==========

Name： 李桂欽 ID：R04725050 Department：資訊管理 碩一 Homework： 1

==========

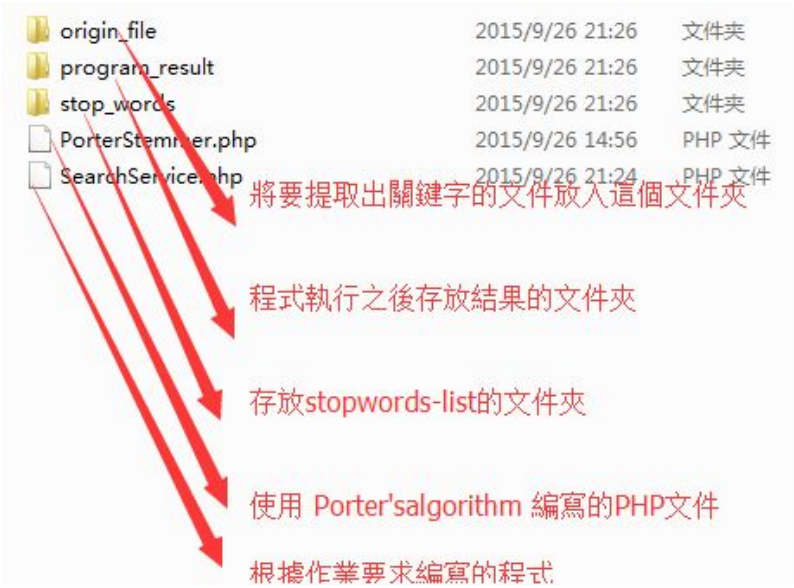
Programming Assignment 1:

You have to do:
Tokenization.
Lowercasing everything.
Stemming using Porter’s algorithm.
Stopword removal.
Save the result as a txt file.

My program result:

yugoslav
author
plan
arrest
coal
miner
opposit
politician
suspicion
sabotag
connect
strike
action
presid
slobodan
milosev
listen
bbc
news
world

My program organization:



My program work flow:

依次讀取指定檔夾下的所有檔案名

->

依次讀取指定檔夾下的所有檔的內容

RESULT:

```
string 'ORIGIN:  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 news.' (length=270)
```

->

Tokenization 原理：根據句子結構標示符破除句子結構

將句子結構標示符去掉，如 ，．！：；？

進階考慮：r13579@ntu.edu.tw r13579, ntu, edu, tw 關於 ″.′的處理問題

RESULT:

```
string 'TOKEN:  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 news' (length=266)
```

->

Lowercasing ：將所有大寫字母變成小寫字母

進階考慮：專有名詞的大小寫匹對準確性問題

變化參考：http://www.jb51.net/article/49629.htm

RESULT:

```
string 'LOWERCASE:  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 news' (length=270)
```

Normalization : 去除() <> [] {} 's,並將字串拆分成單詞數組
進階考慮：組合詞的連接格形式，如 ^ ` - _
the hold-him-back-and-drag-him-away manner.

```
array (size=39)
  0 => string 'and' (length=3)
  1 => string 'yugoslav' (length=8)
  2 => string 'authorities' (length=11)
  3 => string 'are' (length=3)
  4 => string 'planning' (length=8)
  5 => string 'the' (length=3)
  6 => string 'arrest' (length=6)
  7 => string 'of' (length=2)
  8 => string 'eleven' (length=6)
  9 => string 'coal' (length=4)
  10 => string 'miners' (length=6)
  11 => string 'and' (length=3)
  12 => string 'two' (length=3)
  13 => string 'opposition' (length=10)
  14 => string 'politicians' (length=11)
  15 => string 'on' (length=2)
  16 => string 'suspicion' (length=9)
  17 => string 'of' (length=2)
  18 => string 'sabotage' (length=8)
  19 => string 'that' (length=4)
  20 => string 'in' (length=2)
  21 => string 'connection' (length=10)
  22 => string 'with' (length=4)
  23 => string 'strike' (length=6)
  24 => string 'action' (length=6)
  25 => string 'against' (length=7)
  26 => string 'president' (length=9)
  27 => string 'slobodan' (length=8)
  28 => string 'milosevic' (length=9)
  29 => string 'you' (length=3)
  30 => string 'are' (length=3)
  31 => string 'listening' (length=9)
  32 => string 'to' (length=2)
  33 => string 'bbc' (length=3)
  34 => string 'news' (length=4)
  35 => string 'for' (length=3)
  36 => string 'the' (length=3)
  37 => string 'world' (length=5)
```

->

Stemming : Porter's algorithm
演算法實現過程:

第一步，處理複數，以及 ed 和 ing 結束的單詞。
第二步，如果單詞中包含母音，並且以 y 結尾，將 y 改為 i。
第三步，將雙尾碼的單詞映射為單尾碼。
第四步，處理-ic-，-full，-ness 等等尾碼。
第五步，在<c>vcvc<v>情形下，去除-ant，-ence 等尾碼。
第六步，也就是最後一步，在 m()>1 的情況下，移除末尾的“e”。

演算法使用說明:

傳入的單詞必須是小寫

RESULT:

```
array (size=39)
  0 => string 'and' (length=3)
  1 => string 'yugoslav' (length=8)
  2 => string 'author' (length=6)
  3 => string 'are' (length=3)
  4 => string 'plan' (length=4)
  5 => string 'the' (length=3)
  6 => string 'arrest' (length=6)
  7 => string 'of' (length=2)
  8 => string 'eleven' (length=6)
  9 => string 'coal' (length=4)
  10 => string 'miner' (length=5)
  11 => string 'and' (length=3)
  12 => string 'two' (length=3)
  13 => string 'opposit' (length=7)
  14 => string 'politician' (length=10)
  15 => string 'on' (length=2)
  16 => string 'suspicion' (length=9)
  17 => string 'of' (length=2)
  18 => string 'sabotag' (length=7)
  19 => string 'that' (length=4)
  20 => string 'in' (length=2)
  21 => string 'connect' (length=7)
  22 => string 'with' (length=4)
  23 => string 'strike' (length=6)
  24 => string 'action' (length=6)
  25 => string 'against' (length=7)
  26 => string 'presid' (length=6)
  27 => string 'slobodan' (length=8)
  28 => string 'milosev' (length=7)
  29 => string 'you' (length=3)
  30 => string 'are' (length=3)
  31 => string 'listen' (length=6)
  32 => string 'to' (length=2)
  33 => string 'bbc' (length=3)
  34 => string 'news' (length=4)
  35 => string 'for' (length=3)
  36 => string 'the' (length=3)
  37 => string 'world' (length=5)
```

Stopword removal,並消除重複的關鍵字

思路 1：根據 stopword list 去除 stopword ，為提高準確度，stopword list 盡可能設置很小

思路 2：根據 stopword list 設置 stopword 的 weight，在匹配的時候根據權重設置返回結果

RESULT

```
array (size=20)
  0 => string 'yugoslav' (length=8)
  1 => string 'author' (length=6)
  2 => string 'plan' (length=4)
  3 => string 'arrest' (length=6)
  4 => string 'coal' (length=4)
  5 => string 'miner' (length=5)
  6 => string 'opposit' (length=7)
  7 => string 'politician' (length=10)
  8 => string 'suspicion' (length=9)
  9 => string 'sabotag' (length=7)
  10 => string 'connect' (length=7)
  11 => string 'strike' (length=6)
  12 => string 'action' (length=6)
  13 => string 'presid' (length=6)
  14 => string 'slobodan' (length=8)
  15 => string 'milosev' (length=7)
  16 => string 'listen' (length=6)
  17 => string 'bbc' (length=3)
  18 => string 'news' (length=4)
  19 => string 'world' (length=5)
```

Save the result as a txt file.