Report

1 執行環境

Jupyter Notebook

2 程式語言

Python 3

3 執行方式

使用 Jupyter Notebook 執行，並且把 data/ 跟 stopwords.txt 放到跟程式碼同個資料夾直接按執行即可。

4 處理邏輯

這份程式碼先把所有需要的檔案以及 stopwords 讀進來，在讀檔的同時就對每個文本做 preprocess，接著計算 df 並且把結果存到 dictionary.txt 這個檔案裡，t_index、term、以及 df 之間用 tab 隔開。下一步就是計算 tf-idf，把之前計算 df 的結果直接拿來用，計算完之後再把每個文本的 tf-idf unit vector 存到 output/ 這個資料夾底下，t_index 以及 tf-idf 之間一樣用 tab 隔開。最後計算 cosine similarity，`cosine(Docx, Docy)` 這個函式會把 output/ 裡面的 `Docx.txt` 和 `Docy.txt` 檔案讀進來計算 cosine similarity，最後把結果 print 出來。