Introduction to Information Retrieval and Text Mining

Programming Assignment 1

R12725026 秦孝媛

1. 執行環境：VS code
2. 程式語言：python 3.10
3. 執行方式：
   I. 使用以下指令安裝 NLTK 套件

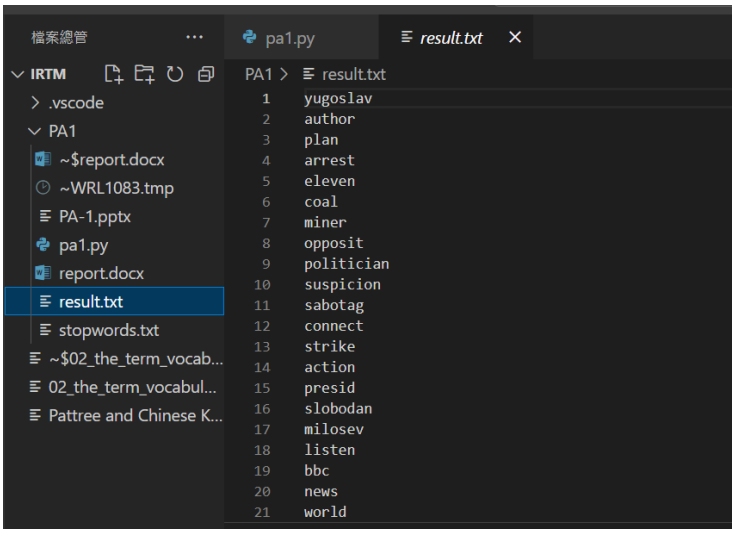
   ```
   pip install nltk
   ```

   II. 確保 stopwords.txt 在相同的目錄中，在命令提示字元（cmd）或終端機中，移動到包含 Python 檔案的目錄，並使用以下指令執行：

   ```
   python3 pa1.py
   ```

   III. 程式輸出結果

   IV. 會在同一目錄下產生一個名為 result.txt 的文件，裡面即為包含處理後的結果。



4. 作業處理邏輯說明：

I. 載入 stopwords 和 input text 資料

首先從網路上查詢的 stopwords.txt 文件中（參考自：

https://gist.github.com/larsyencken/1440509 ）讀取 stopwords，並

將其儲存為一個 set。接著，定義作業中指定要處理的 text

collection。

```
# Load the stopwords from a file
# https://gist.github.com/larsyencken/1440509
with open('stopwords.txt', 'r') as f:
    stop_words = set(f.read().splitlines())

# Load the input text
text = """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."""
```

II. Tokenization

移除標點符號，接著用 `split()` 的方式進行 tokenization。

```
# Step 1: Tokenization using basic Python functions (without external library)
# Remove punctuations and split the text into words
translator = str.maketrans('', '', '''!"#$%&'()*+,-./:;<=>?@[\\]^_`{|}~''')
tokens = text.translate(translator).split()
```

III. Lowercasing

將所有單詞轉換為小寫。

```
# Step 2: Lowercasing
tokens = [word.lower() for word in tokens]
```

IV. Stopwords Removal

只保留不在停用詞列表中的單詞。

```
# Step 3: Stopword Removal using the loaded stopwords set
tokens = [word for word in tokens if word not in stop_words]
```

V. Stemming

我們使用 NLTK library 中的 `PorterStemmer` 來對每個單詞進行

stemming。

```
# Step 4: Stemming using Porter's algorithm
ps = PorterStemmer()
tokens = [ps.stem(word) for word in tokens]
```

VI. Result

將處理後的 token list 保存到一個名為 result.txt 的文件中，每行一個 token。

```
# Step 5: Save the result as a txt file
with open('result.txt', 'w') as f:
    for token in tokens:
        f.write("%s\n" % token)

# Print out the tokens for verification
print(tokens)
```