HW1 - Report

資管碩二 R10725046 江雨柔

作業一統整

本次作業主要是針對一段英文新聞稿做 斷詞與斷詞的整理。透過轉乘小寫，以 Porter’s algo 去除不一樣的字根，最後去除stopwords，來抓取此文章中的詞彙。

1. 環境

執行環境 : VScode

2. 程式語言：python 3.8.2

3. 執行方式

```
# 需要下載的環境
pip install gensim

# stopwords list
# 使用的是 princeton 提供的stopwords.txt

# 執行方式
# 可以透過 VS code 一格一格跑（run below）
# 或是直接按 Run 全部一起執行
```

4. 作業處理邏輯說明

1. Import Files

   將txt檔案裡面的文字使用 readlines() 一行一行讀取進來，變成一個list.

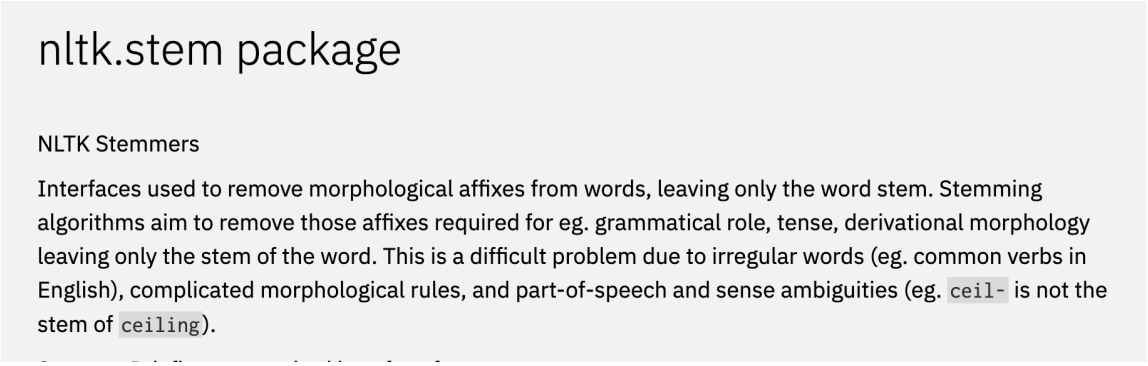
   ```
   ['And Yugoslav authorities are planning the arrest of eleven coal miners \n', "and two opposition politici
   ans on suspicion of sabotage, that's in \n", 'connection with strike action against President Slobodan Mil
   osevic. \n', 'You are listening to BBC news for The World.']
   ```

2. Lowercasing everything.

   因為讀取的時候是一行一行讀取，所以每一個句子都有換行符號。使用 string.lower()將每個句子換成小寫，並且去掉換行符號。

```
['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']
```

3. Stemming using Porter’s algorithm



使用的是 nltk 提供的 PorterStemming，因為裡面有提供將多個句子做 stemming 處理的函數，因此將doc_list 中的文字一個一個輸入去做 stemming，即可得。

```
['and yugoslav author ar plan the arrest of eleven coal miner',
 "and two opposit politician on suspicion of sabotage, that' in",
 'connect with strike action against presid slobodan milosevic.',
 'you ar listen to bbc new for the world']
```

4. Tokenization

整理過後，由於英文是由空白格來分割每個字，因此直接使用split來將句子做分割得到一個單詞的list。

```
[['and',
  'yugoslav',
  'author',
  'ar',
  'plan',
  'the',
  'arrest',
  'of',
  'eleven',
  'coal',
  'miner'],
 ['and',
  'two',
  'opposit',
  'politician',
  'on',
  'suspicion',
  'of',
  'sabotage,',
  "that'",
  'in'],
 ['connect',
  'with',
  'strike',
  'action',
  'against',
  'presid',
  'slobodan',
  'milosevic.'],
 ['you', 'ar', 'listen', 'to', 'bbc', 'new', 'for', 'the', 'world']]
```

5. Stopwords remove

使用的是 princeton 提供的stopwords list 來對 stemming 後的token 做過濾

來源 ：https://algs4.cs.princeton.edu/35applications/stopwords.txt

a. 清理詞彙

由上一個步驟的結果中可以看到，經過stemming之後，有些字的尾巴會有標點符號，因此還需要先把符號去除。有觀察到 ‘ . ’ ，‘ \’ ‘ 以及 ‘,’， 因此先將這兩個符號去掉之後。

b. stopwords removal

將剛剛處理好的tokens 比對 stopwords之後，若不在stopwords list 裡面就將該詞彙放置 final_token list裡面，同時也比對該詞彙是否已經存在在 final_token list，最後就會得到一個乾淨的 詞彙 list。