1. **執行環境**
   - Jupyter Notebook

2. **程式語言&版本**
   - 語言：Python
   - 版本：3.6.6

3. **執行方式**

打開 cmd，cd 到檔案位置

```
C:\Users\hp>cd C:\Users\hp\Desktop\IR\HW1_r07725044

C:\Users\hp\Desktop\IR\HW1_r07725044>
```

再輸入以下指令，即可執行並顯示 output：

```
C:\Users\hp\Desktop\IR\HW1_r07725044>C:\Users\hp\Anaconda3\python HW1_r07725044.py
```

藍色部分填入 python.exe 及其路徑，如下圖

紅色填入檔名 HW1_r07725044.py

得到 output：

```
Text:
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.
Tokenized Result:
['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']
Lowercase:
['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']
Stemming:
['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', 'milose
v', 'you', 'are', 'listen', 'to', 'bbc', 'news', 'for', 'the', 'world']
Result:
['yugoslav', 'author', 'plan', 'arrest', 'eleven', 'coal', 'miner', 'two', 'opposit', 'politician', 'suspicion', 'sabotag', 'connect',
 'strike', 'action', 'presid', 'slobodan', 'milosev', 'listen', 'bbc', 'news', 'world']
```

另一種執行方式：將腳本檔 HW1_r07725044.ipynb 放到 jupyter notebook 執行，在 kernel 按下 Restart&Run All

4. 作業處理邏輯說明

前置處理：爬蟲抓取文本

In [1]:
```
from bs4 import BeautifulSoup
import requests
from lxml import html
from bs4 import BeautifulSoup
headers = {'User-Agent' :'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/69.0.3497.100 S
```

In [2]:
```
url = 'https://ceiba.ntu.edu.tw/course/35d27d/content/28.txt'
txt = requests.get(url = url, headers = headers)
bs = BeautifulSoup(txt.text, 'lxml')
c_result = bs.select("p")
for s in c_result:
    string1 = s.text
string1
```

Out[2]:
```
"And Yugoslav authorities are planning the arrest of eleven coal miners \r\nand two opposition politicians on suspicion of sabo
tage, that's in \r\nconnection with strike action against President Slobodan Milosevic. \r\nYou are listening to BBC news for T
he World."
```

- Tokenization

  先將標點符號替換成 whitespace，便於 tokenize

  tokenize 有 2 種方法，一種是使用 nltk 套件，但需要先安裝（指令如下）；另一種則是用 split 方法，根據 whitespace 斷詞

  pip install nltk

  nltk.download('punkt')

In [3]:
```
import nltk
import string

re = string1.translate(str.maketrans(string.punctuation, ' '*len(string.punctuation))) #標點符號換成whitespace，便於tokenize

res = nltk.word_tokenize(re)  #套件的方法
#res = "".join([stri for stri in string1 ]).split(" ") #用whitespace做分割
result = list()
for r in res:
    if ',' not in r or '.' not in r:
        result.append(r)

print("Tokenized Result:")
print(result)
```

```
Tokenized Result:
['And', 'Yugoslav', 'authorities', 'are', 'planning', 'the', 'arrest', 'of', 'eleven', 'coal', 'miners', 'and', 'two', 'opposit
ion', 'politicians', 'on', 'suspicion', 'of', 'sabotage', 'that', 's', 'in', 'connection', 'with', 'strike', 'action', 'agains
t', 'President', 'Slobodan', 'Milosevic', 'You', 'are', 'listening', 'to', 'BBC', 'news', 'for', 'The', 'World']
```

本次作業採用套件的分割方法 nltk.word_tokenize()

- Lowercasing everything

  用迴圈將每個 element 做 lowercase，再存取到新的 list 中

In [4]:
```
new = list()  #用以存取lower case的結果
for r in result:
    rlow = r.lower()
    new.append(rlow)

print('Lowercase:')
print(new)
```

```
Lowercase:
['and', 'yugoslav', 'authorities', 'are', 'planning', 'the', 'arrest', 'of', 'eleven', 'coal', 'miners', 'and', 'two', 'opposit
ion', 'politicians', 'on', 'suspicion', 'of', 'sabotage', 'that', 's', 'in', 'connection', 'with', 'strike', 'action', 'agains
t', 'president', 'slobodan', 'milosevic', 'you', 'are', 'listening', 'to', 'bbc', 'news', 'for', 'the', 'world']
```

- Stemming using Porter’s algorithm
  import nltk 套件底下的方法，進行 stem，把字尾去除

```
In [5]: from nltk.stem.porter import PorterStemmer #import porter algorithm的套件
porter = PorterStemmer()  #定義方法
stemmer = [ porter.stem(element) for element in new]  #stemming

print('Stemming:')
print(stemmer)
```

```
Stemming:
['and', 'yugoslav', 'author', 'are', 'plan', 'the', 'arrest', 'of', 'eleven', 'coal', 'miner', 'and', 'two', 'opposit', 'politi
cian', 'on', 'suspicion', 'of', 'sabotag', 'that', 's', 'in', 'connect', 'with', 'strike', 'action', 'against', 'presid', 'slob
odan', 'milosev', 'you', 'are', 'listen', 'to', 'bbc', 'news', 'for', 'the', 'world']
```

- Stopword removal

```
LookupError:
**********************************************************************
  Resource stopwords not found.
  Please use the NLTK Downloader to obtain the resource:

  >>> import nltk
  >>> nltk.download('stopwords')

  Searched in:
    - 'C:\\Users\\hp/nltk_data'
    - 'C:\\nltk_data'
    - 'D:\\nltk_data'
    - 'E:\\nltk_data'
    - 'C:\\Program Files (x86)\\Microsoft Visual Studio\\Shared\\Anaconda3_64\\nltk_data'
    - 'C:\\Program Files (x86)\\Microsoft Visual Studio\\Shared\\Anaconda3_64\\share\\nltk_data'
    - 'C:\\Program Files (x86)\\Microsoft Visual Studio\\Shared\\Anaconda3_64\\lib\\nltk_data'
    - 'C:\\Users\\hp\\AppData\\Roaming\\nltk_data'
**********************************************************************
```

先使用 nltk.download()載入 stopwords，否則可能會上面的 error

```
In [6]: from nltk.corpus import stopwords
stop = set(stopwords.words('english'))
final = []
for s in stemmer:
    if s not in stop:
        final.append(s)

print('Result：')
print(final)
```

```
Result：
['yugoslav', 'author', 'plan', 'arrest', 'eleven', 'coal', 'miner', 'two', 'opposit', 'politician', 'suspicion', 'sabotag', 'co
nnect', 'strike', 'action', 'presid', 'slobodan', 'milosev', 'listen', 'bbc', 'news', 'world']
```

將前面 stem 完剩下的詞和 stopwords 做比對，若不在 stopwords 中，則視爲結果

- Save the result as a txt file
  建立新檔案 result.txt
  用迴圈將 list 中每個項目寫入檔案，且在每個項目後面加上\n 換行
  全部寫入後將檔案關閉
  在 py 檔的同個目錄下可找到 result.txt

```
In [8]: f = open('result.txt', 'w') #建立新檔案，名爲result.txt
for element in final:
    f.write(element + '\n') #寫入list的每個項目，並換行
f.close() #寫完，關閉檔案儲存
```