1. 組員：

101820302 施帛辰

101820324 林哲琛

101820330 楊昇翰

101820329 林家文

101820307 施彙康

1. 程式優點：Low Memory、High Speed
2. 遭遇到的困難：讀取大的warc時，時不時會莫名其妙的卡住，無法繼續進行，原因好像是因為visual studio的html物件，會執行html檔中的java script，然後就停下來了…
3. 使用說明 :
   1. 使用visual studio開啟專案，然後執行程式
   2. 在title為IndexingForm的視窗上，點選File按鈕，選擇要輸入的.warc檔案，只能選擇一個
   3. 在同一個視窗上，點選Dictionary按鈕，選擇要輸出的「資料夾」
   4. 選擇token的設定，可以選擇none、case folding to upper或是case folding to lower
   5. 接著按下NEXT按鈕，這時候，就會產生inverted index等等
   6. 到QueryForm上，點選OpenDictionary按鈕，選擇在第三步所選擇的檔案輸出資料夾
   7. 接著在下面的Textbox中，輸入要query的字串，然後按下Query按鈕
   8. 最後就會在下面的ListBox中出現document編號，與對應的similarity score
4. 備註：
   1. global dictionary儲存的是每個term出現的文章ID，而article dictionary則是儲存在那篇文章中出現的位置
   2. 若使用visual studio執行專案時，出現錯誤，表示缺少某個dll時，在information retrieval的專案中，把HTMLparserLibDotNet20加入參考之中，這樣應該就可以執行了
   3. 輸入的warc檔中，每一個html的標籤不能有其他attribute，要純粹的<html>，如果是<html lang="en">，我們的專案會無法讀到
   4. 可能是上次助教您沒收到我們寄給您的信，或是各種原因沒有收到我們的上次作業，可否麻煩助教順便檢查作業一呢？Inverted Index的生成，Inverted Index存在「編號.dictionary.txt」中，編號是，各個html的編號，而mainDictionary.dictionary.txt則是儲存每個term出現在那些文章中。mainDictionary.dictionary.txt就是備註第一點所提到的global dictionary，而「編號.dictionary.txt」則是article dictionary