IRTM - Programming Assignment 4 Report

資管碩一 R10725018 郭宇雋

1. 執行環境：

Command Line

2. 程式語言：

Python3

3. 執行方式：

需要進行 pip install 套件 nltk，而後直接透過指令 python pa4.py 執行，即可產生分群檔案

4. 作業處理邏輯說明：

- 首先透過與前幾次作業相同的方式產生 Document TF-IDF Vector
- 並計算兩兩文件之間的相似度，儲存成 Matrix (C)
- 建構並初始化I與A向量
- 建構 HAC，並使用 Complete-Link 做為 Cluster 之間的相似度指標
- 因為是 Complete-Link，因此合併 i 與 m 群並計算新的群間相似度時，要取兩者與其他新群相似度的 Min 作為新的相似度，因為 Complete-Link 是把最遠的兩點當作群間的相似度，再更新相似度矩陣 C
- 每次合併的歷史紀錄會存在 A 中，最後寫入檔案時，逐步讀取並記錄每個時刻的合併與群組狀態