1. 執行環境

Visual Studio Code

1. 程式語言

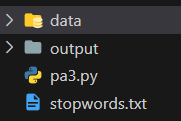
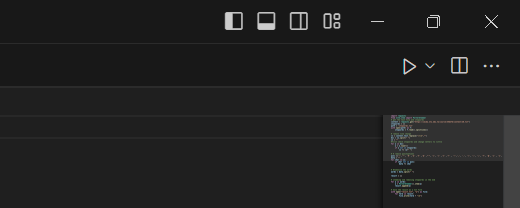
Python 3.10.6

1. 執行方式

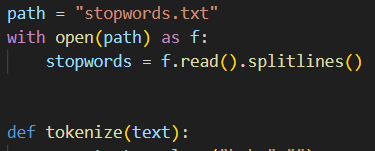
確認python已經安裝並可由vscode執行

在cmd中使用pip install pandas, pip install nltk來安裝套件，額外import的套件為os, math, string(應為python原生，不須額外install)

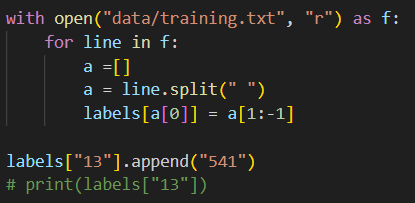
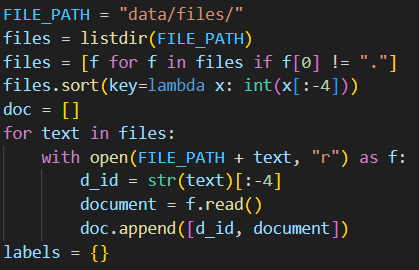
確認data資料夾(將訓練集跟training.txt放在一起)，output資料夾(開始前應為空)，stopwords.txt與pa2.py在同個資料夾中，在vscode打開該資料夾點擊run即可執行(必須打開整個資料夾避免找不到stopwords.txt)

  執行完應會發現原本為空的output資料夾中有output.csv，點開即可看見結果

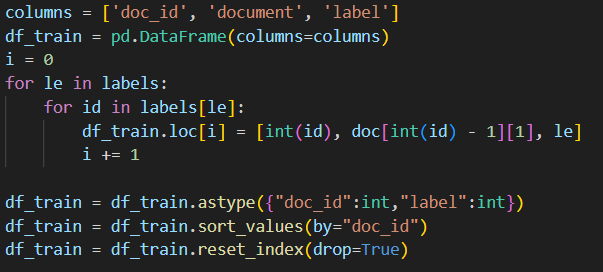
1. 邏輯說明
2. Read in stopwords and define function doing preprocessing for documents.



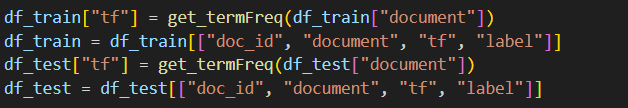
1. Read in training and testing files. First read in all documents, put those in training text with labels, since in with open(), it can not read the last element, so I put it in the labels manually.



1. Building data frame for training, use pd.Dataframe to represent the labeled documents. Columns represent its document\_id, text and class label respectively.



1. Like testing Dataframe, by the same way create testing dataframe, but this time set the column label to none.
2. Make a function which will return an input text’s dictionary. Put every documents’ term frequency dictionary in the training and testing Dataframe.



1. Define a function to return the unique words in a document
2. Define a function to use log likelihood ration to extract features (words).
3. Define functions to do training and testing NB classifier using training, testing Dataframe and the features selected.
4. Write the labels I got by apply testing in out/output.csv.