IR HW1 Report

B05705025 資管四 陳漢威

1. 執行環境:

VSCode, terminal

2. 程式語言

Python3

3. 執行方式

程式會讀入在同一資料夾內檔名為”input.txt”的文字檔，並產生出”output.txt”。
Output.txt 中會列出所有 stemming 過後且不包含 stopwords 的單字們，並用逗號隔開
執行程式前需要先 pip install 好以下套件:

- nltk (for Stemming and stopwords)
- re (for filtering the symbol)

4. 作業處理邏輯說明:

讀入.txt 檔後，會將所有的內容存為同一個字串並轉成全部小寫。在此之後會用 re 套件找出所有非字母的標點符號並將它去除。
清理過的字串再用.split()做 tokenized 轉成單字形成的 list，最後再比對每個單字是否為 stopwords，如果不是，就用 Porter Stemming method 修剪單字同時加入 output 字串，並用逗點隔開每個單字。
最後再將字串匯出成名為 output 的 txt 檔。