1. 為何要用Selenium?

**BeautifulSoup**、**Selenium**、**Scrapy**在網頁爬蟲中各有不同的用途，以下是它們的比較：

1. BeautifulSoup

* **適用場景**：適合處理靜態 HTML 頁面，從網頁中提取特定內容。
* **優點**：
  + 輕量、易於使用，適合新手快速上手。
  + 解析 HTML 結構非常靈活，可搭配 requests 取得網頁內容。
  + 適用於處理小型爬蟲項目，例如從靜態頁面提取文字、連結等。
* **缺點**：
  + 無法處理 JavaScript 動態生成的內容。
  + 不支援自動化操作（例如點擊按鈕、填寫表單等）。
* **最佳用途**：適合爬取靜態網站，例如新聞文章、部落格內容、表格資料等。

1. Selenium

* **適用場景**：適合處理動態網頁，尤其是需要模擬用戶行為（如點擊、滾動、輸入表單等）的網站。可以設計反爬蟲機制。
* **優點**：
  + 可控制瀏覽器，適用於 JavaScript 渲染的網頁。
  + 能夠執行用戶操作，例如登入、翻頁、無限滾動等。
  + 支援多種瀏覽器（如 Chrome、Firefox）。
* **缺點**：
  + 執行速度較慢，因為需要啟動瀏覽器。
  + 需要安裝瀏覽器驅動（如 ChromeDriver）。
  + 不適合大規模爬取，因為效率較低。
* **最佳用途**：適合爬取需要與網頁互動的網站，例如社交媒體、需要登入的網站、動態載入內容的電商網站等。

1. Scrapy

* **適用場景**：適合處理大規模的網頁爬蟲，提供完整的爬取框架。
* **優點**：
  + 高效、支援並行請求，適合爬取大量數據。
  + 內建爬取機制，可自動處理請求與回應，提高開發效率。
  + 可儲存數據為 JSON、CSV、XML 等格式。
  + 可透過 **Scrapy-Splash** 或 **Scrapy-Playwright** 處理 JavaScript 渲染內容。
* **缺點**：
  + 學習曲線較陡，需熟悉 Scrapy 的架構與運作方式。
  + 不適合需要大量互動的網站（如填寫表單、模擬點擊）。
* **最佳用途**：適合大規模爬取網站，例如電商網站的商品資訊、新聞網站的多個頁面等。

**總結**

| **工具** | **適用場景** | **優勢** | **劣勢** |
| --- | --- | --- | --- |
| **BeautifulSoup** | 靜態網頁、小型爬蟲 | 簡單易用，適合解析 HTML | 無法處理 JavaScript |
| **Selenium** | 需要用戶互動的動態網頁 | 可模擬用戶操作 | 速度較慢，資源消耗大 |
| **Scrapy** | 大規模爬取、多頁面爬蟲 | 高效、支援並行請求 | 學習曲線較陡，不適合互動式網站 |

如果你的需求只是爬取靜態內容，**BeautifulSoup** 是最佳選擇；  
如果需要爬取動態網頁內容，**Selenium** 是比較好的選擇；  
如果是大規模爬取，且不需要模擬人為操作，**Scrapy** 是最適合的工具。

1. 爬蟲如何翻頁?

遇到像 Dcard 這種的文章列表或留言列表，它是往下滾，就會送出新請求來取得下一頁資料。

如何「翻頁」呢？ 三種常遇到的「翻頁」方式：

1. 頁數 (page)
2. 偏移 (limit & offset)
3. 指定ID (pid)
4. Selenium 可處理JS按鈕。當使用JS按鈕時，上述方法都不適用，可以用Selenium來替你爬資料。

[參考](https://blog.jiatool.com/posts/web_crawler_page_method/)

1. How to find “next page” button click() event on CNA news website?

問題:連上中央社新聞網頁之後，怎麼知道 “下一頁按鈕” 呼叫甚麼函數?

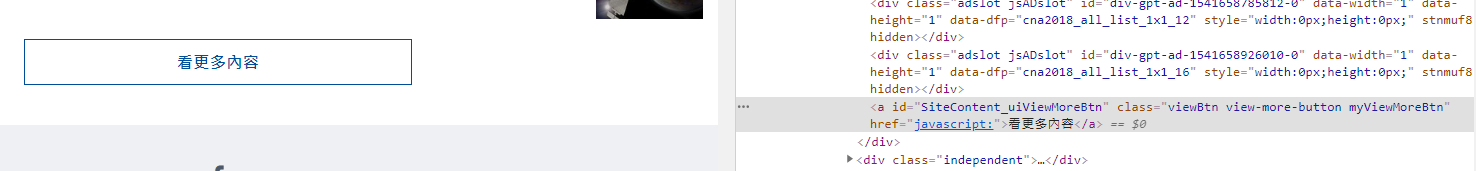
問題:許多網頁是用JavaScrip依據使用者的點選，動態產生內容，如何爬取這樣的網頁?

1. 從網頁程式碼中找線索



Here is the id for click envent

Right click this text



This is what we want to get:

<a id="SiteContent\_uiViewMoreBtn" class="viewBtn view-more-button myViewMoreBtn" href="javascript:">看更多內容</a>

1. F12 功能鍵進入開發者環境:Network XHR

從F12 Network XHR Header中可得到按鈕的事件名稱。

有些網頁可以透過此方式找到，但是CNA網頁不適合，找不到任何事件名稱。

1. Katalon Selenium IDE for Chrome

You can install Selenium IDE for Chrome. But Selenium IDE cannot export python code, as far as I know. Selenium IDE沒有Export為python的功能

Kadalon Selenium IDE can export python code. So we install Kadalon Selenium IDE.

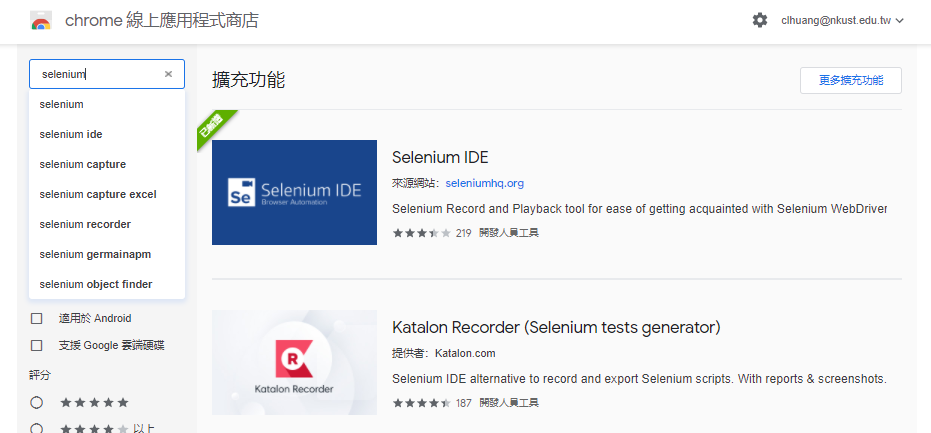
Install Kadalon Selenium IDE

Please google:

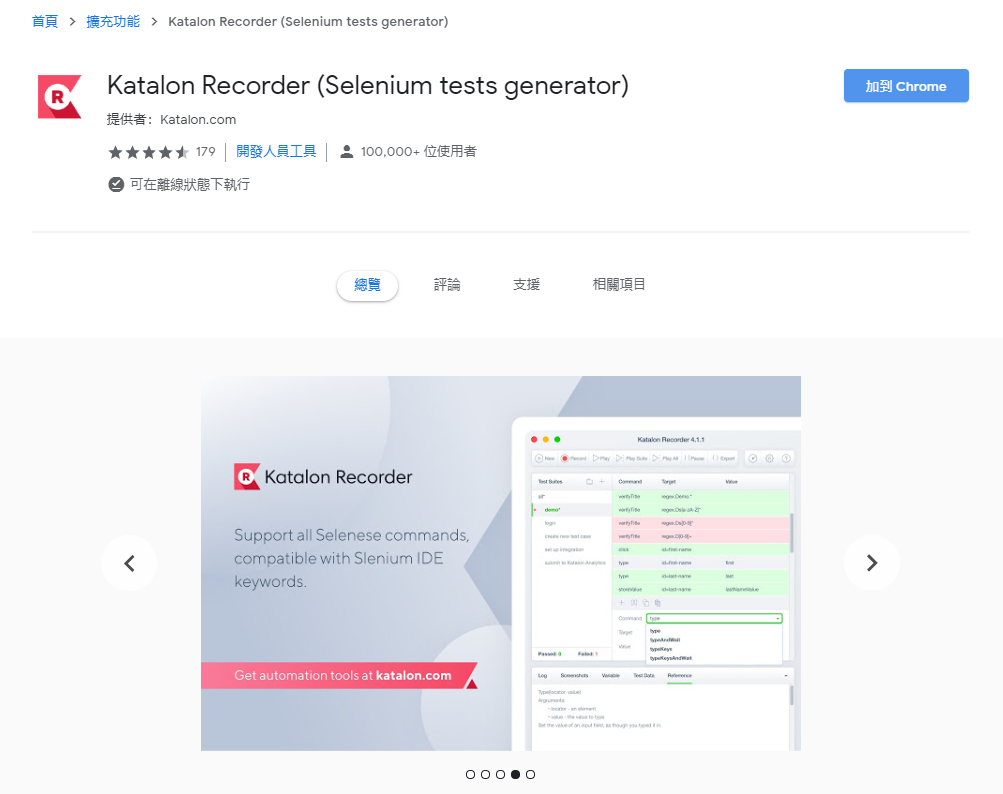
google chrome selenium plugin

You will find several selenium plugins.

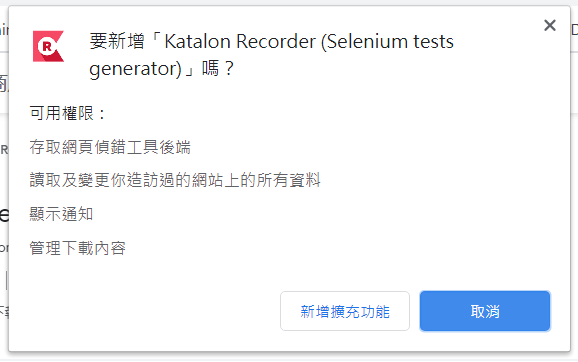
1. 在Chrome安裝Katalon Selenium IDE外掛

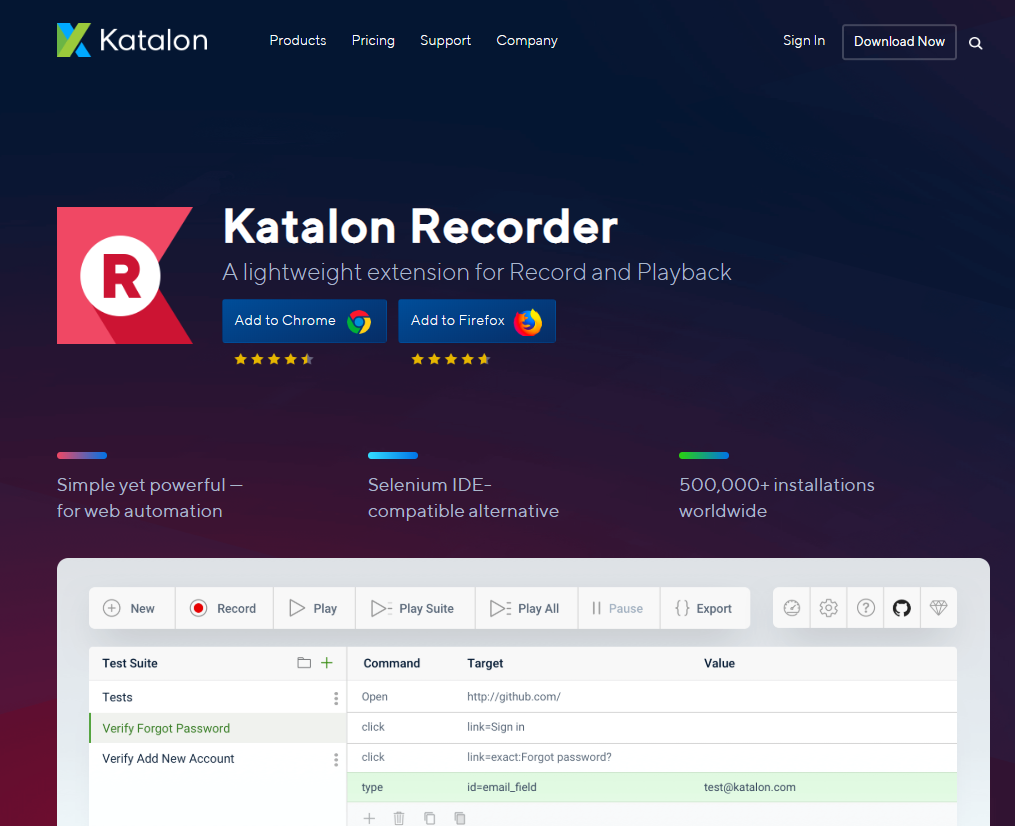


Kadalon Selenium IDE



[Source link](https://chrome.google.com/webstore/detail/katalon-recorder-selenium/ljdobmomdgdljniojadhoplhkpialdid)





1. Start Selenium IDE on Chrome

Link to CNA News website. Select a news category, e.g., “國際”.



Press here to start Selenium IDE.



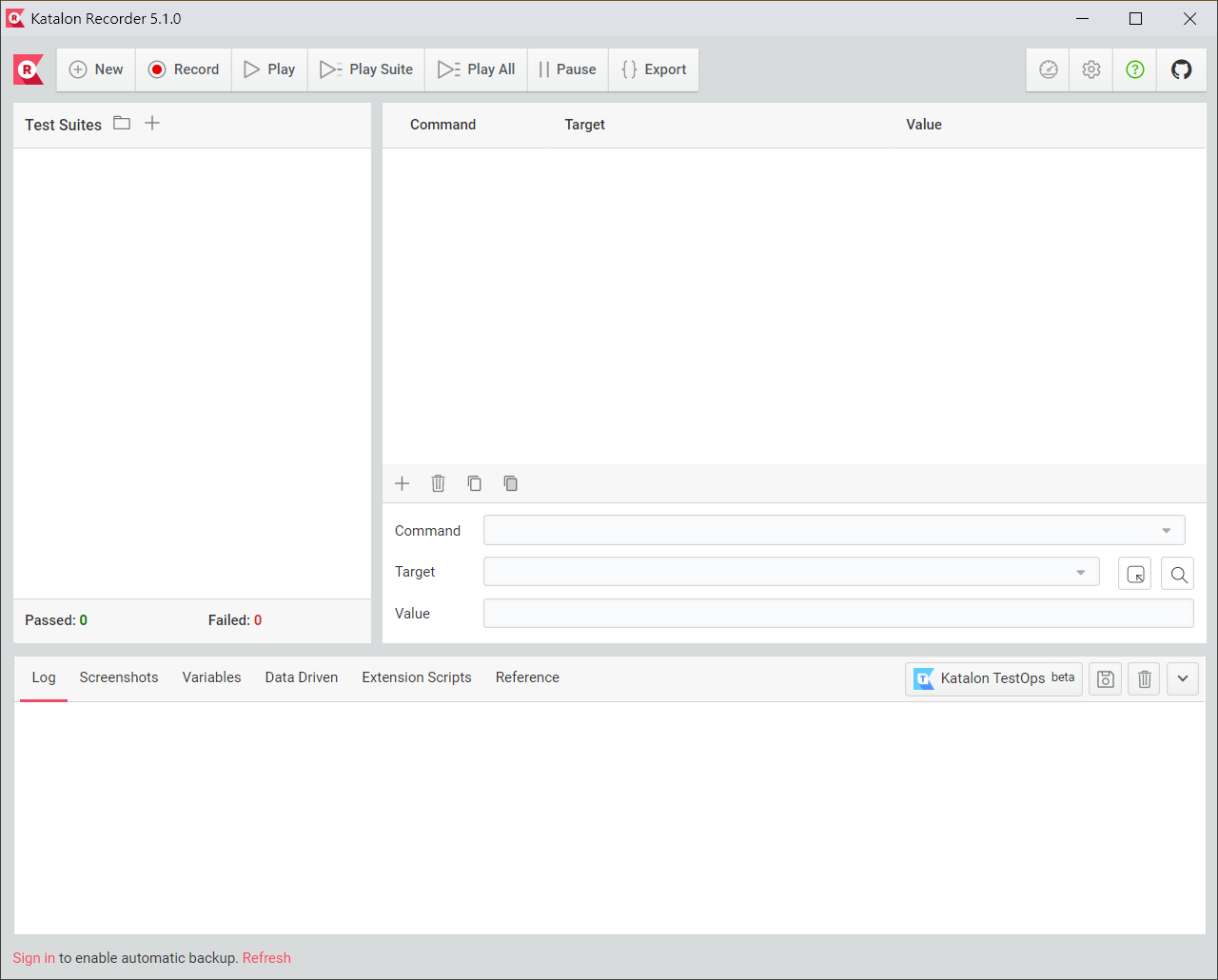
Start Selenium IDE

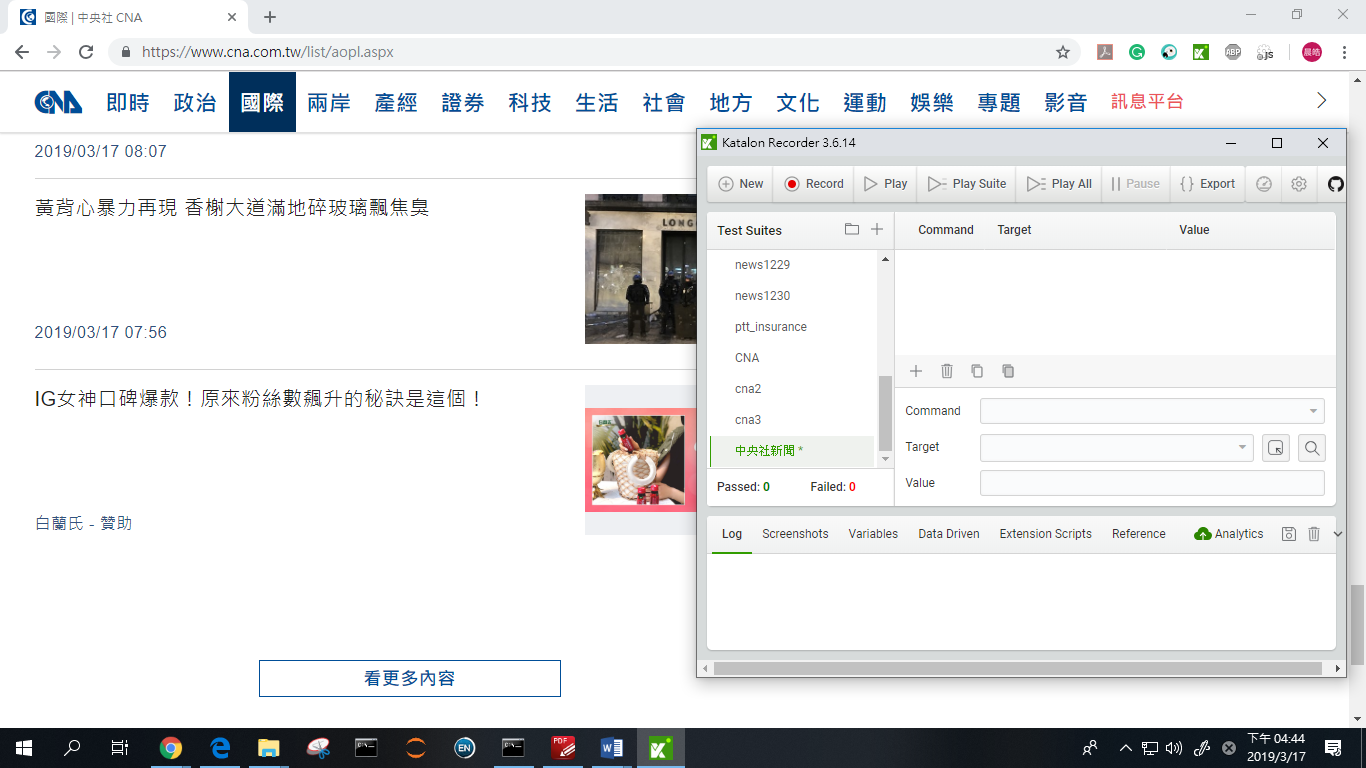
1. 在Chrome啟動Selenium IDE外掛

啟動Selenium IDE



1. Record your browsing steps
2. Press “New” to add a new “Test Case” 新增一筆測試案例
3. Press “Record” 開始錄製
4. Press “看更多內容”



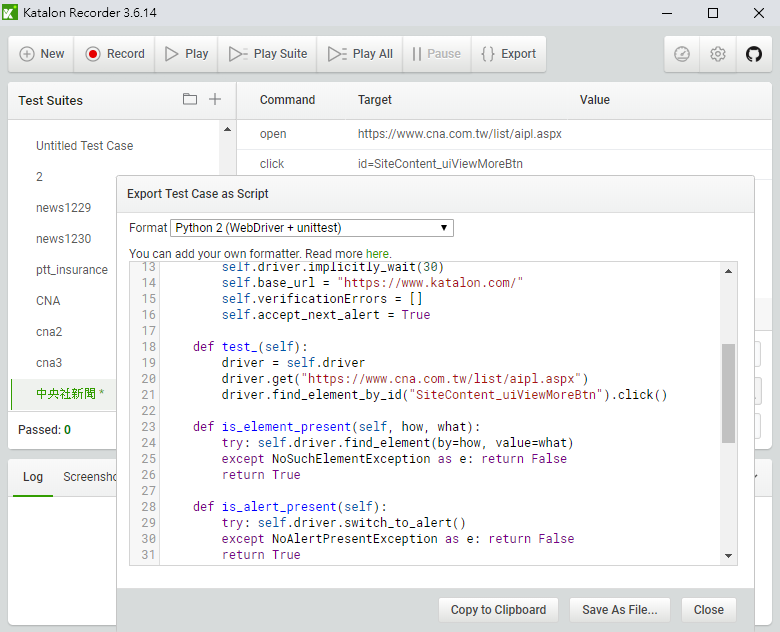


Press “New”

Press “Record”

Press “看更多內容”

1. Press “Stop” 停止錄製
2. Press “Export” to export Test Case 匯出程式碼
3. Press “Python 2 (WebDriver + unittest)”



This is what we need!

Export to Python code

1. Find the following codes. That’s the python code that we need.

“看更多內容”按鈕按下去 的 事件 在哪裡?

Where is the event function for the button “看更多內容”?

“看更多內容”按鈕執行的動作會在以下函式裡:

def test\_(self): ….

driver.get("https://www.cna.com.tw/list/aopl.aspx")

driver.find\_element\_by\_id("SiteContent\_uiViewMoreBtn").click()

1. Selenium IDE外掛

Selenium IDE (沒有Export為python的功能，指令須自己寫，初學者會覺得較困難)

Kadalon Selenium IDE

