資訊檢索與文字探勘導論

作業一

會計四
B06702064
林聖硯

1. 執行環境

Anaconda Spyder

2. 程式語言

Python 3.6.8

3. 執行方式

(1) 需要 import nltk 套件，使用 pip install nltk

(2) 打開 py 檔並執行 89-91 行即可得到 result.txt

(txt 檔及 stopword 檔已經寫死在程式裡)

```
87
88  #需要執行的部分
89  text = get_text()
90  text_final = text_preprocessing(text)
91  save_result(text_final)
92
```

4. 作業處理邏輯說明

(0)import 套件

由於後面需要使用 PorterStemmer，故在此先從 nltk 導入

```
1  from nltk.stem.porter import PorterStemmer
2
```

(1)讀檔

將文本從 txt 檔讀入並回傳

```
3  def get_text():
4      '''
5      read content from txt
6      '''
7      f = open('content.txt')
8      text = []
9      for line in f.readlines():
10         text.append(line)
11     f.close()
12     return text
```

(2)切文本

將文本去除換行符號(使用 strip)，並且用空白符號加上 python 內建的 split 將字與字分開，最後再將一些簡寫先還原(如’s 變成 is 等等)

```
def tokenize_text(text):
    '''
    去除換行符號、切開並將一些字還原
    '''
    replace_dictionary = {"'s": " is", "'ve": " have", "'re": " are"}
    text_tokenized = []
    for sentence in text:
        sentence = sentence.strip()
        for key, value in replace_dictionary.items():
            sentence = sentence.replace(key, value)
        words = sentence.split(' ')
        for word in words:
            text_tokenized.append(word)
    return text_tokenized
```

(3)刪除標點符號、將子母縮成小寫

此處函數在將文本中的詞去除標點符號並且將全部字母縮成小寫

```
def delete_delimiters(text_tokenized):
    '''
    刪除標點符號
    '''
    delimiter = [',', "."]
    text_final = []
    for token in text_tokenized:
        word = ""
        for element in token:
            if element not in delimiter:
                word += element
        word = word.lower()
        text_final.append(word)
    return text_final

def lowercase_words(text_tokenized):
    '''
    將字母全部縮成小寫
    '''
    text_tokenized = [word.lower() for word in text_tokenized]
    return text_tokenized
```

(4)stemming

將显處裡的文本使用 nltk 中的 porter’s algorithm 套件做 stemming

```
def stemming_word(text_tokenized):
    '''
    Stemming using Porter's algorithm
    '''
    porter_stemmer = PorterStemmer()
    text_tokenized = [porter_stemmer.stem(word) for word in text_tokenized]
    return text_tokenized
```

(5)將停用字去除

```
def remove_stopwords(text_tokenized):
    '''
    Stopword removal
    '''
    f = open('stopwords.txt')
    stopwords = []
    for stopword in f.readlines():
        stopwords.append(stopword.strip())
    f.close()
    text_final = [word for word in text_tokenized if word not in stopwords]
    return text_final
```

(6)將上述所有的前處理步驟打包並且回傳最後的文本(型態是 list，每一個元素代表一個處理完的詞)

```
def text_preprocessing(text_tokenized):
    text_tokenized = tokenize_text(text)
    text_no_del = delete_delimiters(text_tokenized)
    text_lower = lowercase_words(text_no_del)
    text_stemmed = stemming_word(text_lower)
    text_final = remove_stopwords(text_stemmed)
    return text_final
```

(7)將處理好的文本存成 txt 檔

```
def save_result(text_final):
    '''
    Save the result as a txt file
    '''
    with open('result.txt', 'w') as f:
        for item in text_final:
            f.write("%s\n" % item)
```