**IR HW2 Report**

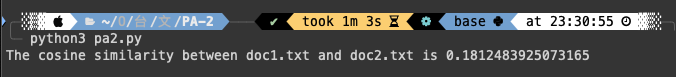
**R10725035 呂文楷 資管碩一**

1. **執行環境：** Python 3.8.8 @ Anaconda
2. **程式語言：** Python 3.8.8 @ Anaconda
3. **執行方式：**

程式會去data的資料夾中讀取資料，tokenize 後依序計算tf, idf, tfidf，最後算出並印出1.txt 和 2.txt 的cosine similarity。

該次使用的套件有：

* nltk (for stemming and stopwords)
* numpy (for vector operation)
* collections (for term count)
* typing (for function return type)
* os (for read directory)
* math (for calculating log)



1. **作業處理邏輯說明：**

**Text

Description automatically generated**

步驟如下：

1. 讀取路徑中所有的文字檔，紀錄檔名、地址、DocID。

Text

Description automatically generated

1. 將每個文件的文字tokenize、lowercase、normalized、stem成term（pa1的方法），紀錄freqeuncy後，建立一個對應的instance把資料存進去。

Text

Description automatically generated

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Text

Description automatically generated

1. 計算整個dictionary的idf，會先從所有文件中找出dictionary，然後計算dictionary中每個term的document frequency，然後計算idf。

Text

Description automatically generated

1. 有個每個文件的tf，以及idf後，計算tfidf

Text

Description automatically generated

Text

Description automatically generated

1. 最後，把dictionary, doc1.txt, doc2.txt的tfidf存起來，

Text

Description automatically generated

1. 然後用cosine\_similarity這個function計算這兩者的cosine similarity。

Text

Description automatically generated