SUBJECT NO-CS60092, SUBJECT NAME- INFORMATION RETRIEVAL

LTP- 3-0-0,CRD- 3

SYLLABUS :-

Introduction to Information Retrieval: The nature of unstructured and semi-structured text. Inverted index and Boolean queries.Text Indexing, Storage and Compression: Text encoding: tokenization, stemming, stop words, phrases, index optimization. Index compression: lexicon compression and postings. lists compression. Gap encoding, gamma codes, Zipfs Law. Index construction. Postings size estimation, merge sort, dynamic indexing, positional indexes, n-gram indexes, real-world issues.Retrieval Models: Boolean, vector space, TFIDF, Okapi, probabilistic, language modeling, latent semantic indexing. Vector space scoring. The cosine measure. Efficiency considerations. Document length normalization. Relevance feedback and query expansion. Rocchio.Performance Evaluation: Evaluating search engines. User happiness, precision, recall, F-measure. Creating test collections: kappa measure, interjudge agreement.Text Categorization and Filtering: Introduction to text classification. Naive Bayes models. Spam filtering. Vector space classification using hyperplanes; centroids; k Nearest Neighbors. Support vector machine classifiers. Kernel functions. Boosting.Text Clustering: Clustering versus classification. Partitioning methods. k-means clustering. Mixture of Gaussians model. Hierarchical agglomerative clustering. Clustering terms using documents.Advanced Topics: Summarization, Topic detection and tracking, Personalization, Question answering, Cross language information retrievalWeb Information Retrieval: Hypertext, web crawling, search engines, ranking, link analysis, PageRank, HITS, XML and Semantic web.References1.Manning, Raghavan and Schutze, Introduction to Information Retrieval, Cambridge University Press.2.Baeza-Yates and Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley.3.Soumen Charabarti, Mining the Web, Morgan-Kaufmann.4.Survey by Ed Greengrass available in the Internet.