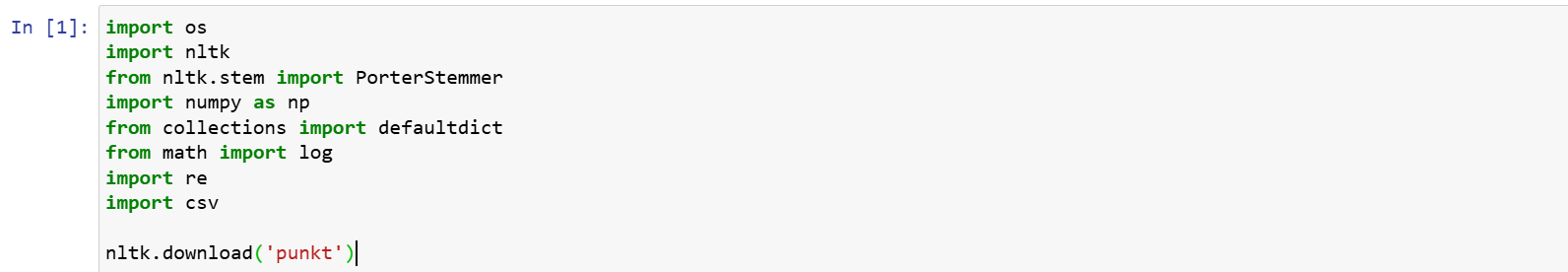
文字探勘HW3

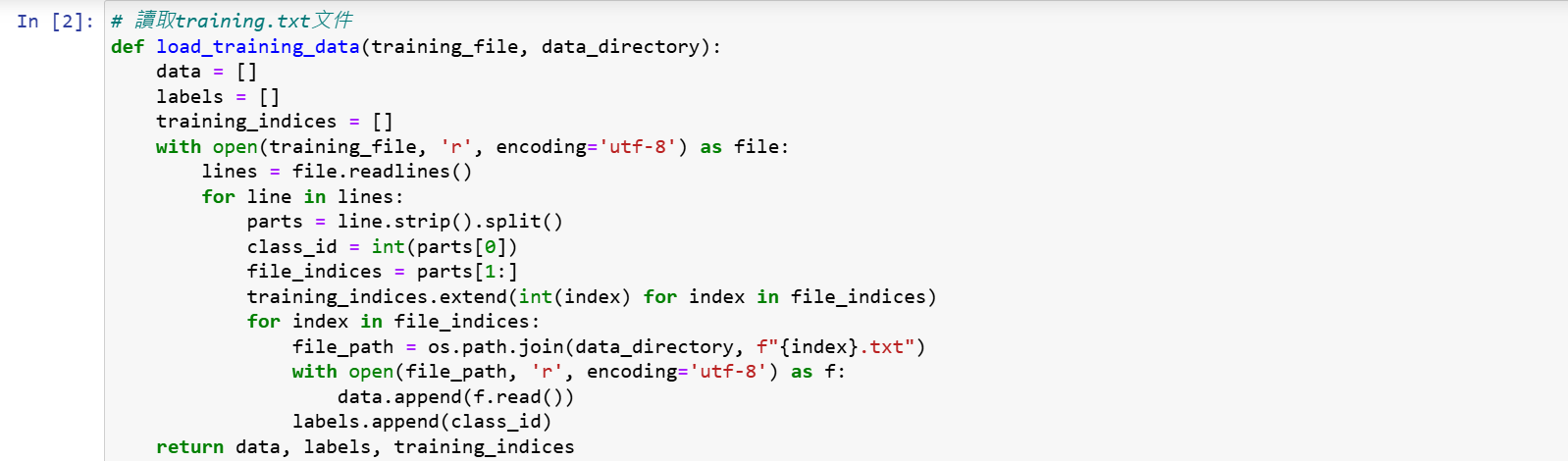
資管三 B11705033 江睿宸

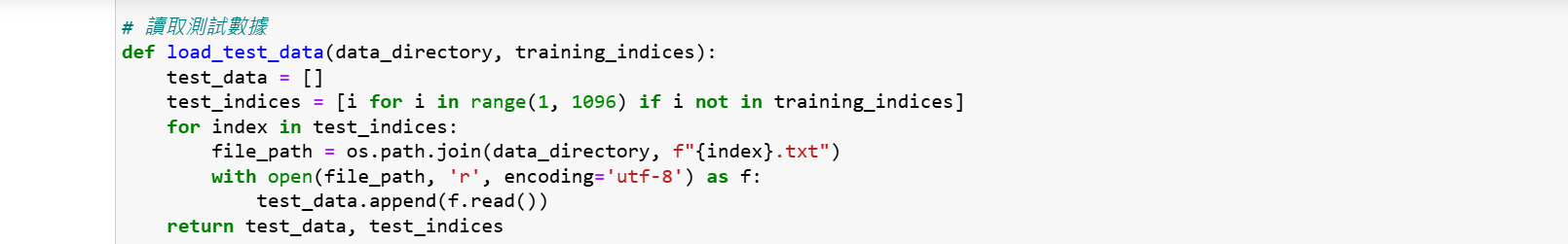
1. 執行環境：Jupyter Notebook
2. 程式語言：Python3
3. 程式碼說明

匯入需要的套件

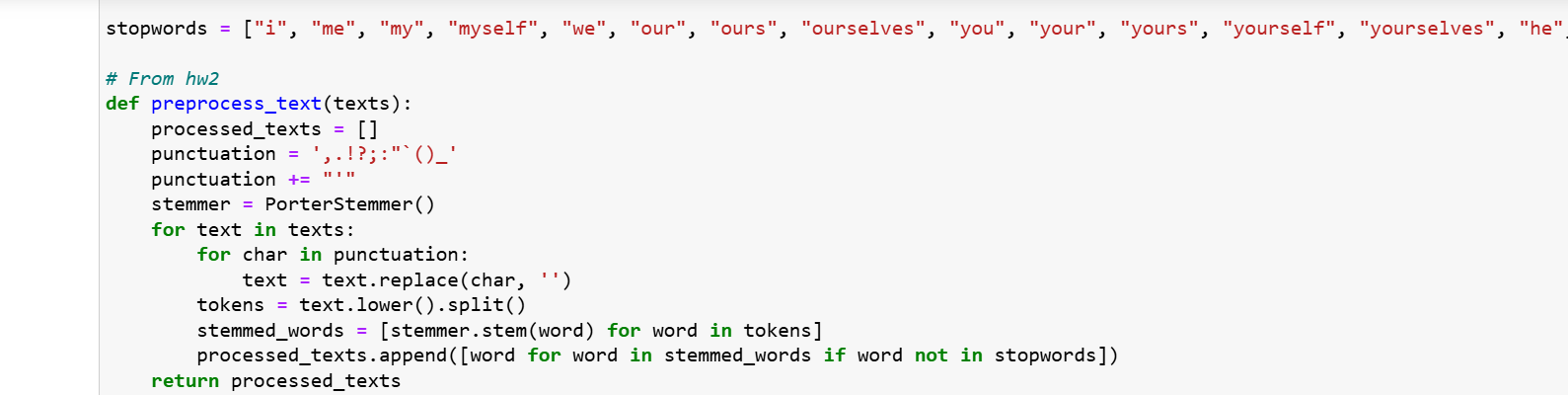


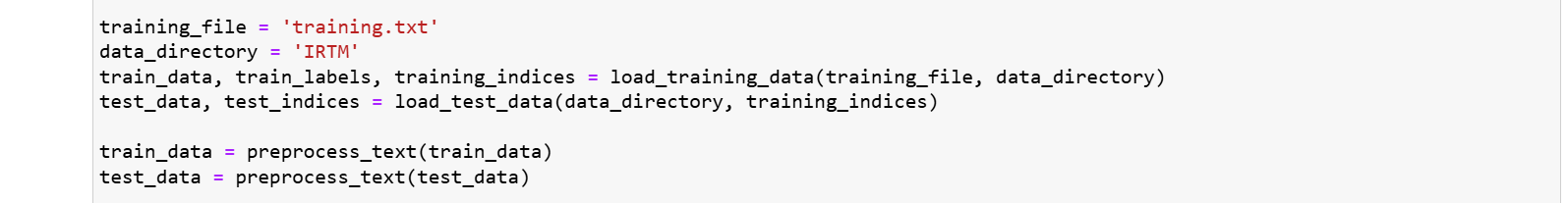
讀取1095個檔案並根據training.txt 分類



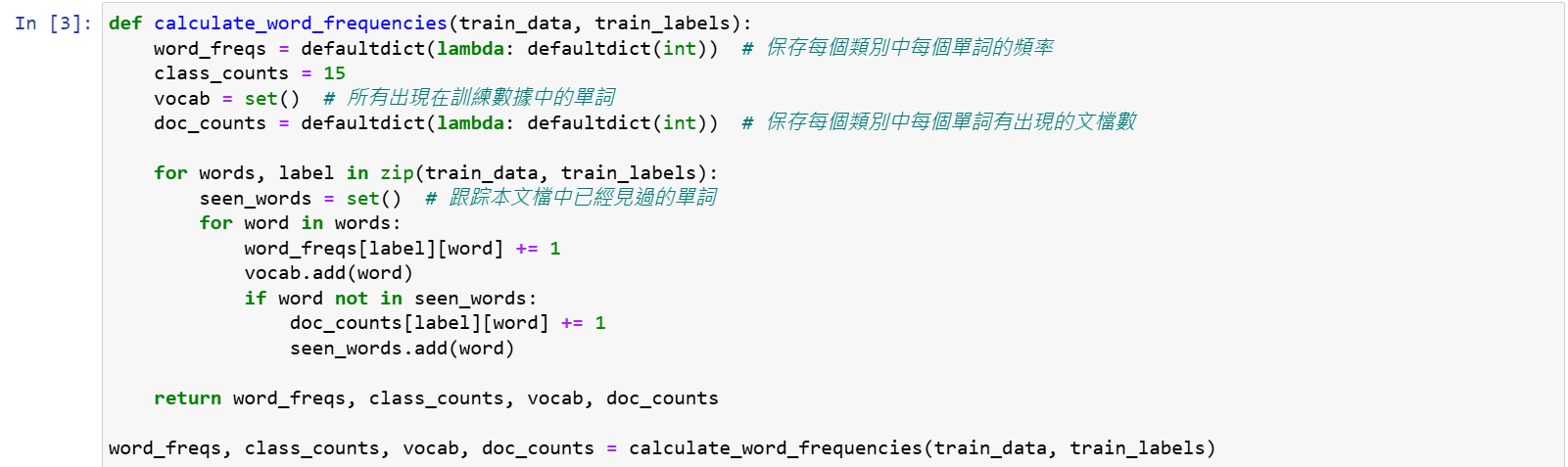


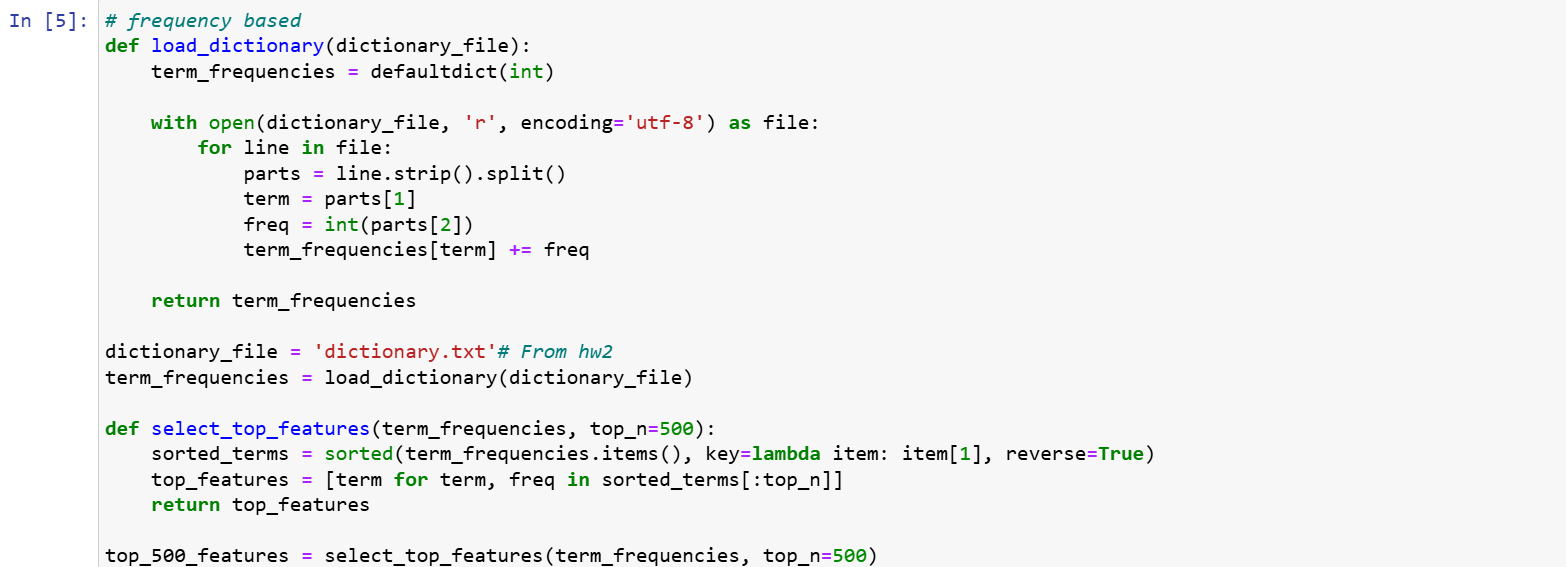
使用hw2實作過的功能提取tokens



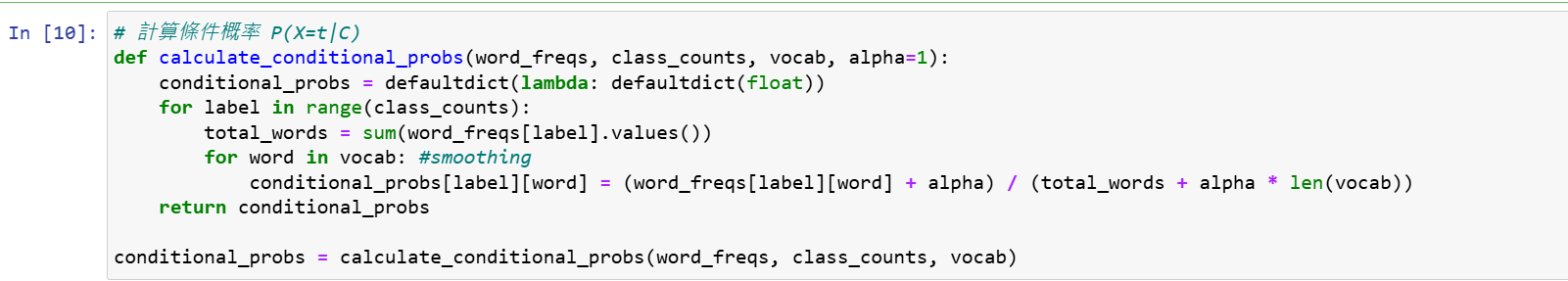


計算Naïve Bayes會用到的資料

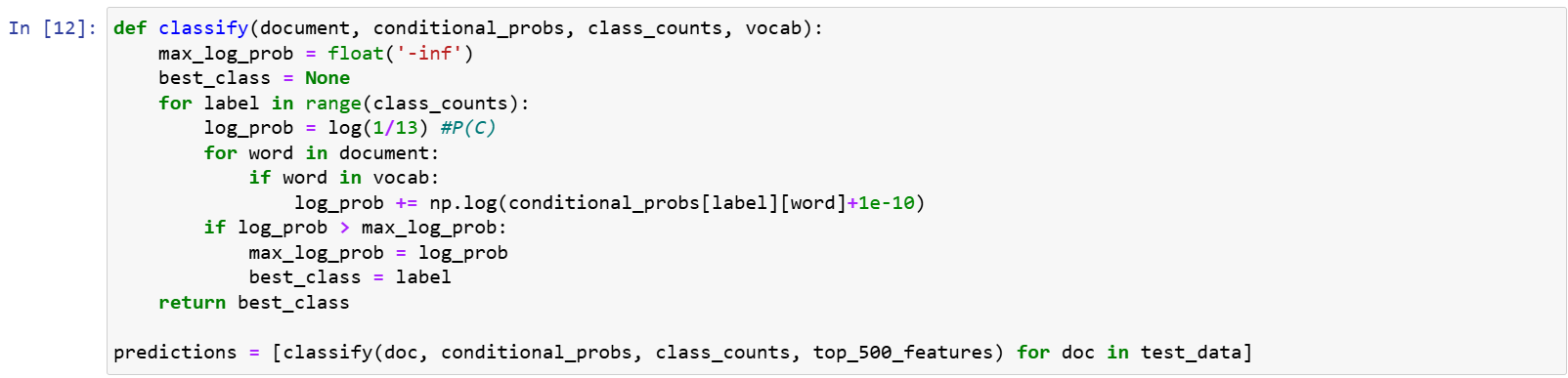


使用frequency-based method計算出前500個term，計算中使用hw2統計過的term frequency(在交出的作業已經將這段註解掉，但還是附上dictionary.txt)

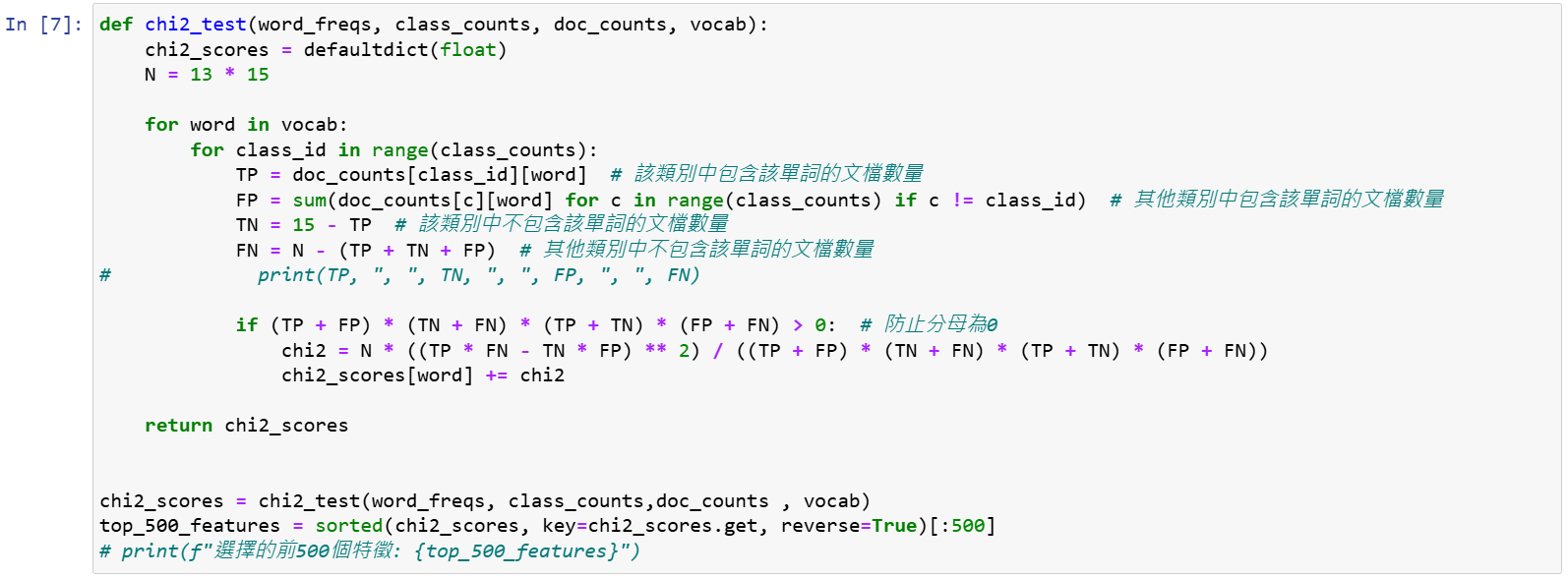
計算P(X=t|C)



使用NB計算P(C)\*，為了減少計算負擔先取log後相加



因為上次已經計算過term frequency，因此使用frequency-based method計算出前500個term，但在開始前先試過一次不選擇500個term而是整個單字集都一起計算的，認為相較之下結果不是很好，因此重新使用chi-square計算



結果比整個資料集算出的模型還要更好