NLZ Rule

Update date: 2016/03/14

Cyberon Corporation
Software solution provider for embedded system
http://www.cyberon.com.tw/

About NLZ Rule

NLZ Rule 是 Cyberon CReader SDK 用來提供文字正規化(Text Normalization)的一個機制，使用字串替換的方式來修正 CReader SDK 中不正確的發音，以 regular express 的方式呈現，可指定想要尋找的 pattern 和想要置換的部分，在 CReader SDK 中，NLZ Rule 被儲存在一個文字檔裡面，其格式為 UTF8，可經由工具轉換成 binary file 或 hardcode 成 integer array，再傳入 CReader。

NLZ Rule 文字檔格式說明

- 一行為一個 Rule，由上到下循序尋找符合的 Rule，第一行未被使用可略過或填入任意內容，每行的第一個字元若為#則表示此行為註解，不會被加入
- 若發音字串中某部份已符合某一個 Rule，則不會再符合另一個 Rule，但是發音字串的其他部分可以符合某個 Rule
- 每個 Rule 由兩個欄位組成，兩個欄位間以**一個 Tab** 隔開，第一個欄位為 Pattern(以 regular express 方式呈現)，第二個欄位為置換的部份，下面為一個例子，它會將"曾雅妮"替換為"增雅妮"

  (曾雅妮)　"增雅妮"

- Pattern 欄位說明：
  - Patten 以 regular express 的方式來撰寫，欲置換的部份以(…)框起來，不想置換的部份可以放在(…)之外，下面為一個例子，它會將字串”建仔”或”殷仔”中的”仔”置換為”宰”

    [建殷](仔)　"宰"
  - 可以在整個 Pattern 中再指定 Pattern，如下例，它會將"數字後面有跟著百分比符號的數字(例如 45%)"置換為"百分之四十五"的字串(在這個例子中置換欄位裡面的"?0"表示整個 Pattern，"?1"則為([\d]+)這個 Pattern)

    (([\d]+)[%|%])　"百分之?xn:1"
  - Regular express 說明：
    - [x]表示某個字元，[]裡面可列出全部可能字元
    - [^x]表示非某個字元，[]裡面可列出全部可能字元
    - \d 表示數字 0~9
    - 對於某些字元，因為被用於 regular express，因此

若要使用這些字元必須在其前面加上"\"，例如"\+"、"\-"、"\?"…

- x|y 表示 x 或 y
- x+表示 x 出現 1 次以上
- x*表示 x 出現 0 次以上
- x?表示 x 出現 0 次或 1 次
- x{n,m}表示 x 出現 n~m 次
- x{n,}表示 x 出現 n 次以上
- x{n}表示 x 出現 n 次
- \b 表示字元邊界
- \A 表示字首
- \s 表示空白字元

◆ 置換欄位說明：

■ 以"…"框起來，其中除了可使用一般字元之外還可使用"?0"、"?1"、 "?2"…來指定 Pattern 欄位中第 0 個 Patten、第 1 個 Patten、第 2 個 Patten...

■ 此欄位中目前支援數字轉換為 Number 或 Digit 說法的功能(目前只支援中文)，"xn"為 Number 說法，"xnd"為 Digit 說法，它一般搭配上面的 Patten 使用，使用方法如下例:

"?xn:1"

NLZ Rule 轉換工具

為了提升載入 NLZ Rule 的效率，Cyberon 也提供了一個 PC 軟體工具 *regex-comp.exe*，可將 NLZ Rule 轉換為 binary data，來傳入 CReader engine。其 binary data 也會以兩種形式同時輸出，一個是 binary file，讓應用程式在 runtime 依需要載入需要的 rule；另一個是將內容 hard code 成 integer array 的 header file，讓應用程式在編譯時便可直接 include。

regex-comp.exe 是一個 conole mode(command line)的執行程式，使用方式如下：

regex-comp <NLZRuleFile> <OutArrayName> [<OutBinFileName>] > <OutHeaderFileName>

參數說明：

- <NLZRuleFile>: 為輸入的 NLZ Rule 文字檔案，**請注意其文件格式必須為 UTF8。**
- <OutArrayName>: 給定輸出的 header file 裡 integer array 的名稱。
- <OutBinFileName>: 為產生的 binary file 名稱。如果省略這項參數，便不會產生 binary file。
- <OutHeaderFileName>: Hardcode array 會以 stdout 的方式傳出來，因此可以將其導向，存成一個指定檔案。

相關使用範例，請參考 NLZRuleTool 目錄。