Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. What are the basic characteristics of web?

2. Distinguish between Information Retrieval and Web Search.

3. What are steps involved in TF-IDF weighting?

4. Is it necessary to do query expansion always? Why?

5. List the parameters to identify spams.

6. State Heap's law.

7. How does relevance scoring works in web search.

8. What is Collaborative filtering? Give an example.

9. Define Text Mining. Give a suitable example.

10. Write the pros and cons in using classification algorithms over clustering approaches in text mining.

PART B — (5 × 13 = 65 marks)

11. (a) Explain the process of Information Retrieval and the components involved in it with a neat architecture.

Or

(b) Explain the process of web search and the components of a search engine.

12. (a) Describe Boolean retrieval model with an example.

Or

(b) (i) Write short notes on Latent Semantic Indexing with clear example. (7)

(ii) What is meant by Inverted Indices? Explain. (6)

13. (a) Describe in detail about Web Crawling with a neat architecture diagram.

Or

(b) Explain about the vector space model in XML retrieval. Write in detail about the evaluation of the data retrieved.

14. (a) (i) Explain the working of Hyperlink – Induced Topic search (HITS) algorithm with an example. (7)

(ii) How to implement a Map Reduce function through Hadoop? (6)

Or

(b) Explain the use of snippet generation, summarization and Question answering in web search and retrieval. (13)

15. (a) Explain in detail the various steps involved in finding the root node and its leaf nodes through information gain value by means of an decision tree with suitable algorithms.

Or

(b) Explain the application of Expectation – Maximization (EM) algorithm in text mining. (13)

PART C — (1 × 15 = 15 marks)

16. (a) Design a simple sentiment analysis system considering a database of your choice with minimal records. Sketch the architecture and highlight the functions of the modules identified.

Or

(b) Demonstrate the use of K means clustering algorithm in grouping different documents. Assume a simple database with 3 types of documents. State the assumptions made.

———————