GUJARAT TECHNOLOGICAL UNIVERSITY

Bachelor of Engineering
Subject Code: 3170718
INFORMATION RETRIEVAL
7th Semester

Type of course: Elective

Prerequisite: Basic mathematics background is also required. You are supposed to be familiar basic concepts of probability (e.g., Bayes's theorem), linear algebra (e.g., vector, matrix and inner product).

Rationale: Information Retrieval (IR) systems give access to large amounts of online information stored as text, images, speech or video, e.g., Web documents. IR systems should only retrieve those documents that are relevant to a user's interest but have to deal with the uncertainty of describing what a document is about and what a user is actually interested in.

Teaching and Examination Scheme:

Teaching Scheme			Credits	Examination Marks				Total Marks
				Theory Marks		Practical Marks		
L	T	P	C	ESE (E)	PA (M)	ESE (V)	PA (I)	
3	0	0	3	70	30	0	0	100

Syllabus:

Sr. No.	Content	Total Hrs
1	**Introduction to Information Retrieval:** The nature of unstructured and semi-structured text. Inverted index and Boolean queries.	5
2	**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's Law. Index construction. Postings size estimation, merge sort, dynamic indexing, positional indexes, n-gram indexes, real-world issues.	7
3	**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.	7
4	**Performance Evaluation:** Evaluating search engines. User happiness, precision, recall, F-measure. Creating test collections: kappa measure, interjudge agreement.	4
5	**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.	5
6	**Text Clustering:** Clustering versus classification. Partitioning methods. k-means clustering. Mixture of Gaussians model. Hierarchical agglomerative clustering. Clustering terms using documents.	6
7	**Advanced Topics:** Summarization, Topic detection and tracking, Personalization, Question answering, Cross language information retrieval	6
8	**Web Information Retrieval:** Hypertext, web crawling, search engines, ranking, link analysis, PageRank, HITS.	5
9	**Retrieving Structured Documents:** XML retrieval, semantic web	3

Suggested Specification table with Marks (Theory):

Distribution of Theory Marks					
R Level	U Level	A Level	N Level	E Level	C Level
10	15	30	20	20	5

Legends: R: Remembrance; U: Understanding; A: Application, N: Analyze and E: Evaluate C: Create and above Levels (Revised Bloom's Taxonomy)

Note: This specification table shall be treated as a general guideline for students and teachers. The actual distribution of marks in the question paper may vary slightly from above table.

Reference Books:

1. Introduction to Information Retrieval. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2007.
2. Search Engines: Information Retrieval in Practice. Bruce Croft, Donald Metzler, and Trevor Strohman, Pearson Education, 2009.
3. Modern Information Retrieval. Baeza-Yates Ricardo and Berthier Ribeiro-Neto. 2nd edition, Addison-Wesley, 2011.
4. Information Retrieval: Implementing and Evaluating Search Engines. Stefan Buttcher, Charlie Clarke, Gordon Cormack, MIT Press, 2010.

Course Outcome:

After learning the course, the students should be able to:

Sr. No.	CO Statement	Marks % Weightage
1	understand the theoretical basis behind the standard models of IR (Boolean, Vector-space, Probabilistic and Logical models)	35%
2	apply appropriate method of text classification or clustering.	30%
3	use performance evaluation metric for IR	15%
4	understand the standard methods for Web indexing and retrieval	20%