

GANPAT UNIVERSITY									
FACULTY OF TECHNOLOGY									
Programme	Bachelor of Technology				Branch/Spe c.	Computer Science Engineering (BDA)			
Semester	V				Version	1.0.1.0			
Effective from Academic Year		2018-19			Effective for the batch Admitted in			June 2016	
Subject code	2CSE50E18	Subject Name			Information Retrieval				
Teaching scheme					Examination scheme (Marks)				
(Per week)	Lecture(DT)		Practical(Lab.)		Total		CE	SEE	Total
	L	TU	P	TW					
Credit	3	0	1	0	4	Theory	20	30	50
Hours	3	0	2	0	5	Practical	40	60	100
Pre-requisites:									
Database Management System									
Learning Outcome:									
<ul style="list-style-type: none"> • Explain the concepts of indexing, vocabulary, normalization and dictionary in Information Retrieval • Define a boolean model and a vector space model, and explain the differences between them • Explain the differences between classification and clustering • Discuss the differences between different classification and clustering methods • Choose a suitable classification or clustering method depending on the problem constraints at hand • Implement classification in a boolean model and a vector space model • Implement a basic clustering method • Give account of a basic spectral method • Evaluate information retrieval algorithms, and give an account of the difficulties of evaluation • Explain the basics of XML and Web search. 									
Theory syllabus									
Unit	Content								Hrs
1	Introduction Basics of Information Retrieval and Introduction to Search Engines; Boolean Retrieval-: Boolean queries, Building simple indexes, Processing Boolean queries								5
2	Term Vocabulary and Posting Lists Choosing document units, Selection of terms, Stop word elimination, Stemming and lemmatization, Skip lists, Positional postings and Phrase queries; Dictionaries and Tolerant Retrieval: Data structures for dictionaries, Wildcard queries, Permuterm and K-gram indexes, Spelling correction, Phonetic correction								6
3	Index Construction Single pass scheme, Distributed indexing, Map Reduce, Dynamic indexing; Index Compression - Statistical properties of terms, Zipf's law, Heap's law, Dictionary compression, Postings file compression, Variable byte codes, Gamma codes								6
4	Vector Space Model Parametric and zone indexes, Learning weights, Term frequency and weighting, Tf-Idf weighting, Vector space model for scoring, variant tf-idf functions								6
5	Computing Scores in a Complete Search System Efficient scoring and ranking, Inexact retrieval, Champion lists, Impact ordering, Cluster pruning, Tiered indexes, Query term proximity, Vector space scoring and query operations								6

6	Evaluation in Information Retrieval Standard test collections, unranked retrieval sets, Ranked retrieval results, Assessing relevance, User utility, Precision and Recall, Relevance feedback, Rocchio algorithm, Probabilistic relevance feedback, Evaluation of relevance feedback	5
7	Probabilistic Information Retrieval Review of basic probability theory, Probability ranking principle, Binary independence model, Probability estimates, probabilistic approaches to relevance feedback. Text Classification-Rocchio classifier, KNearestneighbor classifier, Linear and nonlinear classifiers, Bias-variance tradeoff, Naïve Bayes and Support Vector machine based classifiers	6
8	Text Clustering Clustering in information retrieval, Evaluation of clustering, KMeans and Hierarchical clustering. Introduction to Linear Algebra, Latent Semantic Indexing	5
Practical content		
<ul style="list-style-type: none"> • Labs will be implemented so as to gain the practical learning of dictionaries and tolerant Retrieval: Data structures for dictionaries and wildcard queries. • Practical will be based on Further labs will focus on concepts like distributed indexing, map Reduce, dynamic indexing and vector space model. • Practical will be based on Single pass scheme, • Practical will be based on Distributed indexing, • Practical will be based on Map Reduce, • Practical will be based on Dynamic indexing; • Practical will be based on Index Compression - Statistical properties of terms, • Practical will be based on Zipf's law, • Practical will be based on Heap's law, • Practical will be based on Dictionary compression, • Practical will be based on Postings file compression, • Practical will be based on Variable byte codes, Gamma codes 		
Text Books		
1	C. D. Manning, P. Raghavan, and H. Schütze, An Introduction to Information Retrieval, Cambridge University Press, 2009.	
Reference Books		
1	1. R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, Pearson Education, 1999.	