巨量資料與統計分析 Fall 2017

第四次作業（2017/12/15繳交）

1. 本次作業著重於文字資料的處理與分析，請各組分析中華民國歷屆總統（共十四任）的就職演講稿，
2. 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；
3. 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；
4. 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。）

1. 章回小說是明清兩朝的代表作品，其文體明顯與常用於書面的文言文不同。本題以小說《紅樓夢》第一回至第五回、《西遊記》第一回至第五回為比較對象，尋找這兩部小說用字的主要差異。（註：本題的重點為「監督學習」。）
2. 仿造我的網站有關《新青年》的文章，以虛字(Functional Words)為判斷依據，比較兩部小說的差異；
3. 除了虛字外，尋求可區隔文體差異的特色（例如：文言文、白話文），但不考慮內容相關用詞（例如：賈寶玉、曹操等人名），探討這兩部小說的寫作風格是否有顯著差別。
4. 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」。本題以寫作時間和故事背景較為類似的《射雕英雄》、《神鵰俠侶》（皆為南宋年間）為題材，尋找能夠區隔兩者風格的可能變數。（註：本題的重點為「非監督學習」。）
5. 透過探索性資料分析(EDA)，尋找兩部小說的差異，不限定文體或內容上的量化變數，但不考慮人名、地名。
6. 以(a)選擇的變數為依據，先將上述各部小說分成10等分，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。