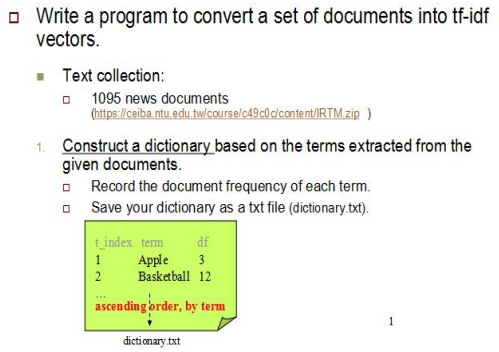
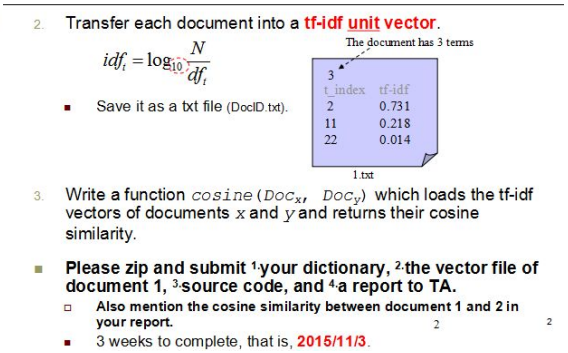
==========

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

==========

Programming Assignment 2:





My program result:

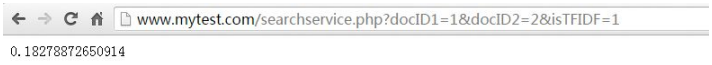
Step1：部署 Hw2

Step2：在流覽器輸入參數

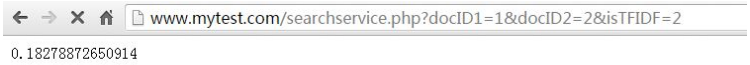
參數說明：docID1 與 docID2 指的是要求相似度的兩個檔案名

isTFIDF 是指使用最基礎的 TFIDF 作為檔向量中 term 的指標，還是使用修改後的 WFIDF 作為檔向量中 term 的指標，本程式支持兩種計算方法

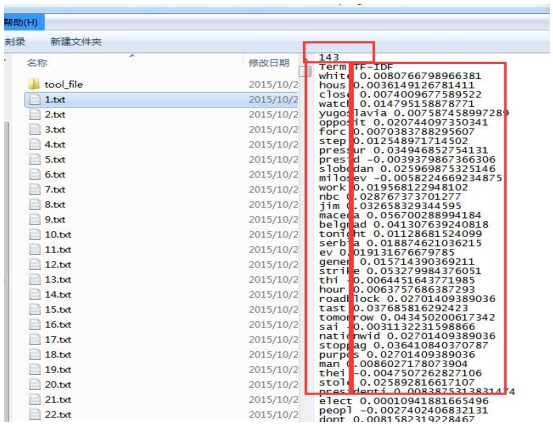
方法 1：使用最基礎的 TFIDF 作為檔向量中 term 的指標，結果如下

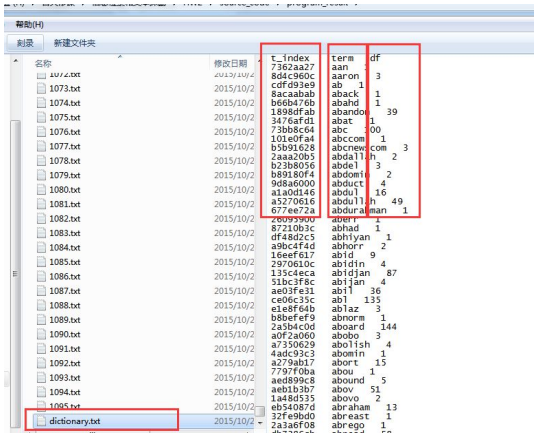


方法 2：使用修改後的 WFIDF 作為檔向量中 term 的指標，結果如下：

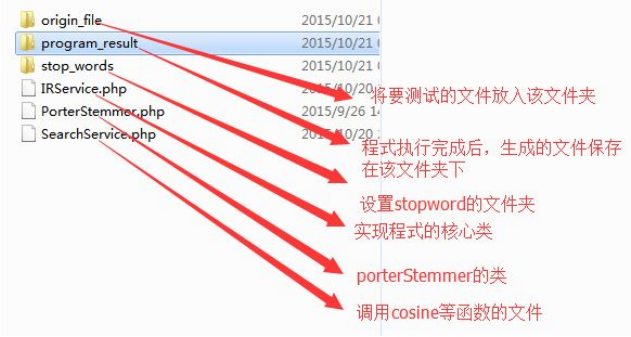


生成的文檔，詳見 program_result 檔夾，大致如下：





My program organization:



My program core class:

PorterStemmer 類說明：

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:

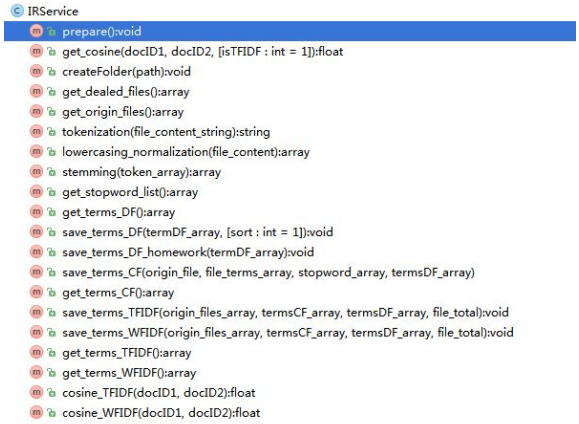
傳入的單詞必須是小寫

參考學習網站:

http://tartarus.org/~martin/PorterStemmer/

http://snowball.tartarus.org/algorithms/english/stemmer.html

http://blog.csdn.net/noobzc1/article/details/8902881



IRService 類說明:

類的主要函數:

類的關鍵函數 1:

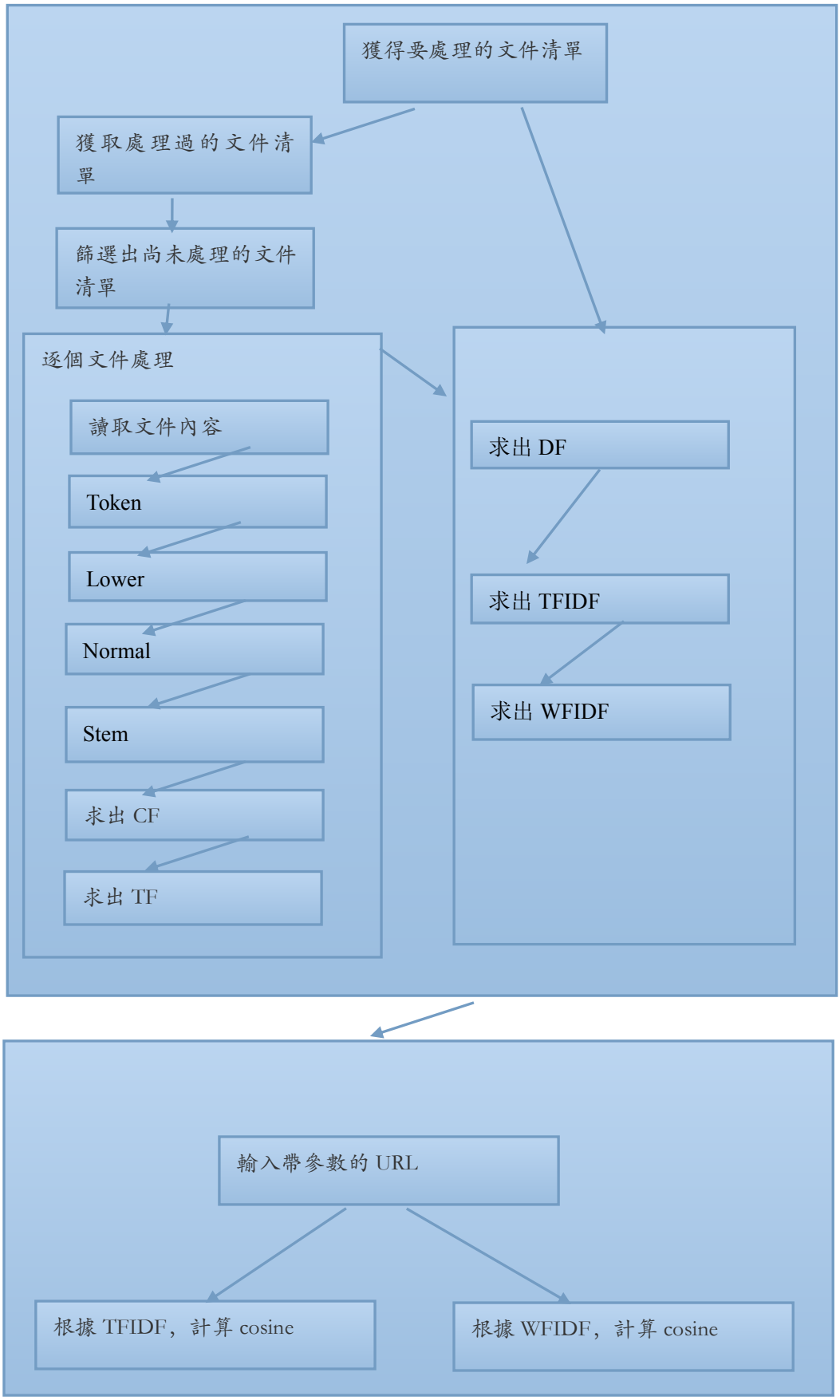
```
public function prepare()
{
    //獲取處理過的檔案名單
    $dealed_files_array = $this->get_dealed_files();
    //    var_dump($dealed_files_array);
    //獲取全部的檔案名單
    $origin_files_array = $this->get_origin_files();
    //var_dump( $origin_files_array);
    //獲取 stopword 清單
    $stopword_array = $this->get_stopword_list();
    //var_dump($stopword_array);
    //獲取已有的 termsDF
    $termsDF_array = $this->get_terms_DF();
    //var_dump($termsDF_array);
    $this->createFolder(dirname(__FILE__) . "\\program_result\\tool_file");
    $handle_file_deal = fopen(dirname(__FILE__) .
"\\program_result\\tool_file\\doc_dealed.txt", "a");
    //逐個處理檔
    foreach ($origin_files_array as $origin_file) {
        //篩選出要處理的檔
        if (!in_array($origin_file, $dealed_files_array, true)) {
            //var_dump($origin_file);
            //讀取檔內容
            $file_path = dirname(__FILE__) . "\\origin_file\\" . $origin_file;
            $file_content_string = file_get_contents($file_path);
            //var_dump( $file_content_string);
```

```
                //token
                $file_content_string = $this->tokenization($file_content_string);
                //var_dump( $file_content_string);
                //lower and normal
                $file_terms_array = $this->lowercasing_normalization($file_content_string);
                //var_dump($file_terms_array);
                //stem
                $file_terms_array = $this->stemming($file_terms_array);
                //var_dump($file_terms_array);
                //計算並保存 CF
                $termsDF_array = $this->save_terms_CF($origin_file, $file_terms_array,
$stopword_array, $termsDF_array);
                //將處理過的檔登記
                $line_content = $origin_file . "\r\n";
                fwrite($handle_file_deal, $line_content);
            }
        }
        fclose($handle_file_deal);
        //計算並保存 DF
        $this->save_terms_DF($termsDF_array);
        $this->save_terms_DF_homework($termsDF_array);
        $file_total = count($origin_files_array);//文章總數
        $termsCF_array = $this->get_terms_CF();//獲取所有檔所有特異單詞 CF
        //var_dump($termsCF_array);
        //計算並保存 TFIDF
        $this->save_terms_TFIDF($origin_files_array, $termsCF_array, $termsDF_array, $file_total);
        //計算並保存 WFIDF
        $this->save_terms_WFIDF($origin_files_array, $termsCF_array, $termsDF_array, $file_total);
}
```

類的關鍵函數 2:

```
public function get_cosine($docID1, $docID2, $isTFIDF = 1)
{
        $this->prepare();
        if ($isTFIDF == 1) {
            //獲取所有文章的所有單詞的 TFIDF
            $doc_terms_TFIDF = $this->get_terms_TFIDF();
            return $this->cosine_TFIDF($doc_terms_TFIDF[$docID1],
$doc_terms_TFIDF[$docID2]);
        } else {
            //獲取所有文章的所有單詞的 WFIDF
            $doc_terms_WFIDF = $this->get_terms_WFIDF();
            return $this->cosine_WFIDF($doc_terms_WFIDF[$docID1],
$doc_terms_WFIDF[$docID2]);
        }}
```

程式的主要流程:


獲得要處理的文件清單
獲取處理過的文件清單
篩選出尚未處理的文件清單
逐個文件處理
讀取文件內容
Token
Lower
Normal
Stem
求出 CF
求出 TF
求出 DF
求出 TFIDF
求出 WFIDF
輸入帶參數的 URL
根據 TFIDF，計算 cosine
根據 WFIDF，計算 cosine