Information Extraction and Retrieval.

Theory L/T (Hours per week): 4/0, Credit: 4

Module-I:

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, Zipf'sLaw.Index construction. Postings size estimation, merge sort, dynamic indexing, positional indexes, n-gram indexes, real-world issues.

Module-II:

Retrieval ModelsBoolean, vector space, TFIDF, Okapi, probabilistic, language modeling, latent semantic indexing.Vector space scoring.Thecosine measurEfficiencyconsiderations. Document length normalization.Relevance feedback and query expansion.Rocchio.
Performance EvaluationEvaluating search engines.User happiness, precision, recall, F-measure. Creating test collections: kappa measure, interjudge agreement.

Module-III:

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 ClusteringClustering versus classification.Partitioningmethods.k-means clustering. Mixture of gaussiansmodel.Hierarchical agglomerative clustering.Clustering terms using documents.

Module-IV:

Advanced TopicsSummarization, Topic detection and tracking, Personalization, Question answering, Cross language informtion retrievalWeb Information RetrievalHypertext, web crawling, search engines, ranking, link analysis, PageRank, HITS.Retrieving Structured DocumentsXML retrieval, semantic web

Textbooks:

Introduction to Information Retrieval Manning, Raghavan and Schutze, Cambridge University Press, draft.
Modern Information Retrieval Baeza-Yates and Ribeiro-Neto, Addison Wesley, 1999.
A comprehensive survey by Ed GreengrassMining the Web, SoumenCharabarti, Morgan-Kaufmann, 2002.