13 * <th>內容3</th>

CSS

1

JS

1

標題一	標題二	標題三
內容1	內容2	內容3

	9am	10am	11am	12am
Monday	Geography	Math	Art	
Tuesday	gym		Home Etc	

Get Started 3

<https://reurl.cc/V6Ypzb>



1.3.17 常用網頁標籤整理

標籤名稱	用法舉例
<div> </div>	<div class= "apple" >內容</div>
 + 	 項目1 項目1
<p> </p>	<p>內容</p>
<a> 	內容
	
<h1>、 <h2>、 ...、 <h6>	<h1>第一章</h1>
<input>	<input type= "text" name= "user_name" >
<button> </button>	<button id= "banana" onclick= "myFinction()" >點我 </button>



1.3.18 網頁的詳細資訊 - meta

```
<!DOCTYPE html>
<html>
  <head>
    <title>Information About Your Pages</title>
    <meta name="description"
      content="An Essay on Installation Art" />
    <meta name="keywords"
      content="Installation, art, opinion" />
    <meta name="robots"
      content="nofollow" />
    <meta http-equiv="author"
      content="Jon Duckett" />
    <meta http-equiv="pragma"
      content="no-cache" />
    <meta http-equiv="expires"
      content="Fri, 04 Apr 2014 23:59:59 GMT" />
  </head>
  <body>
  </body>
</html>
```

網站描述

網站關鍵字

爬蟲控制項

作者

快取控制項

快取期限控制項

[https://blog.xuite.net/kajidojl/mymin/19670583-\[HTML\]+META+標籤筆記](https://blog.xuite.net/kajidojl/mymin/19670583-[HTML]+META+標籤筆記)

02

網頁的美術材料 - CSS





2.1 class屬性與id屬性

```
<!doctype html>
<html>
  <head>...</head>
  <body>
    <div id="topbar-container">...</div>
    <div id="navigation-container">...</div>
    <div id="main-container">
      <div id="main-content" class="bbs-screen bbs-content">
        <div class="article-meteline">
          <span class="article-meta-tag">作者</span>
          <span class="article-meta-value">zhtw (劍走龍蛇 以血作畫)</span>
        </div> == $0
      <div class="article-meteline-right">
        <span class="article-meta-tag">看板</span>
        <span class="article-meta-value">LoL</span>
      </div>
      <div class="article-meteline">
        <span class="article-meta-tag">標題</span>
        <span class="article-meta-value">[閒聊] 弗力貝爾美術設計師是腦袋去撞到?</span>
      </div>
      <div class="article-meteline">
        <span class="article-meta-tag">時間</span>
        <span class="article-meta-value">Sat May 9 14:57:15 2020</span>
      </div>
    </div>
  </body>
</html>
```

同一種標籤，在同一個網頁重複使用

要針對特定部分的標籤去美化、調整



class屬性與id屬性



2.1 class屬性與id屬性

層次分類的概念



tag:

class: 'cat'

id: 'cat_1'



tag:

class: 'dog'

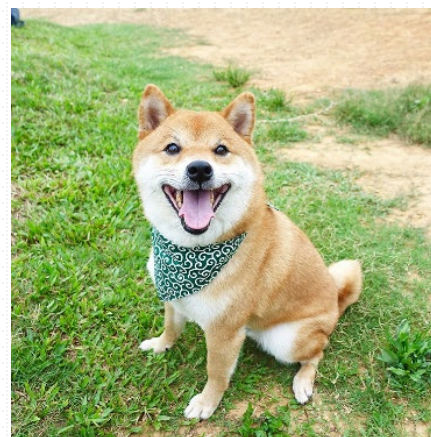
id: 'dog_Tim'



tag:

class: 'cat'

id: 'cat_2'



tag:

class: 'dog'

id: 'dog_Willy'

```
<img class = '種類' id = '名字' src = 'https://aaa/img/kittydoggy.jpg' >
```



2.1 class屬性與id屬性

標籤(tag)、class、id的優先順序

id

>

class

>

tag



2.1 class屬性與id屬性



tag:

class: 'cat'

id: 'cat_1'



tag:

class: 'dog'

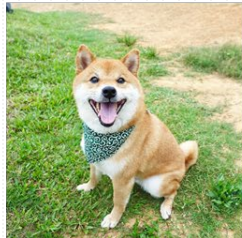
id: 'dog_Tim'



tag:

class: 'cat'

id: 'cat_2'



tag:

class: 'dog'

id: 'dog_Willy'

對img標籤(tag)去做調整

```
img{  
  width:300px;  
  height:200px;  
}
```

CSS語法



2.1 class屬性與id屬性



tag:

class: 'cat'

id: 'cat_1'



tag:

class: 'cat'

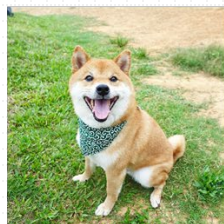
id: 'cat_2'



tag:

class: 'dog'

id: 'dog_Tim'



tag:

class: 'dog'

id: 'dog_Willy'

class優先於tag

同時對img標籤(tag)與class屬性去做調整

```
img{  
  width:300px;  
  height:200px;  
}
```

```
.cat{  
  width:600px;  
  height:400px;  
}
```



2.1 class屬性與id屬性



tag:

class: 'cat'

id: 'cat_1'



tag:

class: 'cat'

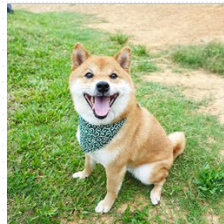
id: 'cat_2'



tag:

class: 'dog'

id: 'dog_Tim'



tag:

class: 'dog'

id: 'dog_Willy'

id優先於class

同時對class屬性與id屬性去做調整

```
.dog{  
  width:300px;  
  height:200px;  
}
```

```
#dog_Tim{  
  width:600px;  
  height:400px;  
}
```



2.1 class屬性與id屬性

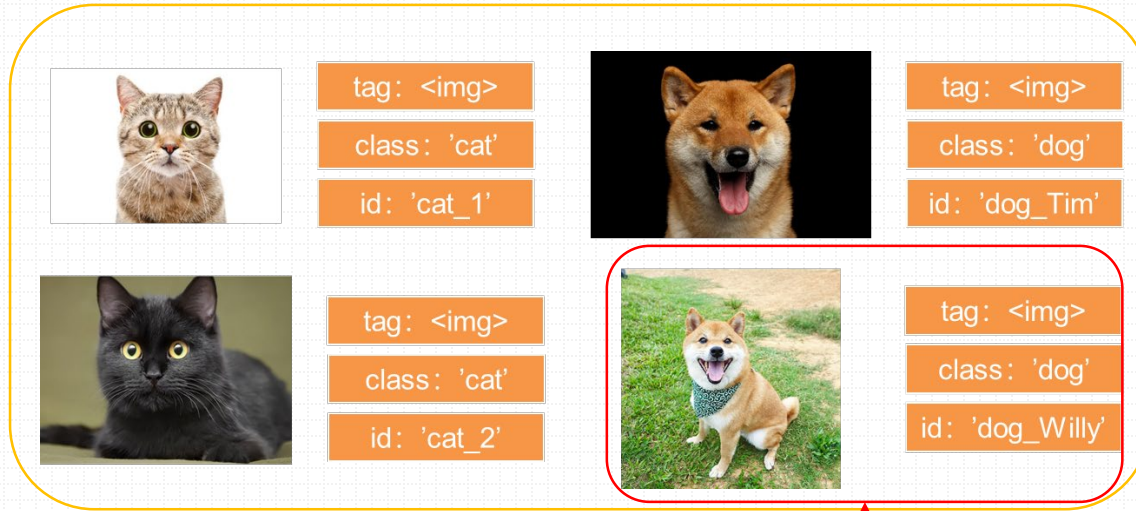
id優先於tag

對tag屬性去做調整

```
img{  
  width:300px;  
  height:200px;  
}
```

對id屬性去做調整

```
#dog_Willy{  
  width:400px;  
  height:300px;  
}
```



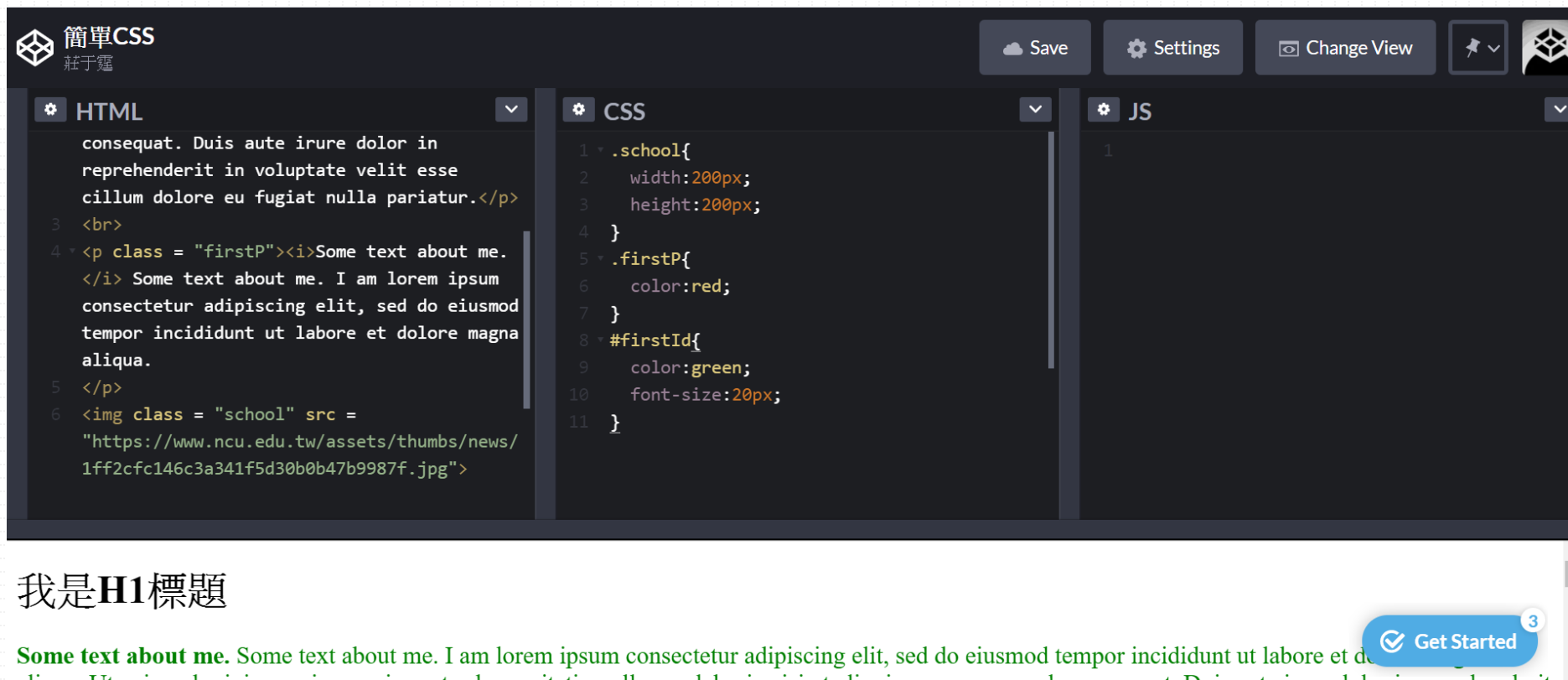


2.2 常見的css屬性

屬性名稱	用法舉例	屬性名稱	用法舉例
width	width:100px;	text-align	text-align:center;
height	height:100px;	float	float:left;
background-color	background-color:#FFC000;	background-image	url("../img/a.png")
color	color:white	padding	padding:10px;
font-size	font-size:24px;	margin	margin:10px;
font-family	font-family:Arial;	text-decoration	text-decoration:none
font-weight	font-weight:bold;	line-height	line-height:32px;
border	border:2px solid gray;	display	display:none;



2.3 自己試試看 - Codepen



<https://reurl.cc/d0QzE2>



讓我們來看看今天要爬的網頁!

批踢踢實業坊 > 看板 PokemonGO 聯絡資訊 關於我們

看板 精華區 最舊 < 上頁 下頁 > 最新

搜尋文章...

27	[閒聊] 0.175程式碼 bmw3633	5/07 ...
35	Re: [情報] 近期將進行寶可幣系統的小測試 success0409	5/07 ...
25	[閒聊] 團體戰有bug,別用飛鬼龍開場 tracheids	5/07 ...
爆	[閒聊] 世代大挑戰2020:城都 任務 lycs0908	M 5/08 ...
12	[情報] 三星活動代碼 Quentin5566	5/08 ...
8	[閒聊] 布魯限定調查日開始嘍 altcd	5/09 ...
8	[情報] 布魯限定調查 開始了	

<https://www.ptt.cc/bbs/PokemonGO/index.html>

03

簡單爬蟲實作



Contents



1、何謂爬蟲

2、爬蟲的流程與常用套件

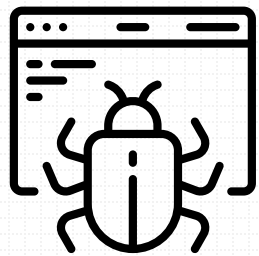
3a、爬蟲實作 - PTT板標題爬取

3b、爬蟲實作 - PTT板多頁面標題爬取

3c、爬蟲實作 - PTT板文章內文爬取



3.1 何謂爬蟲



網路爬蟲是可以自動化替你蒐集網頁上資訊的程式。

爬蟲原理就是藉由你的爬蟲程式去進行請求(Request)，在跟你的程式說接收到回傳(Response)後，該怎麼解析出你想要的資訊，最後再把他們存起來。

常見開發與應用：

電商
數據

比價
網站

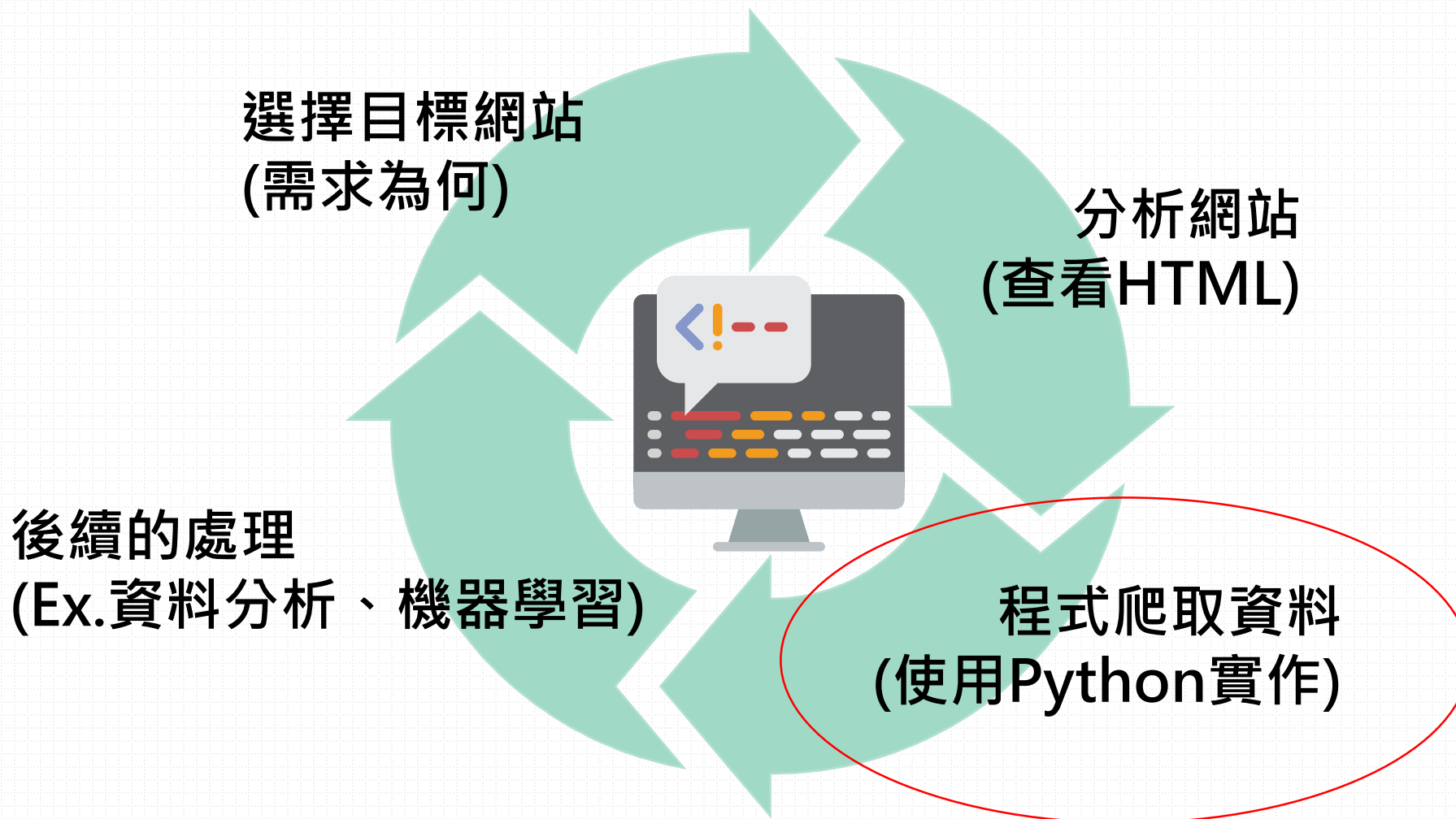
股票
追蹤

搜尋
引擎



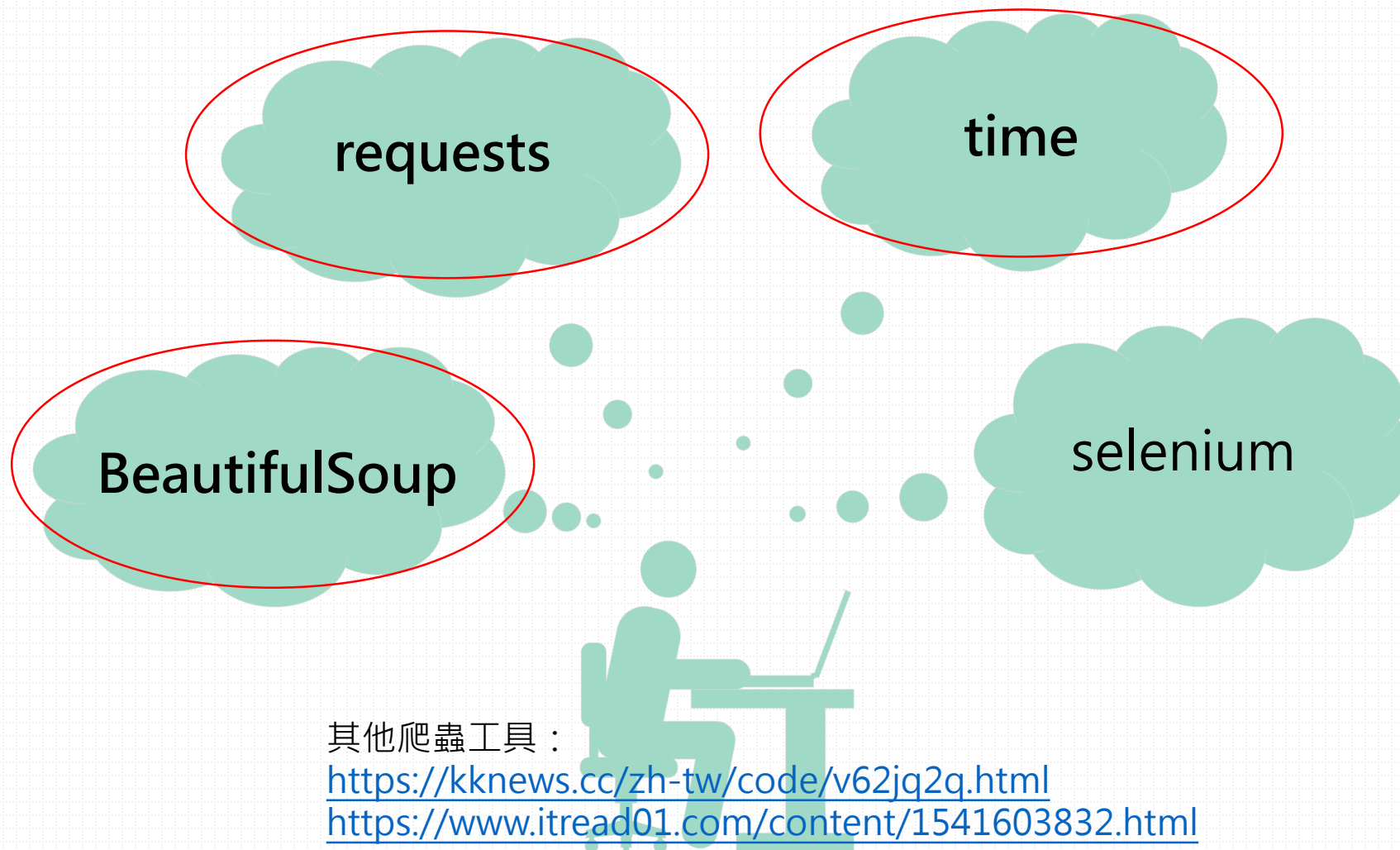
小考

3.2 爬蟲的流程





3.3 爬蟲的套件





小考

3.4 常用函數、屬性或方法

- `ptt_html = requests.get("網頁的網址") # 爬取網頁`
- `pttbs4 = BeautifulSoup(ptt_html.text, 'html.parser')`

`# 這個pttbs4是一種衍生性的物件`
- `pttbs4.find_all("標籤名") # 回傳「所有」符合條件的標籤資料`
- `pttbs4.select("標籤名")`

`#回傳「所有」符合條件的標籤資料，特別擅用於id、class`



3.5 爬蟲實作

批踢踢實業坊 > 看板 PokemonGO

看板 精華區 最舊 < 上頁 下頁 > 最新

搜尋文章... <https://www.ptt.cc/bbs/PokemonGO/index.html>

(已被bingtsien刪除) <wusun0609> 2-3 5/06

(已被bingtsien刪除) <wusun0609> 2-3 5/06

7 [發問] 皮卡丘不能進化??? tinchu 5/07 ...

78 [情報] 近期將進行寶可幣系統的小測試 kong221024 M

13 [發問] 2016/8 寶可夢去留QQ Moonmoonling

27 [閒聊] 0.175程式碼 bmw3633 5/07 ...

35 Re: [情報] 近期將進行寶可幣系統的小測試 success0409 5/07 ...

(已被bingtsien刪除) <ohoh880613> 請到置底

25 [閒聊] 團體戰有bug,別用飛鬼龍開場

實作開始!

START

Inspector Console Debugger Network Style Editor

Q a 15 of 383

```
<!DOCTYPE html>
<html>
  <head>
    <script>
      (function() {
        <div id="topbar">
          <a id="logo" href="/bbs/">批踢踢實業坊</a>
          <span></span>
          <a class="board" href="/bbs/PokemonGO/index.html"></a>
          <a class="right small" href="/about.html">關於我們</a>
          <a class="right small" href="/contact.html">聯絡資訊</a>
        </div>
        <div id="main-container">
          <script>
            <script src="//ajax.googleapis.com/ajax/libs/jquery/2.1.1/jquery.min.js">
            <script src="//images.ptt.cc/bbs/v2.27/bbs.js">
            <script type="text/javascript">
```

html > body > div#topbar-container > div#topbar.bbs-content

Layout Computed Changes Fonts Anim

Flexbox

Select a Flex container or item to continue.

Grid

CSS Grid is not in use on this page

Box Model

margin 0 0 0 0

border 0 0 0 0

padding 0 0 0 0

908.8x40

static

#topbar {
margin: 0 auto;
width: 100%;
height: 40px;
box-sizing: border-box;
color: #aaa;
}

...:13 @screen and (min-width: 768px)
.bbs-content, #article-polling {
font-size: 24px;
}

.bbs-content {
font-family: "細明體", "AR PL

程式碼：<https://reurl.cc/L3KpEX>

THANK YOU!

網頁爬蟲 主題結束!

有任何問題可聯繫：w70024@gmail.com