

作業一 李妮燁

1. 執行環境 & 作業系統

本次採用的作業系統為 win 10 64 位元，執行環境為 Visual Studio Code，需要預先安裝 [Natural Language Toolkit tool](#)，他是一個 Python library，常用在 Textming 的工作。首先要先在電腦用 cmd 安裝 NLTK 的工具包，用以下指令：

```
pip install nltk
```

再來是在 IDLE 安裝處理資料要用的 library 或其他可以用來測試工具的文件集。

```
import nltk  
nltk.download()
```

2. 程式語言, 版本

使用的程式語言和版本為 python 3.7。

3. 執行方式

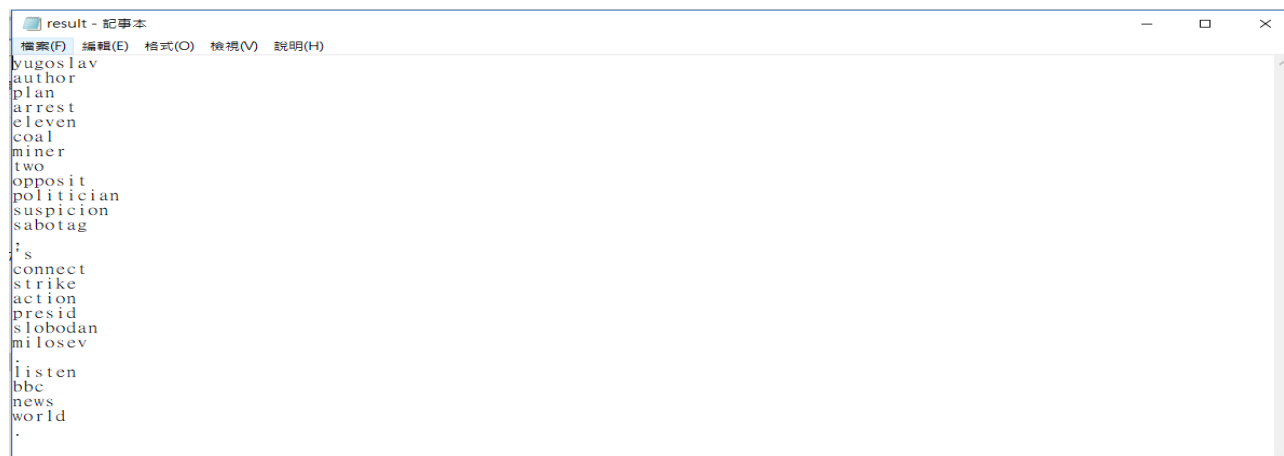
先將要處理的文字存成 **HW1.txt** 檔案並且和 python source code 存放在同一個資料夾，使用 Visual Studio Code 編寫完後，cmd 執行 python 程式。

● 執行畫面



```
D:\download\hw207371_r07725016_5c3a282f5121639_2>python textming_hw1.py  
['yugoslav', 'author', 'plan', 'arrest', 'eleven', 'coal', 'miner', 'two', 'opposit', 'politi  
cian', 'suspicion', 'sabotag', 's', 's', 'connect', 'strike', 'action', 'presid', 'slobodan'  
, 'milosev', '.', 'listen', 'bbc', 'news', 'world', '.']  
  
D:\download\hw207371_r07725016_5c3a282f5121639_2>
```

HW1	2018/9/21 上午 0...	文字文件	1 KB
result	2018/9/28 上午 0...	文字文件	1 KB
textming_hw1	2018/9/27 上午 1...	Python 來源檔案	1 KB
作業一	2018/9/27 上午 1...	Microsoft Word	22 KB
作業一	2018/9/27 上午 1...	PDF Document	194 KB



```
result - 記事本  
yugoslav  
author  
plan  
arrest  
eleven  
coal  
miner  
two  
opposit  
politician  
suspicion  
sabotag  
s  
connect  
strike  
action  
presid  
slobodan  
milosev  
listen  
bbc  
news  
world  
.
```

作業一 李妮燁

4. 作業處理邏輯說明

首先要從 nltk tool import 多項套件，例如：

```
from nltk.stem import PorterStemmer
from nltk.tokenize import WordPunctTokenizer
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
```

- Tokenize

先讀取文字檔，這裡編碼如果採用 utf-8 編碼，執行時會出現“\ufeff”字節順序標記，所以需要採用替代編碼方案“utf-8-sig”，再將檔案內的文字 tokenize，語言設定為英文。

- Lowercase

將 tokenize 後的文字轉換為全部小寫。

- Stemming

使用 nltk 套件中的 PorterStemmer() 對文字進行 Stemming。

- Stopwords

Set stopwords 語言為英文，將 Stemming 後的文字以迴圈的方式讀取，如果遇到 stopwords 將不存入新的 result document。

大部分的 tokenize 的過程採用三元條件運算的概念，簡潔 code 繁雜度，最後再建立新的文字檔，將處理過後的文字檔以迴圈、換行的方式存入新的結果檔。

5. 任何在此作業中的心得

本次作業在準備編寫的時候，我首先遇到的是**環境問題**，nltk 工具安裝過程中，找網路資料時發現 nltk 工具包含很多大型檔案或可供分析的文件，因此如果只有特定分析需求，只要針對特定細項下載來減少架設環境的時間；再來是**編碼問題**，從一開始的文字檔就必須存成 utf-8 的格式，到後來的新檔案也一樣需要注意，否則就會出現相關 bug，在編寫的過程中也查了不少編碼相關的問題與資料，對於編碼也算是有進一步的了解。除了上述作業需要繳交的功能之外，我也嘗試過去**除逗號、非字母的文字處理**，整體來說可以讓文字變得更加簡潔和清楚。

6. 參考資料

[編碼去除“\ufeff”參考解決方法](#)

[NLTK 工具安裝相關問題](#)

[Porter algorithm stemming 使用說明](#)