# Web Crawler plus

**Lecturer: Willy Fang 方聖瑋**

**Student ID: 108423020**

**School: NCU IM/MIS**

**Email: w70024@gmail.com**

# 目錄



- ✓ 靜態網頁
- ✓ 動態網頁
- ✓ Selenium套件
- ✓ 其他套件
- ✓ ……

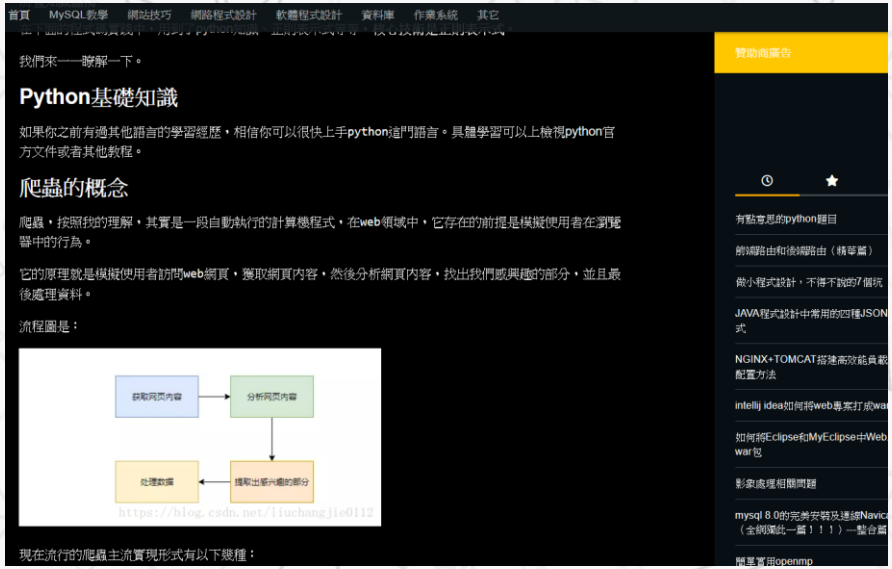
colab

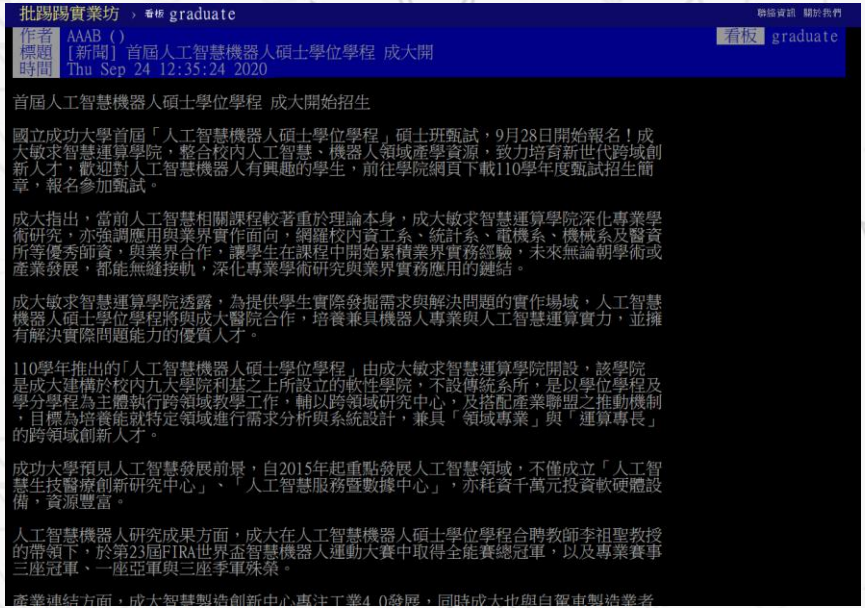
## 簡介

- ✓ 設計給User**純閱讀的網頁**。主要適用於簡單、更新不頻繁的內容
- ✓ **HTML**是基礎技術，常與**CSS**與**JavaScript**組合成一個內容網頁，讓瀏覽器去讀取
- ✓ 一般判斷方式是網頁**副檔名為html或htm**大多皆為靜態網頁，它的優勢在於容易為搜尋引擎所接受
- ✓ 現今較多網頁是動靜態混合型的設計，**純靜態的已經為少數**
- ✓ 所以很多動態網頁會將動態網頁轉變成靜態方式，就是所謂的**【偽靜態網頁】**來**提高搜尋引擎的友善度**達到排名優化的成效

# 1. 靜態網頁

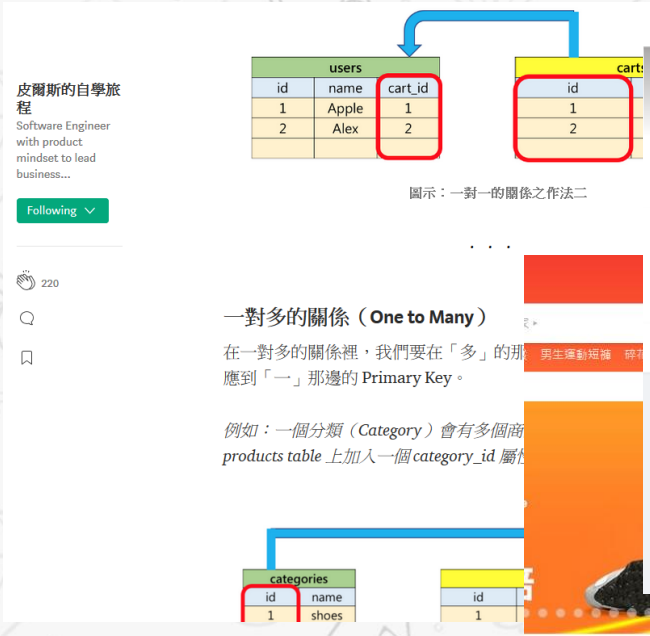


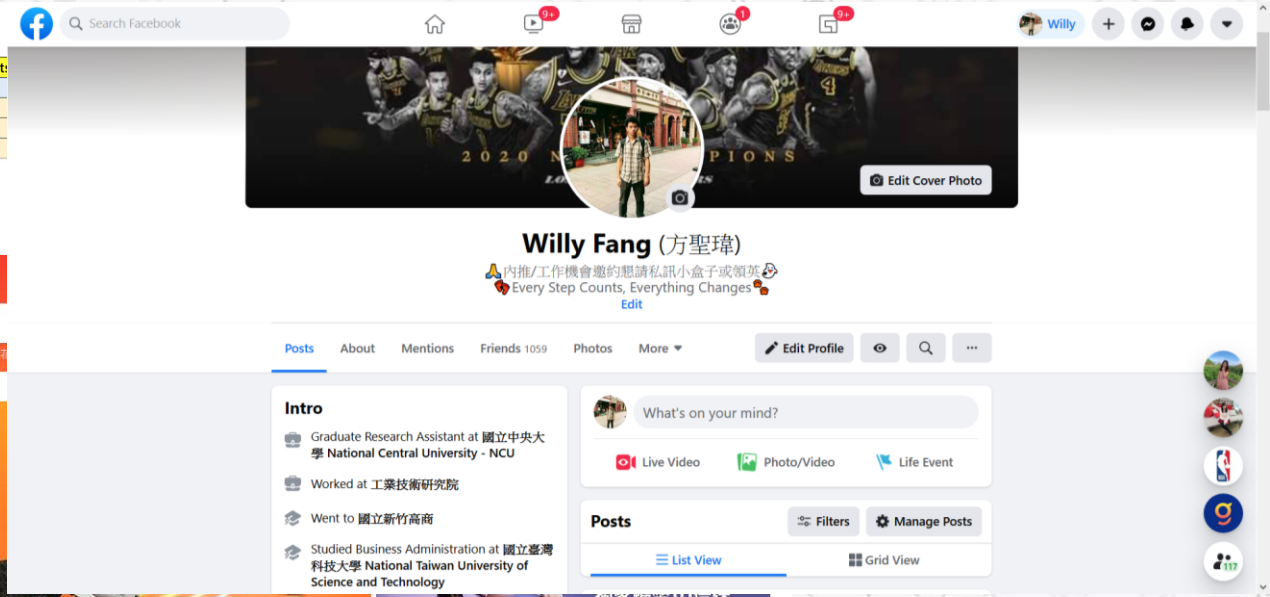

首頁 MySQL教學 網站技巧 網路程式設計 軟體程式設計 資料庫 作業系統 其它
我們來一一瞭解一下。
Python基礎知識
如果你之前有過其他語言的學習經歷，相信你可以很快上手python這門語言。具體學習可以上檢視python官方文件或者其他教程。
爬蟲的概念
爬蟲，按照我的理解，其實是一段自動執行的計算機程式，在web領域中，它存在的前提是模擬使用者在瀏覽器中的行為。
它的原理就是模擬使用者訪問web網頁，獲取網頁內容，然後分析網頁內容，找出我們感興趣的部分，並且最後處理資料。
流程圖是：
獲取網頁內容
分析網頁內容
處理數據
提取出感興趣的部分
https://blog.csdn.net/liuchanglie0112
現在流行的爬蟲主流實現形式有以下幾種：
贊助商廣告
有點意思的python題目
前端路由和後端路由（精華篇）
做小程式設計，不得不說的7個坑
JAVA程式設計中常用的四種JSON式
NGINX+TOMCAT搭建高效能負載配置方法
intellij idea如何將web專案打成war
如何將Eclipse和MyEclipse中Web war包
影象處理相關問題
mysql 8.0的完美安裝及連線Navicat（全網獨此一篇！！！）一整合篇
簡單實用openmp


批踢踢實業坊 › 看板 graduate
聯絡資訊 關於我們
作者 AAAB ()
標題 [新聞] 首屆人工智慧機器人碩士學位學程 成大開
時間 Thu Sep 24 12:35:24 2020
看板 graduate
首屆人工智慧機器人碩士學位學程 成大開始招生
國立成功大學首屆「人工智慧機器人碩士學位學程」碩士班甄試，9月28日開始報名！成大敏求智慧運算學院，整合校內人工智慧、機器人領域產學資源，致力培育新世代跨域創新人才，歡迎對人工智慧機器人有興趣的學生，前往學院網頁下載110學年度甄試招生簡章，報名參加甄試。
成大指出，當前人工智慧相關課程較著重於理論本身，成大敏求智慧運算學院深化專業學術研究，亦強調應用與業界實作面向，網羅校內資工系、統計系、電機系、機械系及醫資所等優秀師資，與業界合作，讓學生在課程中開始累積業界實務經驗，未來無論朝學術或產業發展，都能無縫接軌，深化專業學術研究與業界實務應用的鏈結。
成大敏求智慧運算學院透露，為提供學生實際發掘需求與解決問題的實作場域，人工智慧機器人碩士學位學程將與成大醫院合作，培養兼具機器人專業與人工智慧運算實力，並擁有解決實際問題能力的優質人才。
110學年推出的「人工智慧機器人碩士學位學程」由成大敏求智慧運算學院開設，該學院是成大建構於校內九大學院利基之上所設立的軟性學院，不設傳統系所，是以學位學程及學分學程為主體執行跨領域教學工作，輔以跨領域研究中心，及搭配產業聯盟之推動機制，目標為培養能就特定領域進行需求分析與系統設計，兼具「領域專業」與「運算專長」的跨領域創新人才。
成功大學預見人工智慧發展前景，自2015年起重點發展人工智慧領域，不僅成立「人工智慧生技醫療創新研究中心」、「人工智慧服務暨數據中心」，亦耗資千萬元投資軟硬體設備，資源豐富。
人工智慧機器人研究成果方面，成大在人工智慧機器人碩士學位學程合聘教師李祖聖教授的帶領下，於第23屆FIRA世界盃智慧機器人運動大賽中取得全能賽總冠軍，以及專業賽事三座冠軍、一座亞軍與三座季軍殊榮。
產業連結方面，成大智慧製造創新中心專注工業4.0發展，同時成大也與自駕車製造業者

## 簡介

- ✓ 設計與User互動的網頁。比較適用於資料內容較大、更新快速的內容，如社群平台、購物車、電商網站、討論區等，因此會需要規劃好伺服器、程式與資料庫間的共同運作
- ✓ 網頁的內容隨著User的輸入和互動而有所不同，有Perl、PHP、ASP、JSP等編譯方式
- ✓ 通常具有後台管理員系統，讓維護人員透過帳密登入後，可以更方便地管理網站，以大幅降低維護成本
- ✓ 新增最新消息、發佈參展資訊、修改商品說明、上傳產品圖片…等，只要會一般的電腦文書處理便可以輕鬆完成


皮爾斯的自學旅程
Software Engineer with product mindset to lead business...
Following
220
users
id name cart_id
1 Apple 1
2 Alex 2
id
1
2
圖示：一對一的關係之作法二
一對多的關係（One to Many）
在一對多的關係裡，我們要在「多」的那
應到「一」那邊的 Primary Key。
例如：一個分類（Category）會有多個商
products table 上加入一個 category_id 屬
categories
id name
1 shoes
id
1


Search Facebook
Willy
Edit Cover Photo
Willy Fang (方聖瑋)
內推/工作機會邀約懇請私訊小盒子或領英
Every Step Counts, Everything Changes
Edit
Posts About Mentions Friends 1059 Photos More
Edit Profile
Intro
Graduate Research Assistant at 國立中央大學 National Central University - NCU
Worked at 工業技術研究院
Went to 國立新竹高商
Studied Business Administration at 國立臺灣科技大學 National Taiwan University of Science and Technology
What's on your mind?
Live Video
Photo/Video
Life Event
Posts
Filters
Manage Posts
List View
Grid View


搶購去
9/18 - 9/30
逛逛去
物金
蝦皮電子館
蝦時尚美妝館
男人館$1零寶戰
蝦皮團購
海外直送
蝦皮超便宜
刷卡&活動


千萬$補貼 x 超級 3 日

## 簡介

- ✓ 在網頁App開發時，測試UI是相當麻煩的工作，若以手動測試進行不僅會因為人力時間而受到限制，且容易出錯。Selenium解決了此問題，它能藉由指令自動操作網頁，達到自動化測試的功能。
- ✓ 由於Selenium可以直接以程式碼操控瀏覽器，因此成為網路爬蟲必備的工具/套件之一
- ✓ 支援多種程式語言，包含Java、JavaScript、Python等

## 前提

- ✓ 啟用Selenium之後，尼指定的瀏覽器就會開啟，並依照你所寫的程式指令依序執行，所有網頁的操作，包含: **輸入資料(如帳密)、點選按鈕、滾動頁面(js)、放大視窗**等，都可以使用程式碼進行
- ✓ 由於是**偽裝**成真正的瀏覽器在運作，大致上多數網站都可以輕鬆的突破，但是**運行速度相對慢很多、佔用較多電腦效能**
- ✓ 一般而言，還是會先使用**requests**獲取網頁原始碼，但若無法突破對方網站伺服器的阻隔以及**需要有條件**時，就要用Selenium

## 操作

- ✓ 將在本機端IDE-Spyder進行，需先下載WebDriver程式，放置於程式碼檔案之同個目錄下
- ✓ 需額外安裝套件至本機端

  相信大家都裝好了（？
- ✓ from selenium import webdriver，即可用webdriver.Chrome()建立Google Chrome之瀏覽器物件→Selenium Webdriver API

# 方法/屬性/函數

- ✓ 需先宣告好物件：**browser = webdriver.Chrome()**
- ✓ 一般常用的方法：

  **.get(“url”)、.maximize_window()、.click()、.clear()、.send_keys()、.page_source、……(族繁不及備載)**

- ✓ 取得HTML網頁元素的方法：

  **等到講解程式碼在講唄！**

```
# 取得第一個符合的元素
'''browser.find_element_by_xxxxx()，跟上週教的soup.find()方法很類似

1. find_element_by_id()：用id屬性查詢符合的元素
2. find_element_by_class_name()：用class屬性查詢符合的元素（class屬性若有空格要加.）
3. find_element_by_link_text()：用"超連結"文字查詢符合的元素
4. find_element_by_partial_link_text()：用"超連結"的部分文字查詢符合的元素
5. find_element_by_xpath()：以xml的路徑查詢（全名為XML Path），xpath就是利用節點的樹狀關係，以及每個節點的特性來查詢符合的元素（簡單來說，類似就是利用絕對位置及相對位置去找，也可以用手刻的，方便使用但可能抓失敗，未來操作也可能會跑錯）

6. find_element_by_tag_name()：用標籤名稱查詢符合的元素
7. find_element_by_name()：用name屬性查詢符合的元素
8. find_element_by_css_selector()：用css選擇器定位（高階使用，比較難操作）

PS. 若element加個s，則會是取得所有符合的元素，跟上週教的soup.find_all()或soup.select()方法很類似
'''
```

# 其它套件



- ✓ requests
- ✓ bs4的Beautifulsoup
- ✓ time的sleep
- ✓ openpyxl
- ✓ json
- ✓ os
- ✓ ……

# 主題大綱

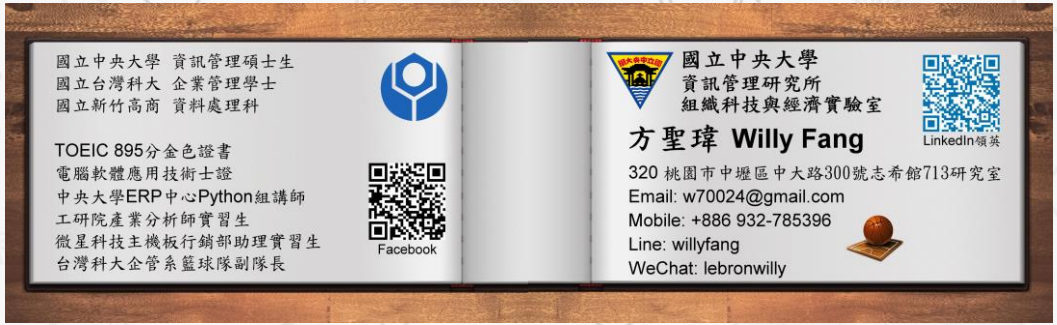


- ✓ （複習）PTT多頁標題
- ✓ Google Maps評論
- ✓ 登入Google&搜尋
- ✓ YouTube留言
- ✓ 政府Open Data
- ✓ Zara圖片
- ~~✓ FB粉專文章~~

*“If You’re Afraid to Fail,*

*You’re Probably Going to Fail.”*

*– Kobe Bryant*




LinkedIn: linkedin.com/in/lebronwillyfang/

Github: github.com/LeBronWilly/