HTML 和 CSS 基礎

- 網路爬蟲的資料來源
  - HTML 5 網頁：
    - 網站如水果園
    - 每一頁 HTML 網頁為水果園中的一顆水果樹
    - 水果樹上的水果為 HTML 標籤
      - 水果：文字內容 HTML 標籤
      - 水果：圖片 HTML 標籤
      - 水果：超連結標籤
    - CSS 為 HTML 標籤的化妝師
      - CSS 並非目標資料，但可定位水果樹上的水果在何處。

- HTML 標籤語法與結構
  - HTML：
    - HyperText Markup Language
      - 超文字標示語言
      - 文件內容的格式編排語言
      - 瀏覽器中顯示的網頁內容

HTML 和 CSS 基礎

- HTML 標籤語法與結構
  - 瀏覽器檢查網頁內容 HTML 標籤：
    - 於瀏覽器開啟網頁(如Chrome)
    - 於網頁內容中『網路爬蟲課程』文字上按右鍵>檢查
      - 顯示此文字的 HTML 標籤 <h3>
      - 下方會顯示標籤的階層結構：
        - html>body>h3#title

HTML 和 CSS 基礎

- HTML 標籤語法與結構
  - HTML 標籤語法：
    - 語法格式：
      - <標籤名稱 屬性1=屬性值1 屬性2=屬性值2 > 文字內容</標籤名稱>
        - 多個屬性之間以空格隔開
        - 通常利用屬性標示不同的功能標籤
          - 在同一份網頁中可能會有多個相同的標籤名稱(如同水果樹上的水果都長得一樣)

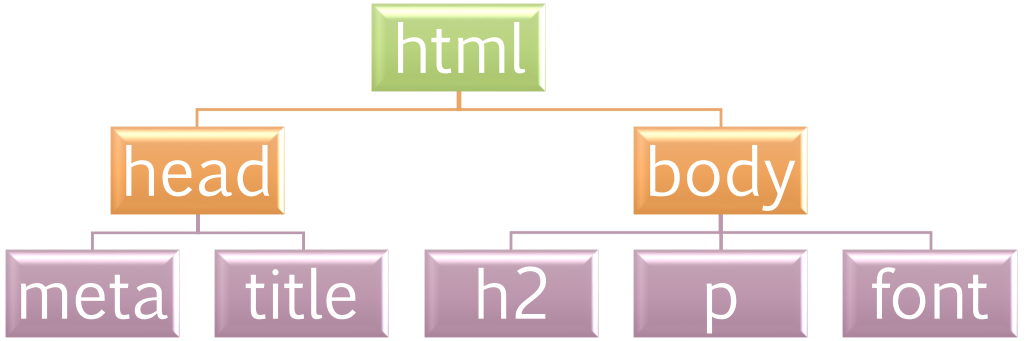
- HTML 標籤語法與結構
  - HTML 標籤語法：
    - 語法格式：
      - <標籤名稱 屬性1=屬性值1 屬性2=屬性值2 > 文字內容</標籤名稱>
        - 常用屬性：

屬性	功能說明
id	HTML 標籤的身分證號，是唯一值。(使用此屬性即可定位目標標籤)
class	HTML 標籤套用的樣式類別，其值為 CSS 選擇器。

HTML 和 CSS 基礎

- HTML 標籤語法與結構
  - HTML 網頁的標籤結構：
    - HTML 標籤可以構成多層次的巢狀結構：
      - <標籤1 屬性1=屬性值1 屬性2=屬性值2 >
        <標籤2 屬性1=屬性值1 屬性2=屬性值2 >
        ……</標籤2>
        <標籤3 屬性1=屬性值1 屬性2=屬性值2 >
        ……</標籤3>
        ……
        文字內容</標籤1>

- HTML 標籤語法與結構
  - HTML 網頁的標籤結構：



HTML 和 CSS 基礎

- HTML 標籤語法與結構
  - <head> 子標籤：
    - 功用：
      - 描述 HTML 網頁本身
      - 常用子標籤：

標籤	功能說明
<title>	顯示瀏覽器視窗標題列或標籤頁的標題文字
<meta>	提供 HTML 網頁的 metadata 資料，如網頁描述、關鍵字、作者、最近修改日期...等資訊。
<script>	標籤內容為客戶端的 script(腳本)程式碼，如 JavaScript 程式碼。
<style>	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。

HTML 和 CSS 基礎

- HTML 標籤語法與結構
  - <body> 子標籤：
    - 功用：
      - 瀏覽器看到的網頁內容，對網路爬蟲而言，此標籤的子標籤內容為要擷取的目標資料。
    - 在 HTML 標籤中，標籤的寫法有以下方式：
      - 只有開始標籤：如<hr>
      - 空內容標籤：如<hr></hr>
      - 開始結束合一：如<hr/>

HTML 和 CSS 基礎

- CSS 基礎
  - CSS：
    - Cascading Style Sheets
      階層式樣式表
    - 為樣式語言
    - 功用：
      - 描述標籤語言的格式
      - 重新定義 HTML 標籤的外觀

- CSS 基礎
  - CSS 樣式：
    - <p> 標籤在 HTML 中為段落標籤，並沒有定義色彩與字體大小...等格式。
    - 可以使用 CSS 重新定義 <p> 標籤的樣式。
    - <style type="text/css">
      p.author { font-size: 10pt;
      color: red; }
      </style>

資料標籤：文字和圖片標籤

- 文字內容標籤
  - 標題文字標籤：
    - <h1>~</h1>
      <h2>~</h2>
      <h3>~</h3>
      <h4>~</h4>
      <h5>~</h5>
      <h6>~</h6>
    - 功用：
      - 將文字顯示為標題文字
      - <h1>字體最大，<h6>字體最小。
      - 字體為粗體效果

資料標籤：文字和圖片標籤

- 文字內容標籤
  - 段落標籤：
    - HTML 網頁內容按下『Enter』不會換行
    - 換行標籤：
      - <p>：
        - 在該位置換行並增加一行距
      -
：
        - 在該位置換行但不增加行距

HTML網頁的文字以 p 標籤換行

HTML網頁的文字以br 標籤換行
HTML網頁的文字

資料標籤：文字和圖片標籤

- 文字內容標籤
  - 容器標籤：
    - ：
      - 換行並定義區塊顯示文字內容
    - ：
      - 定義單行元素(非區塊)，不換行。

西方人很多都是棕色眼睛
跟東方人不同， 西方人很多也都是淡藍色眼睛，跟東方人不同

數據資料爬取與分析

- 靜態網頁擷取：
  - 解析HTML網頁檔案
  - HTML檔案架構→樹狀結構→定義網頁元素
  - 元素架構：
    - 標籤(Tag)
    - 屬性(Attribute)
    - 內容(Content)

- 靜態網頁擷取：
  - 資料擷取中常用標籤：
    - <header>：
      - 網頁標頭資訊
    - <body>：
      - 網頁主體內容
    - ：
      - 網頁的一個區塊，區塊中包含許多元素
    - <title>：
      - 網頁標題

數據資料爬取與分析

- 靜態網頁擷取：
  - 資料擷取中常用標籤：
    - <h1>：
      - HTML 內文標題1
    - <a href>：
      - 網頁超連結
    - <form>：
      - 網頁表單
    - <tr>/<td>：
      - 表格列/表格欄

- 靜態網頁擷取：
  - 資料擷取中常用屬性：
    - id：
      - 代表唯一的元素
    - class：
      - 類別，類似元素的分類

數據資料爬取與分析

- 動態網頁擷取：
  - 動態網頁主要重點在網頁操作方式，要擷取資料，必須先了解網頁如何操作。
  - 模擬網頁與伺服器之間的操作互動，整個操作動作分解出個別步驟。

數據資料爬取與分析

- requests 模組
  - 功能：
    - 文件自動解碼
    - 文件自動解壓縮
    - 支援基本與摘要式認證
    - 分塊傳出編碼請求(Chunked編碼)
    - 連接池功能，對動態資料庫可以不必重新連線。
    - 連線超時處理機制。
    - key:value 結構的 cookie
    - 代理 IP

- requests 模組
  - 功能：
    - 國際化域名
    - 保持持續連線，直到一方中斷。
    - 文件分塊上傳
    - cookie 內含時間終了的運作
    - 流媒體文件下載
    - Unicode 格式的伺服器回應文件
    - .netrc 文件

數據資料爬取與分析

- requests 模組
  - 可以使用 Python 程式發出HTTP的請求，取得指定網站的內容。
  - 使用前必須先安裝，Anaconda 中已經內建。
  - 安裝指令：
    - pip install –U requests

- requests 模組
  - 發送 GET 請求
    - 當開啟瀏覽器輸入網址送出，指定的網站伺服器接收到要求後回應內容。
    - 即可在瀏覽器中看到網頁，這樣的請求方式稱為 GET。
    - requests模組可以不透過瀏覽器就能完成GET請求：
      - import requests
Response 物件 = requests.get('網址')