

1. 執行環境：Visual Studio Code

2. 程式語言：Python 3.8

3. 執行方式：

A. Pip install nltk 以便於接下來使用 Porter's algorithm。

B. 輸入本次作業之文本

4. 作業處理邏輯說明：

根據作業說明所提供之指示，本次作業主要分為五個步驟，以下將依序說明。

A. Tokenization

首先將資料讀進程式中，接著用空格來分割每個詞成為 token。

```
# Tokenization
token = []
for i in range(4):
    a = input().split(' ')
    for j in range(len(a)):
        token.append(a[j])

for i in range(len(token)):
    try:
        token.remove('')
    except:
        break
```

B. Lowercasing everything

在將 token 變為小寫前，我先對我的資料做了兩件事情，第一為刪除 list 中重複的 element，第二為刪除部分 element 存在標點符號的問題。完成上述兩步驟最後再小寫化所有的 token。

```

# remove duplicates
def my_function(x):
    return list(dict.fromkeys(x))

token = my_function(token)

# remove punctuations
token = [i.strip(".",")" for i in token]

for i in range(len(token)):
    if "" in token[i]:
        p = token[i].find("")
        token[i] = token[i][:p] + token[i][p+1:]

# Lowercasing everything
token = [i.lower() for i in token]

```

C. Stemming using Porter's algorithm

由於先前已 import 完成，因此即可使用套件來執行 Porter's algorithm。

```
from nltk.stem import PorterStemmer
```

```

# Stemming using Porter's algorithm.
ps = PorterStemmer()
pstoken = []

for w in token:
    pstoken.append(ps.stem(w))

```

D. Stopword removal

由於作業要求不能引用套件，因此我直接參考網路上寫好的 list 在自己的程式碼建一個 list (由於 list 過長因此僅截圖部分)

```

# Create a stop words list and eliminate them
stopwordlist = ["i", "me", "my", "myself", "we", "our", "ours"]
for i in range(len(stopwordlist)):
    for j in range(len(pstoken)):
        try:
            if stopwordlist[i] == pstoken[j]:
                pstoken.remove(pstoken[j])
        except:
            continue

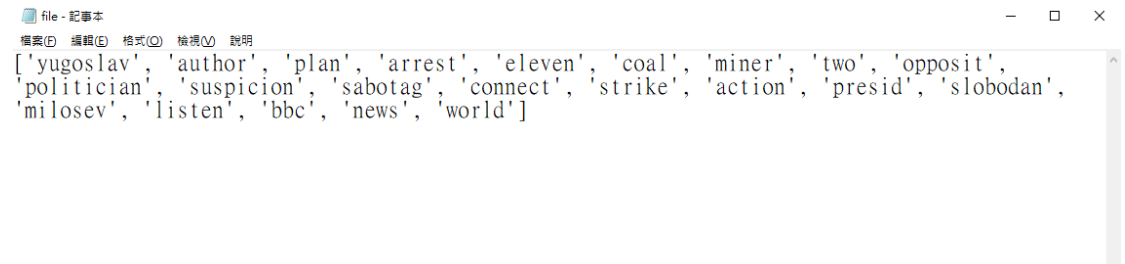
```

E. Save the result as a txt file

完成上述的步驟即可輸出為一個 txt 檔。

```
with open("result.txt", "w") as output:  
    output.write(str(pstoken))
```

以下為最後結果。



A screenshot of a Notepad window titled "file - 記事本". The window contains a list of words in single quotes, separated by commas: ['yugoslav', 'author', 'plan', 'arrest', 'eleven', 'coal', 'miner', 'two', 'opposit', 'politician', 'suspicion', 'sabotag', 'connect', 'strike', 'action', 'presid', 'slobodan', 'milosev', 'listen', 'bbc', 'news', 'world']