IR HW4 Report

B05705025 資管四 陳漢威

1. 執行環境:

VSCode, terminal

2. 程式語言

Python3

3. 執行方式

程式會讀入在同一資料夾內路徑為”IRTM”的資料夾，並開啟裡面所有的檔案作為計算的依據，程式最終會產出選定的 txt 檔案
該次使用的套件有(編譯前須先 pip install):

- nltk (for Stemming and stopwords)
- re (for filtering the symbol)
- numpy (for vector operation)
- collections (for terms count)
- pattern (for lemmatization)
- heapq(for minimum heap)

4. 作業處理邏輯說明:

此次 clustering 會利用作業二計算的 tf-idf 與 cosine similarity 作為依據去做 clustering，本次我採用的方法是 Centroid，在一開始會先將每篇文章互相計算 cosine similarity 並 * -1 方便我們建立 priority queue，接者會進入迴圈:
每次 iteration 會 pop 出 queue 裡面相關性最大的 sets 作合併，但會先檢查這兩個 set 是否分別已經被其他 set 合併了，如果有就跳到下一個最大的 sets，如果沒有就建立新的 cluster，並找出該 cluster 中心與其他所有 cluster 中心的相似度，並將這些結果更新到 heap 內，並同時刪掉合併前的兩個 cluster 資料，並記錄這兩個 sets 已經有合併過，避免之後的 iteration pop 到舊資料。
最後當 cluster 數量小於給定的 k 時就跳出迴圈，並 dump txt 結果。