1. 執行環境

Visual Studio Code

1. 程式語言

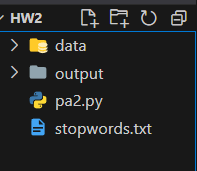
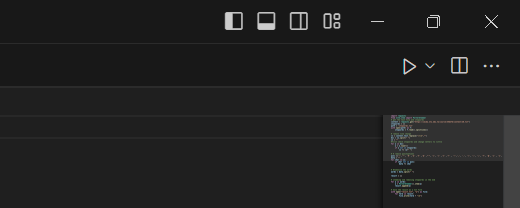
Python 3.10.6

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 numpy, 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)

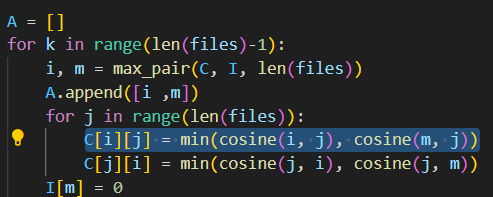
確認data資料夾(存解壓縮後的文字檔)， stopwords.txt與pa4.py在同個資料夾中，在vscode打開該資料夾點擊run即可執行(必須打開整個資料夾避免找不到stopwords.txt)

  執行完應會發現原本為空的8.txt, 13.txt, 20.txt.

1. 邏輯說明

直接複製貼上pa2.py，已經定義了cosion()函式來計算similarity，對每一對document pair算出similarity。

接著進行n-1次的iteration，每一次利用max\_pair()來選出最大similarity的cluster，在進行合併、更新。使用complete link，所以更新時取最小值，再把得到的pair存入A。



接著輸出分成20、13、8個cluster的結果。diction代表clustering的結果，一開始預設每篇文章都是cluster，從A裡面取得每一次合併的編號，按照編號合併diction裡的cluster，直到cluster數 = 20、13、8。