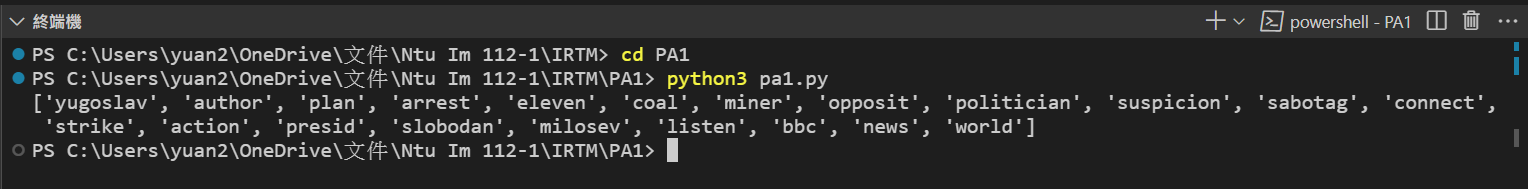
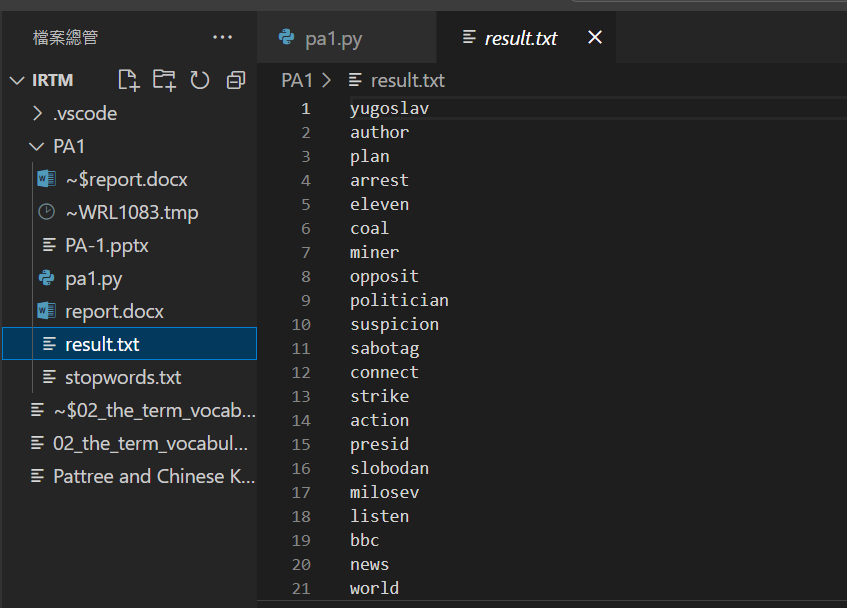
Introduction to Information Retrieval and Text Mining

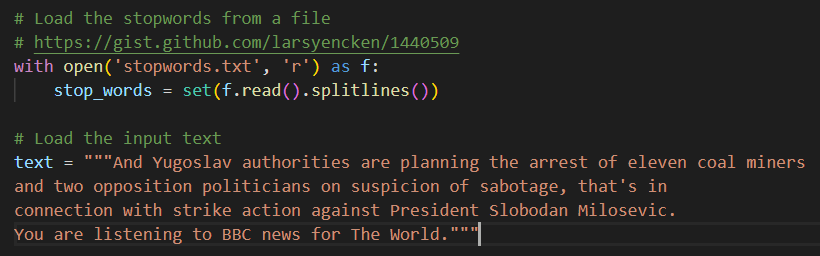
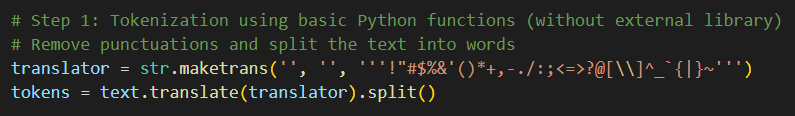
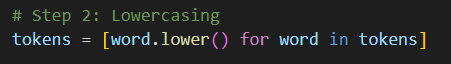
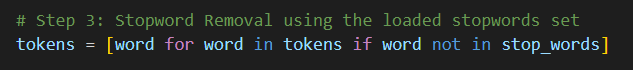
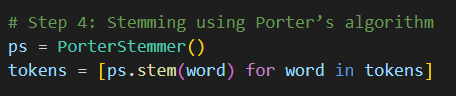
**Programming Assignment 1**  R12725026 秦孝媛

1. **執行環境**：VS code
2. **程式語言**：python 3.10
3. **執行方式**：
   1. 使用以下指令安裝 NLTK 套件

pip install nltk

* 1. 確保stopwords.txt在相同的目錄中，在命令提示字元（cmd）或終端機中，移動到包含 Python 檔案的目錄，並使用以下指令執行：python3 pa1.py
  2. 程式輸出結果  
     
  3. 會在同一目錄下產生一個名為 result.txt 的文件，裡面即為包含處理後的結果。



1. **作業處理邏輯說明**：
   1. 載入stopwords和input text資料  
      首先從網路上查詢的stopwords.txt文件中（參考自：<https://gist.github.com/larsyencken/1440509> ）讀取stopwords，並將其儲存為一個set。接著，定義作業中指定要處理的 text collection。  
      
   2. Tokenization  
      移除標點符號，接著用 split() 的方式進行 tokenization。  
      
   3. Lowercasing  
      將所有單詞轉換為小寫。  
      
   4. Stopwords Removal  
      只保留不在停用詞列表中的單詞。  
      
   5. Stemming  
      我們使用NLTK library中的PorterStemmer來對每個單詞進行 stemming。  
      
   6. Result  
      將處理後的 token list 保存到一個名為result.txt的文件中，每行一個 token。  
      