CSE225 Data Structures, 2020(FALL)

Project #1 (Deadline: 7.12.2020, 24.00)

Text Representation with Binary Trees

In text classification studies, textual materials are represented with the frequencies of the words. Suppose that Table 1 gives the frequencies of 50 words in a document.

Table 1. Frequencies of words in the document

Number	Word	Frequency
1	people	6
2	country	10
3	city	15
4	news	2
5	population	1
6	society	28
7	university	35
8	sports	62
9	economics	4
10	book	89
11	library	3
12	computer	7
13	mouse	16
14	memory	27
15	game	50
16	student	60
17	club	70
18	text	83
19	algorithm	46
20	compiler	44
21	excel	49
22	name	51
23	department	56
24	head	54
25	faculty	22
26	teacher	33
27	professor	100
28	room	201
29	lab	92
30	kitchen	94
31	clock	97
32	class	93
33	board	64

34	pencil	65
35	window	61
36	team	19
37	software	13
38	group	14
39	grade	26
40	meeting	88
41	bag	99
42	television	205
43	visit	300
44	Ankara	74
45	New York	77
46	Dubai	76
47	plane	41
48	traffic	42
49	car	43
50	bus	75

a) Build a BST with the key “Word”.

b) Suppose that the number of accesses to word in your tree is directly the frequency of the word given in the table.

Calculate Total Access Time in the tree you build in (a).

c) Suppose that the number of accesses to word in your tree is directly the frequency of the word given in the table. Construct a BT to keep these records in the main memory so as to **minimize the total access time**, where one time unit is the time taken to compare the key of a tree with the key searched!

d) Calculate Total Access Time in the tree you build in (c).

e) Discuss your results in (b) and (d).

VERY IMPORTANT

The main goal of this project is to be familiar with trees. So, use of arrays/linked lists instead of trees is not acceptable.

In this project, you are expected to develop an algorithm that is capable of finding a solution to the above problem and ***implement this algorithm in ANSI C that runs under either UNIX or Windows***.

CODE SUBMISSION:

You should use Google Classroom in order to submit your code:

Your any submission after the project submission due date, will not taken into consideration.

1.) **Your c file**
naming standart:
name_surname.c

2.)**Your report**
naming standart:
name_surname.doc/pdf

Answers to a,b,c,d and e. In (a) and (c) print wour tree.

DO NOT COMPRESS FILES, JUST SUBMIT TWO FILES SEPERATELY. SUBMIT THE REPORT FOR REPORT PART IN GOOGLE CLASSSROOM, SUBMIT C FILE IN C CODE PART IN GOOGLE CLASSROOM.

*****IF YOU COMPRESS YOUR FILES YOU WILL LOSS 20 POINTS.**

GRADING:

When grading your project, first we will run your code and compare your results of the output with the results in your report. If you did not get the result by your algorithm, then do not write any answer for that part in your report. Because this is a programming homework.

Then, we will run your program by entering arbitrary words. We want to see if they are working correctly or not, and your project grade will be calculated based on number of your test cases, which are working correctly.

Individual effort is required and strictly expected in your projects. Your source code will be thoroughly checked both automatically by a cross check software tool and by your TA. Any attempt to any type of plagiarism will result in ALL involved students directly failing the class; their names appear in the black list of department and necessary disciplinary action taken by the instructor.

Good luck!!!