資管四 B06705048 王佩琳

邏輯：

對 training data 進行切詞後建立一 classToken 為 13 * 15 * 文章 tokens (list*list*list)，再對每個 token 依據 log likelihood ratio 公式計算出對每個 class 下的 p_t, p_1, and p_2 和 -2log($L(H_1) / L(H_2)$)，將 token 作為 key，value 為對 13 個 class 的 LLR value，以 dictionary 形式儲存，對每個 token 求出 13 個 class 的 LLR 的平均，取出平均 LLR 前 500 高的 token 作為 terms 依據，忽略不在那 500 terms 裡的 token，建出 newClassToken （13 * 15 * 在 500 terms 的文章 tokens），training 依照公式 P(X=t|c) by using add-one smoothing，算 500 字各自出現在每個 class 的機率，testing 不在 training data 裡的文件，取最高 score 的 class 作為預測該文件的 class

執行：

Python3 pa3_NB.py

輸出 result.csv