III B. Tech I Semester	INFORMATION RETRIEVAL LAB	L	T	P	C
		0	0	3	1.5

Course Outcomes: On completion of this course, the student will be able to

- Compute the similarity bet ween text documents
- Apply all pre-processing steps for text-data
- Implement classification of text documents.
- Perform document clustering using different algorithms.
- Implement PageRank algorithm for any network.

Programming Language: Python/R

Lab Experiments:

1. Representation of a Text Document in Vector Space Model and Computing Similarity between two documents.
2. Pre-processing of a Text Document: stop word removal and stemming
3. Construction of an Inverted Index for a given document collection comprising of at least 50 documents with a total vocabulary size of at least 1000 words.
4. Classification of a set of Text Documents into known classes (You may use any of the Classification algorithms like Naive Bayes, Max Entropy, Rochio's, Support Vector Machine). Standard Datasets will have to be used to show the results.
5. Text Document Clustering using K-means. Demonstrate with a standard dataset and compute performance measures- Purity, Precision, Recall and F-measure.
6. Crawling/ Searching the Web to collect news stories on a specific topic (based on user input). The program should have an option to limit the crawling to certain selected websites only.
7. To parse XML text, generate Web graph and compute topic specific page rank
8. Implement Matrix Decomposition and LSI for a standard dataset.
9. Mining Twitter to identify tweets for a specific period (and/or from a geographical location) and identify trends and named entities.
10. Implementation of PageRank on Scholarly Citation Network.