1. 執行環境

Visual Studio Code

1. 程式語言

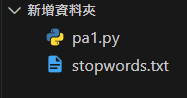
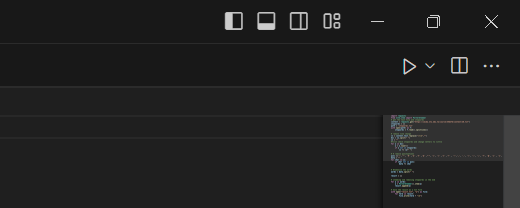
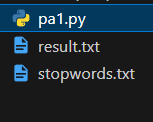
Python 3.10.6

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 requests, pip install nltk來安裝套件

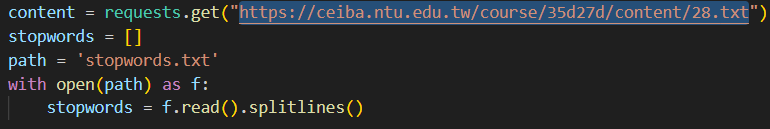
確認stopwords.txt與pa1.py在同個資料夾中，在vscode打開該資料夾點擊run即可執行(必須打開整個資料夾避免找不到stopwords.txt)

  可發現result.txt出現，點開result.txt即可看見結果

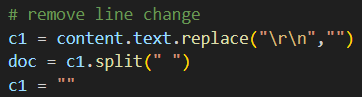
1. 邏輯說明
2. Package import: import requests and import PorterStemmer from nltk.stem



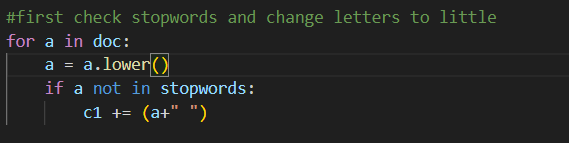
1. Read data: use requests.get() method to get the text file I need to process at <https://ceiba.ntu.edu.tw/course/35d27d/content/28.txt> and read the stopwords.txt file for stop words, the stop words I use are based on this website, <https://www.ranks.nl/stopwords> .



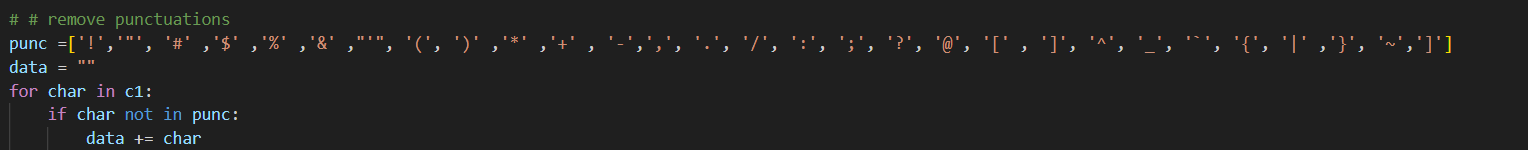
1. Data preprocessing: the data originally contained change line symbol, represented by \r\n, I first delete them and we split the sentence by white space.



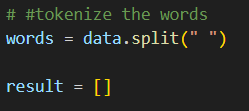
1. Lowercase and remove stopwords: use lower() method to change all letters to lowercase and check if they are in the stopwords. If they are, remove them, otherwise keep them. The reason I’d like to do stopwords removal first is that the following steps are punctuations removal and stemming. If I don’t do removal first, words like “ill” and “I’ll” will become non-distinguishable and words like” you’ve” may become “youv” after stemming, so I chose to do removal first



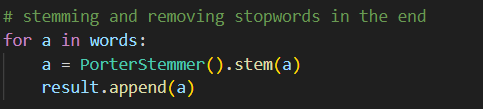
1. Delete punctuations: if we find punctuations, replace it with empty strings. String.puntuation is a good fit, but since I can’t import String, I just list them in a list to do the task.



1. Tokenization: since I now throw away all the punctuations and stop words, I can easily tokenize the words in the sentence by split with the white space.



1. Stemming using Porter’s algorithm: for every token, I stem it using Porter’s algorithm (by the imported function PorterStemmer).



1. Save the result as a txt file:

