Course Objectives: To have an advanced level of understanding of common and emerging methods of organizing, summarizing, and analyzing large collections of unstructured and lightly-structured text.


Informational Retrieval: Query processing models. Probabilistic models (Binary independence model, Robertson/Spark Jones weighting formula, Two-Poisson model), Relevance feedback (Term selection, Pseudo relevance feedback); Language models: Unigram, Bigram language models, Generating queries from documents, Language models and smoothing, Ranking with language models, Kullback-Leibler divergence, Divergence from randomness, Passage retrieval and ranking.


Types of information retrieval systems: Web retrieval and mining, Semantic web, XML information retrieval, Recommender systems and expert locators, Knowledge management systems, Decision support systems, Geographic information system (GIS).

Indexing: Inverted indices, Index components and Index life cycle, Interleaving Dictionary and Postings lists, Index construction, Query processing for ranked retrieval, Compression: General-purpose data compression, Symbol-wise data compression, Compressing posting lists, Compressing the dictionary.

Information categorization and filtering: Classification, Probabilistic classifiers, linear classifiers, Similarity-based classifiers, Multi category ranking and classification, learning to rank, Introduction to the clustering problem, Partitioning methods, Clustering versus classification, Reduced dimensionality/spectral methods.

Sentiment Analysis: Introduction to sentiment analysis, Document-level sentiment analysis, Sentence-level sentiment analysis, Aspect-based sentiment analysis, Comparative sentiment analysis, baseline algorithm, Lexicons, Corpora, Tools of Sentiment analysis, Applications.

Laboratory Work: In Laboratory Assignments students can learn search engines and common open-source software to perform common methods of exploratory and predictive analysis and apply text analysis techniques discussed in class to solve problems of data analysis.

Recommended Books