Page No.
Date

Name : Dhiraj Bodake
Roll No : 18141216

TUTORIAL NO : 08

Q.1 Write short note on document clustering.

→ Document clustering is a operation of groueping together similar document is classes. In this record, document clustering is not reply on any operation on the text but an operation on the collection of document.

- The operation of clustering document is usually of two types - global and local.
- In the global clustering strategy, the document are grouped accordingly to their occurence in the whole collection.
- In the local clustering strategy, the grouping of document is affected by the context defined by the current query and its local set of document.
- Clustering method are usually used in IR to transform the original query in an attempt to better represent the user information need. From this perpective clustering is an operation which is more related to the transformation of the user query.

Q.2 Explain huffman coding and arithmetic coding as statistical coding strategies.

→ Huffman coding :

- Huffman coding is a classiers data compression algorithm. The idea is to design variable - length codes to input characters, lengths of assigned codes are on the frequencies of corresponding characters.
- the most frequent character gets the smallest code and the least frequent characters gets the largest code.
- The variable - length codes assigned to input characters are prefix codes, means the codes are assigned in such way that the code assigned to one character is not the prefix of code assigned to any other character.
- this is how huffman coding makes sure that there is no ambiguity

Arithmetic coding :

- In a popular compression algorithm after huffman coding and it is particularly useful for a relatively small and skewed alphabet.
- In this, a message is represented by an interval of real numbers between 0 and 1. As the message becomes larger, the interval needed to represent it becomes smaller and the number of bits needed to specify that interval grows.

- successive symbols of the message reduces the size of interval in accordance with symbol probabilities generated by the model.

Q3 compare arithmetic coding, character-based huffman coding, word-based huffman coding and zive lempel coding w.r.t. compression ratio, compression speed, decompression speed, memory space, compression pattern matching, random access

→

	Arithmetic	character huffman	word huffman	ziv-lempel
Compression ratio	very good	poor	very good	good
compression speed	slow	fast	fast	very fast
Decompression speed	slow	fast	very fast	very fast
memory space	slow	low	high	moderate
compressed matching	no	yes	yes	yes
Random access	no	yes	yes	no