資訊檢索語文字探勘導論 PA1

1. 執行環境：Jupyter Notebook
2. 程式語言：Python 3.10.9
3. 執行方式

   （1）在 terminal 使用 pip install nltk 和 re 套件

   （2）使用 Jupyter Notebook 執行 pa1.ipynb

4. 處理邏輯

   （1）讀入 1.txt 並進行前處理

```
with open("1.txt", "r") as r:
    f = r.read()
doc = f.replace("\n", "")
doc = re.sub(r"[^\w\s]", "", doc)
```

   第二行去除換行符號，第三行去除非英文字母和空格的元素

   （2）進行 tokenize 和 lowercase

```
tokenization = [word.lower() for word in doc.split(" ")]
```

   將文章以空格分割，並將所有字母轉換成小寫，存入 tokenization 陣列

   （3）用 Porter’s Algorithm 進行 stemming

```
from nltk.stem import PorterStemmer
ps = PorterStemmer()
stemming = [ps.stem(word) for word in tokenization]
```

   使用 nltk.stem 中的 PorterStemmer 對 tokenization 中的所有元素進行 stemming，並將結果依序存入 stemming 陣列

   （4）讀入 stopwords.txt 並用裡面的 stopwords list 進行 stopword removal

```
r = open("stopwords.txt")
stopwords = r.read()
result = [word for word in stemming if word not in stopwords]
```

   在 stopwords.txt 中定義所有 stopword，比對 stemming 中的所有元素，若不存在 stopwords.txt 中，存入 result 陣列

（5）將結果寫入 result.txt

```
with open("result.txt", "w") as f:
    for word in result:
        f.write(word + "\n")
```

將 result 陣列中的所有元素寫入 result.txt 中，即為最終結果