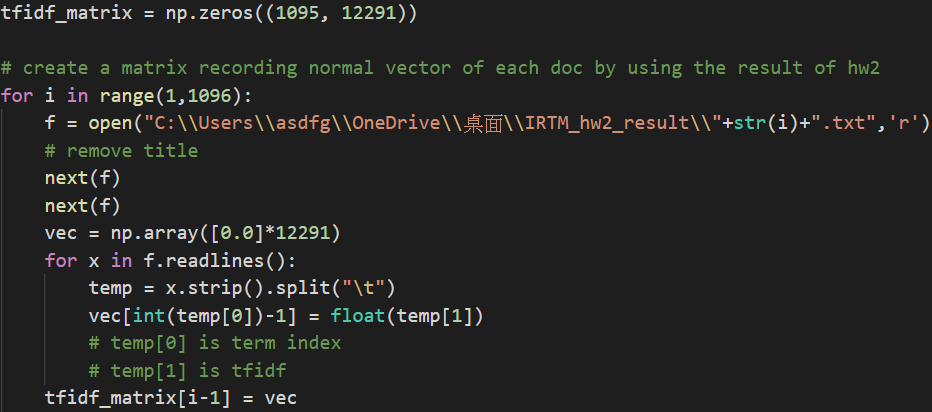
Introduction-to-Information-Retrieval-and-Text-Mining, HW 4

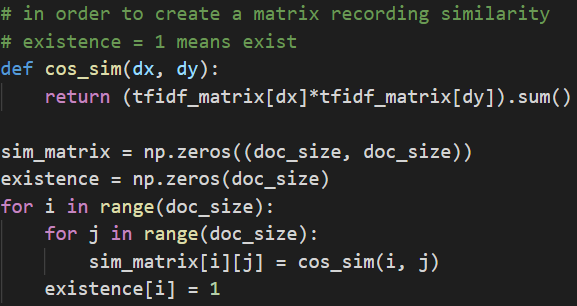
B07302230, 財金三 王博奕

1. 執行環境：Visual Studio Code
2. 程式語言：Python 3.8
3. 執行方式：
   1. Pip install np 來計算cosine similarity
4. 作業處理邏輯說明：

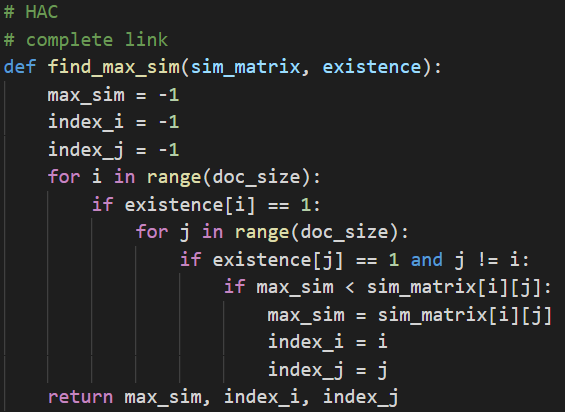
Step1: 讀取作業二的解答，並且裝進一個1095(文章數)\*12291(字典)的矩陣。其中每個文章都已變成tfidf的向量。

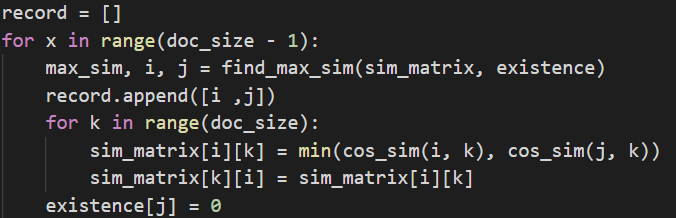


Step2: 創出一個矩陣sim\_matrix，裡面放各文章彼此的相似度，此外紀錄該文章是否被併入。



Step3: 定義出一個找出sim\_matrix裡最大值的函數，同時尚未被併，並且記錄兩者的編號。之後以complete link的方式來更新sim\_matrix的值。





Step4: 依據作業要求定義寫檔的方式。先定義一個 dictionary 其中讓每一篇文章 ID 作為 key，而 value 就是目前的分群結果，初始值即為每一篇文章自己的 ID。將 record 中的每一個 pair 照順序讀入，一一將他們合併起來，當合併結果的群組共有 20、13、8 時就寫檔。

