期末報告

A、B、C 三個選題擇一

A. 校園 E-mail 整理與分析

透過自動化方式蒐集、整理與分析校園 E-mail，並運用資料庫技術進行分類、結構化與統計分析，以了解校園公告郵件的特性與模式

撰寫爬蟲，爬下至少2,000封郵件，建立資料庫儲存。

根據進信件類別、寄件人、主題、內容做分類整理，分析校園公告信的特徵，供未來制定校園公告信件寄送規範參考

資料蒐集

1. 撰寫爬蟲程式
   - 實作一支可自動下載校園公告信件內容的爬蟲
   - 使用 Python（requests、BeautifulSoup、Selenium 等工具）

2. 郵件數量要求
   - 需蒐集至少 2,000 封校園公告郵件
   - 對資料進行清洗（清除資料缺失、不完整的數據，清洗完後至少有2,000封）

資料庫建立

- 資料庫可使用 MySQL、Firebase
- Email 基本資料表（例如: 寄件人、收件人、主旨、日期、內容…）

1. 郵件特性分析

- 主要寄件單位有哪些？
- 郵件常見主題類型？
- 是否有季節性（如開學前大量郵件）？

2. 適用對象分析

- 受眾分群（大一學生？全校？特定系所？）

3. 時間分布分析

- 一天/一週/一學期的郵件量變化。
- 哪些時間點最常發布公告？

4. 建議的郵件發送規則

B. 論壇鄉民活躍度分析

社群平台人數快速成長，論壇中出現更多假帳號、分身帳號與洗文章行為。部分使用者可能透過大量發文、推文或刻意操作特定議題來影響輿論。

建立「論壇使用者活躍度分析系統」，以特定論壇（如 PTT、Dcard、Mobile01 等）為資料來源，分析看板、作者帳號、文章與推文行為，並建立互動式查詢與帳號活躍度偵測功能。

資料蒐集

- 收集至少**10,000 篇文章**（文章+推文+回文數據 大於 10,000筆）
- 可跨看板
- 必須排除「文章量極少」的冷門看板，以免造成統計失真
- 對資料進行清洗（清除資料缺失、不完整的數據，清洗完後至少有10,000篇）

資料庫建立

- 資料庫可使用 MySQL、Firebase
- 每篇文章至少需包含：
  - 作者帳號
  - 看板（子論壇）
  - 標題
  - 日期時間
  - 內文/回文/推文

看板篩選功能

- 可選擇特定看板
- 使用者可選擇，某一年或月份的時間區間
- 顯示該期間的文章統計與分析結果。

鄉民活躍度分析

- 發文量、回文（推文）量
- 例如：直方圖（Histogram）、折線圖（時間序列）、長條圖（Top N 作者）

指定帳號「歷史行為查詢」

- 輸入：帳號名稱 + 時間區間
- 回傳：
  - 該使用者的發文摘要
  - 回文摘要
  - 該期間內的看板分布（在哪些板最活躍）
  - 發文/回文量時間序列

偵測「疑似 轉手帳號」

- 根據使用者行為模式做偵測
- 可用的偵測指標（至少需用 3 項）：
  - 發文時間異常（例如：每天只在 03:00～04:00）
  - 看板分布突變（突然大量換板發文）
  - 語言風格相似度
  - 自訂…
- 列出可疑帳號清單

C.新聞分析器

建立「新聞分析系統」，透過大量新聞資料進行關鍵字、記者、新聞類別等面向的分析，並提供互動式查詢功能，以深入了解新聞產製的模式與新聞工作者的活動分布。

許多新聞可能包含 AI 生成內容、未經查證的資訊或再製文章，建立新聞分析系統，可用於觀察內容來源、記者真實性、新聞類別的一致性，並協助偵測可疑或品質低落的新聞資料。

資料蒐集

需收錄至少 10,000 筆新聞資料

新聞資料來源不限（API / 網頁爬蟲 / 資料庫）

對資料進行清洗（清除資料缺失、不完整的數據，清洗完後至少有10,000筆）

資料庫建立

- 資料庫可使用 MySQL、Firebase
- 每則新聞至少包含以下欄位：
  - 標題
  - 內容
  - 日期
  - 新聞類別（如：政治、財經、體育…）
  - 記者署名

時間選擇功能

- 使用者可選擇：
- 某一年或月份的時間區間
- 顯示該期間的新聞統計與分析結果。

新聞關鍵詞活躍分布圖

- 針對選定時間區間，繪製「關鍵詞活躍度變化圖」
- 折線圖、直方圖、熱力圖皆可

各類新聞的活躍記者

- 每種類別中撰寫新聞最多的記者
- 前 10 名記者的新聞量統計

針對記者署名做屬性推測

- 輸入記者名稱，查詢該記者
- 該記者主要關注哪些新聞類型?
- 他的大部分新聞來自哪個媒體?
- 是否有跨版性（跨政治、跨娛樂等）?
- 活躍期間（年份、月份）
- 是否為大量產新聞的記者

報告

每組 3~4 人一組 (組員人數也可以維持跟期中報告一樣)

12 分鐘

組員分工佔比

繳交:

- 投影片 (請包含資料庫數據統計圖或表)
- 所有的程式 (.py 檔案)
- ~~數據 (不用上傳數據)~~

上傳e-Learning的資料格式說明:

程式.zip (資料夾): 請壓縮

投影片 (PDF 或 pptx 格式)

分組名單
https://docs.google.com/spreadsheets/d/1ude6KmtkhbrrtLNHXeAXbWHI-Wnsr_IbQ9MiQDJYSwM/edit?usp=sharing