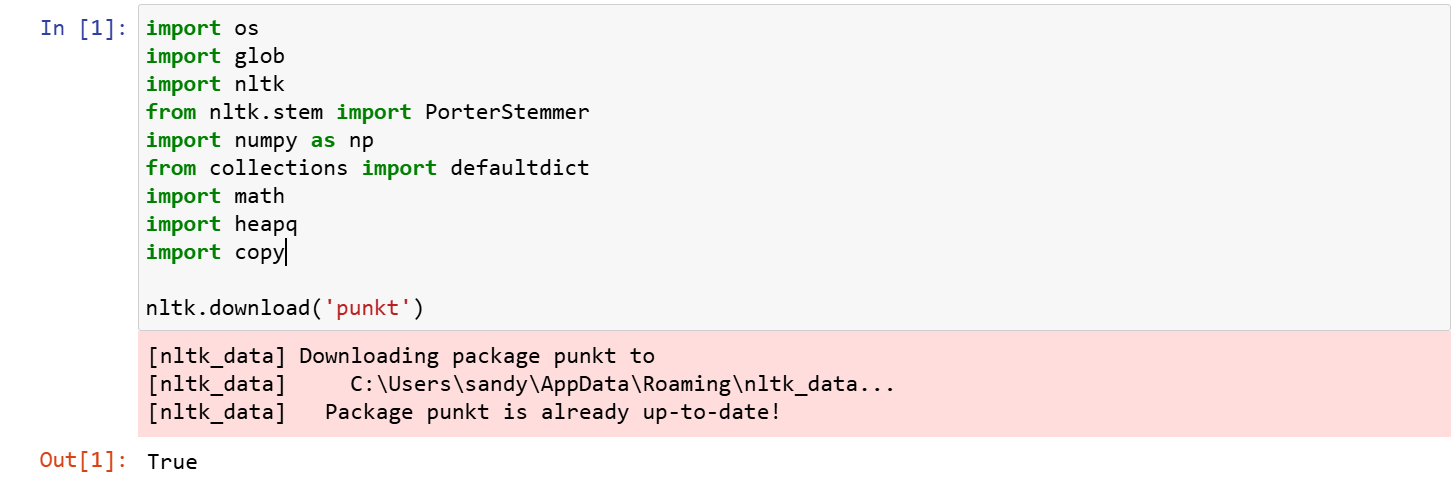
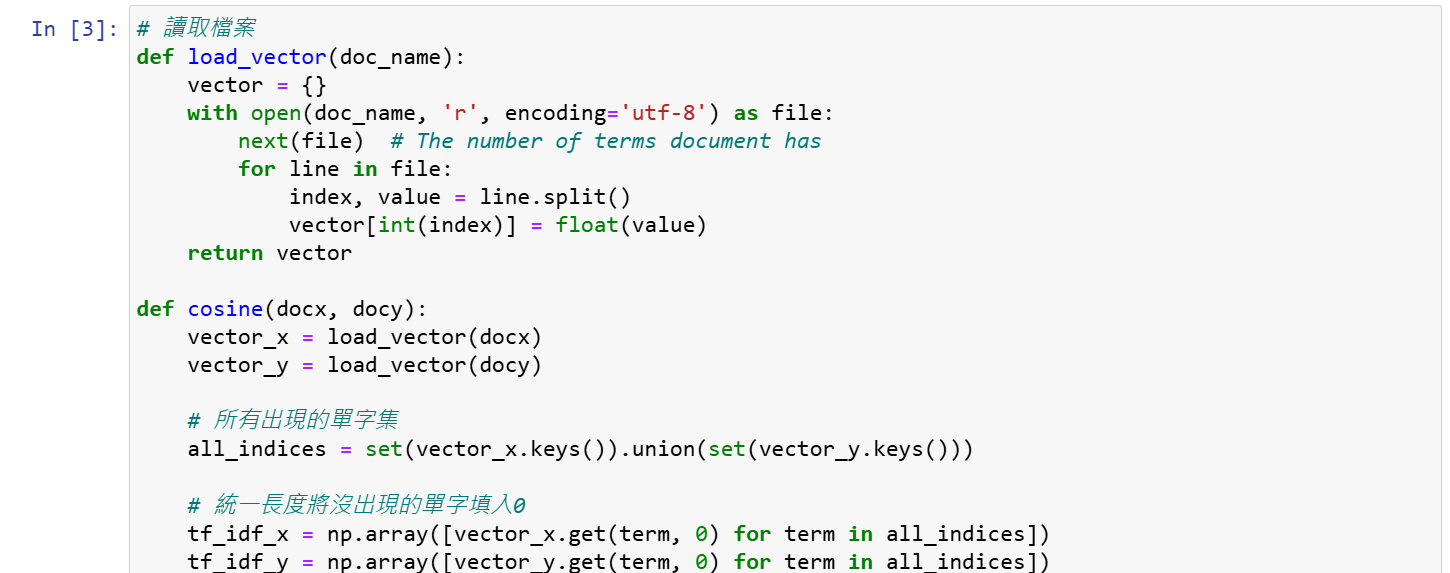
文字探勘HW3

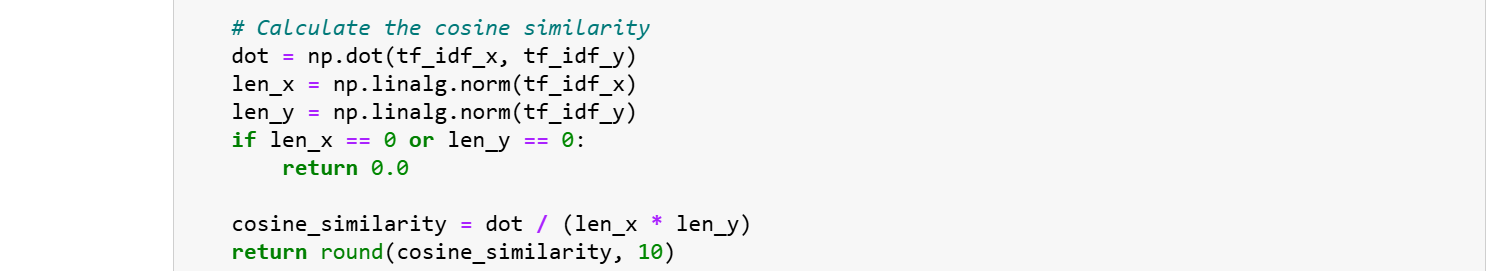
資管三 B11705033 江睿宸

1. 執行環境：Jupyter Notebook
2. 程式語言：Python3
3. 程式碼說明

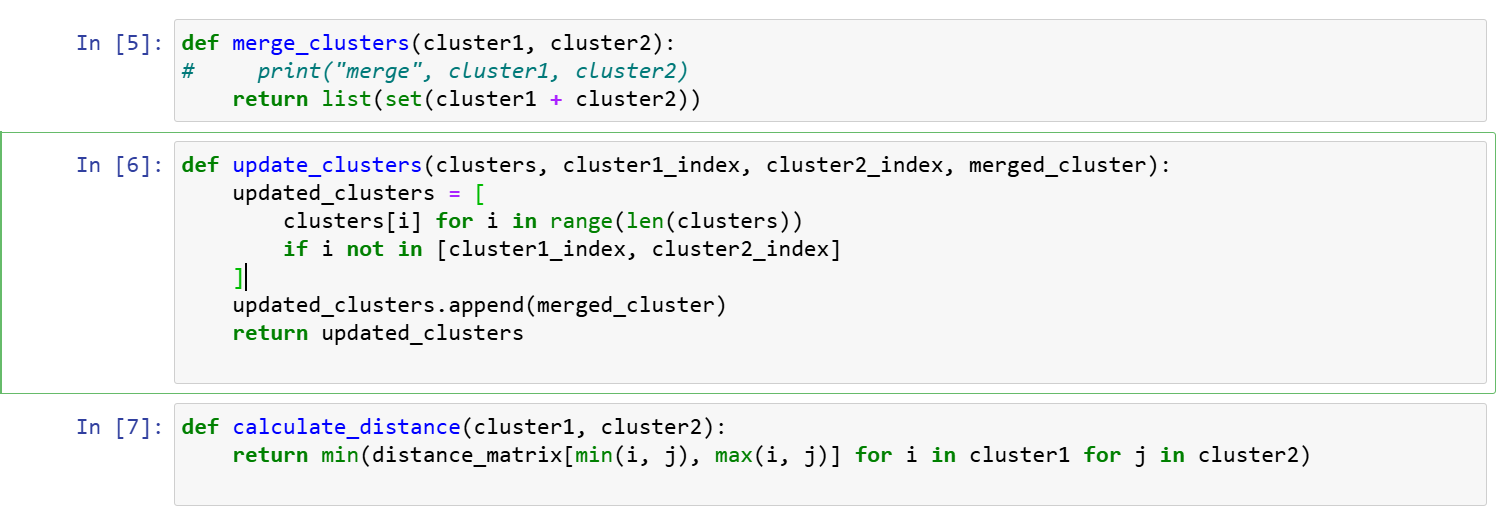
匯入需要的套件

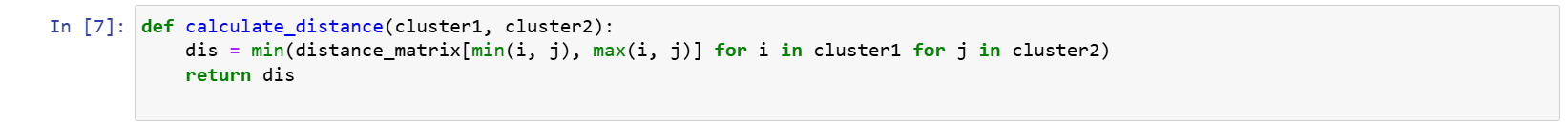


使用作業二計算出的vector (位於output中)計算cosine similarity



使用merge\_clusters合併cluster

使用upadate\_clusters將合併前的cluters刪除後新增新的cluster

Calculate\_distance則重新計算cluster之間的距離，本題使用single，使用complete會發現最後幾筆最大距離都是1，也就是相似性0，因此取距離最小

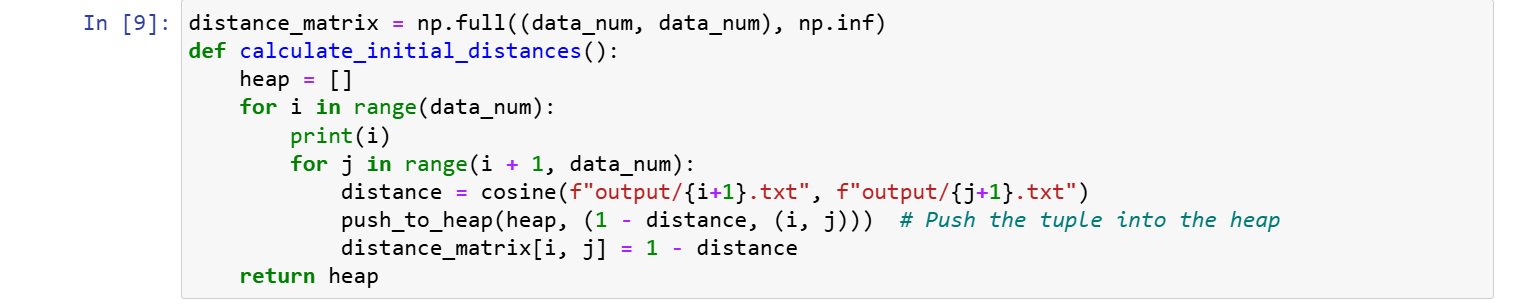
實作Heap的功能

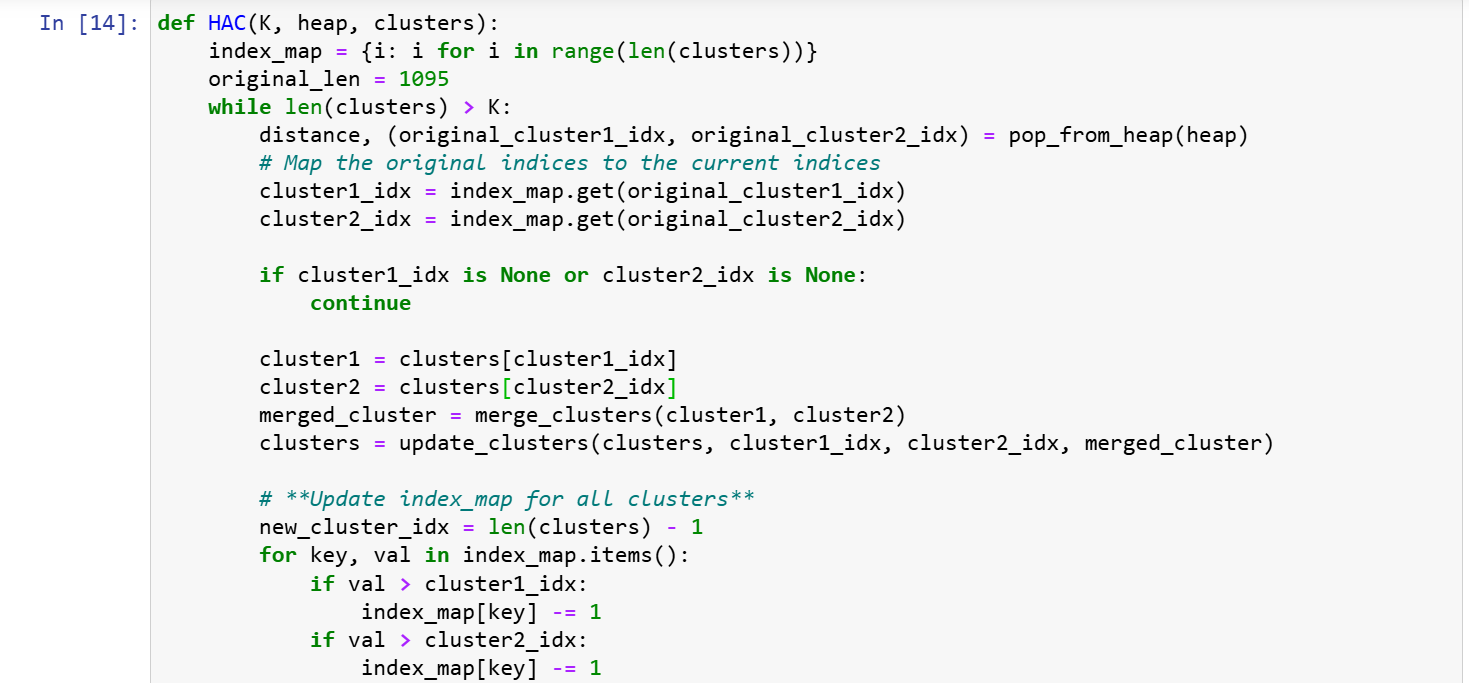


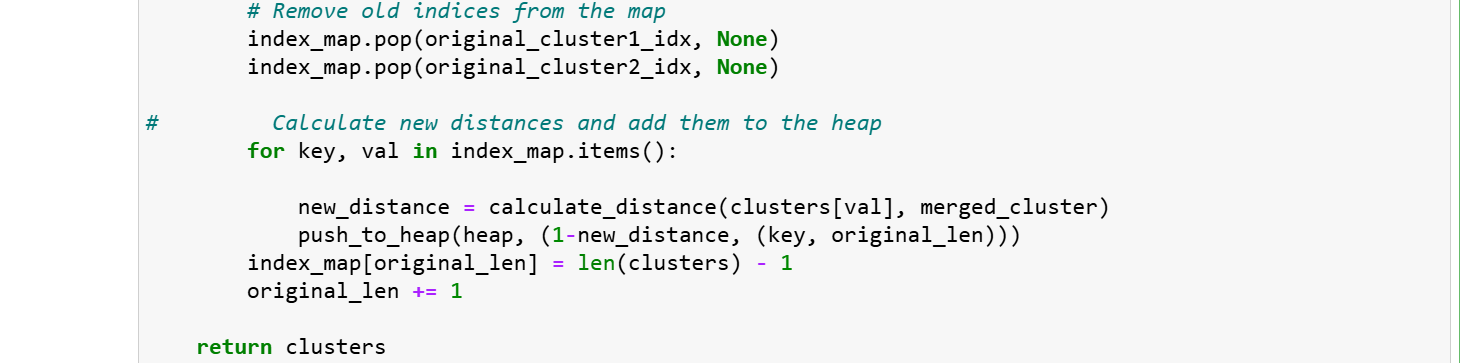


計算初始文件的相似度，並用1-similarity作為距離，為了避免I/O使用distance\_matrix將一開始計算的cosine\_similarity紀錄，不用反覆開檔案。

實作min\_heap將結果存在heap中



當cluster被合併刪除時會產生位移，因使用index\_map確保合併過後編號還指向正確的位置，並根據結果重新計算距離



執行並將結果存為指定檔案格式



