伍、語料庫處理的基本的計算工具

UNIX 與 LINUX 作業系統具備相當多的工具可以用來處理語料庫資料。我們建議讀者在 PC 上安裝 LINUX 作業系統，如此可以使用 UNIX 作業系統所具備的工具，並撰寫 PERL 程式抽取語料庫的訊息。[4]

一、字串轉換程式 tr

將英文檔案 datafile 中所有的大寫變成小寫並存到新檔案 output 裡面

tr 'A-Z' 'a-z' < datafile > output

將檔案中所有不是大寫或小寫的符號轉成空白行

tr –sc 'A-Za-z' '\012' < datafile > output

二、排序程式 sort

按照 ASCII 順序排序 sort datafile > output

按照數字從小到大排序 sort –n datafile > output

按照數字從大到小排序 sort –nr datafile > output

按照第三欄位從大到小排序 sort +2 –nr datafile > output

三、處理連續重覆行的程式 uniq

連續重覆的行只保留一行 uniq datafile > output

計算各行資料連續重覆數 uniq –c datafile > output

排序後去掉重覆的行 sort datafile | uniq > output

排序後計算頻率 sort datafile | uniq –c > output

將檔案中所有不是大寫或小寫的符號轉成空白行，排序，再計算頻率

[4] LINUX 作業系統內附 PERL 語言的轉譯器(Interpreter)可以自網路免費取得。此外 PERL 語言的轉譯器也有安裝在 Window 作業系統的版本。

tr –sc ‘A-Za-z’ ‘\012’ < datafile | sort | uniq –c > output

計算行數，字數，字元數的程式 wc

計算行數 wc –l　datafile

計算字數 wc –w　datafile

計算字元數 wc –c　datafile

四、從檔案中將包含某一字串或形式的行列出來 grep

從檔案中將包含 pattern 這個字串的行列出來 grep ‘pattern’ datafile > output

五、awk 一種簡單但功能強大的程式語言

從檔案中將包含 pattern 這個字串的行列出來

awk ‘/pattern/ { print }’ datafile > output

列印 datafile 第二欄與第一欄中間以 tab 間隔

awk ‘{ print $2 “\t” $1}’ datafile > output

如果 datafile 第三欄的值大於 1.65 列印第二欄與第三欄，中間並以 tab 間隔

awk ‘{ if ($3 >= 1.65) { print $2 “\t” $3} }’ datafile > output

六、perl：結合了 C 語言，awk，sed，shell programing，功能比 awk 更強大，非常適合處理語料。