Huffman Coding

Theory

Huffman Coding is a technique of compressing data to reduce its size without losing any of the details.

It was first developed by David Huffman.

Huffman Coding is generally useful to compress the data in which there are frequently occurring characters.

Algorithm

create a priority queue Q consisting of each unique character.

sort then in ascending order of their frequencies.

for all the unique characters:

    create a newNode

    extract minimum value from Q and assign it to leftChild of newNode

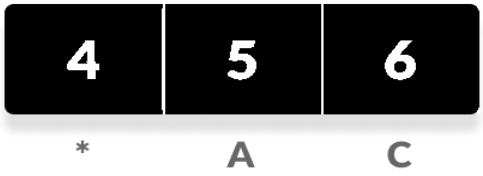
    extract minimum value from Q and assign it to rightChild of newNode

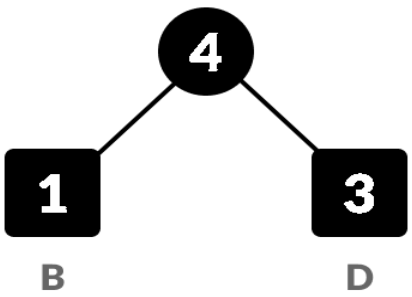
    calculate the sum of these two minimum values and assign it to the value of newNode

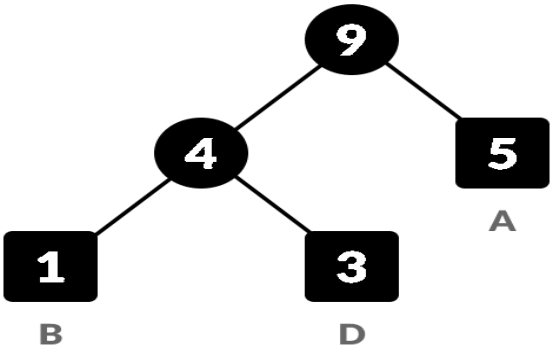
    insert this newNode into the tree

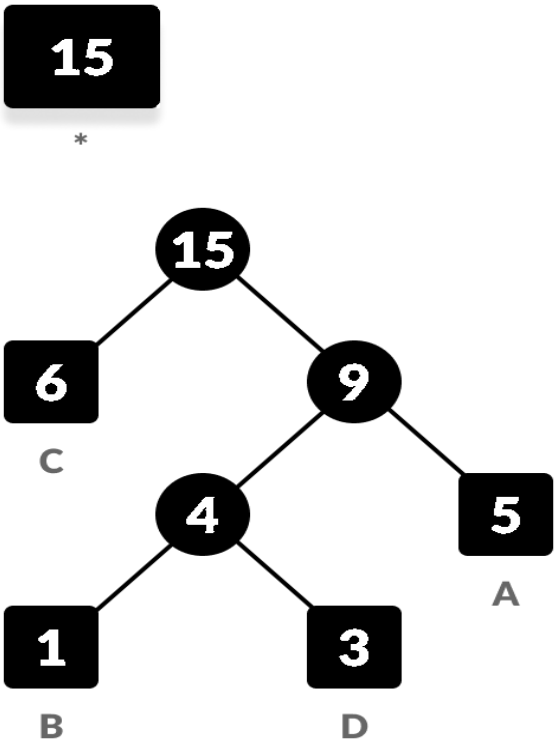
return rootNode

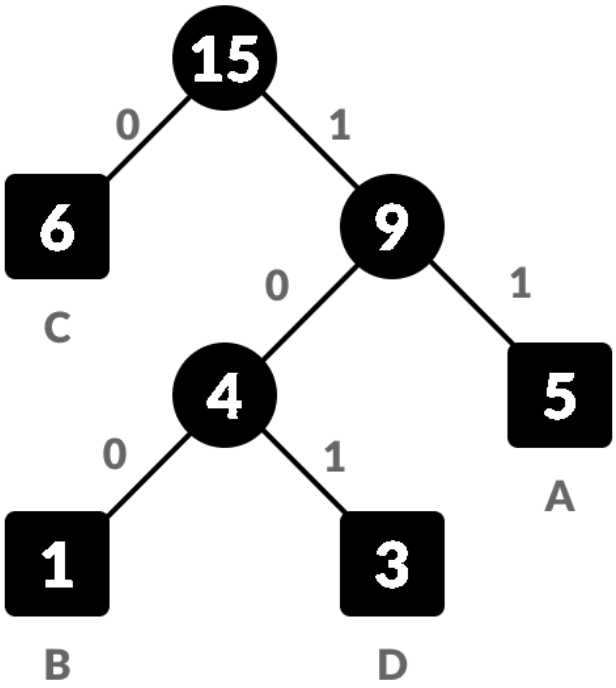
Explanation With Example











Character	Frequency	Code	Size
A	5	11	5*2 = 10
B	1	100	1*3 = 3
C	6	0	6*1 = 6
D	3	101	3*3 = 9
4 * 8 = 32 bits	15 bits		28 bits

Each character occupies 8 bits. There are a total of 15 characters in the above string. Thus, a total of 8 * 15 = 120 bits are required to send this string.

Without encoding, the total size of the string was 120 bits. After encoding the size is reduced to 32 + 15 + 28 = 75.

Huffman Coding Complexity

The time complexity for encoding each unique character based on its frequency is O(nlog n).

Extracting minimum frequency from the priority queue takes place 2*(n-1) times and its complexity is O(log n).

Thus the overall complexity is O(nlog n).

Huffman Coding Applications

Huffman coding is used in conventional compression formats like GZIP, BZIP2, PKZIP, etc.

For text and fax transmissions.