R12725049徐尚淵 作業四

1. 執行環境：Jupyter Notebook
2. 程式語言：Python (版本3.11.4)
3. 執行方式：

直接使用Jupyter Notebook GUI Run code

一張含有 文字, 字型, 白色, 標誌 的圖片

自動產生的描述

1. 作業處理邏輯說明
   1. 讀取需使用的package

一張含有 文字, 字型, 螢幕擷取畫面, 白色 的圖片

自動產生的描述

* 1. 創建一個空的List名為total\_tfidf\_df\_list，裡面將用List儲存每個文件的用字intex以及其tfidf值，並且每個文件用dataframe的格式來儲存

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

* 1. 定義normalize function

計算unit vector，並定義cosine function計算任兩文件的cosine similarity (同作業二)

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

* 1. 用名為C的dataframe來儲存兩兩文件所計算出的similarity結果，因為對稱矩陣所以只須計算上半三角，下半三角可直接對稱

一張含有 文字, 字型, 白色, 螢幕擷取畫面 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 字型, 行 的圖片

自動產生的描述

* 1. 定義 sim\_matrix function

用以更新此Cluster演算法中Similarity matrix的function，每次迭代都實作一次，首先確保收到的id為int格式，而後利用np.minimium找出兩個cluster中的最小值，更新到cluster1的欄位，並更新對稱矩陣以及設定cluster1自己跟自己比的similarity為負無限大，cluster2亦同以免日後受到其干擾

一張含有 文字, 字型, 螢幕擷取畫面 的圖片

自動產生的描述

* 1. 定義Max\_pos function

能夠找到Test\_C這個餘弦相似矩陣中的最大元素的index，也就是最相似的兩個文件or分群

一張含有 文字, 字型, 行, 螢幕擷取畫面 的圖片

自動產生的描述

* 1. 定義 Apply\_HAC Function

大致上遵照講義的虛擬碼、以及complete-link的邏輯來實作，在達到k類之前都會持續循環，找到最相似的兩個cluster的索引並合併，之後清空k\_max表示其已經被合併，最後更新相似度矩陣並return結果

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

* 1. 定義 Write\_file Function

用以寫入文件

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

* 1. 主程式部分

For 迴圈條件設定為8、13、20三群，先初始化日後會用到的output\_k、test\_C以及clusters，並將C複製起來以便儲存，再直接利用apply\_HAC實作，並排序後寫入文件中。

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述