UNIT - III

Text & Image compression.

Compression principle:-

* Some of the compression algorithms in widespread use, it will be helpful if we first build up an understanding of the principles on which they are based.

Source encoders & destination decoders

Lossless and lossy compression.

Entropy Encoding.

Source Encoding.

Source Encoders & destination decoders.

* As have just indicated, prior to transmitting the source information relating to a particular multimedia application, a compression algorithm is applied to it.

* This implies that in order for the destination to reproduce the original source information - or, in some instances, a nearly exact copy of it - a matching decompression algorithm must be applied to it.

*The application of the compression algorithm is the main function carried out by the ~~s~~. source encoder and the decompression algorithm is carried out by the destination decoder.

Lossless & lossy compression:-

* compression algorithms can be classified as being either lossless or lossy.

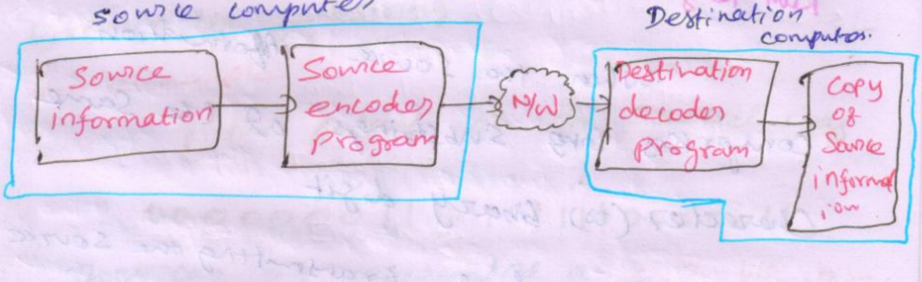
* In case of a lossless compression algorithm the aim is to reduce the amount of source information to be transmitted in such a way that, when the compressed information is decompressed, there is no loss of information.

* Lossless compression is said, therefore to be reversible.

* An example application of lossless compression is for the transfer over a network of a text file since, in such applications, it is normally imperative that no part of the source information is

lost during either the compression or decompression operations.

* In contrast the aim of lossy compression algorithms is normally not to reproduce an exact copy of the source information after decompression but rather a version of it which is perceived by the recipient as a true copy.



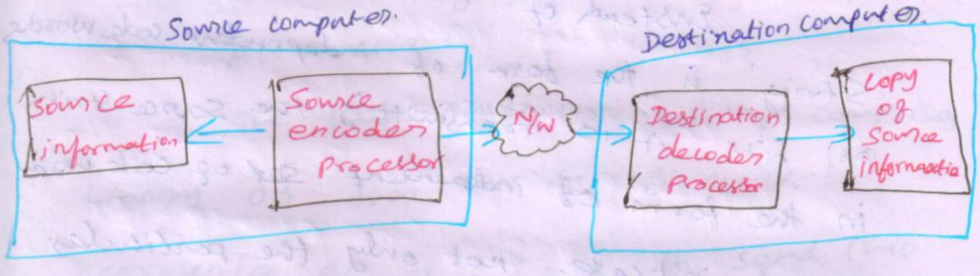


Fig: Source encoder / destination decoder alternatives
a) Software only b) special processor/hardware.

* In general, with such algorithms the higher the level of compression being applied to the source information the more approximate the received version.

Entropy Encoding:-

Entropy encoding is lossles and independent of the type of information that is being compressed.

It is concerned solely with how the information is represented.

Run length encoding.

When the source information comprises long substrings of the same character (or) binary digit

Instead of transmitting the source string in the form of independent code words (or) bits it is transmitting the source string in the form of independent set of code words which indicate not only the particular charactor (or) bit being transmitted but also an indication of the number of characters, bits in the substring.

Example

In an application that involves the transmission of long strings of binary bits that comprise a limited number of substrings each sub string can be assigned a separate code word.

The total bit string is then transmitted in the form of a string of code wordes selected from the codeword set.

each indicates both the bit in binary form (0&1) and the number of bits in the substring.

For example

if the output of the scanned was

0000000 1111111111 000001 1...

This can be represented as 0,7 1,10, 0,5 1,2, ... Alternatively

The first substring always comprises binary 0's then the string could be represented as 7, 10, 5, 2, ... to send this in a digital form, the individual decimal digits would be sent in their binary form.

Statistical encoding

Many applications use a set of code words to transmit the source information.

A set of ASCII code words are often used for the transmission of strings of cracter.

All the code words are often used for the transmission of strings of a comprise a fixed number of binary bits

7 bits in the case of ASCII

Statical encoding exploits this property by using a set of variable length code words. with the shortest code words used to represent the most frequently occuring symbols.

A code word set that avoids this happening is said to posses the prefix property and an example of encoding scheme.

The theoritical minimum Average number of bits that are required to transmit a particular source stream is known as Entropy of the source and can be computed using a formula attributed to Shannon.

$$\text{Entropy } H = -\sum_{i=1}^{n} P_i \log_2 P_i$$

where n is the number of different symbols in the source stream

P_i is the probability of occurrence of symbol "i"

$$\text{Average number of bits per code word} = \sum_{i=1}^{n} N_i P_i$$

Source Encoding :-

It exploits a particular property of the source information in order to produce an alternative form of representation that is either a compressed version of the original form (or) is more amenable to the application of compression.

Differential encoding.

It is used extensively in application whose the amplitude of a value (or) signal covers a large range but the difference in amplitude b/w successive value / signal is relatively small.

If the number of bit used is sufficient to cater for the maximum difference value then it is lossless.

Transform coding:-

It involves transforming the source information from one form into another, the other form lending itself more readily to the application of compression.

The rate of change in magnitude as one tranverse the matrix gives rise to a term known as spatial frequency.

The equivalent matrix of spatial frequency components is known as co-efficent.

Hence interms of compression if we can transform the original spatial frequency components are represented.

Image compression:

Graphic interchange format (GIF)

Tagged image file format (TIFF)

GIF

It is used extensively with the internet for the representation and compression of graphical images.

In a compression ratio of 3:1 the table of colors can relate either to the whole image in which case it is referred to as the global color table.

TIFF

It support pixel resolutions of up to 48 bits -16bits each for R, G & B and is intended for the transfer of both images and digitized documents

These use the same compression algorithms that are used in facsimile machines.

TEXT compression:

These are three different types of text as

1. unformatted Text
2. Formatted Text.
3. Hyper text.

This all represented as strings of charactered selected from a defined set.

The different types of text uses and interpret the latter in different ways.

These are two types of statistical encoding methods which are used with text,

Two examples of the former are the Huffman and arithmatic coding algorithms and the later is the Lempel-ZIV (LZ) algorithm

1. Static coding.
2. dynamic (or) adaptive coding.

Static - Huffman coding:-

The degree of imbalance is a function of the relative frequency of occurance of the characters, larger the spread, the more unbalanced is the tree. The resulting tree is known as the Huffman code tree.

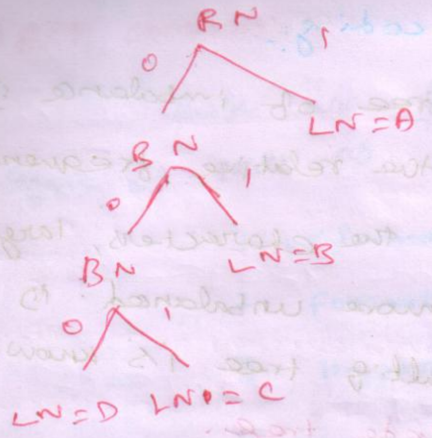
A Huffman (code) tree is a binary tree with branches assigned the value 0 or 1.

The base of the tree is knowns as root node and the point at which a branch divides a branch node.

Example:

A huffman code tree is considered and the corresponds to the string of characters

AAAABBCD



RN = root node

BN = branch node

LN = Leaf node

A = 1

B = 01

C = 001

D = 000

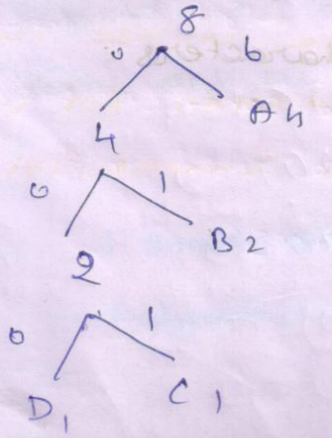
a) $A_4 \rightarrow A4 \rightarrow A4(1)$ $\qquad$ $A4(1) \rightarrow 1$

$B2 \rightarrow B2(1)$ $\qquad$ $\rightarrow 4(0)$ $\qquad$ $B2 = (1)(0) \rightarrow 01$

$C1(1) \rightarrow 2(0)$ $\qquad$ $C1 = (1)(0)(0) \rightarrow 00$

$D1(0)$ $\qquad$ $D1 = (0)(0)(0) \rightarrow 00$

frequency of occurrence. $\qquad$ Starting at leaf node $\qquad$ Star... at roo... node



weight order

$= D1\ C1\ 2\ B2\ 4\ A4$

Problem :-

A series of messages is to be transferred between two computers over a PSTN. The message comprise just the characters A through H. Analysis has shown that the probability of each character is as follows.

A and B = 0.25, C and D = 0.14, E, F, G and H = 0.555.

a) use shannons formulla to derive the minimum average number of bits per character.

b) use huffman coding to derive a code word set and prove this is the minimum set by constructiong the corresponding Huffmann code tree.

c) Derive the average number of bits per character for ur code word set and compare with

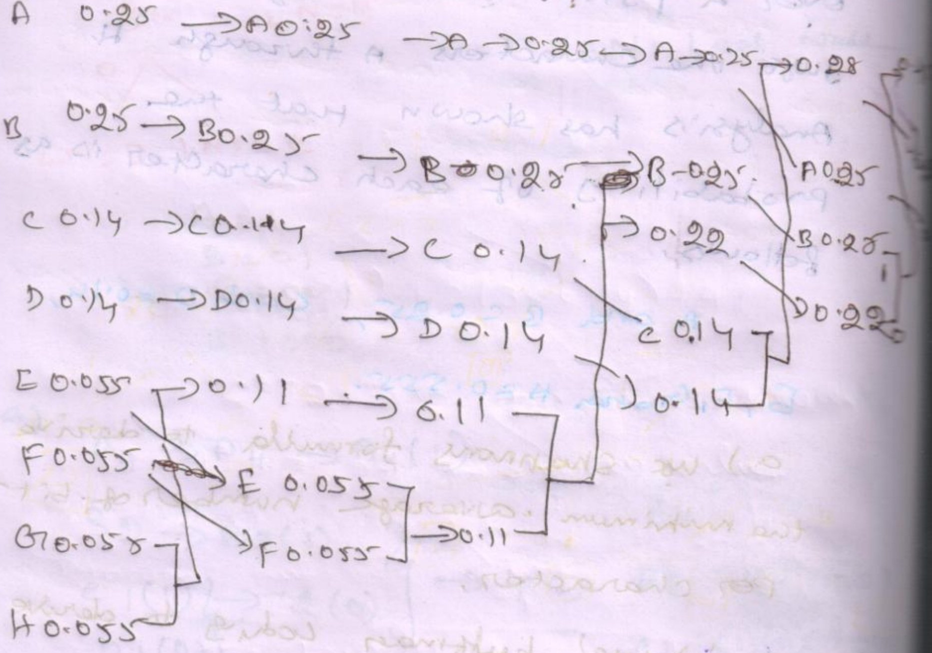
i) The entropy of the message (shannon's value

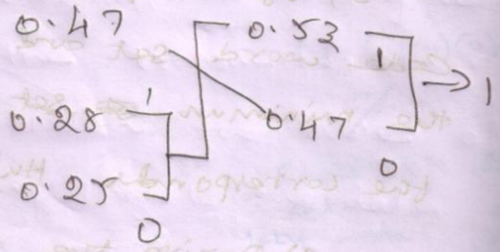
ii) fixed length binary code words

iii) 7-bit ASCII code words,

Solution:-

a) Code word generation of Huffman coding.





A = (0) (1) → 10

B = (1) (0) → 01

C = (1) (1) (1) = 111

D = (0) (1) (1) → 110

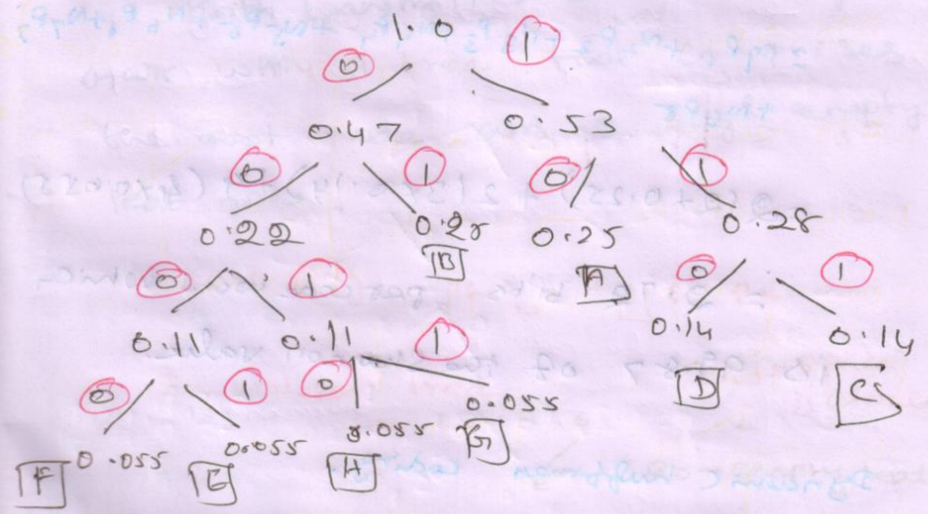
E = (1) (0) (0) (0) → 0001

$F = (0)(0)(0)(0) \rightarrow 0000$

$G = (1)(1)(0)(0) \rightarrow 0011$

$H = (0)(1)(0)(0) \rightarrow 0010.$

b) Huffman code tree:



Weigh order $= 0.055 \quad 0.055 \quad 0.055$

$0.055 \quad 0.11 \quad 0.11 \quad 0.14$

$0.14 \quad 0.14 \quad 0.22 \quad 0.25$

$\cancel{0.28} \quad 0.47 \quad 0.53.$

Shannon's fromula states

Entropy $H = -\sum_{i=1}^{8} p_i \log_2 p_i$ bits per code word

$\therefore H' = -\left(2\,(0.25 \log_2 0.25) + 2\,(0.14 \log_2 0.14) + 4\,(0.55 \log_2 0.055)\right)$

$= 1 + 0.794 + 0.921$

$= 2.175$ bits per code words.

$= \sum_{i=1}^{n} N_i P_i$

$= N_1 P_1 + N_2 P_2 + N_3 P_3 + N_4 P_4 + N_5 P_5 + N_6 P_6 + N_7 P_7 + N_8 P_8$

$= 2(2 \times 0.25) + 2(3 \times 0.14) + 4(4 \times 0.055)$

$= 2.72$ bits per code word which is 99.87 of the Shannon value.

Dynamic huffman coding.

The basic huffman coding method requires both the transmitter and the receivers to known the table of code words relating to the data being transmitted.

If the caracter to be transmitted is currently present in the tree its codeword is determined and sent in the normal way.

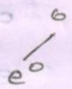
Assume that the data to be transmitted starts with the following character string

This is simple.

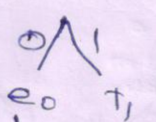
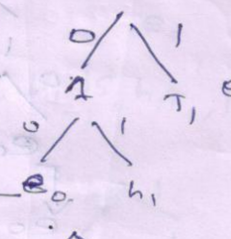
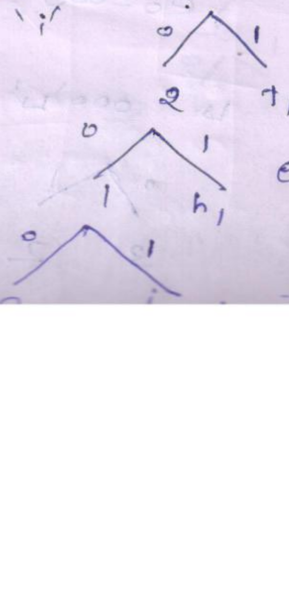
Both transmitter and receiving start with a tree that comprises the root node and a single empty leaf node.

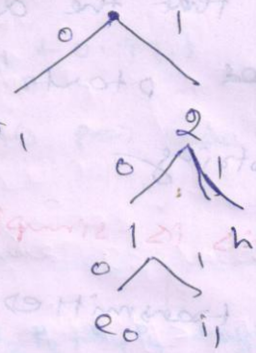
input string = This␣ is␣ simple.

Initialized tree. ␣ = space character.



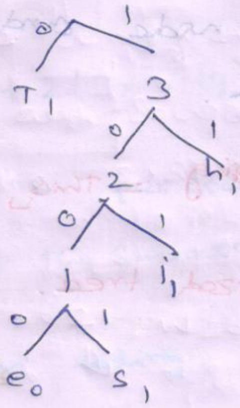
e_0 = empty leaf list.

a)	Character	output	
	T	"T"	 $e_0 T_1$
b)	h	0 'h'	 $e_0 h_1 1 T_1$
c)	i	00 'i'	 $e_0 i_1 h_1 2 T_1$

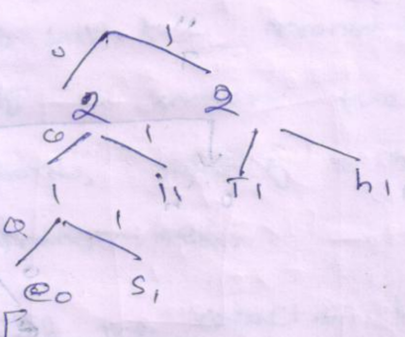


$e_0, i_1, 1, h_1, T_1, 2$

d) S 100'S'

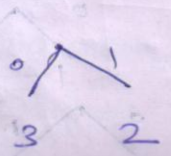


$e_0 S_1, 1 i_1, 2 h_1, T$
3

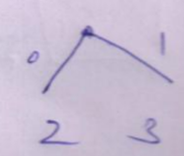


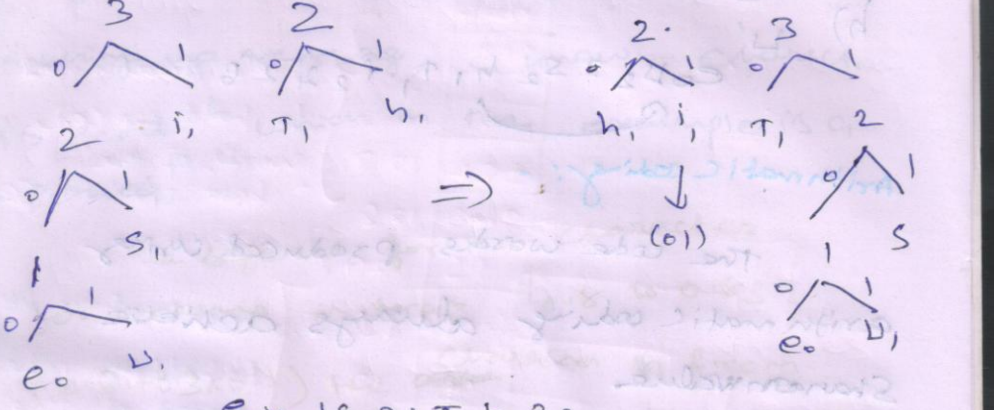
$e_0 = 000$

e) u 000'u'



$\Rightarrow$



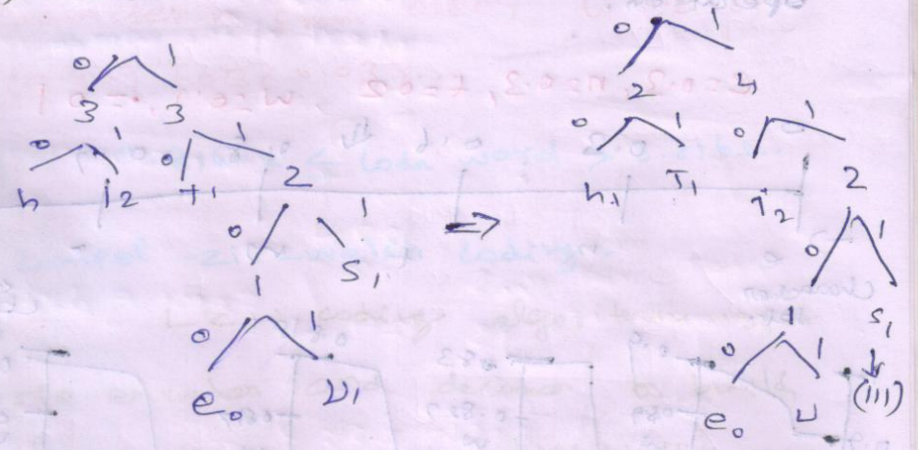


$e_0 \nu$, 1S, 2i, T h, 32

$e_0 \nu$, 1S, h, 1, T, 2

23

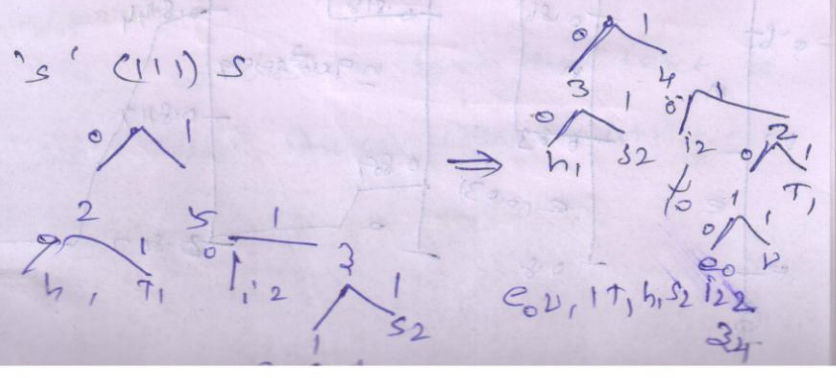
f) 01:



$e_0 \nu$, 1S, h, i_2 T, , 233

$e_0 \nu$, 1S, h, T, , i_2 2, 24.

g) 's' (111) S



$e_0 \nu$, 1T, h, S_2 i_2 2

34

h) etc '␣'

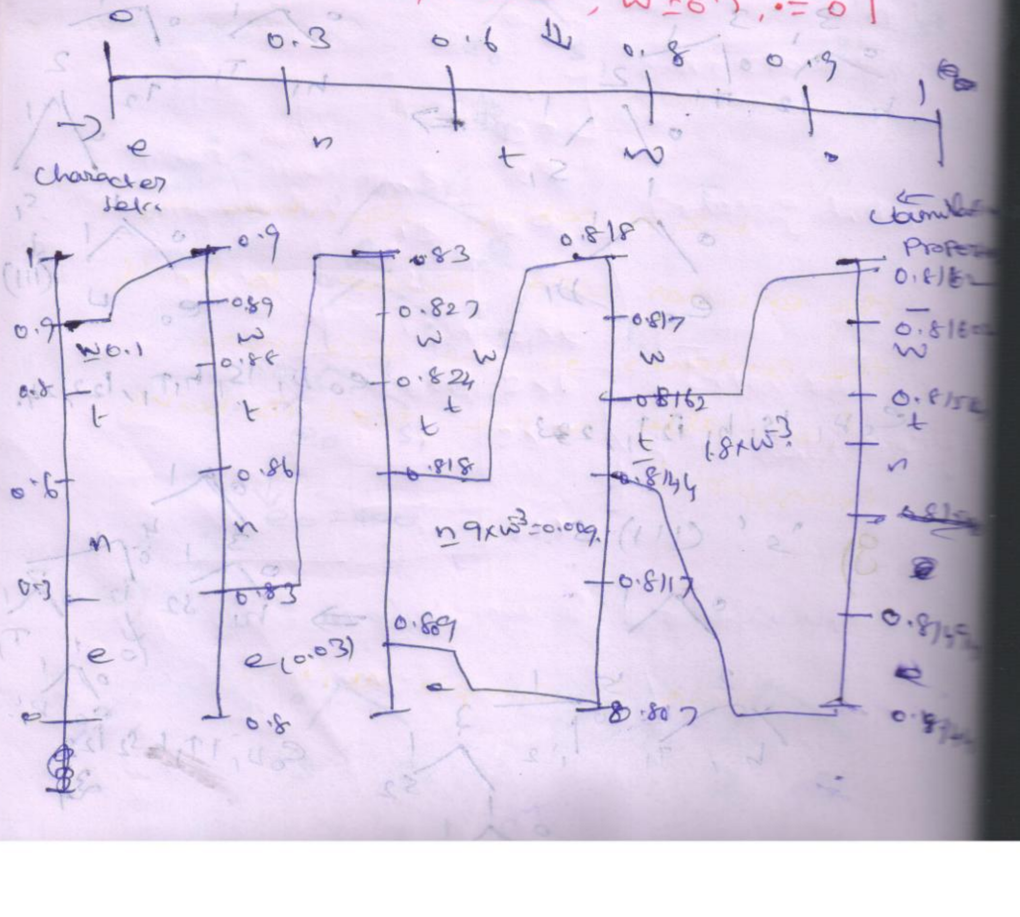
$e_0 ␣_2 \; 2 \; s_2 \; h_1 \; t_1 \; i_2 \; 4 \; 2 \; 6$

Arithmatic coding :-

The code words produced using arithmatic coding always achieve Shanon value

It is more complicated than Huffman coding and source shall limit our basic static coding mode operation,

e = 0.3, n = 0.3, t = 0.2, w = 0.1, . = 0.1



At the end of each character string making up message, a known character is sent which in the example is a period.

The segment for the character for example is from 0.8 to 0.83 (0.8 + 0.3 × 0.1) the ~~[illegible]~~ character n from (0.83) to (0.86) (0.83 + 0.3 × 0.1) and so on.

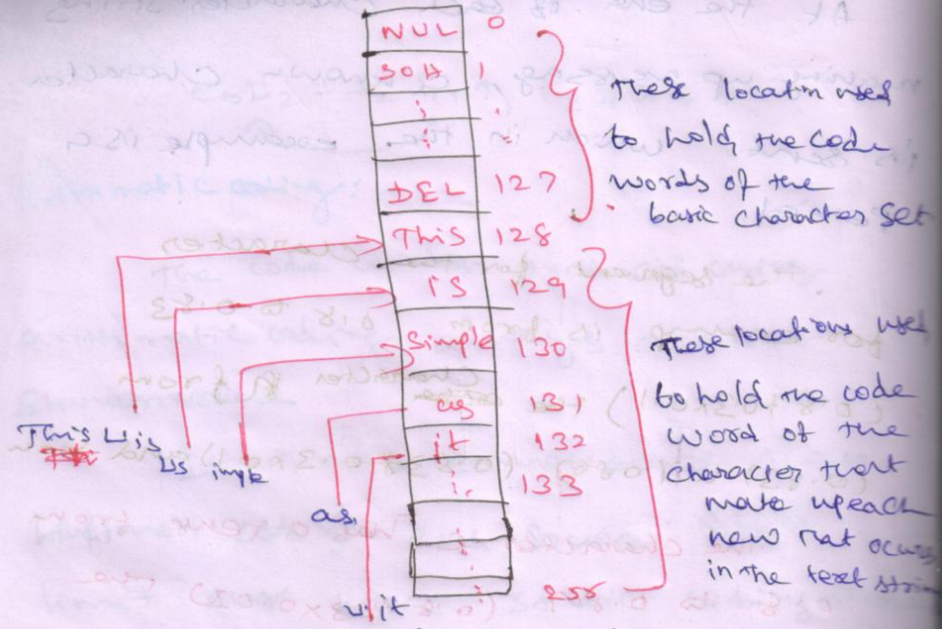
The character e has a range from 0.8 to 0.809 (0.8 + 0.3 × 0.03) the character n from 0.809 to 0.818 (0.809 + 0.3 × 0.03)....

0.81602 ≤ code word > 0.8162.

Lampel-ziv-welsh coding:-

LZW coding algorithm is for the encoder and decoder to build the contents of the dictionary dinamically as the text is being transferred.

Let us assume that the text to be compressed starts with the string this is simple as it is.



(This is sent using the index of the words is (129)

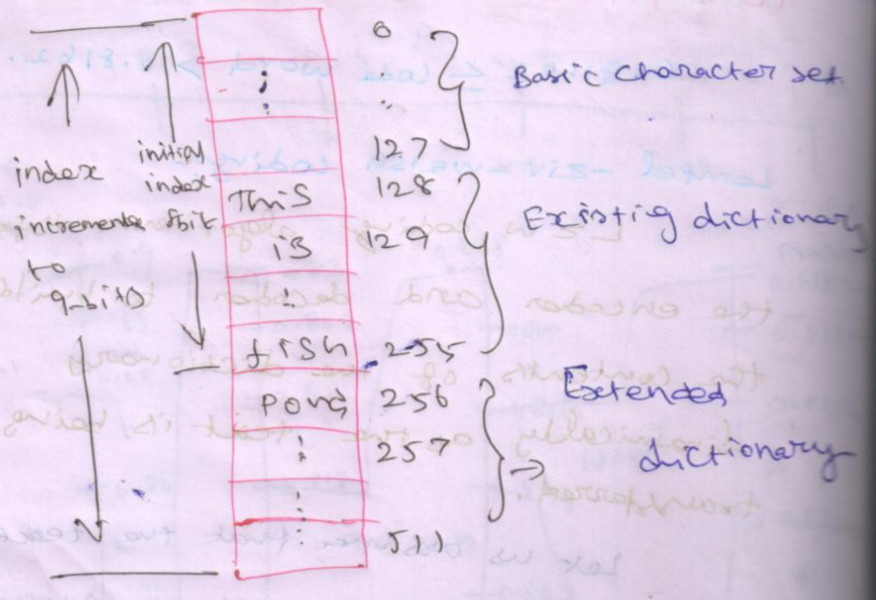


Fig. Basic operations of LZW compression algorithm.