Programming Assignment #1 - NOTE

Deadline: 15:00 PM, **May 27, 2018**

Submission: Place all of your files into a folder and zip it. Name the zip file after your student ID. E.g. if your student ID is B123456789, the file name should be "B123456789.zip". Upload the zip file to "ceiba". The folder must contain the following.

- **Program files** (source, object…).
- **A simple report** that describes your methods (e.g. data-structure, algorithm, …), including performance analysis (e.g. time-complexity…), etc. This report should not exceed two pages.
- **Pagerank**:
  For every d = 0.25, 0.45, 0.65, 0.85 and DIFF = 0.1, 0.01, 0.001, create a txt file with a list of the pages in page rank order (highest to lowest), for each page listing its outbranching (how many pages it connects to) and the page rank for that page (to 8 significant digits).
  If d = 0.45 and DIFF = 0.01, the file name should be "45_01.txt". The rest are given in the same way.
- **Reverse Index**:
  Output the reverse index into a file named "reverseindex.txt". See the details given in "prog1.pdf". (Note: please use ASCII order, i.e., word "NASA" will be listed before word "dog".)
- **Search Engine**:
  Each line of file "list.txt" is a list of words. With respect to each of the above combinations of d (say, 0.x) and DIFF (say, 0.y), search the lists of words line by line and output the top ten pages (sorted by page rank) that contain them (if less than 10, then list all of them) into file "result_x_y.txt". For instance, for d=0.45, DIFF=0.01, output the file "result_45_01.txt". When a single word is read in, just output the pages for that word. When multiple words are read in, output two lists - the top ten pages with all the words (AND semantics) and the top ten pages with any of the words (OR semantics). If none of the pages contain the read-in word output the word "none" into the file.
  The output format must follow the rules of the following example. (Note: Uppercase letters and lowercase letters should be treated differently. Space character and newline character are different from each other.)

Example

Content of file “list.txt”:

Dog
cat walk
NASA
men’s book three
…

Content of file “result.txt”:

page1000 page200 page2 page4 page5 page59 page3 page11 page10 page13
AND none
OR page15 page200 page7 page4 page43 page9 page3 page11 page77 page13
none
AND page200 page7 page9
OR page200 page23 page7 page9 page77 page13
…

See “prog1.pdf” for details. Should you have questions, please send your inquiries to the TA at bottle1116@hotmail.com.

程式作業 #1 繳交說明

繳交期限: **May 27, 2018**
繳交方式: 上傳至 ceiba (為避開系統尖峰時段，繳交時間為當日的 15:00。)

繳交格式:
請將所有檔案壓縮到一個 zip 檔，並以你的學號作為檔案名稱。例:假設你的學號為 B123456789，則上傳的檔案為 “B123456789.zip” 。其中，須包含下述檔案。

- 程式檔 (source, object…)。
- 說明檔。簡單說明你所使用的方法(資料結構，演算法…)，及分析(時間複雜度…)，或其他(如，你對此問題的看法)。說明請勿超過 2 頁。
- Pagerank、 Reverse Index、及字串搜尋結果之輸出。
  - Pagerank: 分別將 d = 0.25, 0.45, 0.65, 0.85，及 DIFF = 0.1, 0.01, 0.001 的所有 12 種組合 pagerank 的計算結果分別輸出到不同的 txt 檔。若 d = 0.45 及 DIFF = 0.01，則所輸出的檔案名稱為 “45_01.txt” 。其他檔案命名方式依此類推。
  - Reverse Index: 依 “prog1.pdf” 之說明，將 reverse index 結果輸出到一命名為 “reverseindex.txt” 的檔案。(注: 請依 ASCII order 排列輸出。故，字串 “NASA” 會列在字串 “dog” 之前。)
  - 字串搜尋: 對於上述的每組 d 與 DIFF，以行為單位，將 “list.txt” 裏的字串搜尋結果輸出到一命名為 “result.txt” 檔案。如 d=0.45, DIFF=0.01 則對 list.txt 搜尋輸出到命名為 “result_45_01.txt” 的檔案。
    如該行只有一組字串，則輸出包含該字串 pagerank 值最高的 10 個頁面，須依 pagerank 值由高到低排列。若包含該字串的頁面不到 10 個，則列出所有的頁面。一樣，須依 pagerank 值由高到低排列。若沒有任何頁面包含該字串，則輸出結果為 none。
    如該行有兩組以上的字串，則依序輸出 “AND” 及 “OR” 兩種方式的搜尋結果。(注：有可能一個或兩個搜尋結果為 none。)
    輸出之格式，如大小寫、空格、換行等，須遵照以下範例。

    例
    “list.txt”的內容為:

    Dog
    cat walk
    NASA
    men’s book three
    …

“result.txt”的內容為:

page1000 page200 page2 page4 page5 page59 page3 page11 page10 page13
AND none
OR page15 page200 page7 page4 page43 page9 page3 page11 page77 page13
none
AND page200 page7 page9
OR page200 page23 page7 page9 page77 page13
…

其餘細節，參照檔案“prog1.pdf”。
如對上述細節有任何疑問，請寄信至 bottle1116@hotmail.com 與助教聯絡。