IRTM - Programming Assignment 1 Report

資管碩一 R10725018 郭宇雋

1. 執行環境：

Jupyter

2. 程式語言：

Python3

3. 執行方式：

需要進行 pip install 套件包含 requests, re, nltk，而後程式直接全部執行即可。

4. 作業處理邏輯說明：

- 首先透過 requests 套件讀取文本檔案內容，再透過取代函式與正則表達式，將文本中的「換行符號（\r\n）」以及「非文字與空白的標點符號」去除。
- 接下來透過 split 與 lower 函式，對文本進行 Tokenization 與 Lowercasing。
- 透過 NLTK PorterStemmer 物件，對每個單詞進行 Stemming 動作。
- 透過 NLTK stopwords 的 English Stopwords List，過濾所有 Stopwords，至此所得到的結果就是最後要輸出的所有單詞。
- 最後一步便是將所有單詞寫入 result.txt 檔案中（一個單詞一行），即完成所有處理流程。