IRTM HW1 Report

R06725035 資管碩二 陳廷易

執行環境

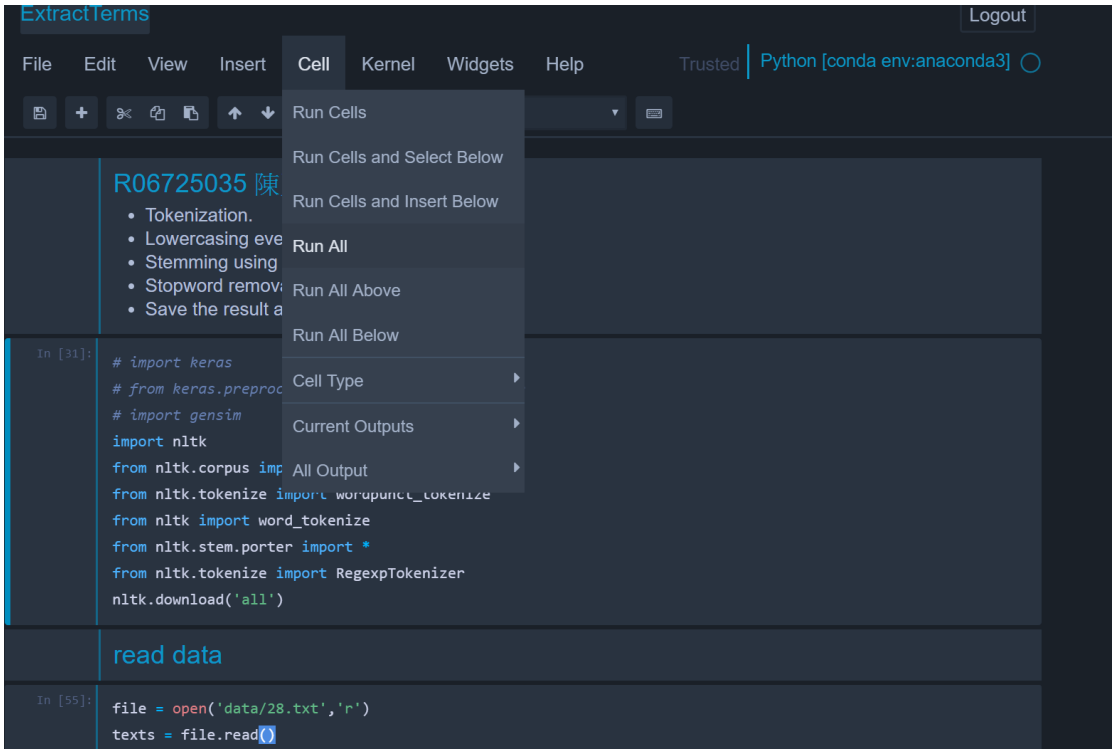
- Ubuntu 16.04
- Jupyter Notebook

程式語言

- Linux Anaconda Python 3.6

執行方式

- 提供 jupyter 版本以及一般 python 版本: *ExtractTerms.*
  - 可利用 jupyter 開啟 ipynb 以後執行全部



  - 或可利用 python3 ExtractTerms.py 直接執行 python 檔案

```
leoqaz12@SuperPC3:~/leoqaz12_python/IR/HW1$ python ExtractTerms.py
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.
yugoslav author plan arrest eleven coal miner two opposit politician suspicion sabotag connect strike
 action presid slobodan milosev listen bbc news world
```

- 需 pip install:
  - nltk 並 download data
- 確保提供的 28.txt 預設放於 data/目錄下
- 產出的結果預設放於 result/目錄下: output.txt

作業邏輯說明

1. 先將 28.txt 讀入

read data

```
In [55]:
file = open('data/28.txt','r')
texts = file.read()
texts
```

```
"And Yugoslav authorities are planning the arrest of eleven coal miners \nand two opposition politicians on suspicion of sabotage, tha
t's in \nconnection with strike action against President Slobodan Milosevic. \nYou are listening to BBC news for The World."
```

2. 利用 nltk 套件初始化 porterstemmer
3. 並宣告相關 stop words 集合
4. 將讀入的文件轉換為小寫，若沒在 stop words 當中的才會保留並進行 tokenize
5. 最後將每個 token 進行 stemming，並加回字串當中

main preprocessing

```
In [64]:
ps = PorterStemmer() # Stemming
stop_words = set(stopwords.words('english')) #Stopword
stop_words.update(['.', ',', '"', "'", '?', '!', ':', ';', '(', ')', '[', ']', '{', '}',
                   '\'s','\'m','\'re','\'ll','\'d','n\'t','shan\'t']) # remove it if you need punctuation


tokens = [i for i in word_tokenize(texts.lower()) if i not in stop_words]  # Tokenization.# Lowercasing
token_result = ''
for i,token in enumerate(tokens):
    if i != len(tokens)-1: # 最後不要空白
        token_result += ps.stem(token) + ' '
    else:
        token_result += ps.stem(token)
```

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

6. 再把最後結果輸出

Output

In [67]:

```
# output=""
# for token in tokens:
#     output+=token+' '
# print(output)
file = open('result/output.txt','w')
file.write(token_result) #Save the result
file.close()
print(token_result)
```

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