

第一次破解爬蟲就上手

— Inspiration from 郭有鴻

好不容易找到喜歡的網站

寫好了爬蟲程式

但是你要的資料網站就是不讓你抓？

[illegible]

試試看新的方法吧！

使用範例：

TibaMe 全方位線上課程與職能學習平台

<https://www.tibame.com/courselibrary>



The banner for TibaMe features a vibrant blue and green background with stylized clouds and a mountain peak. On the left, a woman in a pink shirt and blue pants is running. On the right, two men are climbing a mountain. The text 'TibaMe 提拔你 | 資通訊即學、即戰、即就業' is at the top. Below it, the main message reads '嘿！一起在這裡學習，讓我們為你的人生 增值吧！'. Two yellow buttons are prominent: '立即索取 35 堂超值課程 免費註冊會員' and '和 145,830 會員一起學習 了解最新活動'. The bottom section lists four features with icons: '線上培訓 不限次數不限時間', '直播課程 技術疑問即時解答', '實作課程 一年內免費重聽一次', and '就業養成 就業率高達85%'.

TibaMe 提拔你 | 資通訊即學、即戰、即就業

嘿！一起在這裡學習，
讓我們為你的人生 增值吧！

立即索取 **35** 堂超值課程
免費註冊會員

和 **145,830** 會員一起學習
了解最新活動

線上培訓
不限次數不限時間

直播課程
技術疑問即時解答

實作課程
一年內免費重聽一次

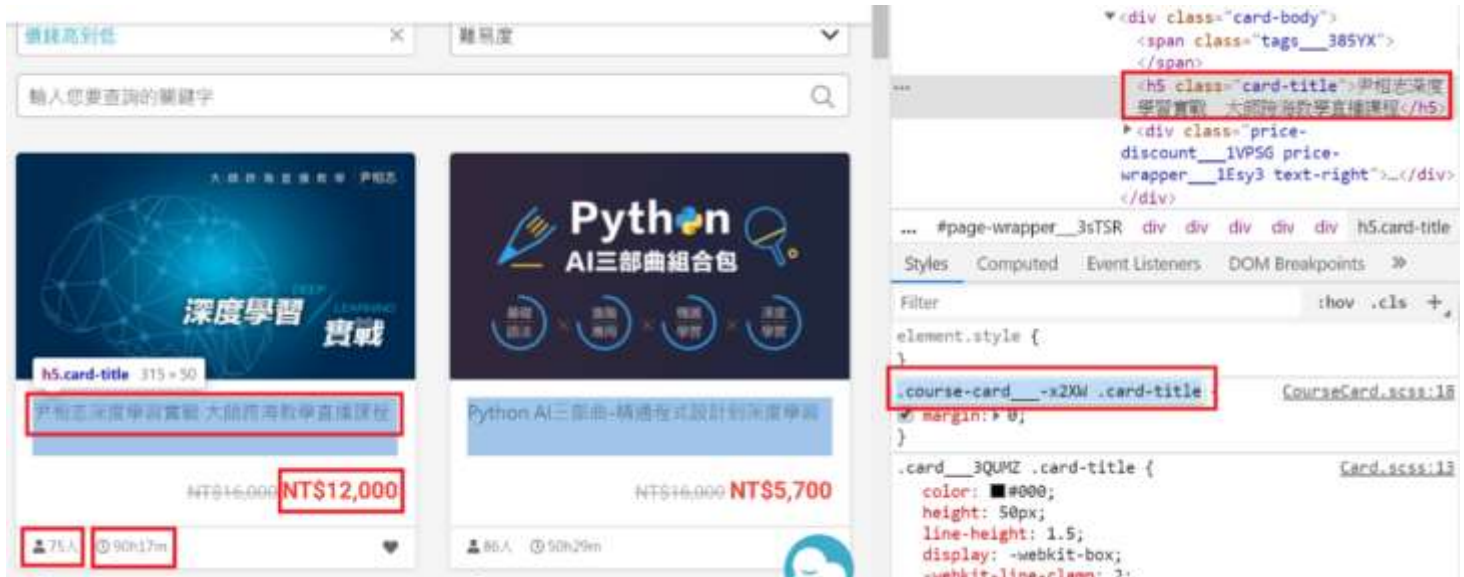
就業養成
就業率高達85%

準備工具：XAMPP



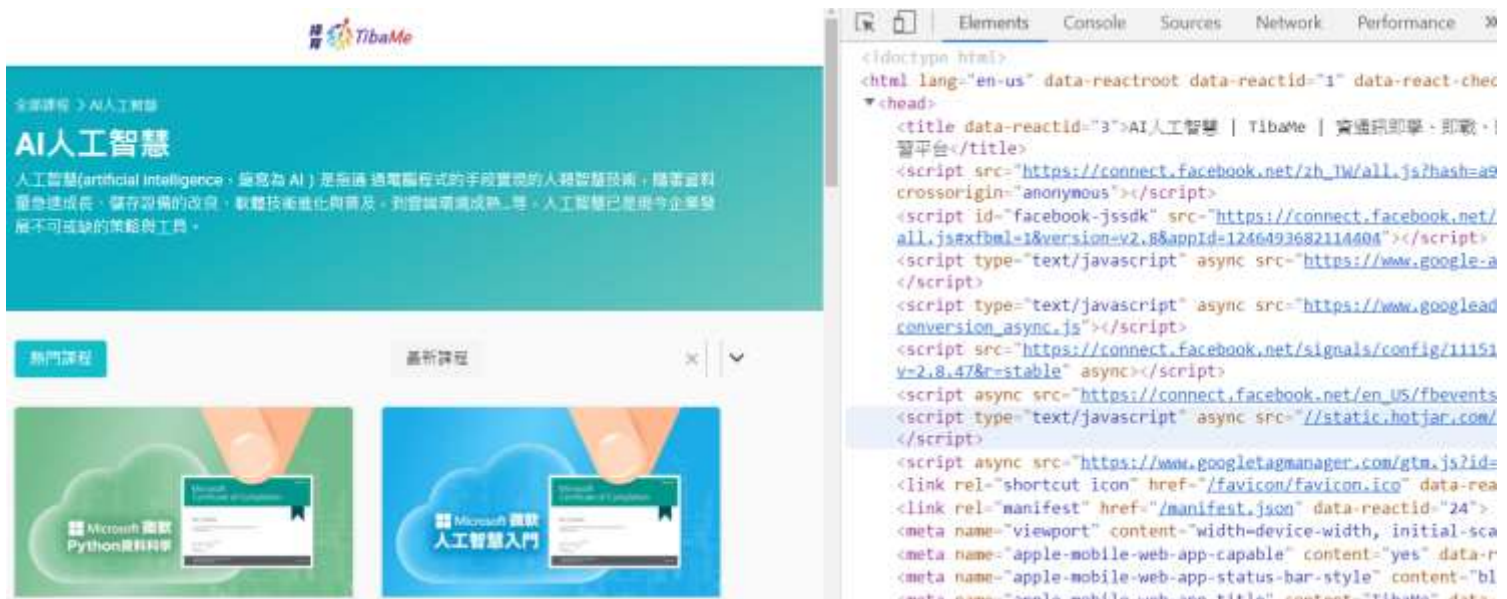
目標：

抓取課程名稱、課程價錢、修課人數、課程總長度，四種標籤



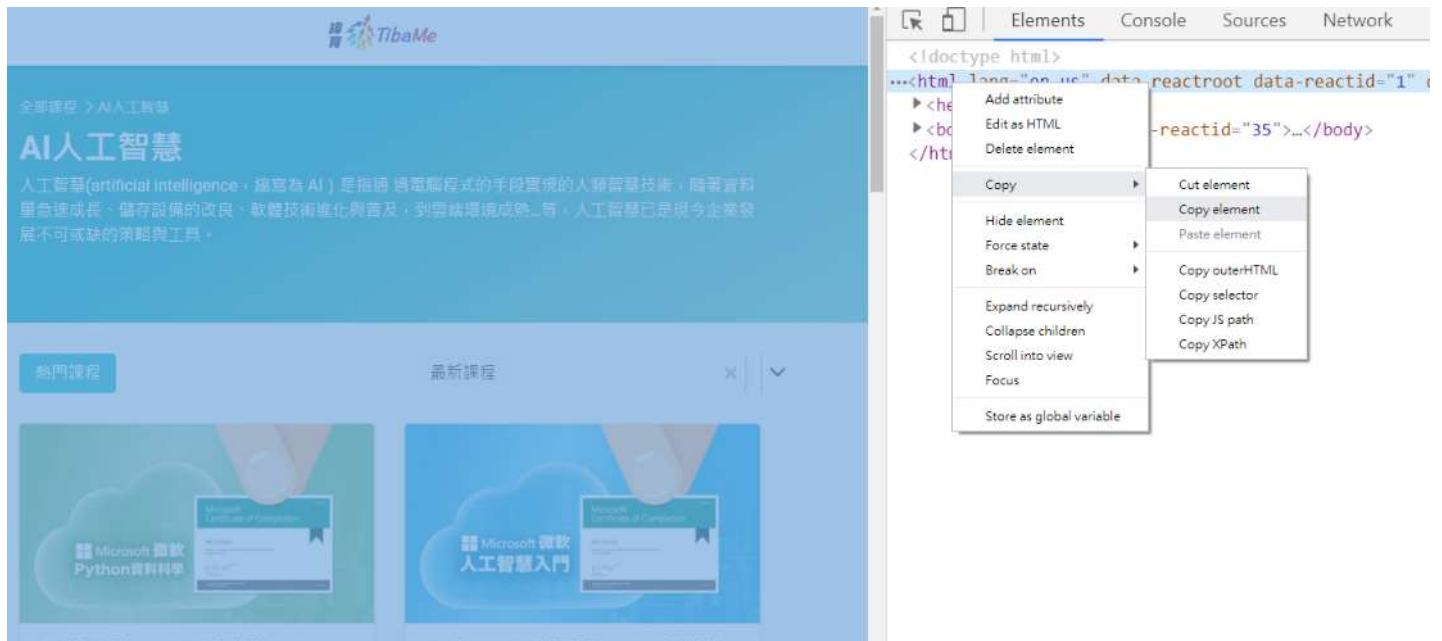
步驟一：

先開啟要爬蟲的網站，按 F12 打開開發人員工具，顯示出一邊網頁、一邊程式碼的模式



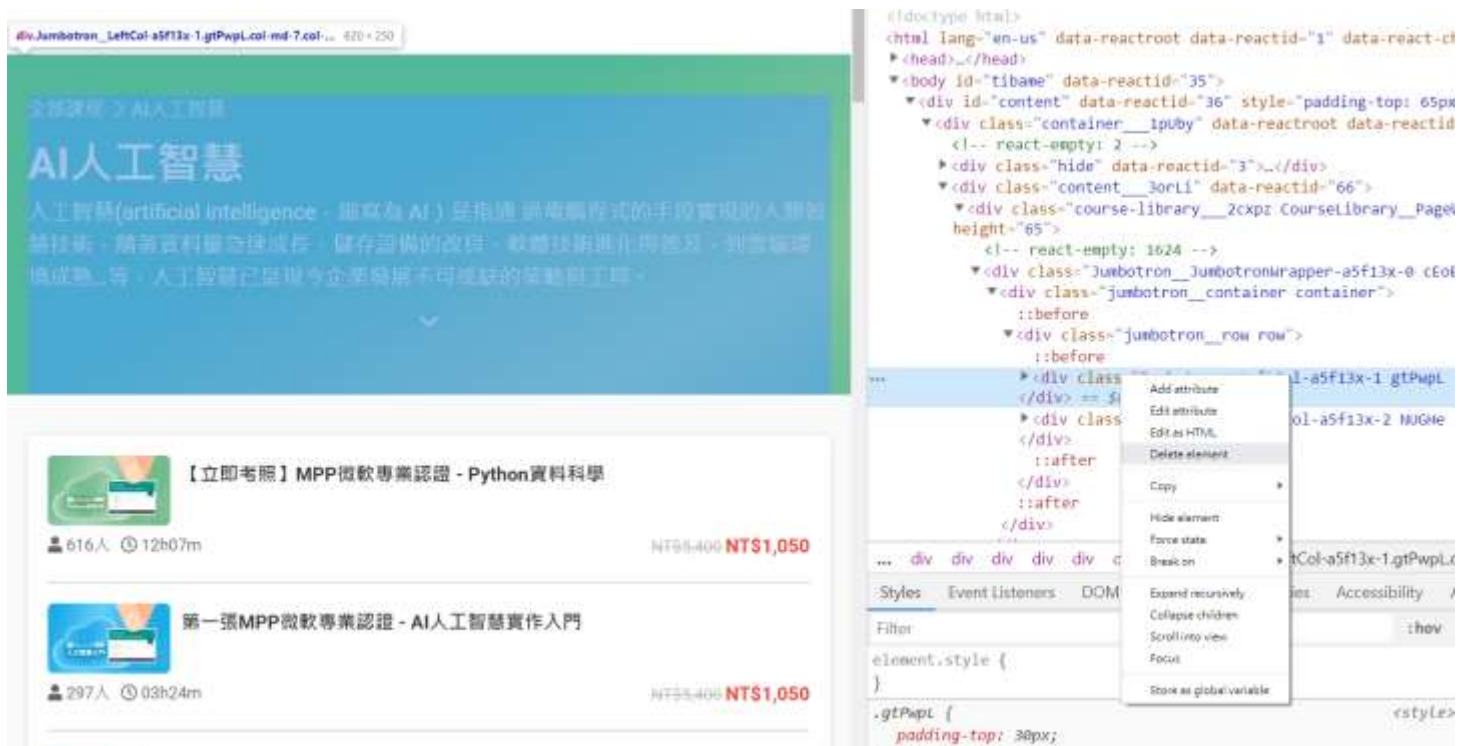
步驟二：

將所有 HTML 縮回去，右鍵點選第一個 HTML 複製所有元素



步驟二-補充：

也可以先刪除不必要元素(如：頁首、頁尾、區塊、Javascript function)，讓畫面保持整潔的同時還能辨識需要的元素，再進行複製，如果想重來只要重新整理畫面就好



步驟三：

開啟文字編輯器後貼上，記得在第一行補上<!doctype html>，儲存在
C:\xampp\htdocs 檔案目錄底下

```
1 <!doctype html>
2 <html lang="en-us" data-reactroot="" data-reactid="1" data-react-checksum="241643696" class=""><head>
3 <title data-reactid="3">AI人工智慧 | TibaMe | 資通訊即學、即戰、即就業 | 全方位線上課程與職能學習平台</title><link
  rel="shortcut icon" href="/favicon/favicon.ico" data-reactid="20"><link rel="manifest" href="/manifest.json"
  data-reactid="21"><meta name="viewport" content="width=device-width, initial-scale=1" data-reactid="22"><meta
  name="apple-mobile-web-app-capable" content="yes" data-reactid="23"><meta name="
  apple-mobile-web-app-status-bar-style" content="black" data-reactid="24"><meta name="
  apple-mobile-web-app-title" content="TibaMe" data-reactid="25"><meta name="theme-color" content="#ffffff"
  data-reactid="26"><link rel="stylesheet" href="https://fonts.googleapis.com/css?family=Roboto:400,700"
  data-reactid="27"><link rel="stylesheet" href="/dist/vendor-0b00dd56.css" data-reactid="28"><link rel="
  stylesheet" href="/dist/main-97e1bb87.css" data-reactid="29"><script src="https://connect.facebook.net/zh_TW/
  all.js?hash=39de0e2452e7c6931460e9131d55f386&ua=modern_es6" async="" crossorigin="anonymous"></script><
  script id="facebook-jssdk" src="https://connect.facebook.net/zh_TW/all.js#xfbml=1&version=v2.8&
  appId=1709181332733246"></script><script type="text/javascript" async="" src="https://www.googleadservices.com
  /pagead/conversion_async.js"></script><script type="text/javascript" async="" src="
  https://www.google-analytics.com/analytics.js"></script><script src="https://connect.facebook.net/signals/
  config/111513262705913?v=2.8.45&r=stable" async=""></script><script async="" src="
  https://connect.facebook.net/en_US/fbevents.js"></script><script type="text/javascript" async="" src="
  //static.hotjar.com/c/hotjar-1161734.js?sv=5"></script><script async="" src="https://www.googletagmanager.com/
  gtm.js?id=GTM-KQLWC8"></script><script src="/config.js?timestamp=1554085996365" data-reactid="30"></script><
  script async="" src="https://cdn-static.tibame.com/js/tinymce/4.9.2/tinymce.min.js" data-reactid="31"></script>
  <style type="text/css"></style><style data-styled-components=""></style><script type="text/javascript"
  charset="utf-8" async="" src="/dist/56-89120764.js"></script><script type="text/javascript" charset="utf-8"
  async="" src="/dist/7-5f6d5643.js"></script><script charset="utf-8" src="https://script.hotjar.com/
  modules.dece760f116806f8e142.js"></script><style type="text/css">iframe#_hjRemoteVarsFrame {display: none
  !important; width: 1px !important; height: 1px !important; opacity: 0 !important; pointer-events: none
  !important;}</style><style type="text/css" data-href="client/css/caas-plugin.css">#easychat-floating-button
  {
  background-color: transparent;
```

步驟四：

透過 localhost 開啟剛剛存的檔案，網址：localhost/你的檔名.html

難易度
[全級別基礎進階](#)
難易度 No Selection ▼
價錢高到低
[最新價錢高到低價錢低到高](#)

排序 No Selection ▼
輸入您要查詢的關鍵詞



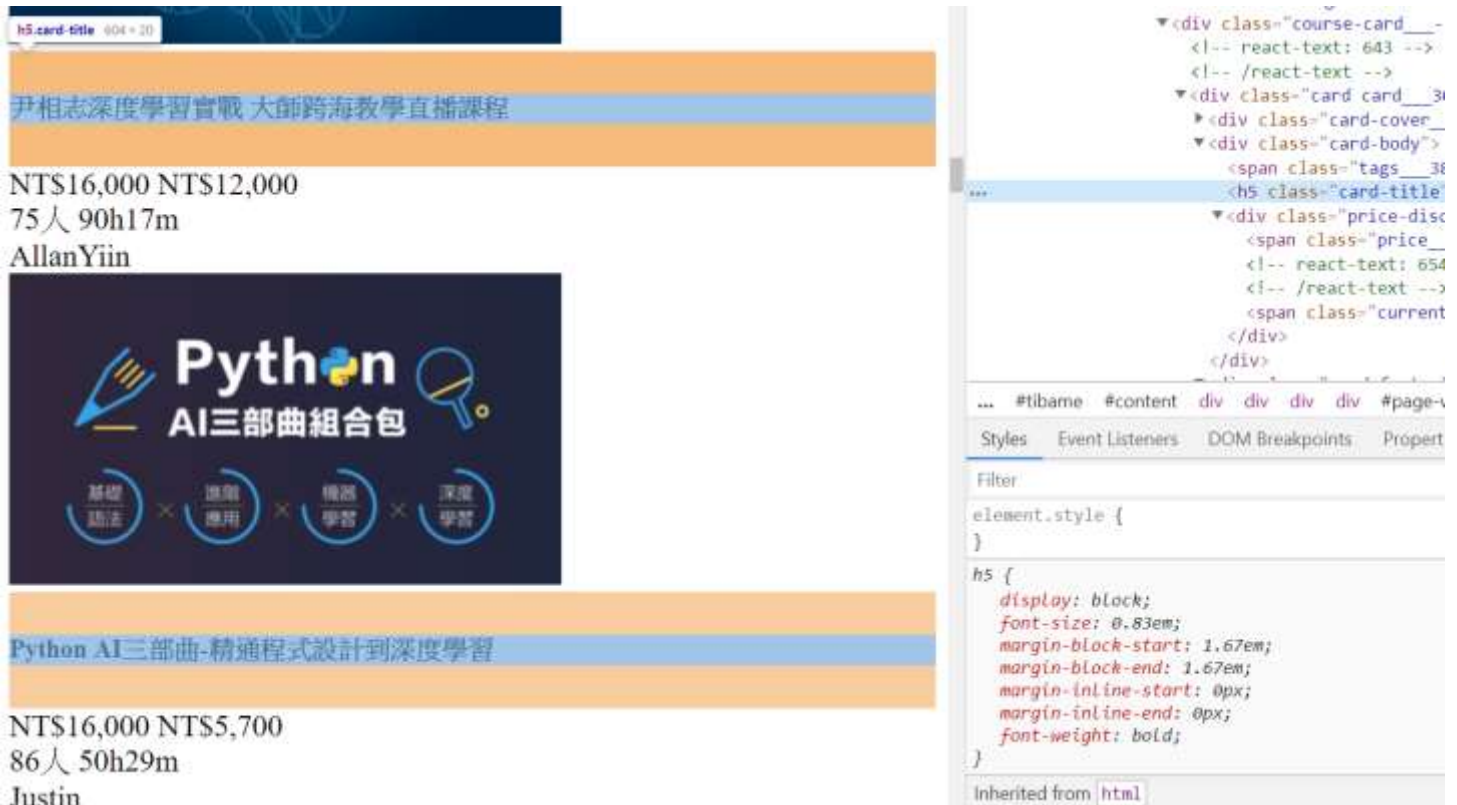
尹相志深度學習實戰 大師跨海教學直播課程

NTS16,000 NTS12,000
75人 90h17m
AllanYiin

注意！因為只有複製最基本的 HTML 網頁原始碼，沒有額外的 Javascript 語法和文件載入，所以用 localhost 打開網頁會較醜，但複製網頁當下原本有的標籤文字都還在。

步驟五：

再來將原本要的標籤選擇起來



The screenshot shows a web browser displaying two course cards. The first card is for '尹相志深度學習實戰 大師跨海教學直播課程' with a price of NT\$16,000 and 75 people. The second card is for 'Python AI三部曲-精過程式設計到深度學習' with a price of NT\$16,000 and 86 people. To the right, the browser's developer tools are open, showing the HTML structure of the first card. The 'card-title' element is selected, and its CSS styles are displayed in the 'Styles' pane.

```
h5.card-title 604 * 20
尹相志深度學習實戰 大師跨海教學直播課程
NT$16,000 NT$12,000
75人 90h17m
AllanYiin
Python AI三部曲組合包
基礎語法 × 進階應用 × 機器學習 × 深度學習
Python AI三部曲-精過程式設計到深度學習
NT$16,000 NT$5,700
86人 50h29m
Justin
```

```
<div class="course-card">
  <!-- react-text: 643 -->
  <!-- /react-text -->
  <div class="card card_3">
    <div class="card-cover">
      <div class="card-body">
        <span class="tags">
          <h5 class="card-title">
            <span class="price">
              <!-- react-text: 654 -->
              <!-- /react-text -->
            <span class="current">
              </div>
            </div>
          </h5>
        </div>
      </div>
    </div>
  </div>
</div>
```

```
... #tibase #content div div div div #page-1
Styles Event Listeners DOM Breakpoints Propert
Filter
element.style {
}
h5 {
  display: block;
  font-size: 0.83em;
  margin-block-start: 1.67em;
  margin-block-end: 1.67em;
  margin-inline-start: 0px;
  margin-inline-end: 0px;
  font-weight: bold;
}
Inherited from html
```

步驟六：

最後就能夠順利爬蟲，進行分析！

```
> ai <- read_html("http://localhost/tibame_ai.html")
> ai_classname <- html_nodes(ai, ".card-title") %>% html_text()
> ai_classprice <- html_nodes(ai, ".current-price__2Zu5l") %>% html_text()
> ai_numofpeople <- html_nodes(ai, ".footer__2mND7 div:first-child span:first-child") %>% html_text()
> ai_classlen <- html_nodes(ai, ".footer__2mND7 div:first-child span:nth-child(2)") %>% html_text()
>
> ai_classname
[1] "尹相志深度學習實戰 大師跨海教學直播課程"
[3] "【解鎖課程】Python人工智慧入門：機器學習到深..."
[5] "地表最好懂的Python網路爬蟲 零基礎也能學會！"
[7] "AI第二部曲-Python機器學習"
[9] "關聯式資料庫設計SQL Server和MySQL資..."
[11] "Python 輕鬆上手學"

"Python AI三部曲-精過程式設計到深度學習"
"Python人工智慧入門：機器學習到深度學習"
"【解鎖課程】地表最好懂的網路爬蟲彩蛋內容！"
"AI第三部曲-Python深度學習"
"AI首部曲-Python從零開始精過程式設計"
"Python 網站擷取與資料分析"
```

補充：

直接用 Ctrl + S 存檔.html 網頁後放入 htdocs 資料夾會更快速

