Advanced Data Structures, the 3rd Assignment

Submit your code and report (report doesn't have to contain anything about code) through Ninova. Any questions: ataka@itu.edu.tr

Deadline: August 10, 2011 23:00 (please don't send your hw via e-mail)

Experimenting Effect of Order in B-Trees

With different order values (m), searching for data in B-trees can take different execution times. In this assignment, for B-Trees, experimenting effects of different order values on searching is the main issue. Tests will be performed on Reuters dataset which is provided with assignment.

Building the Tree

From command prompt, your b-tree builder implementation must take the order value (m). With respect to user-defined m, b-tree is built using Reuters dataset. This dataset contains news articles between <BODY> and </BODY> tags. With only considering the part between <BODY> and </BODY> tags, you must built a B-tree with using words as keys (use words as keys – use strcmp() function for lexical order). Node n which is represented with word w (key of n is w), should contain the list with two integers x and y (x is the order of news article and y is the sentence order in the xth news article). Then program stores b-tree in a file for later use (main experiment part).

For example, suppose that word 'advanced' is placed in the 4th sentence of the 2nd news and 6th sentence of the 5th news. Then the node for key 'Advanced' contains following couples: (2,4), (5,6).

For the first step, define a rule of heuristic for determining sentences (sentences must not be determined with 100% accuracy – it is ok). Split your sentences into words from spaces and then perform following operations on the words

- Make all characters lowercase
- Eliminate all characters except for letters

Then build the B-tree considering the words finally obtained.

Searching for the words

With using the stored B-tree with your B-Tree builder implementation, your searcher implementation searches n words which are given from command prompt. There is a node with given word in B-tree, program outputs the data which is contained by that node. Otherwise program outputs an error message for that search. Finally, program outputs the total execution time.

Report

- Explain your rule or heuristic on determining sentences. What are the advantages and disadvantages of your method?
- Built B-tree for m= 4, 6, 8, 9, 10, 14 and 16. For each tree, search n=17, 19, 21, 23, 25, 27 and 29 words which are selected randomly. Draw a histogram with the results obtained and discuss the results.
- Explain, how did you select random words for using in searcher implementation?
- Discuss, is there any open issue in this experiment at the beginning (caused by us) or at the performing process (caused by you)?

Any strong similarities (we know the level of ‘strong’ here, so don’t be afraid) among submissions or between implementations on the internet and submissions will be punished. Thank you.