WEB AND TEXT MINING (Major Elective-II)

Semester II (Computer Engineering)
SUB CODE: MECE206-A

TEACHING SCHEME (Credits and Hours):

Teaching scheme				Total Credit	Evaluation Scheme					
L	T	P	Total		Theory		Mid Sem Exam	CIA	Pract.	Total
Hrs	Hrs	Hrs	Hrs		Hrs	Marks	Marks	Marks	Marks	Marks
04	00	02	06	05	3	70	30	20	30	150

LEARNING OBJECTIVES:

The objective of this course is

- To enable students to learn and implement various information retrieval models
- To enable students to understand and implement various text mining algorithms
- To enable students to understand and implement link analysis algorithms
- To enable students to understand and implement recommender systems

OUTLINE OF THE COURSE:

Unit No	Topics
1	Information Retrieval and Web Search
2	Web Crawling, index construction, index compression
3	Link analysis: HITS, PageRank
4	Text classification: naive bayes, vector space, support vector machines
5	Tex clustering: k-mean, Hierarchical
6	Recommender systems

Total hours (Theory): 60

Total hours (Practical): 30

Total hours: 90

DETAILED SYLLABUS:

Sr. No	Topic	Lecture Hours	Weight age (%)
1	**Information retrieval and web search** • Basic concepts of IR • IR models: boolean, vector space, probabilistic • Relevance feedback • Text and webpage pre-processing	08	15
2	**Web crawling, index construction** • Crawler construction • Inverted index, posting lists • Latent semantic analysis • Index compression	10	15
3	**Link analysis:** • HITS • PageRank • Community Discovery	10	15
4	**Text classification** • Naive-bayes • Vector space models • Topic models • **Support Vector Machines**	16	25
5	**Text Clustering:** • k-means and other centroid-based schemes • Hierarchical methods	08	20

6	**Recommender Systems** • Collaborative filtering: k-nn • Collaborative filtering: association rules • Collaborative filtering: matrix factorization	08	10

INSTRUCTIONAL METHOD AND PEDAGOGY (Continuous Internal Assessment (CIA) Scheme)

- At the start of course, the course delivery pattern, prerequisite of the subject will be discussed.
- Lectures will be conducted with the aid of multi-media projector, black board, OHP etc.
- Attendance is compulsory in lecture and laboratory which carries 10 marks in overall evaluation.
- One internal exam will be conducted as a part of internal theory evaluation.
- Assignments based on the course content will be given to the students for each unit and will be evaluated at regular interval evaluation.
- Surprise tests/Quizzes/Seminar/tutorial will be conducted having a share of five marks in the overall internal evaluation.
- The course includes a laboratory, where students have an opportunity to build an appreciation for the concepts being taught in lectures.
- Experiments shall be performed in the laboratory related to course contents.

STUDENTS LEARNING OUTCOMES:

On successful completion of the course, the student will:

- Be able to learn and implement various information retrieval models
- Be able to understand and implement various text mining algorithms
- Be able to understand and implement link analysis algorithms
- Be able to understand and implement recommender systems

REFERENCE BOOKS:

1. Introduction to information retrieval – Prabhakar Raghwan, Chris Manning
2. Speech and Natural language processing – Daniel Jurafsky, Martin

LIST OF PRACTICALS:

Sr. No	Name of Experiment
1	Implement a small-scale web information retrieval system
2	Implement recommender systems
3	Implement svm-based text-classification system
4	Implement hierarchical agglomerative clustering for web-pages