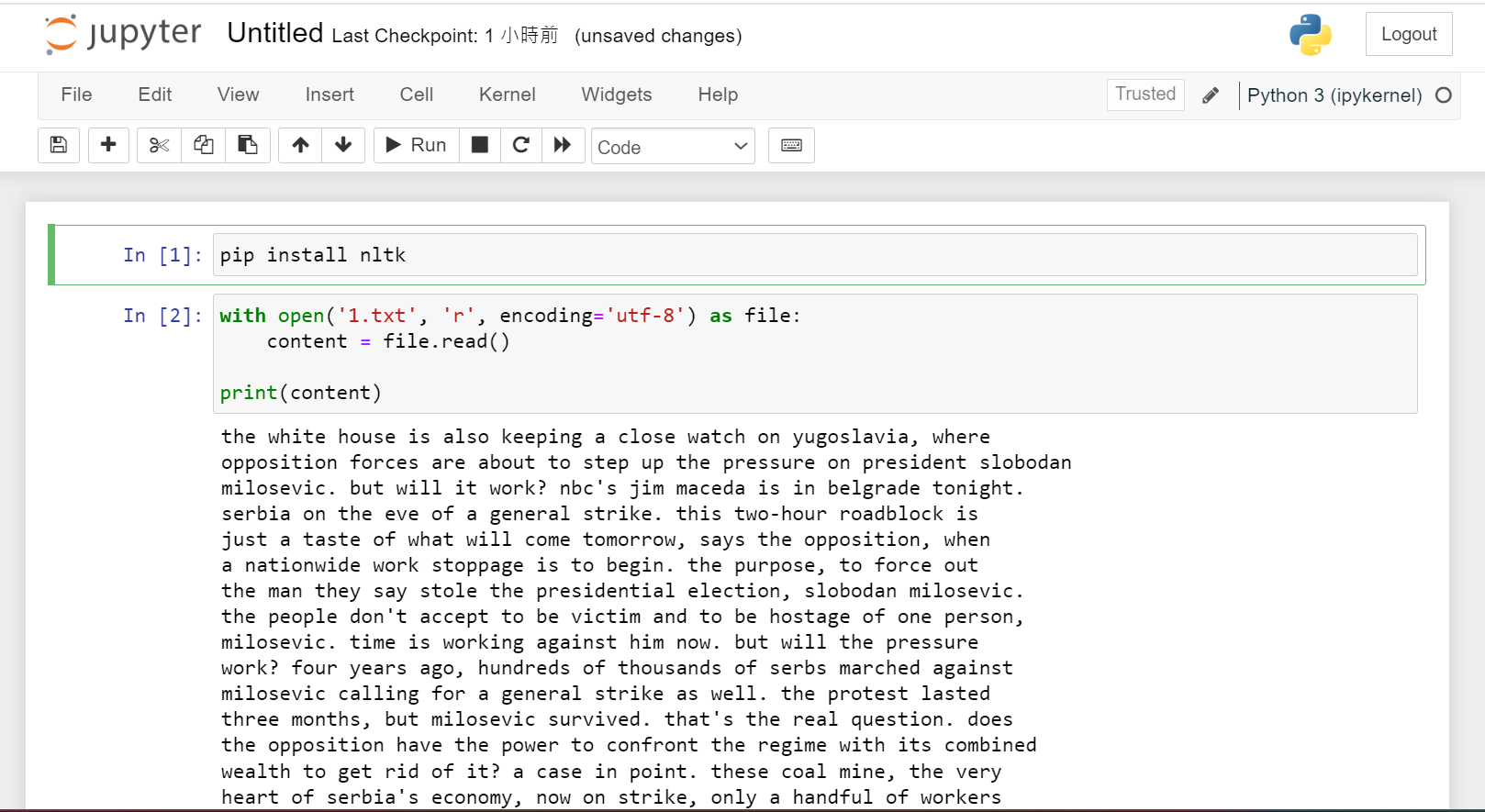
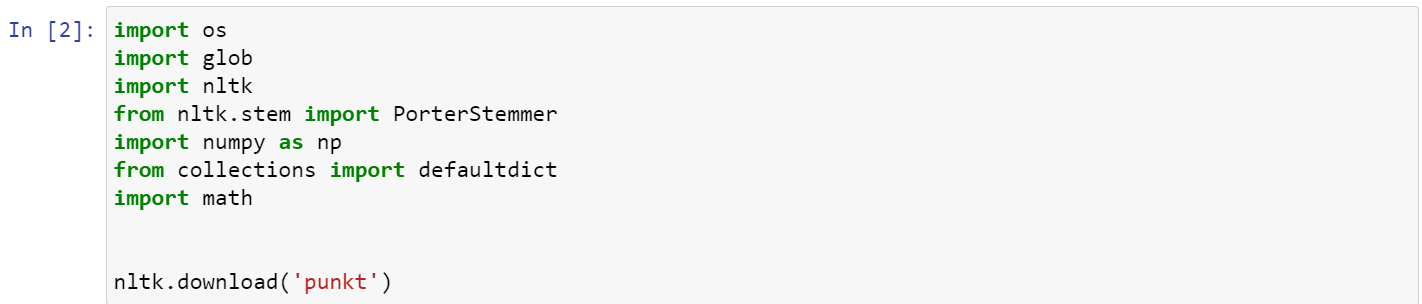
文字探勘HW2

資管三 B11705033 江睿宸

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

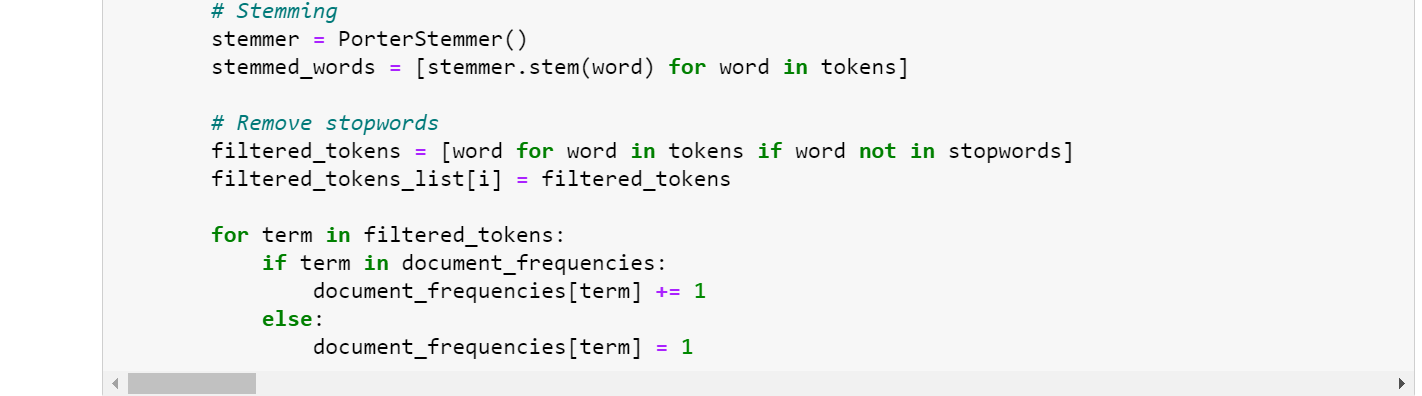
如果沒有ntlk套件需要先安裝(執行第一個cell)

使用jupyter notebook執行全部

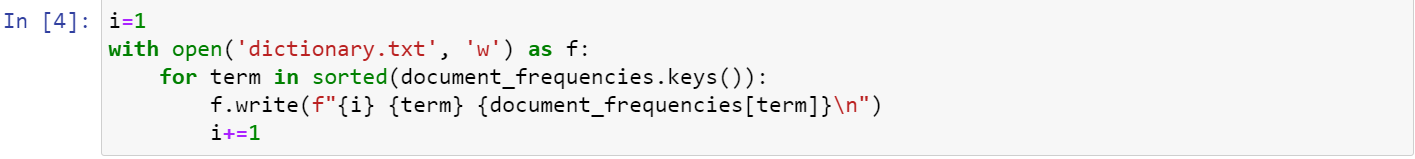
1. 

Import 需要的套件

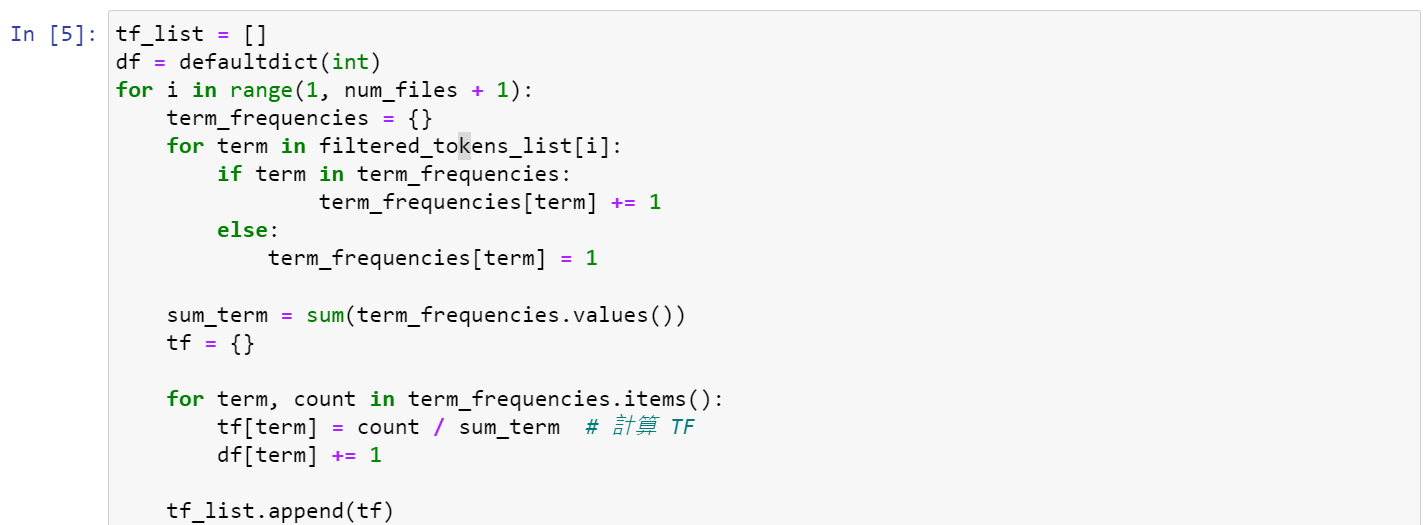


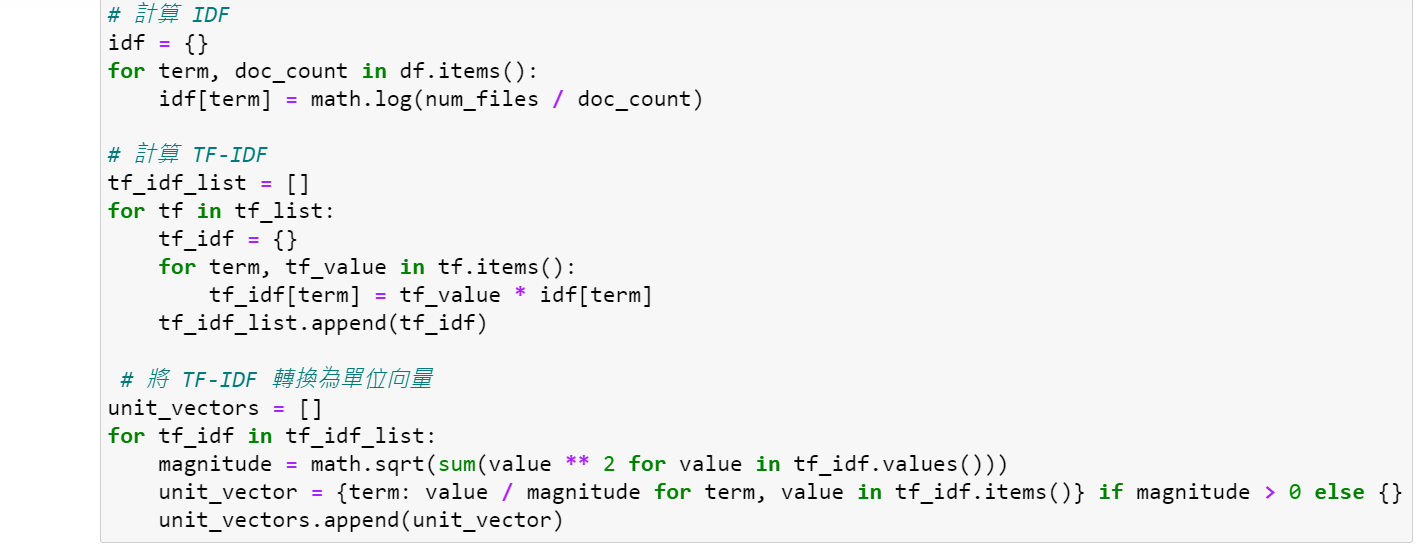


使用作業一的tokenize取出token，並計算每個單字的出現次數



將結果根據單字升序寫入dictionary文件中



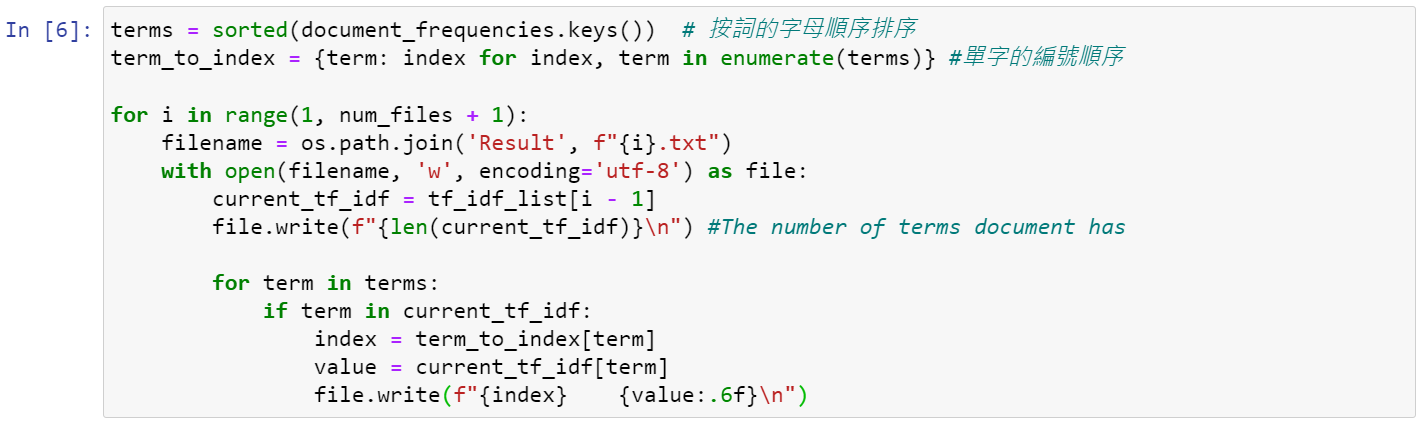


根據公式計算TF-IDF

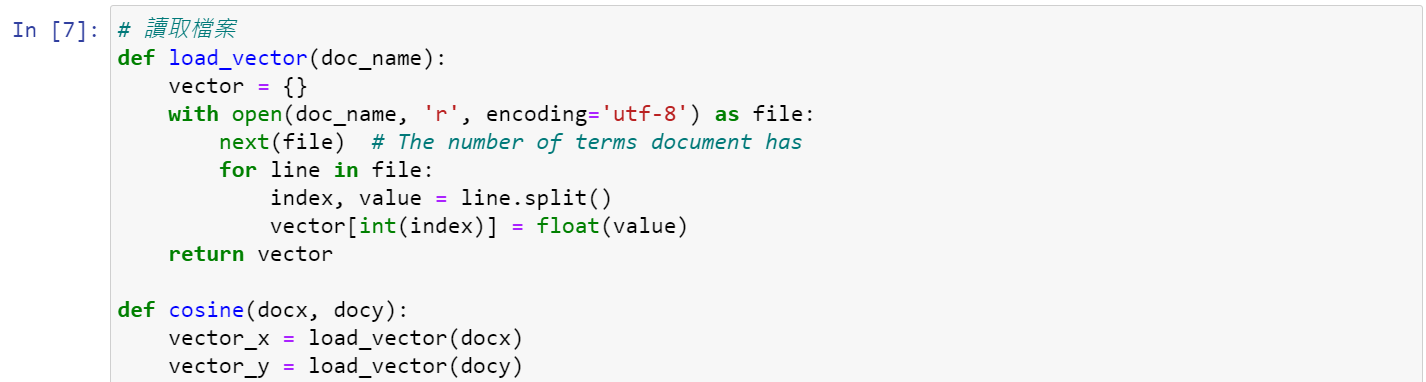
TF = 計算每個單字在該篇的出現次數 / 該篇單字量

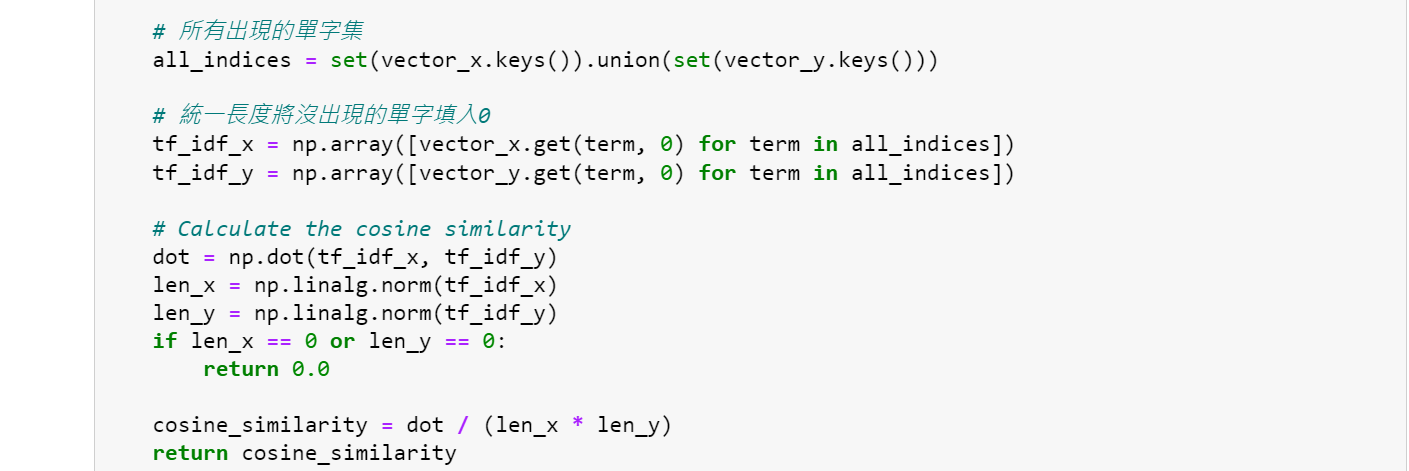
IDF = 根據單字計算總文章數 / 在出現在各幾篇文章中並取log

相乘後除以長度得單位向量

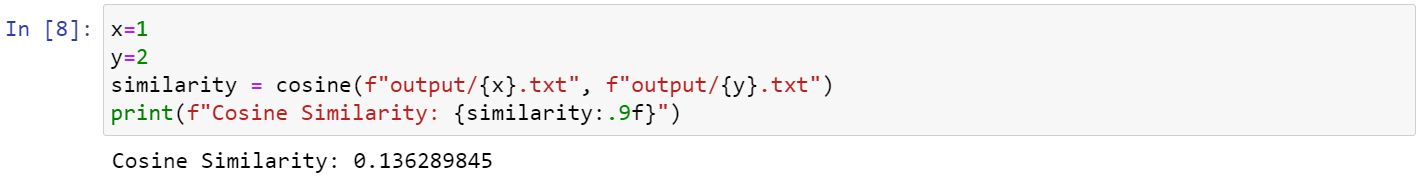


在Result資料夾中根據每個document寫入單字的及其TF-IDF值





使用numpy函數取得內積及兩vector長度，並計算cosine similarity



在x y分別輸入想確認的兩個文件編號，便可得cosine similarity