Unit-III Text categorization: Supervised text categorization algorithms, Naive Bayes, k Nearest Neighbor (kNN), Logistic Regression.

Unit-IV Text clustering: Clustering structure of a corpus of text documents, Hierarchical clustering, Centroid-based clustering. Topic Modeling- Latent Semantic Indexing (LSI).

Readings:

1. Ricardo Baeza – Yates, Berthier Ribeiro – Neto, **Modern Information Retrieval: The concepts and Technology behind Search**, (ACM Press Books), Second Edition, 2011.

2. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze, **Introduction to Information Retrieval**, Cambridge University Press, First South Asian Edition, 2008.

3. Steven Struhl, **Practical Text Analytics: Interpreting Text and Unstructured Data for Business Intelligence**, Kogan Page, 2015.

4. Matthew A. Russell, **Mining the Social Web**, O'Reilly Media, 2013.