

# Programming Assignment 1

---

- Write a program to extract terms from a document.
  - You have to do:
    - Tokenization.
    - Lowercasing everything.
    - Stemming using Porter's algorithm.
    - Stopword removal.
    - Save the result as a txt file.
  - Text collection:
    - One English news document  
(<https://ceiba.ntu.edu.tw/course/35d27d/content/28.txt> ).
  - **Please zip and submit <sup>1</sup>the result of document 1, <sup>2</sup>the source code, and <sup>3</sup>a report to TA.**
    - 2 weeks to complete, that is **2023/9/25**.



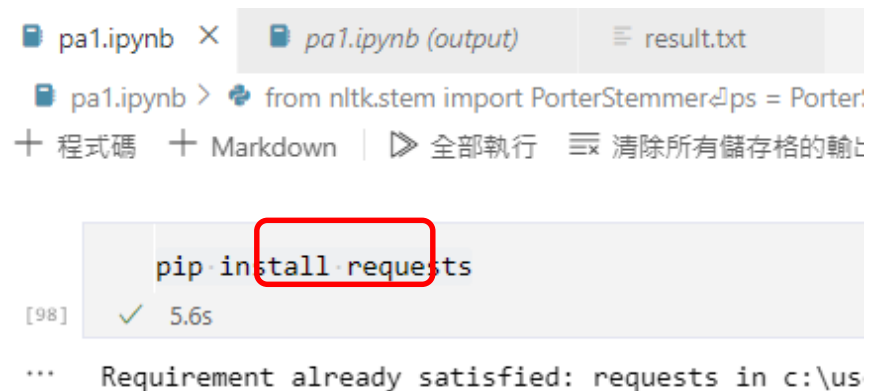
# Report

---

執行環境: Visual Studio Code  
程式語言: Python 3.9.5

# 執行方式

- 使用VS code跑pa1.ipynb檔
- 需要下載的套件有:
  - `pip install requests`: 讀text file所需
  - `pip install nltk`: 使用Porter's algorithm. 所需
- 不需要另外再輸入別的指令，需要下載套件的指令都有打在notebook裡
- 直接按全部執行即可
- Result:



The screenshot shows a Jupyter Notebook with a tab labeled 'pa1.ipynb (output)'. Below the tab, there is a code cell with the command `pip install requests` highlighted by a red box. The output of the cell shows a green checkmark, the time '5.6s', and the message 'Requirement already satisfied: requests in c:\us'.

# 處理邏輯

---

## □ STEP1: read text file

使用requests套件得到English news document的文字並存入str

```
import requests
```

```
file_url = 'https://ceiba.ntu.edu.tw/course/35d27d/content/28.txt'  
response = requests.get(file_url)
```

```
str = response.text  
str
```

✓ 0.7s

MagicPyth

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

# 處理邏輯

## □ STEP2:Tokenization and Lowercasing

將一些換行符號、逗點等等以replace function先清除掉  
再將剩下的字元全部轉換成小寫  
最後用split的方式做tokenization並存入token中

```
str = str.replace(',', '').replace('\n', '').replace('\r', '').replace('.', '').replace('"s', "")
str = str.lower()
token = str.split(' ')
token
```

✓ 0.4s

MagicPython

Output exceeds the [size limit](#). Open the full output data [in a text editor](#)

```
['and',
 'yugoslav',
 'authorities',
 'are',
 'planning',
 'the',
 'arrest',
 'of'.
```

# 處理邏輯

## □ STEP3: Stemming using Porter's algorithm

利用PorterStemmer套件進行stemming  
並將處理完的token放入stemming\_word

```
● from nltk.stem import PorterStemmer  
  ps = PorterStemmer()  
  stemming_word = []  
  for word in token:  
      stemming_word.append(ps.stem(word))
```

✓ 0.1s

# 處理邏輯

---

## □ STEP4: Stopword removal

將stemming\_word中的stopword過濾掉

```
stopwords = ['and', 'are', 'the', 'of', 'on', 'of', 'in', 'that', 'with', 'for', 'to']  
filtered_words = [word for word in stemming_word if word not in stopwords]
```

✓ 0.1s

# 處理邏輯

## □ STEP5: Save the result as a txt file

最後以join的方式將處理乾淨的token串在一起  
並寫入至result.txt

```
result = ' '.join(filtered_words)
result
```

✓ 0.1s

MagicPython

```
'yugoslav author plan arrest eleven coal miner two opposit politician suspicion sabotag connect strike action against presid slobodan milosev
you listen bbc news world'
```

```
path = 'result.txt'
f = open(path, 'w')
f.write(result)
```

✓ 0.1s

MagicPython