Text Mining HW1 Report　　R12723006 財金碩一 林姝延

1. 執行環境: Google Colab

2.程式語言: Python 3

3. 執行方式:

通過使用 Colab 的 drive module，將 Google 雲端硬碟掛載到 Colab 環境中的 /content/gdrive 資料夾，以達到在 Colab 中讀取和寫入 Google 雲端硬碟檔案的目的，並使用 os.chdir 切換到程式碼及 input 資料夾存放的目錄，執行前需調整至執行目錄。如下程式碼及註解說明:

```
# Switch to the working directory path
os.chdir('/content/gdrive/My Drive/Colab Notebooks') # Need to
set the specified folder path
```

而最終各文件的 TF-IDF vec 檔案會輸出到指定的執行目錄下名為 ”output” 的資料夾。

4.邏輯說明

首先，獲取目前程式碼檔案及 input 資料夾所在的目錄路徑，並指定輸入和輸出資料夾的名稱，以組合出輸入和輸出資料夾的路徑。

然後創建一個空 list docs 來儲存每個文件的內容。透過 for 迴圈一個個開啟、讀取每個資料夾中的文件檔，將內容添加到 docs list 當中。 接著下載 nltk 的 stop words 列表，用於稍後的 TF-IDF 向量化過程，過濾掉英文 stop words。

使用 TF-IDF vectorizer ，指定 lowercase 並過濾掉英文 stop words，將 TF-IDF 向量儲存在 tfidf_matrix 中。 接著將 TF-IDF 向量分別寫入各檔案中，也就是在 output 的資料夾中建立 1.vec, 2.vec, ⋯, 1095.vec。

最後，讀取文件 1 和文件 2 的 TF-IDF 向量，使用 cosine_smilarity 函數計算兩項量之間的 cosine similarity，並 print 出計算值。

由結果可看出兩文件之間的 cosine similarity 數值很小 (0.0049)，因此可以推測兩文件相似度不高。