設法搜集網路上對字頻及人名統計的介紹，已有那些分析的結果了？有沒有人已寫出程式了？請你自行整理並重新說明清楚，也記得列出這些資訊的出處。(你如何搜集及撰寫？)

基本上對於文章的分析多半都已經有實作的產品出來了，不管時間字數，詞頻字頻分析，以及有人針對文章的詞性與句法結構都已有了相當的研究了，甚至在Data Mining都已有關鍵字搜索，如何從字句解開說話的語氣還是情緒等等研究，超連結的關鍵字眼也隸屬於對於文字的研究。

資料源自：<http://www.putonghuaworld.com/computer/100625/100625_0101.htm>

<http://203.64.157.175/eo/05EnglishOnline/72FrequencyAnalysis.aspx>

<http://www.rm.spc.ntnu.edu.tw:8080/contents/analy/analy_list.asp?menuID=369>

撰寫一份任務A 的虛擬碼(pseudo-code)出來並仔細說明你的演算法。(你如何撰寫？)

**long** time1, time2 ;

time1 = System.*currentTimeMillis*() ; //時間起始

FileReader Pa = **new** FileReader("C:\\The\_Dream\_of\_the\_Red\_Chamber.txt") ;

**int** c = Pa.read();

**int** total = 0 ; //總字數

**int**[][] Htable = **new** **int**[100000][2] ;//開一個很大的Table[0][1]:次數/字碼

/\*

一開始就先把標點符號的編號將值set為-1，在迴圈判斷時直接跳開(省了之後一系列運算)

\*/

**while**(逐字讀整篇紅樓夢，讀到沒有為止){

**if**(如果是標點符號，再讀)

{

//再讀

}

**else**

{

//直接映射 編碼 = 位置

//非標點符號之中文字計數

//總字數+1

}

c = Pa.read() ; //在讀下一個字

}

**int** s , e ; //讓第一個元素非0

**for**( s = 0 ; Htable[s][0] <= 0 ; s++ ) ; //縮減陣列範圍

**for**( e = 99999 ; Htable[e][0] <= 0 ; e-- ) ;

**int** p , sorted ;

**int** value ; //字的數量

**int** index ; //字的編號

//insert sort

**for**(p=s+1; p<e+1 ; p++){

value = Htable[p][0] ;

index = Htable[p][1] ;

sorted = p - 1 ;

**while**( sorted >= s && Htable[sorted][0] > value){

Htable[sorted+1][0] = Htable[sorted][0] ;

Htable[sorted+1][1] = Htable[sorted][1] ;

sorted = sorted - 1 ;

}

Htable[sorted + 1][0] = value ;

Htable[sorted + 1][1] = index ;

}

FileWriter output = **new** FileWriter("D:\\output.txt" );

PrintWriter out = **new** PrintWriter(output) ;

out.println("字頻序號\t字:出現次數\r\n") ;

s=1 ;

**for**(c = e ; c >= 0 ; c--){ //寫入記事本

**if**(Htable[c][0] > 0){

out.println(s+"\t\t"+(**char**)Htable[c][1] + ":" + Htable[c][0]) ;

s++ ;

}

**if**(!(Htable[c][0] != 0))

**break** ;

}

time2 = System.*currentTimeMillis*() ; //停止時間

out.println("\r\n紅樓夢有"+total+"字而時間共花了"+(time2-time1)+"ms") ;

Pa.close() ;

output.close() ; //印出總字數和時間

撰寫program\_A 的程式出來。請提供你的原始碼、執行檔及如何執行你的程式的說明。原始碼中需有註解輔助說明。(你如何寫程式？)

使用JAVA寫，program\_A與program\_B為不同個專案，我有使用編輯器寫(Eclips)，如果正常下沒有環境的話，要先安裝JDK並設定環境變數將檔案放置你想要的位置去逐一編譯，如果有Eclips就直接載入專案按下執行即可運作。

而關於輸出檔案統一輸出到D槽下面，因為我的C槽不能寫有系統存取權限問題，其輸出字頻檔案名為 output.txt

試分析program\_A 的time complexity。請附上答案及說明。(你如何分析？)

假設總字數為n，因逐一讀字在陣列做O(1)的運算就為**O(n)**

因為我開的陣列有10萬個，為避免insert sorting不要交換次數太多，在前我先做縮減陣列的範圍動作。

**for**( s = 0 ; Htable[s][0] <= 0 ; s++ ) ; //(含0元素太多，大於1才有價值)縮減陣列範圍

**for**( e = 99999 ; Htable[e][0] <= 0 ; e-- ) ;

這兩個迴圈執行次數相加必定<=100000，就假定陣列大小為m，時間複雜至多為**O(m)**

而insert sorting由 s → e之範圍至多為m，時間複雜度**O(m^2)**

最後寫入txt檔至多也是**O(m)**

在最差情況下Total 可能為：**O(n)+ O(m)+ O(m^2)+ O(m) = O(n) + O(m^2)+2\*O(m)**

可能屬於**O(n) + O(m^2)**

請做一些實驗，設法繪出program\_A 的不同n 值(可使用字數不同的小說)範圍的T(n)時間成長統計圖表，並請加以說明成長的函數看來大約為何。注意：請詳細說明你所使用之機器軟硬體規格以及你為何選擇這樣的規格。(你如何估算？)

因為是用載入的人物名稱去判斷，而百家姓對於具有其他屬性(詞意詞性)並不能直接判斷，需要先經過(外用)字典判定，而名字有長有短並不是特別設計程式，而有些姓可以被當作動詞或其他名詞來用，可能會使抓取的名字誤判(抓出來發現不是人名的可能)，所以整個實驗僅能以紅樓夢作為實驗對象。

故利用1~10倍的紅樓夢字數來做實驗，當小說字數越多，花費的時間也會隨之遞增，圖形的成長的趨勢為線性遞增圖形。X軸為n倍紅樓夢，Y軸為時間(ms)

CPU: Intel i5-2500 3.3GHz

RAM: 12GB

作業系統: Windows 7 64位元

硬碟:1000GB

軟體: Eclipse in JAVA

因為只有個人實驗室電腦，之所以要用Java是因為他有些function是已經寫好，我可以套用字串的函數，但在C/C++對於字串處理並不是很強，需要重寫很多函數。

撰寫一份任務B的虛擬碼(pseudo-code)出來並仔細說明你的演算法。你如何找出「人名」？ 這部份你用了甚麼辦法？

FileReader Pb = **new** FileReader("C:\\紅樓夢人物.txt") ;

BufferedReader BufferedStream = **new** BufferedReader(Pb) ;

String[] name = **new** String[1000] ;

**int**[] n\_value = **new** **int**[1000] ;

String tmp = "" ;

**int** c = 0 ;

**int** x = 0 ;

**int** indexofname = 0 ;

**long** time1, time2 ;

time1 = System.*currentTimeMillis*() ; //時間起始點

**do**{

tmp = BufferedStream.readLine() ; //讀紅樓夢人物的每一行寫入陣列name[]

name[c] = tmp ;

c++ ;

**if**(tmp==**null**)

**break**;

}**while**(**true**) ;

Pb.close() ;

FileReader Ac = **new** FileReader("C:\\The\_Dream\_of\_the\_Red\_Chamber.txt") ;

BufferedReader context = **new** BufferedReader(Ac) ;

c = c - 1 ;

String allstr = "" ;

tmp = "" ;

**do**{

tmp = context.readLine() ; //讀紅樓夢內容的每一行

allstr = tmp ; //還原字串用的

x = 0 ;

**if**(allstr==**null**)

**break**;

**while**( x < c )//讀出每一行紅樓夢的內容用陣列name逐一比對 {

**if**((indexofname = tmp.indexOf(name[x]))> 0 ){//比對名字的位置

n\_value[x]++ ; //有就+1

indexofname = indexofname + name[x].length() ;

tmp = tmp.substring(indexofname, tmp.length()) ;//切割部分字串往後比對

}

**else**{

x++ ; //沒有繼續搜尋

tmp = allstr ;

}

}

}**while**(**true**) ;

**int** value,sorted ;

String n\_data ;

//insert sort

**for**(x = 1; x < c ; x++){

value = n\_value[x] ;

n\_data = name[x] ;

sorted = x - 1 ;

**while**( sorted >= 0 && n\_value[sorted] > value){

n\_value[sorted+1] = n\_value[sorted] ;

name[sorted+1] = name[sorted] ;

sorted = sorted - 1 ;

}

n\_value[sorted + 1] = value ;

name[sorted + 1] = n\_data ;

}

FileWriter output = **new** FileWriter("D:\\outname\_sorted.txt" );

PrintWriter out = **new** PrintWriter(output) ;

out.println("人名序號\t人名\t\t出現頻次\r\n") ;

**for**(x=c-1,value=1 ; value < 201 ; x--,value++)//寫入記事本

{

out.println(value+"\t\t"+name[x]+"\t\t"+n\_value[x]) ;

}

time2 = System.*currentTimeMillis*() ; //時間停止點

out.println("\r\n\r\n而時間共花了"+(time2-time1)+"ms") ;

output.close() ;

對於沒有使用hash table在於需要字串編碼時，而又要設計一個collison的發生，我覺得這部分有點難設計，而且設計一個hash function也很難有一定的公式或是方法絕對避免collison，似乎要開一個比現在更大的table可能幾十萬之類的，造成coding的困難度。而我覺得直覺的方式是比較輕易做出來，但為了要實行執行快速通常是在讀取文章內容和排序造成的時間較為高，故只針對排序加快，而原本想嘗試quick sort做出的，但java function似乎不能直接把一個很大的陣列載入，只好用插入排序。

撰寫program\_B 的程式出來。請提供你的原始碼、執行檔及如何執行你的程式的說明。原始碼中需有註解輔助說明。(你如何寫程式？)

使用JAVA寫，program\_A與program\_B為不同個專案，我有使用編輯器寫(Eclips)，如果正常下沒有環境的話，要先安裝JDK並設定環境變數將檔案放置你想要的位置去逐一編譯，如果有Eclips就直接載入專案按下執行即可運作。

而關於輸出檔案統一輸出到D槽下面，因為我的C槽不能寫有系統存取權限問題，其輸出字頻檔案名為 outname\_sorted.txt

試分析program\_B 的time complexity。請附上答案及說明。(你如何分析？)

讀入人名時寫入name[]，假設m個人名就**O(m)**

以bufferreader載入紅樓夢內容所需行數為**O(n)**，逐一名字比對，至少n\*m，但如果同一句有一種人名出現兩次可能也會造成部分行數搜尋增加，假設此種情況發生為k次(但遠比n來的小)，精準一點的算法就是以**O((n+k)\*m)**，然後針對所有人名insert sort排序至多**O(m^2)**次，寫入記事本前兩百名就**O(1)**

**T(n) = O((n+k)\*m)+ O(m^2)+ O(m) +O(n)+ O(1)，如果n>>m是以時間複雜度O((n+k)\*m)**

請做一些實驗，設法繪出program\_B 的不同n 值(可使用字數不同的小說)範圍的T(n)時間成長統計圖表，並請加以說明成長的函數看來大約為何。注意：請詳細說明你所使用之機器軟硬體規格以及你為何選擇這樣的規格。(你如何估算？)

因為是用載入的人物名稱去判斷，而是以線性的方式逐一比對內容是否出現了人物名稱的名字去計算，故整個時間都是線性成長，實驗僅能以紅樓夢作為實驗對象。

故利用1~10倍的紅樓夢字數來做實驗，當小說字數越多，花費的時間也會隨之遞增，圖形的成長的趨勢為線性遞增圖形。

X軸為n倍紅樓夢，Y軸為時間(ms)

CPU: Intel i5-2500 3.3GHz

RAM: 12GB

作業系統: Windows 7 64位元

硬碟:1000GB

軟體: Eclipse in JAVA

因為只有個人實驗室電腦，之所以要用Java是因為他有些function是已經寫好，我可以套用字串的函數，但在C/C++對於字串處理並不是很強，需要重寫很多函數。

請說明你做此作業所踫到的一些狀況及困難。

一開始為了求好心切，不停以hash的概念去想，但想了幾天有幾個想法但實作起來對於不擅coding的我有些難度，有著手去做但發現時間花費得太久而且又做不出來，索性先求有再求好的想法，因為還有其他科目要顧阿，不能一直把時間放在coding上，其實當初遇到的困難是在於排序的地方，一直想用quick sort解決，但是真的無法用副程式載入10萬個陣列，又去找了直敘式的quick sort，概念過於複雜所以放棄，而因為對於java沒有很熟，很多時間都在看function如何使用，和同學討論一些邏輯觀念居多，剩下的時間都在debug，一種自以為邏輯沒錯但語法錯誤的使用，造成debug 難以找出，而且trace感覺很正常的情況下。

注意：此作業要求快速的計算，愈快的程式愈好，而且要求你的程式所使用的記憶體(含程式本身)最多為1G bytes，你需說明你如何估算記憶體所佔用的大小。

Program\_A 因為開了一個很大的table array為100000\*2的大小，而都是int型態4byte，故800000byte換算成781.25KB不到1MB。

Program\_B開了兩個為1000大小的array，一個String另一個int型態，而因為每一個中英文字都以2個byte組成，String:1000\*2\*m(陣列所儲存的字數)，int:1000\*4

總和：2000m+4000，因為人名約略有300多個，所乘起來的字數>800000byte也是約略1MB的大小，不會超過1GB的記憶體。

(12) 本作業沒有規定的部份，你可自由發揮。

姓名： 朱聖池

Email: [ms0529756@gmail.com](mailto:ms0529756@gmail.com)

電話： 0937763697