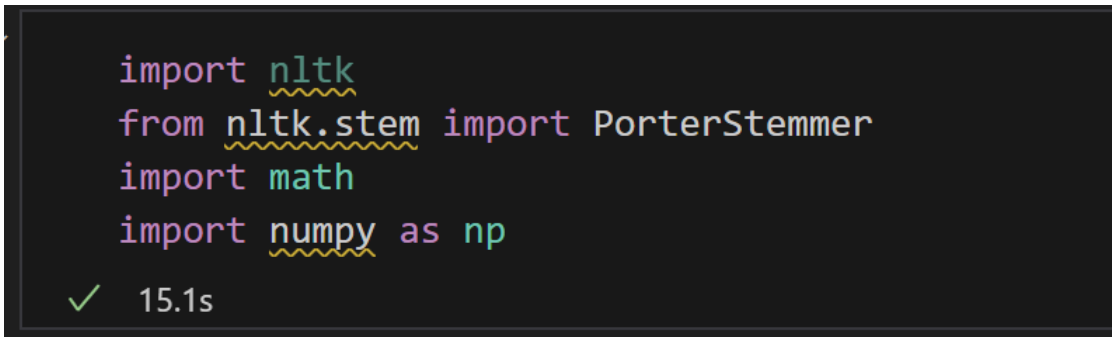
IRTM HW3 report

B11705058 資管三 廖振翔

執行環境：Jupyter Notebook

程式語言：Python3.12.7

執行方式：需 install 的套件



只需要在跟此程式檔同一層級的地方有作業提供的 IRTM 資料夾，就可以 Run All 執行程式碼。

作業處理邏輯說明：此程式碼先用 for 迴圈逐個讀取文檔，同時進行前文檔的前處理，並計算每個文檔的 tf 和 df，在 for 迴圈跑完的時候就已經會有所有文檔的 tf 和 token 的 df，之後再套公式算出每個 term 的 idf，之後就可以開始計算每個文檔中所有 term 的 tf-idf 並 normalize 為 unit vector。接下來寫出題目需要的 dictionary.txt 和 1.txt，最後定義 cosine function 來計算兩個文檔之間的 consine similarity。