程式碼 User Manual

**1. 前置作業**

**txt 部分:**

執行檔：createAllFolders\_txt.sh（在資料夾home/controlFolder內）

目的：將所有資料夾RawFiles\_txt/內的原始檔案，移除所有html tags(<OOXX>)，被移除的tags其中不能含有子tags。如此一來，搜尋關鍵字時，找到的結果不會被txt檔內原含有的html tags所影響。

**HTML 部分:**

執行檔：createAllFolders\_html.sh（在資料夾home/controlFolder內）

目的：將所有資料夾RawFiles\_html/內的原始檔案中，如果有些檔案沒有適當的換行(在html tag </div>後沒有換行符號)，則在每個沒有換行的</div>結尾，加上換行符號。

**2. 搜尋關鍵字**

**txt 部分:**

執行檔：searchForHtml.sh（在資料夾home/controlFolder內）

目的：依照檔案listOfKeywords\_txt.xlsx （在資料夾home/controlFolder內）中的keyword sets進行關鍵字搜尋。

過程：

searchForTxt.sh會在資料夾home/all\_txt/內執行從traverseEachKeywordSet.sh檔（資料夾home/scripts/而來的）。

traverseEachKeywordSet.sh會依據listOfKeywords\_txt.xlsx去做關鍵字的搜尋。

執行traverseEachKeywordSet.sh會用到資料夾home/scripts/內的以下8個檔案：

add\_hyperLink\_index.pl

fromMultipleFileNames\_to\_one.pl

stripOffNewLine.pl

remove\_delimiters.pl

append.sh

addHTML\_headers.pl

addHTML\_headers\_toAllFiles.pl

add\_keyIndexToAllFiles.pl

結果：執行完的結果會以zip檔的方式出現在資料夾home/resultFolder 中。

**HTML 部分:**

執行檔：searchForHtml.sh（在資料夾home/controlFolder內）

目的：依照檔案listOfKeywords\_html.xlsx 中的keyword sets進行關鍵字搜尋。

過程：

searchForHtml.sh會在資料夾home/all\_html/內執行從traverseEachKeywordSet.sh檔（資料夾home/scripts\_html/而來的）。

traverseEachKeywordSet.sh會依據listOfKeywords\_html.xlsx去做關鍵字的搜尋。

執行traverseEachKeywordSet.sh會用到資料夾home/scripts\_html/內的以下9個檔案：

add\_hyperLink\_index.pl

fromMultipleFileNames\_to\_one.pl

stripOffNewLine.pl

clearTags.pl

remove\_delimiters.pl

append.sh

addHTML\_headers.pl

cp\_desired\_html.sh

add\_keyIndexToAllFiles.pl

結果：執行完的結果會以zip檔的方式出現在資料夾home/resultFolder中。

**3. 清除所有執行過程中的資料（不會清除資料夾home/resultFolder中zip檔的結果）**

**txt部分：**

執行檔：clearAllFolders\_txt.sh（在資料夾home/controlFolder內）

目的：目的就是當資料夾home/all\_txt裡面子資料夾很多，或者硬碟快滿時，清理資料。

clearAllFolders\_txt.sh會刪除資料夾home/all\_txt/內所有的資料夾，但不會刪除原始的.txt檔案，亦不會影響資料夾home/resultFolder/內的結果。

**html部分：**

執行檔：clearAllFolders\_html.sh（在資料夾home/controlFolder內）

目的：目的就是當資料夾home/all\_html裡面子資料夾很多，或者硬碟快滿時，清理資料。

clearAllFolders\_html.sh會刪除資料夾home/all\_html/內所有的資料夾，但不會刪除原始的.htm檔案，亦不會影響資料夾home/resultFolder/內的結果。