Middle East Technical University — Department of Computer Engineering

CENG 567 - Second Program

Design and Analysis of Algorithms

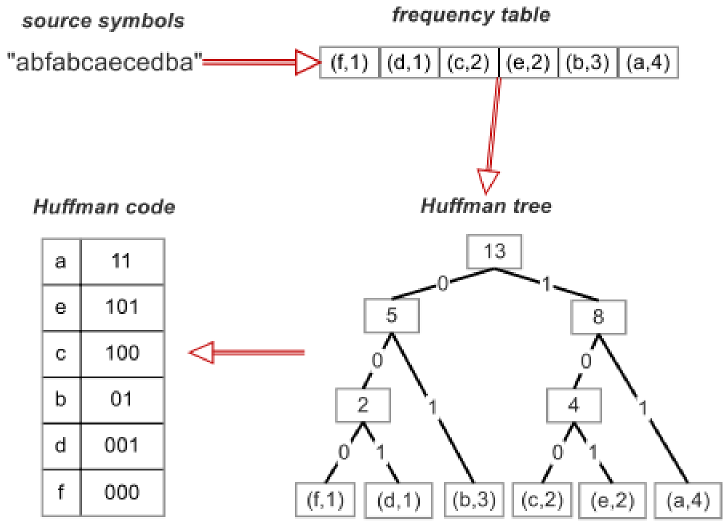
Spring 2019-2020

Take Home Exam 2 (THE-2)

Due date: 24 April 2020 - 23:55

1 Introduction

In this homework you will practice greedy algorithms. You will implement two versions of a lossless data compression technique called **Huffman Coding**. Huffman coding is a way of assigning prefix codes (bit sequences) to characters that contain less memory than the original character. Each prefix code assigned to a character is guaranteed to be unique. In this way, Huffman Coding provides a good way to make lossless data compression and decompress without ambiguities.



a	11
e	101
c	100
b	01
d	001
f	000

Figure 1: Huffman Coding. Total size after compression is 32 bits (1 times 3 bits (f), 1 times 3 bits (d), 2 times 3 bits (c), 2 times 3 bits (e), 3 times 2 bits (b), 4 times 2 bits (a))[image retrieved from - robotics.cs.tamu.edu]

In Huffman Coding roughly you have 3 stages (See Figure-1):

1. Given the source symbols, find the character frequencies.
2. Build Huffman Tree with the following steps:
   (a) Make a leaf node for all unique character.
   (b) Generate min heap of all leaf nodes.
   (c) Get two nodes having minimum frequencies from min heap.

   (d) Make a new internal node and assign its frequency value as the summation of the frequencies of the two nodes chosen.

   (e) Make the first extracted node the left child and the second extracted node right child.

   (f) Repeat steps *b*, *c*, *d* and *e*, until there exists a single node in the developed min heap.

3. Generate huffman codes by traversing the tree, add a 0 to huffman code for every left child and add a 1 to huffman code for every right child.

2 Part-1 - 50 pts

Input: Text file up to 100.000 upper case letters of English alphabet (nothing else, no blank or any other character).

process step 1: Determine the counts of each letter.

process step 2: Construct huffman tree and determine the codes for each letter.

process step 3: Determine the total size after compressing.

Output: Codes for each letter and size of the compressed file (do not consider table for character vs code)

2.1 Example Input/Output

Example Input:

```
AABBABABACACACABAA
```

Example Output:

```
Frequencies:
A: 10
B: 5
C: 3

Codes:
A: 0
B: 10
C: 11

Total Size:
26 bits
```

In this example A becomes one bit, B and C becomes 2 bits. When we replace characters by their huffman codes, there are 10 A's (10x1 bits) there are 5 B's (5x2 bits) and there are 3 C's (3x2 bits) and hence the total size is 26 bits.

3 Part-2 - 50 pts

In this part you will implement the same algorithm in the first part, the only difference is instead of letters you will use letter pairs (bi-grams) [assume total number of characters in the file is even]. All the other input assumptions are same in this part(uppercase letters,nothing else, no blank or any other character)

3.1 Example Input/Output

Example Input:

```
AABBABABACACACABAA
```

Example Output:

```
Frequencies:
AA: 2
BB: 1
AB: 3
AC: 3

Codes:
AA: 000
BB: 001
AB: 01
AC: 1

Total Size:
18 bits
```

Note that another solution:

```
Frequencies:
AA: 2
BB: 1
AB: 3
AC: 3

Codes:
AA: 00
BB: 01
AC: 10
BA: 11

Total Size:
18 bits
```

Note that, you will always start from the beginning and take two by two characters. For example in the example input you will read AA, then BB, then AB, then another AB etc. You cannot skip any of the characters, so there is no CA or BA in the character list.

4 Suggested Sources to Study Huffman Coding

- Link-1: Purdue.
- Link-2: Wikipedia.
- Thomas H. Cormen - Introduction to Algorithms Book.

5 Deliverables

- You can use C++,Java or Python. You will only upload one file named as the1.cpp (you can change the part .cpp to any extension you want according to language you use). Do not submit any other files.
- Submit the file through the ODTUClass system before the given deadline. **24 April 2020 - 23:55**

6 Regulations

- **Cheating:** This homework should be done individually. Copying another one's solution as it is, copying another one's solution and modifying it, or giving your own solution to someone else will be respected as cheating. Taking code directly from internet is also regarded as cheating. Grader of this homework has the right to request defense from those he suspects to be cheaters. Those whose defenses are not accepted will be subject to disciplinary actions according to the university regulations. Please note that defenses will only be taken in written way.
- **Odtuclass Discussions:** You should follow the odtuclass for discussions and possible updates on a daily basis.
- **Late Submission:** Not allowed.