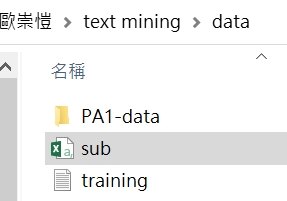
B08310054 資管三 歐崇愷

HW1

Python 版本: python 3.9.7

程式執行方式:執行.py 檔，修改程式碼內file\_path變數為任意絕對路徑，並將 PA1-data 與training.txt放到該路徑的資料夾內

資料夾內，並建立空白資料夾 result

在第一次執行的時候，須執行 以下載 nltk package

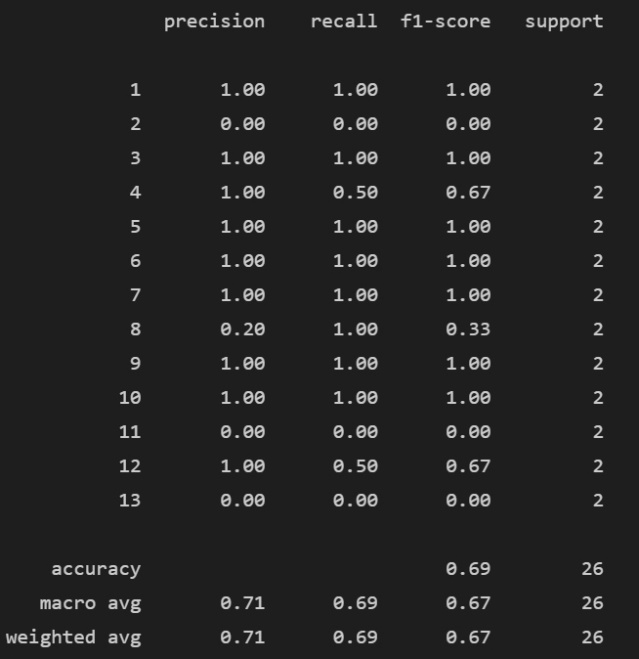
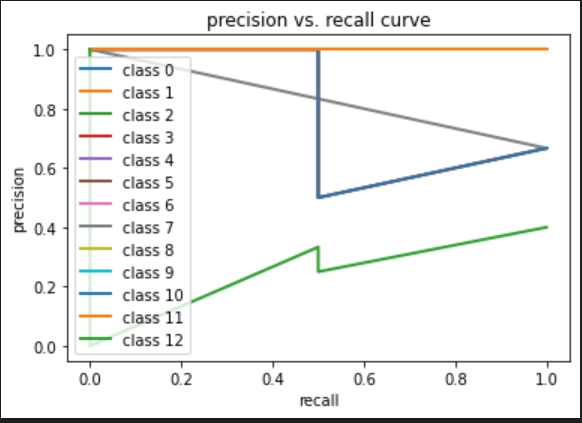
所需要的資料庫

程式邏輯 :

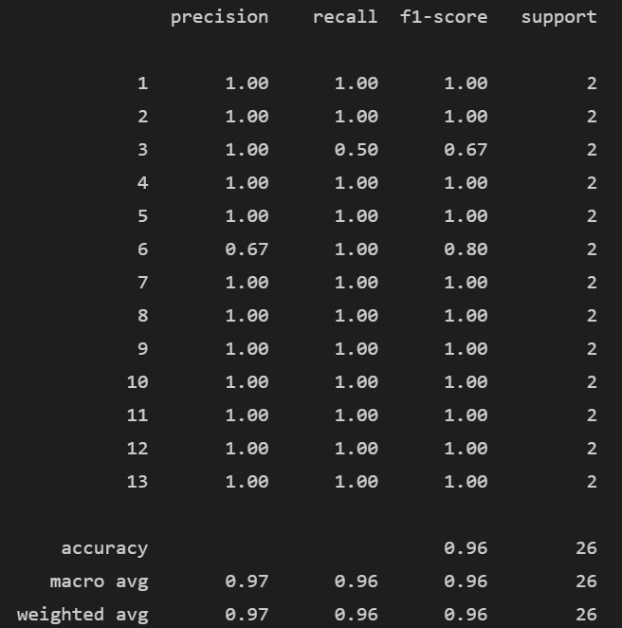
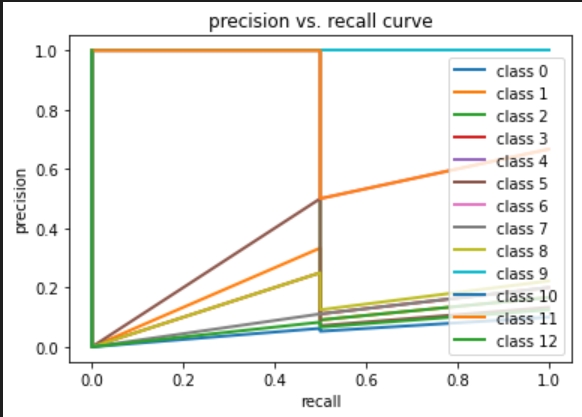
* 首先， 引用 numpy, nltk, sklearn 等 package
* 將 stopword assign 為 nltk 中預設的 stop word
* 將 PA1-data 裡每個文本 tokenize，並 lemmatize 後重新合併成文本
* 將文本合併至成一個 list
* 將有label的資料萃取出來，並把每個class最後兩筆資料作為evaluation，共為26筆資料
* 透過各個model去train，最後選擇表現最好，accuracy 為 1 的svmlinear作為上交kaggle的訓練模型
* 檔案回儲存在一開始指定的資料夾內，檔案名為sub.csv

模型表現:

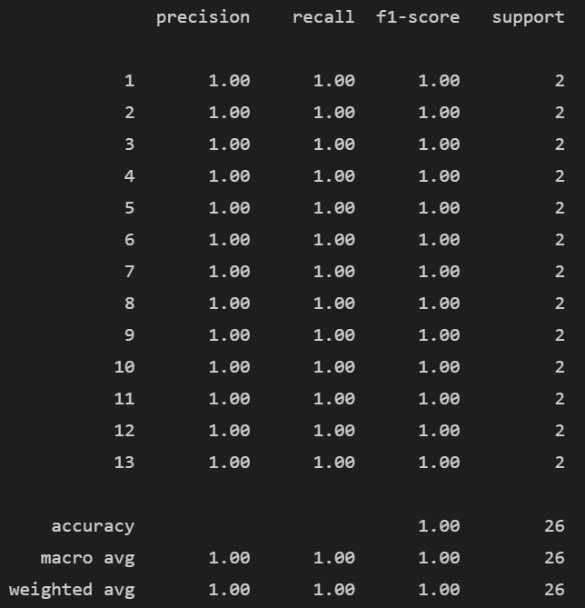
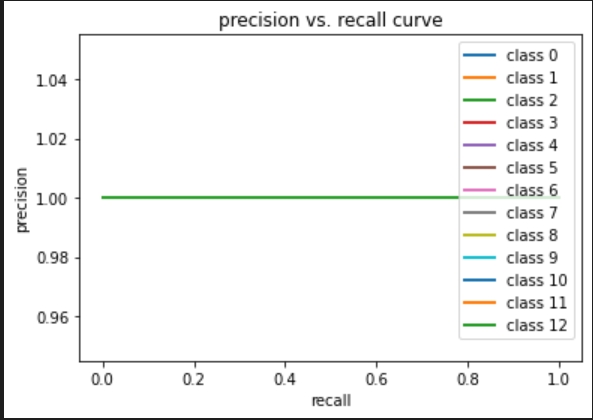
NB



SVM RBF



SVM LINEAR



整體而言

SVM linear表現更好，甚至達到全對的表現，加上lemmatize在kaggle上也得到98.多分，並列第三名