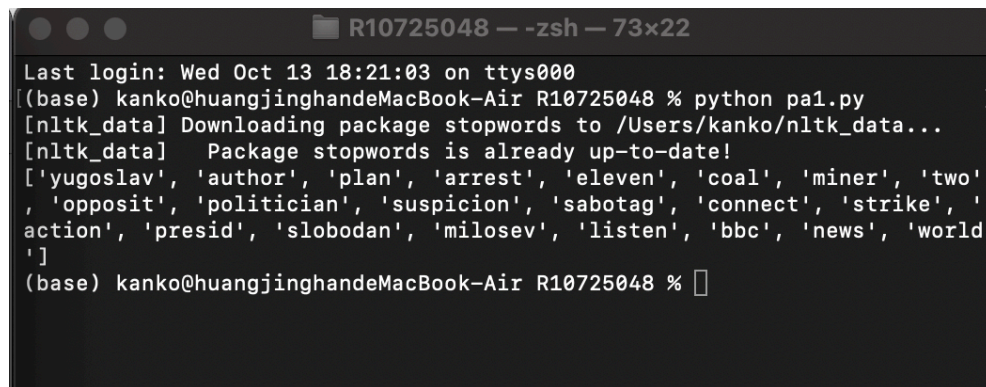
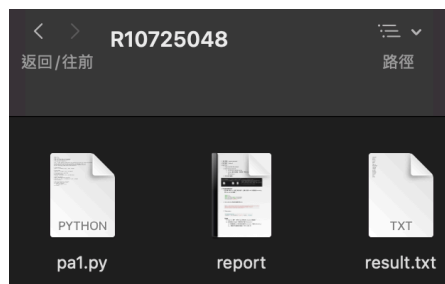


1. 執行環境：Jupyter Notebook
2. 程式語言：Python 3
3. 執行方式：
 1. \$ pip install nltk (for python)
 2. Open 'pa1.py' and run
 - on terminal: \$ python pa1.py
 - Output: 會印出最後一個步驟（移除 stop words）的結果，並輸出成 result.txt。
 3. 執行成功畫面：



```
R10725048 — -zsh — 73x22
Last login: Wed Oct 13 18:21:03 on ttys000
[(base) kanko@huangjinghandeMacBook-Air R10725048 % python pa1.py
[nltk_data] Downloading package stopwords to /Users/kanko/nltk_data...
[nltk_data] Package stopwords is already up-to-date!
['yugoslav', 'author', 'plan', 'arrest', 'eleven', 'coal', 'miner', 'two',
 'opposit', 'politician', 'suspicion', 'sabotag', 'connect', 'strike',
 'action', 'presid', 'slobodan', 'milosev', 'listen', 'bbc', 'news', 'world']
(base) kanko@huangjinghandeMacBook-Air R10725048 %
```



4. 作業處理邏輯說明：
 1. 載入套件
 - Import re 以做正則表達式
 - Import nltk 以便後續 stemming 與 stop words 的處理
 - Import request 以爬取文字

```
import re
import nltk
import requests
from nltk.stem import PorterStemmer
from nltk.corpus import stopwords
nltk.download('stopwords')
```

2. 透過 request.get 將 text collection 存為 text

```
res = requests.get('https://ceiba.ntu.edu.tw/course/35d27d/content/28.txt')
res.encoding = 'utf-8'
text = res.text
text
```

```
"And Yugoslav authorities are planning the arrest of eleven coal miners \r\nand two opposition politicians on suspici
on of sabotage, that's in \r\nconnection with strike action against President Slobodan Milosevic. \r\nYou are listeni
ng to BBC news for The World."
```

3. Tokenization

```
# Tokenization
tokens = re.findall("[\w]+", text) #regex
```

Notes

- A. import re 套件，使用 findall 尋找字 (character) 做斷詞。
- B. 正則表達式 “[\w]+” 用來尋找多個字 (character)：
 - a. 中括號裡的 \w 代表匹配任意字 (character)，等同 [A-Za-z0-9_]
 - b. + 號則表示會尋找匹配前一字元 1 到多次

4. Lowercasing

```
# Lowercasing
lowercase = [x.lower() for x in tokens]
```

Notes

- A. 透過 Python 內建的 lower() 將 tokens 中每個 items 轉換為小寫
- B. 相當於：

```
for i in range(len(tokens)):
    tokens[i] = tokens[i].lower()
```

5. Stemming (Porter's algorithm)

```
# Stemming_Porter's algorithm
ps = PorterStemmer()
stemming = [ps.stem(i) for i in lowercase]
```

Notes

透過 nltk 的 PorterStemmer 來處理 stemming

6. Stopwords removal

```
# Stopword removal
stops = stopwords.words('english')
filtered = [w for w in stemming if w not in stops]
print(filtered)

['yugoslav', 'author', 'plan', 'arrest', 'eleven', 'coal', 'miner', 'two', 'opposit', 'politician', 'suspicion', 'sabotage', 'connect', 'strike', 'action', 'presid', 'slobodan', 'milosev', 'listen', 'bbc', 'news', 'world']
```

Notes

- A. 透過 nltk 移除 stopwords
- B. 建立新的 list ‘filtered’，將 stemming 中不屬於 stop words 的詞存至 filtered
- C. 印出 filtered

7. Save the result as a txt file

```
# Save the result as a txt file
txt = open("result.txt", "w")
for item in filtered:
    txt.write(item + "\n")
txt.close()
```

Notes

- A. 以 write 方式開啟 'result.txt'
- B. for loop: 將過濾完 stop words 的 list 'filtered' 以一詞一行的方式存至 'result.txt'

*各階段處理比較 (此為處理時的截圖，最後僅保留 filtered 輸出)

tokens	lowercase	stemming	filtered
['And', 'Yugoslav', 'authorities', 'are', 'planning', 'the', 'arrest', 'of', 'eleven', 'coal', 'miners', 'and', 'two', 'opposition', 'politicians', 'on', 'suspicion', 'of', 'sabotage', 'that', 's', 'in', 'connection', 'with', 'strike', 'action', 'against', 'President', 'Slobodan', 'Milosevic', 'You', 'are', 'listening', 'to', 'BBC', 'news', 'for', 'The', 'World']	['and', 'yugoslav', 'authorities', 'are', 'planning', 'the', 'arrest', 'of', 'eleven', 'coal', 'miners', 'and', 'two', 'opposition', 'politicians', 'on', 'suspicion', 'of', 'sabotage', 'that', 's', 'in', 'connection', 'with', 'strike', 'action', 'against', 'president', 'slobodan', 'milosevic', 'you', 'are', 'listening', 'to', 'bbc', 'news', 'for', 'the', 'world']	['and', 'yugoslav', 'author', 'are', 'plan', 'the', 'arrest', 'of', 'eleven', 'coal', 'miner', 'and', 'two', 'opposit', 'politician', 'on', 'suspicion', 'of', 'sabotag', 'that', 's', 'in', 'connect', 'with', 'strike', 'action', 'against', 'presid', 'slobodan', 'milosev', 'you', 'are', 'listen', 'to', 'bbc', 'news', 'for', 'the', 'world']	['yugoslav', 'author', 'plan', 'arrest', 'eleven', 'coal', 'miner', 'two', 'opposit', 'politician', 'suspicion', 'sabotag', 'connect', 'strike', 'action', 'presid', 'slobodan', 'milosev', 'listen', 'bbc', 'news', 'world']