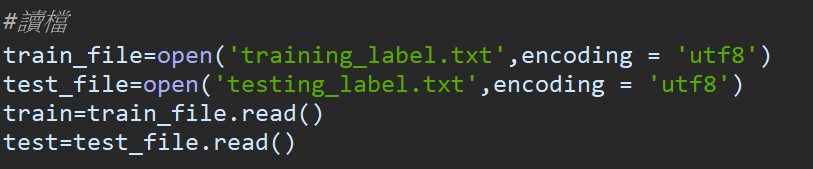
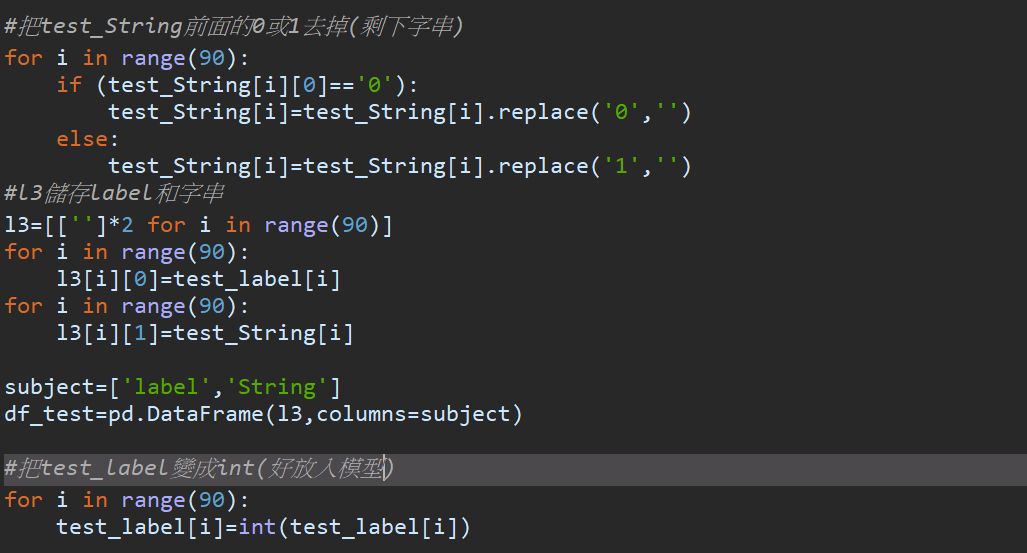
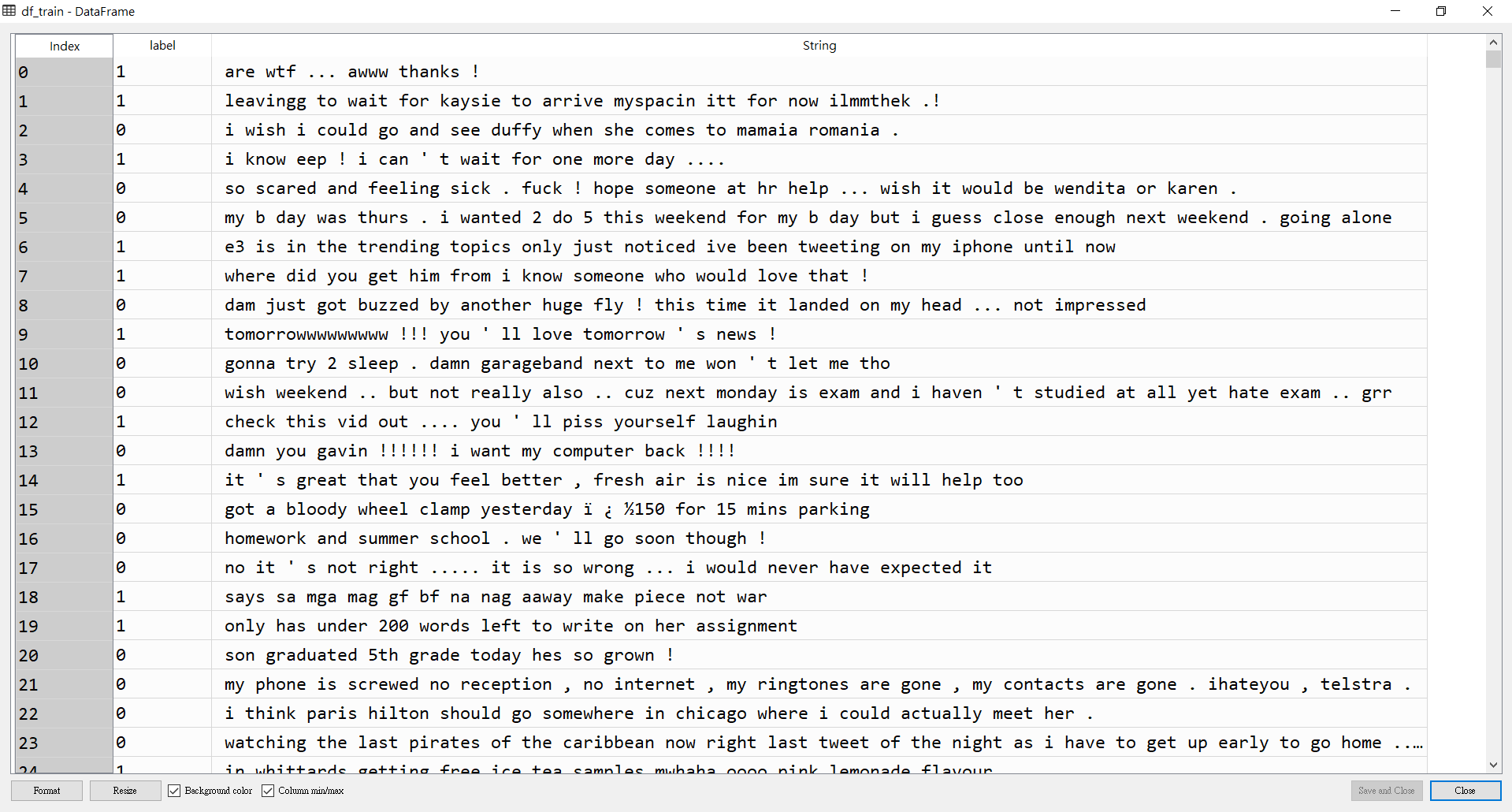
0853422 范姜鈞

1. 文字探勘前處理方法
2. 讀取training\_label.txt、testing\_label.txt檔
3. 利用分割符號切割字串、建立train&test之DataFrame:

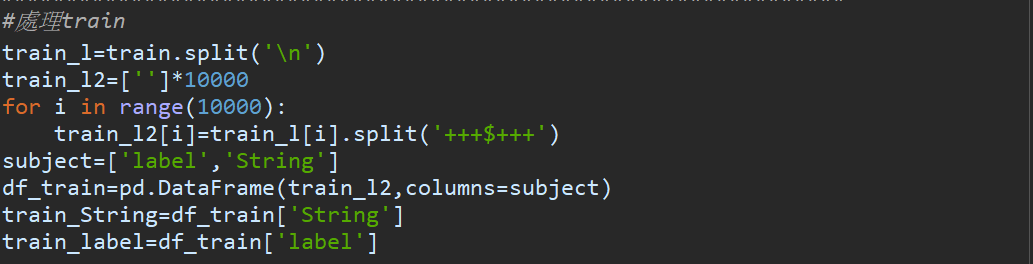
因testing\_label.txt檔中有些資料結尾有兩個”\n”，因此做特別的操作來達成效果(不是利用String.split())

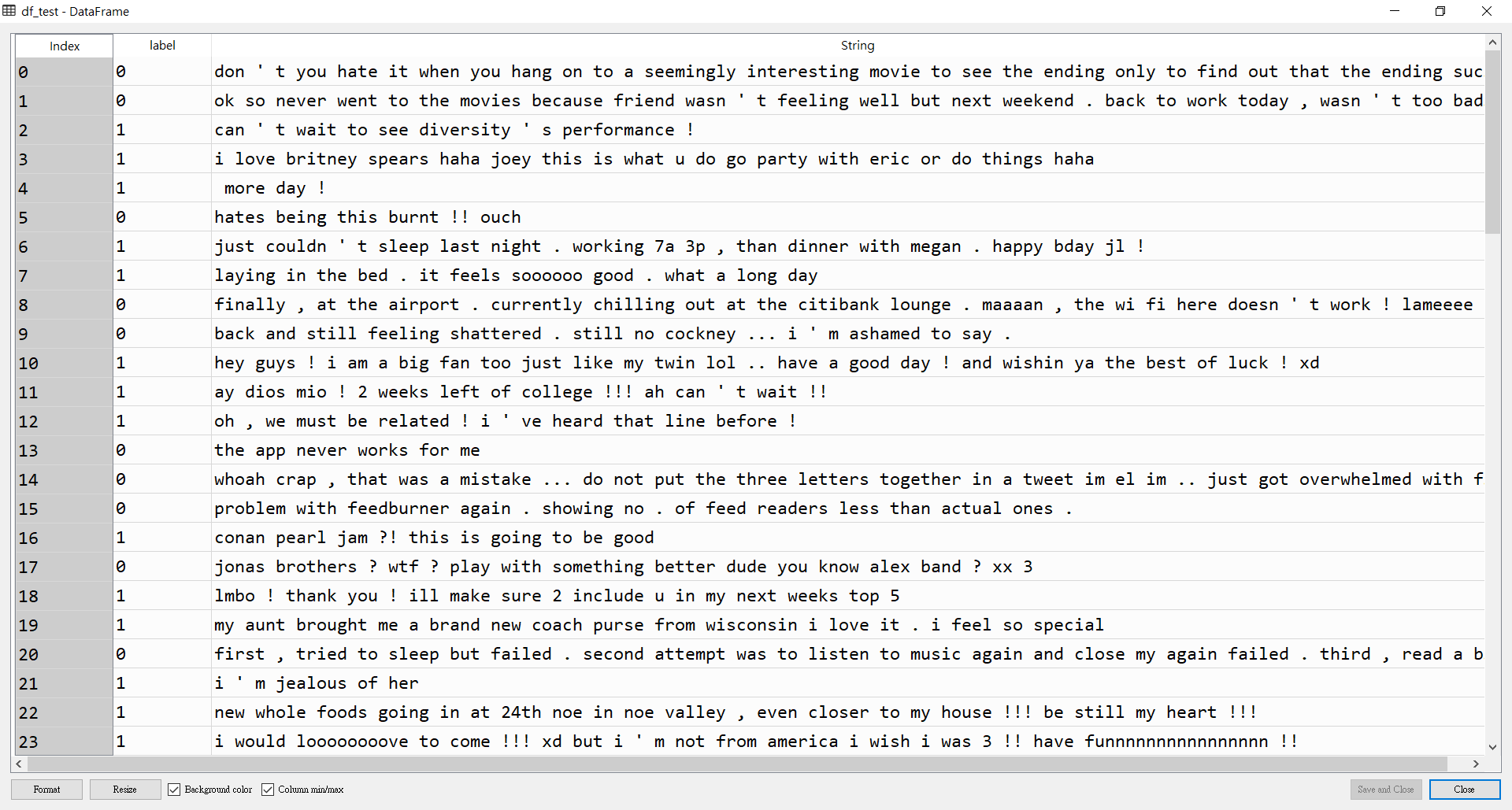






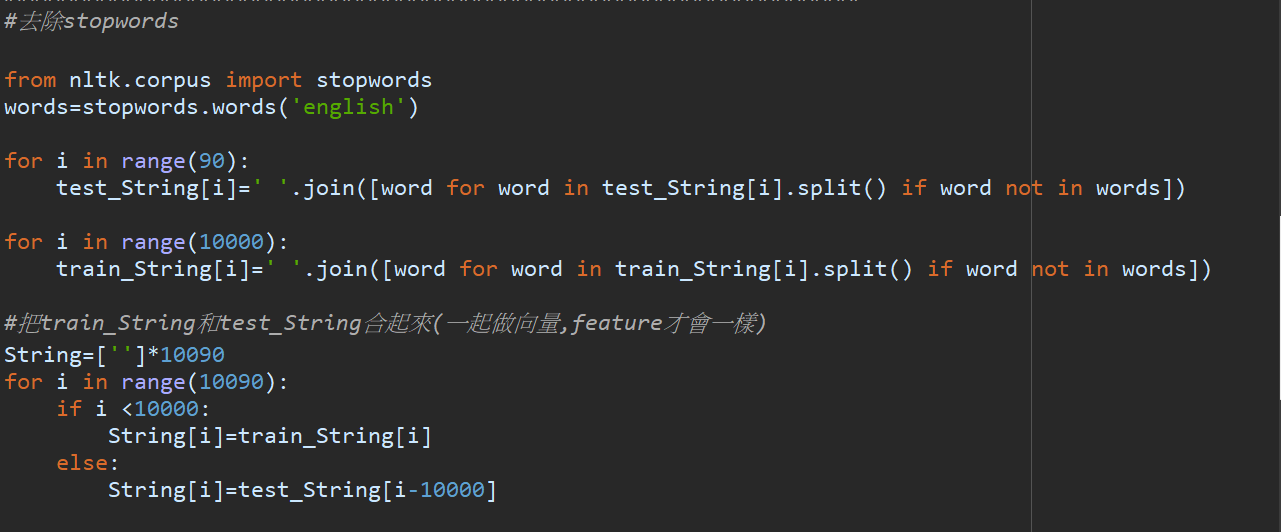
因資料量龐大，後續使用 tf-idf 轉向量後會有記憶體不足的問題，因此取train\_label.txt中的前10000筆資料當作train Data





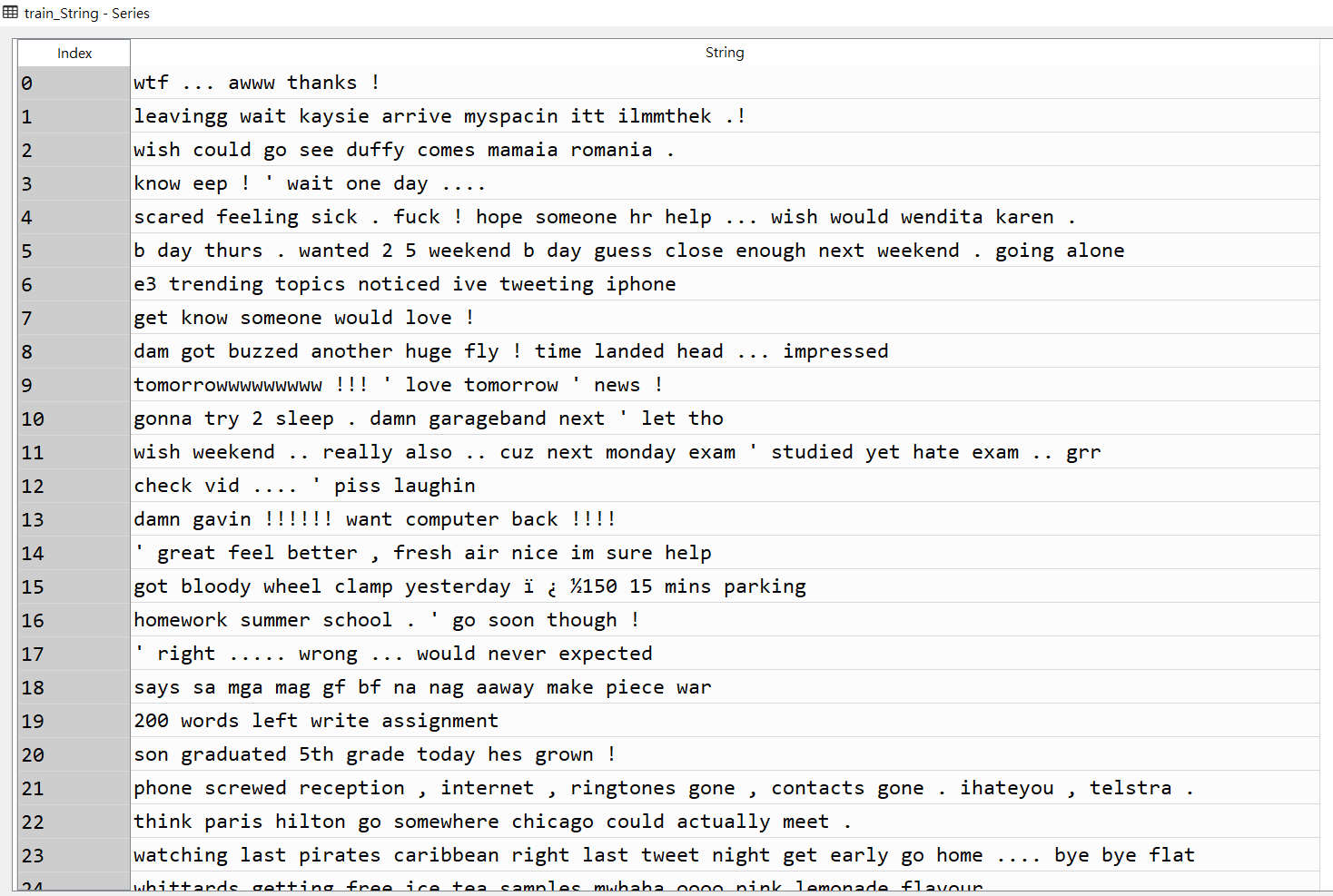
1. 去除stopwords:

針對train data與test data中的文字部分，去除停頓詞(透過import stopwords 中的英文停頓詞庫)

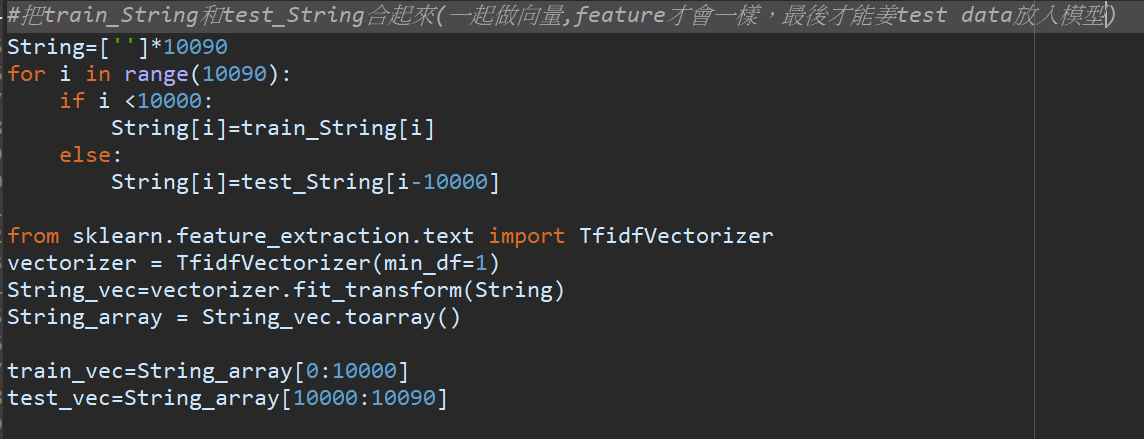


結果:

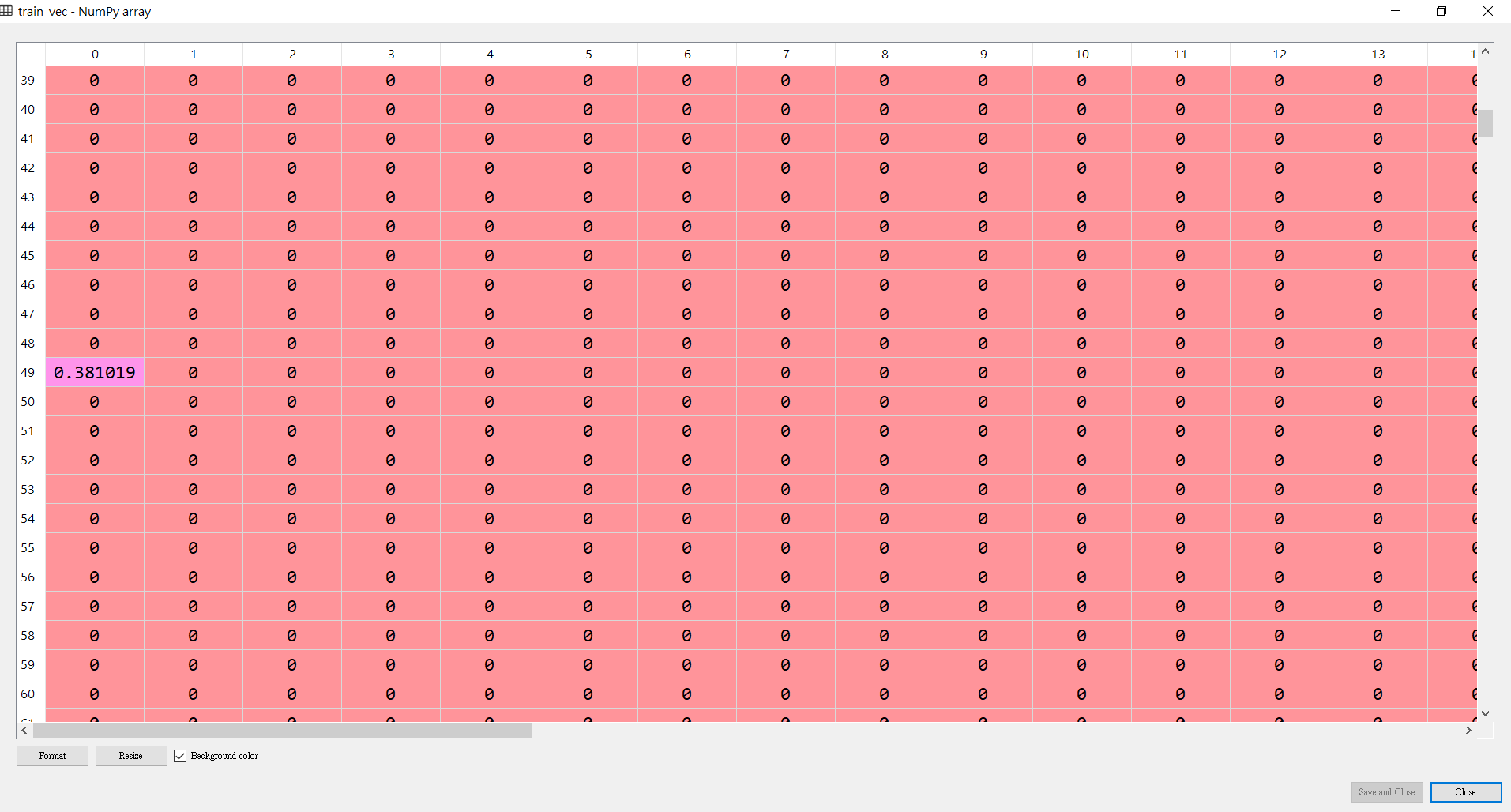




1. 將文字轉換成向量，像是常見的方法 tf-idf

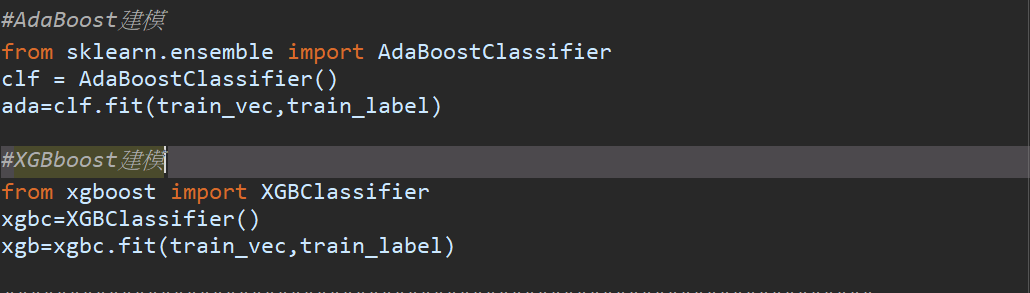


結果:

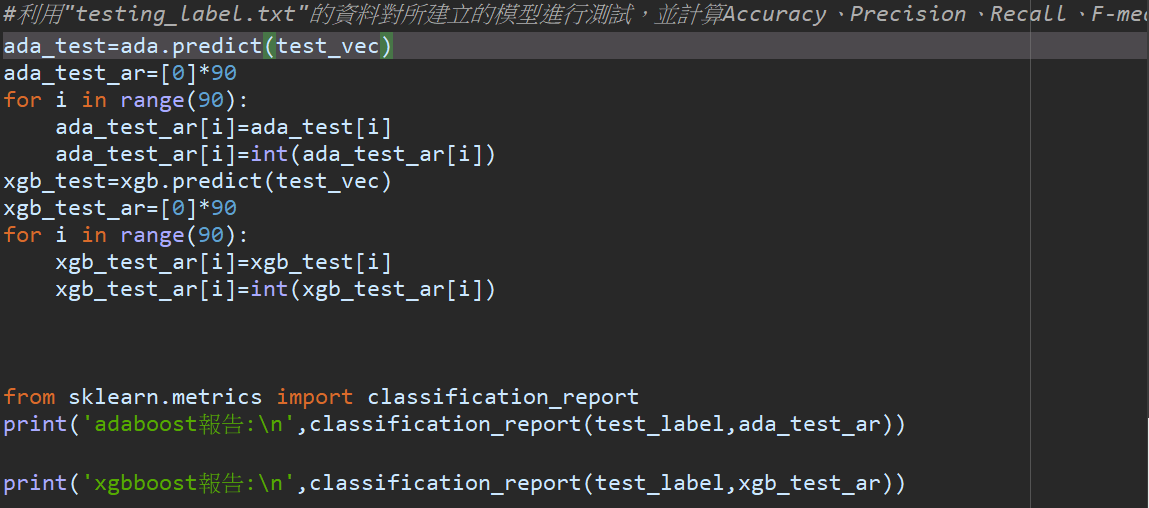




1. AdaBoost與xgboost 模型之結果比較:
2. 利用train data建立AdaBoost與xgboost 模型



1. 利用"testing\_label.txt"的資料對所建立的模型進行測試，並計算Accuracy、Precision、Recall、F-measure



結果:兩種模型的預測效果，XGBBoost較好

Accuracy:XGBBoost較高

Precision: XGBBoost較高

Recall: XGBBoost較高

F1-score: XGBBoost較高

