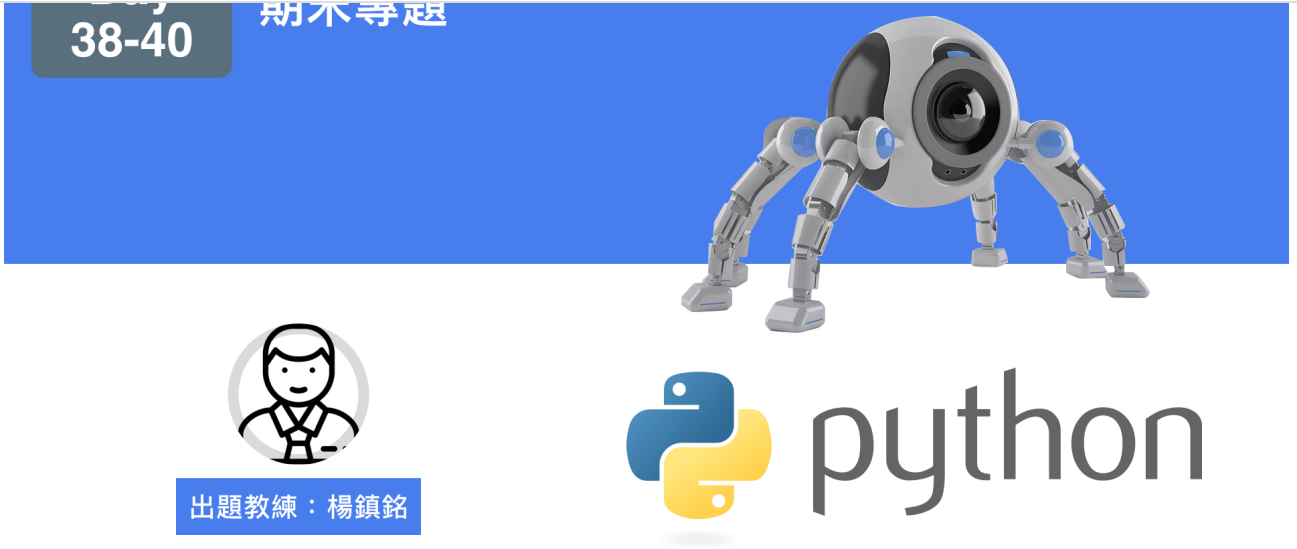
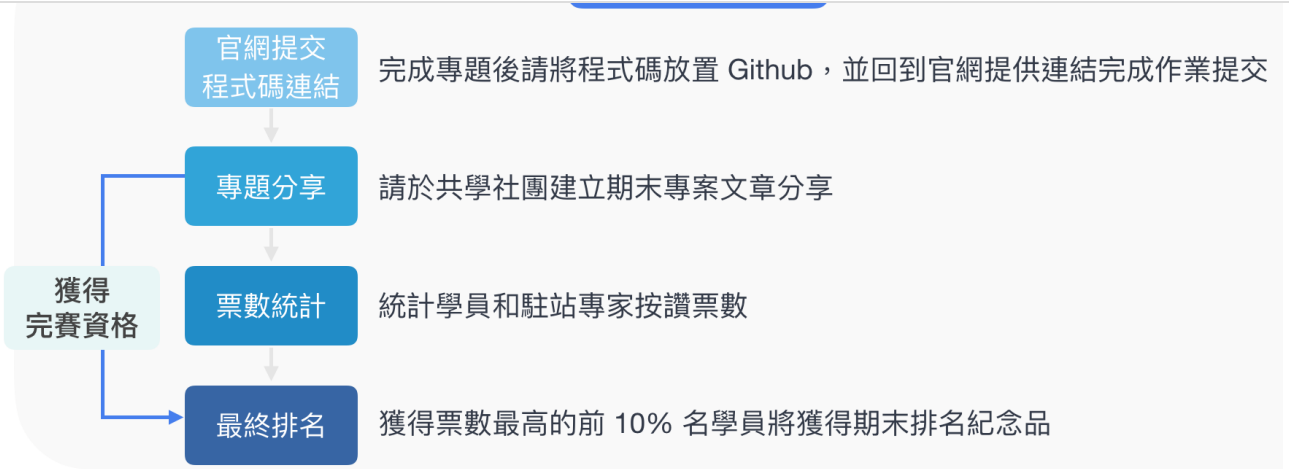
D38 - D40：期末專題

簡報閱讀 範例與作業 問題討論

爬蟲期末專題



期末專題流程


官網提交
程式碼連結
完成專題後請將程式碼放置 Github，並回到官網提供連結完成作業提交
專題分享
請於共學社團建立期末專案文章分享
獲得
完賽資格
票數統計
統計學員和駐站專家按讚票數
最終排名
獲得票數最高的前 10% 名學員將獲得期末排名紀念品

建立期末專題文章


首頁
AI社群
熱門新聞
我的
搜索
AI 人工智慧學習社群 > 論壇首頁 > Python網路爬蟲論壇
Python 網路爬蟲實戰研習馬拉松
首頁
我的學習任務(40)
貼文(59)
問答(380)
共學課程(1)
成員(1.1k)
活動介紹
管理
提問
撰寫文章
分享連結
Python網路爬蟲論壇
Python網路爬蟲實戰研習專屬社團
瀏覽全部
最新公告
2020 7.31
【AI專家開講】資料科學實務經驗談
7.28
【建議閱讀】學習馬拉松上課環境說明
5.29
【上課須知】歡迎新朋友加入 Python網路爬蟲實戰隨到隨跑馬拉松
看更多
Python 網路爬蟲實戰研習馬拉松
學習進度 10%
陪跑專家 11位
參與同學 1.1k位
繼續學習
駐站專家 (11)
立即加入
最新貼文
最新
熱門
搜尋貼文
我要貼文


編輯 插入 格式
1 標題撰寫格式
Python期末專題_百日馬拉松顯示名稱
一、專題摘要 (解釋實作與說明需要解決的問題，限300~500字。)
1. 期末專題主題
2. 期末專題基本目標
二、實作方法介紹 (介紹使用的程式碼、模組，並附上實作過程與結果的截圖，需圖文並茂。)
1. 使用的程式碼介紹
2. 使用的模組介紹
三、成果展示 (介紹成果的特點為何，並撰寫心得。)
2 期末專題文章基本格式內容
四、結論 (總結本次專題的問題與結果)
五、期末專題作者資訊 (請附上作者資訊)
1. 個人Github連結
2. 個人在百日馬拉松顯示名稱

分數計算

期末專題格式

應用所學知識，請學員分享專題實作結果，還能獲得專家回饋唷！

字數300~500字佳，格式不拘請包含下列內容：

專題摘要	限300~500字。
02 實作方法介紹	介紹使用的程式碼、模組，並附上實作過程與結果的截圖，需圖文並茂。
03 成果展示	介紹成果的特點為何，並撰寫心得。
04 結論	總結本次專題的問題與結果。

期末專題主題

期末主題 1 - Cupoy

⛳ **專題目標** ⛳

請任選 Cupoy 新聞服務之某一種分類 (如熱門新聞、科技、商業....)，使用你學習過的爬蟲程式，爬取前 500 篇的文章：

👉👉 https://www.cupoy.com/newsfeed/topstory

⛳ 基礎實作提示 ⛳

TARGET 1 透過開發者工具觀察網站在列出 News Feed 這邊是屬於動態網站還是靜態網站，或是有 API 可以直接送 requests

TARGET 2 根據網站特性選擇 requests / BeautifulSoup / selenium 等工具進行爬蟲整理

TARGET 3 整理成 pandas.DataFrame 後做簡單的統計可以用 matplotlib.pyplot 或是 pandas 內建的 function 畫圖 (histogram / pie chart ...)

⛳ 進階實作提示 ⛳

TARGET 2 可以簡單的計算同樣文字出現的頻率或是透過 TFIDF 的統計方式計算

TARGET 3 將經常出現的 stop words 過濾掉之後對頻率進行排名

TARGET 4 將結果透過 wordcloud 文字雲的方式呈現

期末主題 2 - PTT討論版

⛳ **專案目標** ⛳

根據版的熱門程度跟屬性，可選定以下任一種：

1. 八卦版：https://www.ptt.cc/bbs/Gossiping/index.html
2. 政黑板：https://www.ptt.cc/bbs/HatePolitics/index.html


看板
精華區
搜尋文章…
Re: [新聞] 民眾黨徵助理起薪30K被罵翻 柯文哲：雇少
XSR700
1 Re: [新聞] 年前全漲價！從珍奶到鍋貼 小數點也要賺
popy8789
2 Re: [新聞] 民調：4成5反對蔡英文兼任黨主席 逾7成2
Rrrxddd
1 Re: [新聞] 館長道歉林右昌沒答應展店 「他們可能是
NuclearSnake
3 [新聞] 川普德州取暖之旅 保證美中貿易協議將嘉
CavendishJr

⛳ 基礎實作提示 ⛳

TARGET 2 可以簡單的計算同樣文字出現的頻率或是透過 TFIDF 的統計方式計算

TARGET 3 將經常出現的 stop words 過濾掉之後對頻率進行排名

TARGET 4 將結果透過 wordcloud 文字雲的方式呈現

⛳ **進階實作提示** ⛳

TARGET 1 透過不同帳號，但是相同 IP 且政治用語的詞頻分佈類似的定位成網軍

TARGET 2 進一步分析帳號是否在特定期間 (e.g. 選舉) 有明顯的活動特性

TARGET 3 如果不同帳號但是政治用語的詞頻分佈類似，進一步判斷這些高頻率的單字是 positive / negative 來歸納兩個帳號之間是否具有相同政治立場

期末專題知識點目標

專題結束後你可以學會 😀 😀 😀

1. 了解不同網站實作的爬蟲細節
2. 對於爬蟲流程的分析與判斷有完整的 Overview
3. 可以分析針對不同網站所需的爬蟲複雜度
4. 搭配不同領域知識做出獨特的應用
5. 清楚說明爬蟲流程與作法

開始期末挑戰！

下一步：閱讀範例與完成作業