UNIT V – DOCUMENT TEXT MINING

Part A - Questions

1. **What do you mean by information filtering?**

An information filtering system is a system that removes redundant or unwanted information from an information stream using (semi)automated or computerized methods prior to presentation overload and increment of the semantic signal-to-noise ratio.

2. **What are the characteristics of information filtering?**

- Filtering system involve large amounts of data.
- Information filtering systems deal with textual information.
- It is applicable for unstructured or semi-structured data.

3. **Explain difference between information filtering and information Retrieval.**

Information Filter	Information Retrieval
IF is concerned with the removal of textual information from an incoming stream and its dissemination to groups or individuals.	IR systems are concerned with the collection and organization of texts so that users can then easily find a text in the collection.
Information filtering is concerned with repeated uses of the system by users with long-term, but changing interests and needs.	A query represents a one-time information need.
Filtering is based on descriptions of individual or group interests or needs that are usually called profiles.	Retrieval of information is instead based on user specified information needs in the form of a query.
IF systems deal with dynamic data.	IR systems deal with static databases.

4. **What is text mining?**

- Text mining is understood as a process of automatically extracting meaningful, useful, previously unknown and ultimately comprehensible information from textual document repositories.

- Text mining can be visualized as consisting of two phases: Text refining that transforms free-form text documents into a chosen intermediate form, and knowledge distillation that deduces patterns or knowledge from the intermediate form.

5. What is classification?

Classification is a technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be "sunny", "rainy" or "cloudy".

6. Explain clustering.

Clustering is a process of partitioning a set of data in a set of meaningful subclasses. Every data in the subclass shares a common trait. It helps a user to understand the natural grouping or structure in a data set.

7. What are the desirable properties of a clustering algorithm?

- Scalability
- Ability to deal with different data types
- Minimal requirements for domain knowledge to determine input parameters
- Interpretability and usability

8. What is decision tree?

- A decision tree is a simple representation for classifying examples. Decision tree learning is one of the most successful techniques for supervised classification learning. A decision tree or a classification tree is a tree in which each internal node is labeled with an input features.
- The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes.

9. List the advantages of decision tree.

- Decision tree can handle both nominal and numeric input attributes.
- Decision tree representation is rich enough to represent any discrete value classifier.
- Decision trees are3 capable of handling database that may have errors.
- Decision trees are capable of handling datasets that may have missing values.
- It is self-explanatory and when compacted they are also easy to follow.

10. List the disadvantages of decision tree

- Most of the algorithms require that the target attribute will have only discrete values.
- Most decision-tree algorithms only examine a single field at a time.
- Decision trees are prone to errors in classification problems with much class.
- As decision tree use the "divide and conquer" method, they tend to perform well if a few highly relevant attribute exists, but less so if many complex interactions are present.

11. What is supervised learning?

In supervised learning, both the inputs and the outputs are provided. The network then processes the inputs and compares its resulting outputs against the desired outputs. Errors are then propagated back through the system, causing the system to adjust the weights which control the network.

12. What is unsupervised learning?

In an unsupervised learning, the network adapts purely in response to its inputs. Such networks can learn to pick out structure in their input.

13. What is dendrogram?

Decompose data objects into a several levels of nested partitioning called a dendrogram. A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.