R10725011 陳佑甄

作業一

1. 執行環境

本機開 Anaconda，建立虛擬環境，下 python 去跑.py 檔。

用 Visual Studio Code 撰寫 python 程式碼。

2. 程式語言 (請標明版本)

Python 3.7.11

3. 執行方式

本機開啟 Anaconda Powershell Prompt，Anaconda 版本是 conda 4.10.3，

Anaconda Powershell Prompt (anaconda)

應用程式

- 創建虛擬環境：**(以下我有先創虛擬環境，但也可以不創建)**

```
$conda create --name IR python=3.7
```

下`$conda env list`可以看到剛剛創建的虛擬環境。

```
(base) PS C:\Users\ > conda env list
# conda environments:
#
base                  *  D:\anaconda
AIHW1                    D:\anaconda\envs\AIHW1
IR                       D:\anaconda\envs\IR
pyhon3.7                 D:\anaconda\envs\pyhon3.7
```

用`$ conda activate IR`可以啟動剛剛創好的虛擬環境。

- 安裝 nltk 套件：**(若沒有 nltk 套件一定需要 install)**

然後由於有用到 nltk 套件抓 stopword list ，

所以要下`$conda install -c anaconda nltk`。

然後可以透過`$conda list`看目前環境中載的所有套件。

其他套件是在 create 虛擬環境時有問要不要安裝，我當時選 yes。

```
(IR) PS D:\jupyter> conda list
# packages in environment at D:\anaconda\envs\IR:
#
# Name                    Version                   Build  Channel
ca-certificates           2020.10.14                    0    anaconda
certifi                   2020.6.20                py37_0    anaconda
click                     7.1.2                      py_0    anaconda
joblib                    0.17.0                     py_0    anaconda
nltk                      3.5                        py_0    anaconda
openssl                   1.1.1                 h2bbff1b_0
pip                       21.2.4               py37haa95532_0
python                    3.7.11                h6244533_0
regex                     2020.10.15           py37he774522_0    anaconda
setuptools                58.0.4               py37haa95532_0
sqlite                    3.36.0                h2bbff1b_0
tqdm                      4.50.2                     py_0    anaconda
vc                        14.2                  h21ff451_1
vs2015_runtime            14.27.29016           h5e58377_2
wheel                     0.37.0               pyhd3eb1b0_1
wincertstore              0.2                  py37haa95532_2
(IR) PS D:\jupyter>
```

- 執行程式：

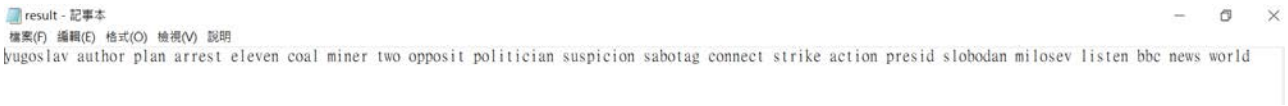
然後去到放 pa1.py 的資料夾，`$python .\pa1.py` 就可以執行程式碼，結果會如下圖，並在相同資料夾下產生一個 result.txt 檔。

```
(IR) PS D:\jupyter\IR_HW> python .\pa1.py
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Annie\AppData\Roaming\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 slob
odan milosev listen bbc news world
(IR) PS D:\jupyter\IR_HW> ls


    目錄: D:\jupyter\IR_HW


Mode                 LastWriteTime         Length Name
----                 -------------         ------ ----
-a----        2021/10/13  下午 11:10           1010 pa1.py
-a----        2021/10/13  下午 11:12            155 result.txt
```

result.txt 點開如下圖。



4. 作業處理邏輯說明

- import

包含用來抓取 url 資料的 urllib.request，用來抓 stopword list 的 nltk 還有用來做 Porter's algorithm 的 PorterStemmer。

```
import urllib.request
import nltk
from nltk.stem import PorterStemmer
```

- 抓取資料

利用 urllib.request.urlopen("url").read()從網頁上讀取 data 並存給變數 contents。抓下來是 bytes 的型態，用 str()並以 utf-8 的方式將型態轉成 string，並將結果給變數 text。

```
# Read text
contents =
urllib.request.urlopen("https://ceiba.ntu.edu.tw/course/35d27d/content/
28.txt").read()
text = str(contents,'utf-8')
```

- 拿掉標點符號與迴車換行

將標點符號給 punc，用 for 迴圈去跑 text，當 text 中有 element 是在 punc 中的標點符號就會被 replace 成空的。然後再同時將\r\n 一起拿掉。

```
# Remove punctuations in text
punc = '''!()-[]{};:'"\,<>./?@#$%^&*_~'''

for e in text:
  if e in punc:
    text = text.replace(e, '')
# Remove \r\n
text = text.replace('\r','').replace('\n','')
```

- Lowercasing

用.lower()將大寫換成小寫做到 Lowercasing。

```
# Lowercasing
text = text.lower()
```

- Tokenization

用.split(' ')以空格作為分割的依據做到 Tokenization，並將結果存給 sptext 這個 list。

```
# Tokenization
sptext = text.split(' ')
```

- Stemming

用 for 迴圈跑每個 sptext 中的元素，並用 PorterStemmer()將元素做 stemming，並將結果 append 到 smtext 中。

```
# Stemming using Porter's algorithm
ps = PorterStemmer()
smtext = []
for t in sptext:
  smtext.append(ps.stem(t))
```

- Remove Stopwords

從 nltk 套件中 download stopwords，只把英文的 stopwords 存給變數 stop_words。用 for 迴圈去跑前面處理好的 smtext，若其中的元素不在 stop_words 中，就將它存到 result 的字串中，並加空格方便最後觀看。

若元素是 stopwords 將他存在 sw 中，方便檢查。

```
# Remove Stopwords
# Using stopwords list from nltk
nltk.download('stopwords')
stop_words = nltk.corpus.stopwords.words('english')
result = ''
sw = []
for t in smtext:
  if not t in stop_words:
    result = result + t + ' '
  else:
    sw.append(t)
print(result)
```

- 將 result 存成 result.txt

```
# Output result txt
with open('.\\result.txt','w') as f:
  f.write(result)
```