UNIT II – INFORMATION RETRIEVAL

Part A - Questions

1. **What do you mean information retrieval models?**

   A retrieval model can be a description of either the computational process or the human process of retrieval: The process of choosing documents for retrieval; the process by which information needs are first articulated and then refined.

2. **What is cosine similarity?**

   This metric is frequently used when trying to determine similarity between two documents. Since there are more words that are in common between two documents, it is useless to use the other methods of calculating similarities.

3. **What is language model based IR?**

   A language model is a probabilistic mechanism for generating text. Language models estimate the probability distribution of various natural language phenomena.

4. **Define unigram language.**

   A unigram (1-gram) language model makes the strong independence assumption that words are generated independently from a multinomial distribution θ

5. **What are the characteristics of relevance feedback?**

   - It shields the user from the details of the query reformulation process.
   - It breaks down the whole searching task into a sequence of small steps which are easier to grasp.
   - Provide a controlled process designed to emphasize some terms and de-emphasize others.

6. **What are the assumptions of vector space model?**

   Assumption of vector space model:

   - The degree of matching can be used to rank-order documents;
   - This rank-ordering corresponds to how well a document satisfying a users information needs.

7. What are the disadvantages of Boolean model?

- It is not simple to translate an information need into a Boolean expression
- Exact matching may lead to retrieval of too many documents.
- The retrieved documents are not ranked.
- The model does not use term weights.

8. Define term frequency.

Term frequency: Frequency of occurrence of query keyword in document.

9. Explain Luhn's ideas

Luhn's basic idea to use various properties of texts, including statistical ones, was critical in opening handling of input by computers for IR. Automatic input joined the already automated output.

10. Define stemming.

Conflation algorithms are used in information retrieval systems for matching the morphological variants of terms for efficient indexing and faster retrieval operations. The Conflation process can be done either manually or automatically. The automatic conflation operation is also called stemming.

11. What is Recall?

Recall is the ratio of the number of relevant documents retrieved to the total number of relevant documents retrieved.

12. What is precision?

Precision is the ratio of the number of relevant documents retrieved to the total number of documents retrieved.

13. Explain Latent semantic Indexing.

Latent Semantic Indexing is a technique that projects queries and documents into a space with "latent" Semantic dimensions. It is statistical method for automatic indexing and retrieval that attempts to solve the major problems of the current technology. It is intended to uncover latent semantic structure in the data that is hidden. It creates a semantic space where in terms and documents that are associated are placed near one another.