CSE102: DSA Lab Assignment Report

Question: 1
Are *min-heap*(*s*) used in your implementation of *Huffman Encoding*?

Answer:
Yes, a *min-heap* is used in my implementation of *Huffman Encoding*. The unique characters from the input file are stored in *Nodes*, which are then stored in a *min-heap*. The two *Nodes* with the minimum frequency are extracted from the *min-heap* using the standard *extractMin*(*heap*) algorithm. A "parent" *Node* is constructed using them and is inserted back into the *min-heap*.

Question: 2
What is the size of the file obtained after compressing the sample input file provided with the Assignment?

Answer:
Size of the sample input file: 10,000 Bytes
Size of the file obtained after compressing the input file: 2921 Bytes

Question: 3
Describe the format of the metadata that is stored in the compressed file.

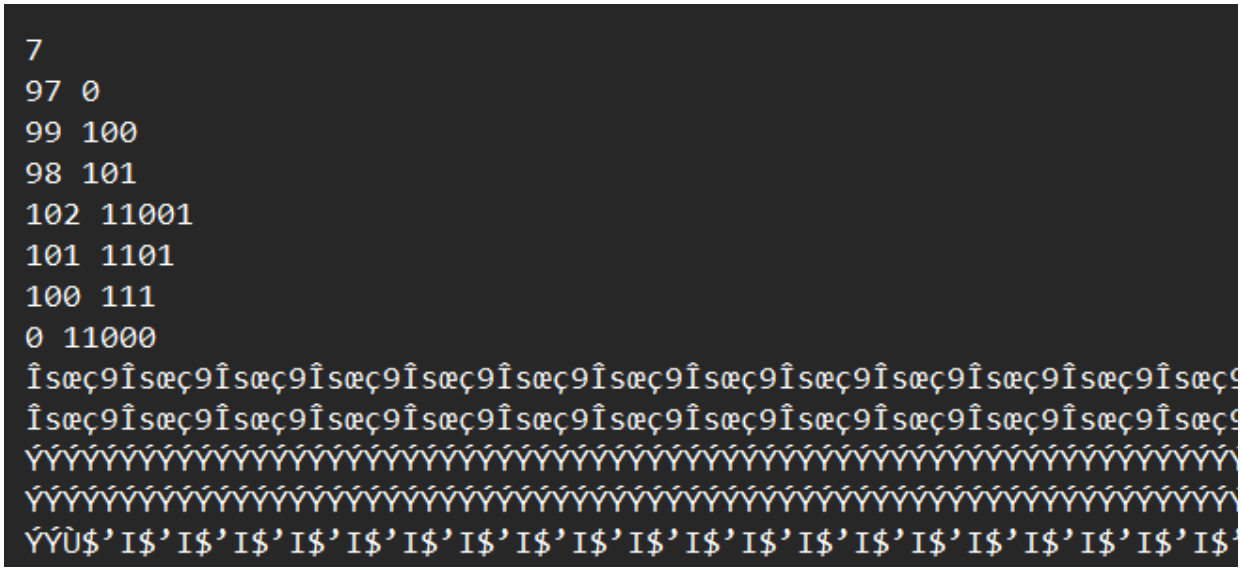
Answer:
Define the terms:

- ***N*:** number of unique characters present in the input file
- ***E*:** a chosen *EOF* character (In this case, the character with ASCII Code 0)
- ***ascii*(*x*):** The ASCII Code of character x
- ***huffmanCode*(*x*):** The *Huffman Code* generated for character x

Then, the generated compressed file contains metadata for the encoding in the following format (for the ease of access):

Line #1:	<*N*+1>
Line #2:	<*ascii*(character-1)> <*huffmanCode*(character-1)>
Line #3:	<*ascii*(character-2)> <*huffmanCode*(character-2)>
...	...
Line #(N+1):	<*ascii*(character-*N*)> <*huffmanCode*(character-*N*)>
Line #(N+2):	<*ascii*(character-*E*)> <*huffmanCode*(character-*E*)>

The compressed data follows this metadata. A picture of the sample output file (generated from the sample input file) is attached below.



For reference, the input file contained the unique characters *'a'*, *'b'*, *'c'*, *'d'*, *'e'*, and *'f'*, i.e. $N = 6$. The first 8 lines of the output file store the metadata.

Divyajeet Singh
2021529