IRTM - Programming Assignment 2 Report

資管碩一 R10725018 郭宇雋

1. 執行環境：

Command Line

2. 程式語言：

Python3

3. 執行方式：

需要進行 pip install 套件包含 re, math, numpy, nltk, os，而後直接在 Command Line 執行以下指令 **python pa2.py** 即可完成執行。

作業要求的 Cosine Similarity 會輸出在 Command Line，其餘輸出檔案則依照作業格式要求輸出至對應的資料夾底下。

4. 作業處理邏輯說明：

- 先設定固定的 Global Variable
- 撰寫函數，包括「文檔前處理」、「計算 TF 與 DF」、「建立 Index Dictionary」、「計算 TF 與 TF-IDF 向量」、「計算 Cosine Similarity」
- 最後一部分為主程式，依序進行讀取檔案、檔案排序整理、計算各文件的 Term Frequency 與整個 Corpus 的 Document Frequency、建立 Index Dictionary，最後進一步計算出各文件的 TF 與 TF-IDF Vector
- 依照作業要求，儲存 Index Dictionary 至 dictionary.txt
- 依照作業要求，儲存各文件的 TF-IDF 至 docID.txt
- 依照作業要求，計算文件 1 與文件 2 的 Cosine Similarity，輸出至 Command Line
- **最終計算出來文件 1 與文件 2 的 Cosine Similarity 為 0.175497**