巨量資料與統計分析　　Fall 20204

第五次作業（2025/01/03 繳交）

1. 本次作業著重於文字資料的處理與分析，請以中華民國歷屆總統（共十六任）的就職演講稿為研究素材，進行以下分析：

   (a) 輸入報導的文字，並統計歷任總統就職演講稿總字數、不同字彙數；

   (b) 統計歷任總統就職演講稿出現頻率前三十名的字彙、雙字詞；

   (c) 以文字雲、Barplot 等圖形繪製出現頻率最高的雙字詞。

   （加分題：以出現頻率最高的字、詞，提出判斷歷任總統的主要特色，可參考網站 https://www.cna.com.tw/project/20200520-inauguraladdress/。）

2. 金庸小說堪稱近年武俠小說的代表，因為題材生動豐富、且與歷史結合等原因，受到老少不同讀者群的喜好，作品依順序為「天、白、俠、飛、倚、射、書、神、笑、連、雪、鹿、越、碧、鴛」。本題以描述武功招式較多的《天龍八部》（北宋年間）、注重故事情節的《鹿鼎記》（初清年間）為題材，尋找能夠區隔兩者風格的可能變數。

   (a) 請以結巴等斷詞軟體，剔除人名、地名等名詞，請以常見字詞（例如：前十、前三十、前一百）比較這兩部小說用詞或寫作風格上的差異。

   (b) 透過探索性資料分析(EDA)，尋找兩部小說的差異。除了上述的常見字詞，也可考慮常見字/詞彙的 Entropy、Simpson Index、齊夫法則、TTR 等指標，但詞彙不考慮人名、地名。根據上述分析結果，各組說明這些結果代表的意義，並提出可區隔兩部小說的可能方式。

   (c) 以(b)選擇的變數為依據（例如：各組可將前 100 雙字詞視為解釋變數），隨機將兩部小說分成 90%及 10%，藉由羅吉士迴歸(Logistic Regression)及交叉驗證(Cross-validation)，討論是否能夠區隔兩部小說的差異。

3. 文字分析可用於描述寫作風格，也可用於探索社會及文化的變遷。本題根據臺灣《蘋果日報》及《自由時報》的 2012 年及 2018 年的頭條報導，比較這兩份報紙的寫作風格，並探討風格是否有明顯變化。（註：本題的重點為「非監督學習」。）

4. 自從電腦科技突飛猛進，人們愈來愈難分辨哪些資訊（文字、圖像、音訊）是偽造或電腦合成。請同學判別提供的影像和文字：

(a) 標註哪些是深偽（Deepfake）影像，說明各組判斷的原則。

(b) 標註哪些摘要是 ChatGPT 生成，並說明根據哪些資訊判斷。