IRTM - Programming Assignment 3 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 numpy, pandas, nltk
- 文章資料放置於 ./data/IRTM/ 路徑底下
- 對應標注資料（training.txt）放置於 ./data/ 路徑底下
- 而後在 Command Line 下指令 python pa3.py 執行
- 輸出的結果會在 ./output/ 路徑底下

4. 作業處理邏輯說明：

- 讀取文章資料（corpus）包含訓練與測試文章
- 讀取對應標注資料（labels）
- 對資料進行前處理，包括移除換行符號、標點符號與數字，並轉換為小寫
- 做 Tokenization 與 Stemming，並移除 Stop words
- 接著計算各文章中，每個 Term 的 TF 值，即得到完整訓練與測試資料
- 進行 Chi-Square Feature Selection，僅保留前 500 重要的 Term
  （Vocabulary Size 從 5445 減少至 500）
- 訓練與測試 Naive Bayes Model
- 產生最終結果 output.csv 檔案（最終的 F-Score = 0.98）