Information Retrieval and Text Mining: Homework #1

R08725008 資管碩二 周若涓

執行環境

- Jupyter Notebook

程式語言

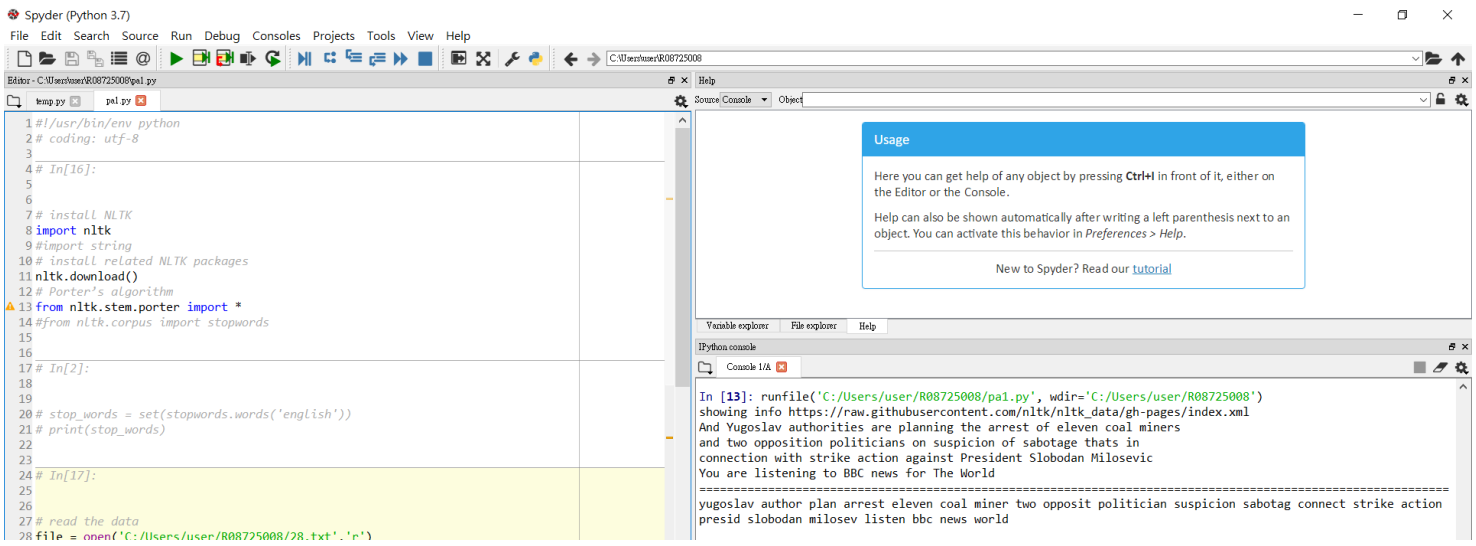
- Python 3

執行方式

- 在執行前需安裝 nltk 套件(command: pip install nltk)

- 以下說明 2 種執行環境:

  1. 可以利用 Spyder 開啟 pa1.py，並執行



2. 或可利用 python3 pa1.py 直接執行 python 檔案

```
(base) C:\Users\user>cd C:\Users\user\R08725008

(base) C:\Users\user\R08725008>python3 pa1.py

(base) C:\Users\user\R08725008>python pa1.py
showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml
And Yugoslav authorities are planning the arrest of eleven coal miners
and two opposition politicians on suspicion of sabotage thats in
connection with strike action against President Slobodan Milosevic
You are listening 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
```

- 確保提供的 28.txt 預設放於 C:/Users/user/R08725008/目錄下
- 產出的結果預設放於 C:/Users/user/R08725008/目錄下: result.txt

作業邏輯說明

1. 先將 28.txt 讀入
2. 並去除 punctuation

```
27  # read the data
28  file = open('C:/Users/user/R08725008/28.txt','r')
29  texts = file.read()
30  texts = texts.translate(str.maketrans('', '', '!\"#$%&'()*+,-./:;<=>?@[\]^_`{|}~"))
31  #print(string.punctuation)
32  #texts = file.read().replace(',', '').replace('.', '').replace(';', '').replace('\'s', '')
33  print(texts)
```

3. Tokenize

```
40  #Tokenization
41  word_tokenize = texts.split()
42  #print(word_tokenize)
```

4. 利用 nltk 套件初始化 PorterStemmer
5. 並宣告相關 stop words 集合(此集合的定義根據 NLTK 提供的 stop words 列表)

```
48  # Stemming using Porter's algorithm
49  ps = PorterStemmer()
50  # Stopword lists
51  stop_words = ['ourselves', 'hers', 'between', 'yourself', 'but', 'again', 'there', 'about', 'once', 'during', 'out', 'very', 'having', 'with', 'they', 'own',
52          'an', 'be', 'some', 'for', 'do', 'its', 'yours', 'such', 'into', 'of', 'most', 'itself', 'other', 'off', 'is', 's', 'am', 'or', 'who', 'as', 'from', 'him',
53          'each', 'the', 'themselves', 'until', 'below', 'are', 'we', 'these', 'your', 'his', 'through', 'don', 'nor', 'me', 'were', 'her', 'more',
54          'himself', 'this', 'down', 'should', 'our', 'their', 'while', 'above', 'both', 'up', 'to', 'ours', 'had', 'she', 'all', 'no', 'when', 'at',
55          'any', 'before', 'them', 'same', 'and', 'been', 'have', 'in', 'will', 'on', 'does', 'yourselves', 'then', 'that', 'because', 'what', 'over',
56          'why', 'so', 'can', 'did', 'not', 'now', 'under', 'he', 'you', 'herself', 'has', 'just', 'where', 'too', 'only', 'myself', 'which', 'those',
57          'i', 'after', 'few', 'whom', 't', 'being', 'if', 'theirs', 'my', 'against', 'a', 'by', 'doing', 'it', 'how', 'further', 'was', 'here', 'than', '.', ',', '"', "'",
58          '?', '!', ':', ';', '(', ')', '[', ']', '{', '}', '\'s','\'m','\'re','\'ll','\'d','n\'t','shan\'t','thats']
```

6. 將讀入的文件轉換為小寫，如果不在 stop words 當中的 word 才會保留
7. 最後將每個 token 進行 stemming，並加回字串當中

```
71  #Lowercasing everything
72  tokens = [i.lower() for i in word_tokenize if i.lower() not in stop_words]  #Stopword removal
73  #print(tokens)
74  token_result = ''
75  for i,token in enumerate(tokens):
76      if i != len(tokens)-1: # not leave empty in the end of file
77          token_result += ps.stem(token) + ' '
78      else:
79          token_result += ps.stem(token)
```

8. 把結果輸出寫入 result.txt

```
85  file = open('C:/Users/user/R08725008/result.txt','w')
86  # Save the result as a txt file
87  file.write(token_result)
88  file.close()
89  print(token_result)
```