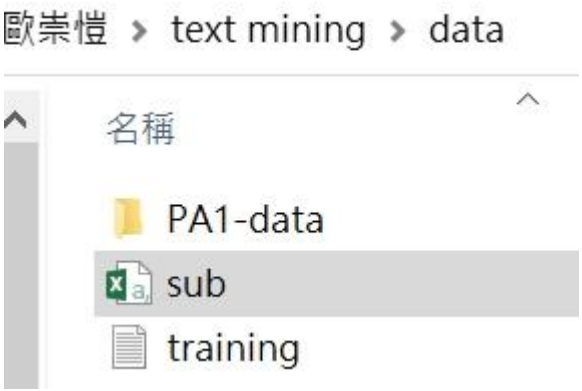
B08310054 資管三 歐崇愷

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

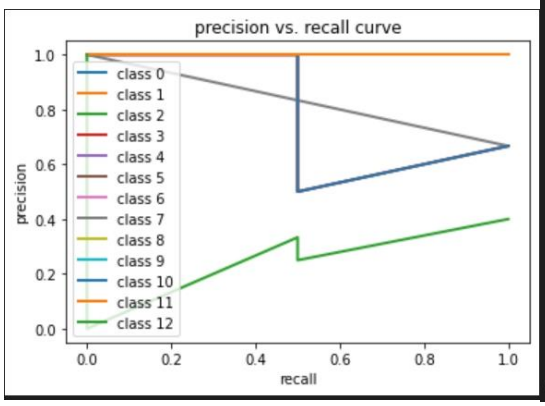
所需要的資料庫

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

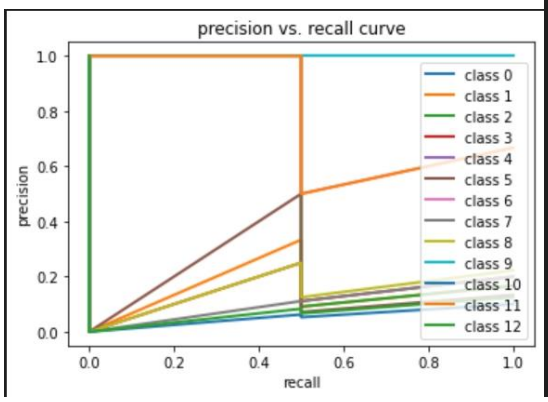
模型表現:

NB



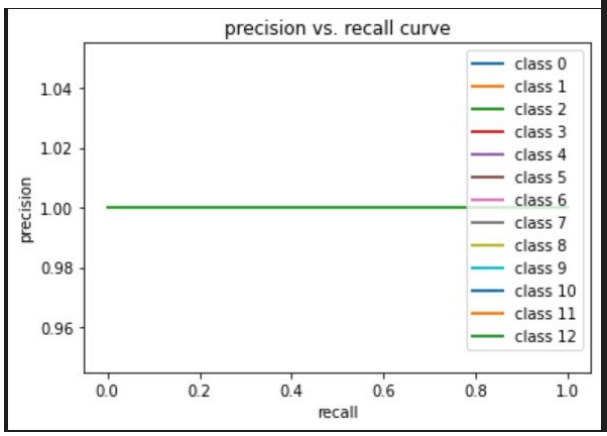
	precision	recall	f1-score	support
1	1.00	1.00	1.00	2
2	0.00	0.00	0.00	2
3	1.00	1.00	1.00	2
4	1.00	0.50	0.67	2
5	1.00	1.00	1.00	2
6	1.00	1.00	1.00	2
7	1.00	1.00	1.00	2
8	0.20	1.00	0.33	2
9	1.00	1.00	1.00	2
10	1.00	1.00	1.00	2
11	0.00	0.00	0.00	2
12	1.00	0.50	0.67	2
13	0.00	0.00	0.00	2
accuracy			0.69	26
macro avg	0.71	0.69	0.67	26
weighted avg	0.71	0.69	0.67	26

SVM RBF



	precision	recall	f1-score	support
1	1.00	1.00	1.00	2
2	1.00	1.00	1.00	2
3	1.00	0.50	0.67	2
4	1.00	1.00	1.00	2
5	1.00	1.00	1.00	2
6	0.67	1.00	0.80	2
7	1.00	1.00	1.00	2
8	1.00	1.00	1.00	2
9	1.00	1.00	1.00	2
10	1.00	1.00	1.00	2
11	1.00	1.00	1.00	2
12	1.00	1.00	1.00	2
13	1.00	1.00	1.00	2
accuracy			0.96	26
macro avg	0.97	0.96	0.96	26
weighted avg	0.97	0.96	0.96	26

SVM LINEAR



	precision	recall	f1-score	support
1	1.00	1.00	1.00	2
2	1.00	1.00	1.00	2
3	1.00	1.00	1.00	2
4	1.00	1.00	1.00	2
5	1.00	1.00	1.00	2
6	1.00	1.00	1.00	2
7	1.00	1.00	1.00	2
8	1.00	1.00	1.00	2
9	1.00	1.00	1.00	2
10	1.00	1.00	1.00	2
11	1.00	1.00	1.00	2
12	1.00	1.00	1.00	2
13	1.00	1.00	1.00	2
accuracy			1.00	26
macro avg	1.00	1.00	1.00	26
weighted avg	1.00	1.00	1.00	26

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