1. 執行環境: VS code
2. 程式語言: Python 3.11.5
3. 執行方式:

   Step1. 將 source.txt 與 stopword.txt 放在名為 IR 的資料夾中

   Step2. 在 powershell 中輸入 pip install nltk，以安裝 nltk 套件

```
PS E:\研究所\code> pip install nltk
```

4. 作業處理邏輯

   Step1. 讀入檔案，命名為 df

```
#read data
with open('.\IR\source.txt','r+') as file:
    df=file.read()
```

   Step2. 將檔案中的大寫字母轉為小寫

```
df=df.lower() #turn uppercase to lowercase
```

   Step3. 建立名為 token 的空 list(後續若指這個 list，會在 token 後面標註)與這時存放字串的 sub 變數。使用 loop 逐一跑過 df 中每個字元，以空格為分割基礎，若字串中含有標點符號或是\n 等描述文件格式的內容就忽略。最後將字元放入 token(list)當中，token(list)的每個成員就為 tokenized token。

```
#tokenization
token = [] #store token
sub="" # store temporary str
for i in df:
    if i !=" ":
        if i in (",",".","'"):
            continue
        sub+=i
        if sub == "\n": #\n 是txt的文字格式，在處理token時就刪除
            sub=""
            continue
    else:
        token.append(sub)
        sub=""
if sub :
    token.append(sub)
```

   Step4. 讀進 stopword.txt，此檔案中存放 stop word。利用 split()函式將 stopword.txt 中的字串分割成 token，存放在名為 sw 的 list 中。若 token(list)中的成員在 sw(list)中有出現，就刪除此成員。

```
#read stop word file
with open('.\IR\stopword.txt','r+') as file_s:
    sw=file_s.read()
sw=sw.split()
for i in range(len(token)-1,-1,-1): #從尾巴檢視是否含有stop word
    if token[i] in sw:
        token.pop(i)
```

Step5. 使用 nltk 套件執行 porter's algorithm，利用 loop 將每個 token 中的成員轉換成單一形式

```
#poter's algorithm
import nltk
from nltk import PorterStemmer
ps=PorterStemmer()
for i in range(len(token)-1,-1,-1):
    token[i]=ps.stem(token[i])
```

Step6. 在相同資料夾中建立一個名為 result.txt 的檔案，並逐一寫入已經清理好的 token(list)內容。

```
#輸出檔案
f=open(".\IR\\result.txt","w")
for i in range(0,len(token)):
    f.write(token[i])
    f.write(" ")
f.close()
```