PA1 Report

r04725040 黃柏睿

說明：
使用 python 2.7.9，主程式為 pa1.py，執行結果存在 result.txt

一些心得：
一開始我直接把文檔做 tokenize, lowercase 跟 stemming 之後，然後拿去跟 stop word list 比較，以為世界就是這麼簡單美好

不過卻發現到有些標點符號會影響 stemming（’．，之類的）

```
➜ PA1  python pa1.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', 'actio
n', 'against', 'president', 'slobodan', 'milosevic.', 'you', 'are', 'listening', 'to', 'bbc', 'news', 'for', 'the', 'wor
ld.']
['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', 'pre
sid', 'slobodan', 'milosevic.', 'you', 'ar', 'listen', 'to', 'bbc', 'new', 'for', 'the', 'world.']
```

所以把他們移除

然後又發現有些字在 stemming 之後會跟 stop word 長得不一樣
ex: are —> ar （上面的 list 是 raw data，下面是做完處理的，可以發現 ar 在那邊沒被砍掉，因為 stop word list 裡面是 are

```
➜ PA1  python pa1.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', 'ar', 'plan', 'arrest', 'eleven', 'coal', 'miner', 'two', 'opposit', 'politician', 'suspicion', '
sabotage', "that'", 'connection', 'strike', 'action', 'president', 'slobodan', 'milosev', 'are', 'listen', 'bbc', 'new',
 'the', 'world']
```

但這樣有點不太舒服，因為 ar 明明就是 are，是個 stopword 卻能存活

所以我加了一個步驟，對 stop word list 也做 stemming，在來比對

```
➜ PA1  python pa1.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', 'act
ion', 'against', 'president', 'slobodan', 'milosevic', 'you', 'are', 'listening', 'to', 'bbc', 'news', 'for', 'the', 'wo
rld']
['yugoslav', 'author', 'plan', 'arrest', 'coal', 'miner', 'opposit', 'politician', 'suspicion', 'sabotag', 's', 'connect
', 'strike', 'action', 'presid', 'slobodan', 'milosev', 'listen', 'bbc', 'new', 'world']
```

這樣看起來好多了！