# 作業一、爬取「古騰堡計劃」中文電子書（20%）

古騰堡計劃 - 中文書

<https://www.gutenberg.org/browse/languages/zh>

作業需求：

1. 取得每一本書的內文，至少 200 本。
2. 依照書名，將內文存在 txt 當中，檔案以書名命名。
3. 將所有 txt 集中放置在同一個目錄當中，不要跟其它程式檔案混在一起，目錄名稱自訂。
4. 參考結果: <https://www.youtube.com/watch?v=gKDBiVvzMfk>
5. 錄製爬取結果的影片（30 秒內），影片名稱為「你的班別\_你的名字\_古騰堡計劃」，例如 111S\_楊○○\_古騰堡計畫。
6. 影片錄製過程中：
   1. 顯示一共有多少檔案的數字（至少 200 個檔案以上，少 1 個扣 1 分）。
   2. 隨機將抓下來的小說檔案打開來進行比對，任意滾動檔案總管的上下捲軸來開啟檔案，一共三個檔案。
   3. 開啟檔案的過程中，將內文當中與檔案名稱一致的文字反白，例如「洛神賦.txt」裡面，會看得到「洛神賦」三個字，將其反白，表示內文正確對應到檔案名稱。

觀察流程：

1. 請依觀察，評估是否使用 requests 或 selenium 來爬取資料，沒有特別限制。
2. 沒有限定要透過「Read this book online: HTML (no images)」、「EPUB (no images)」、「Kindle (no images)」、「Plain Text UTF-8」等格式取得資料，方便剖析資料內容即可。
3. 以 Plain Text UTF-8 連結為例：
   1. 從中文書的連結進去，看到許多中文書的書名，以下以「三字經」為例。
   2. 滑鼠移過去，瀏覽器左下角有該書的超連結(例如 https://www.gutenberg.org/ebooks/12479)。
   3. 同時觀察 b. 裡面的連結，得知連結字串的格式，以及尾端的號碼。
   4. 點進連結後，會看到「Download This eBook」下面的表格，有列出內文的格式與內文連結。
   5. 可以選擇自己方便解析的檔案結構，例如「Plain Text UTF-8」。
   6. 僅將中文內文爬回來即可，頁面前後若有英文文字，則不要爬。
   7. 若是你選擇 Plain Text UTF-8 格式，Regex 請參考:
      1. 正則表達式-全型英數中文字、常用符號unicode對照表  
         <https://reurl.cc/ZrMoWA>
      2. 匹配中文字符的正則表達式： [/u4e00-/u9fa5]  
         <https://www.itread01.com/content/1513168876.html>
   8. 透過正規表達式，有可能配對的句子會變成 list 結構，此時透過 ''.join(listSentences) 將所有配對到的文字變成字串，即是中文全文。
   9. 可以選一篇文章，把前述的正規表達式，一起放到 regex101 裡面檢視。

# 作業二、自動化操作（20%）

1. 請自選一個主題，進行自動化操作，只要上課學到的技術，甚至於是自修學習而來，都可以使用。
2. 主題執行內容不要跟上課內容一樣，可以是網頁遊戲、手機app模擬器、單機或線上互動平台、桌面應用程式等的連續自動化操作；可以是搞笑的梗，或是有助於提升你個人生活品質的功能，甚至是增進公共利益（？）。
3. 錄製執行過程與結果的影片（時間不限），影片名稱為「你的班別\_你的名字\_自動化操作\_你的主題」，例如 111S \_楊○○\_自動化操作\_太鼓達人。

# 繳交方式

1. 上傳到個人 Google Drive 或上架 YouTube，記得分享、提供連結。
2. 期限為 2022-07-31 (日) 17:00:00
3. 完成後，將兩個作業的連結寄至 [darren@darreninfo.cc](mailto:darren@darreninfo.cc)，記得標題寫上班別和你的姓名。

# 補充

1. 基本分 60 分起跳。
2. 我的 IG: darreninfo.cc。（亂入，這不重要 XD）
3. 請勿抄襲，努力過後所累積的經驗，才是自己的；如果作品不會侵權，或許可以考慮列入作品集。