

MINING MASSIVE DATASETS

I Semester: CSE																							
Course Code	Category	Hours / Week			Credits	Maximum Marks																	
BCSC09	Foundation	L	T	P	C	CIA	SEE	Total															
		3	0	0	3	30	70	100															
Contact Classes: 45		Tutorial Classes: Nil		Practical Classes: Nil		Total Classes: 45																	
<p>I. COURSE OVERVIEW: This course is based on text mining of massive data sets and their applications. Topics include map reduce and the new software stack, applications of similarity search, implementation of stream data, link analysis, handling large data set's, clustering, issues in online advertising, recommendation systems and mining social network graphs.</p> <p>II. OBJECTIVES: The course should enable the students to:</p> <ul style="list-style-type: none"> I This course will cover practical algorithms for solving key problems in mining of massive datasets. II This course focuses on parallel algorithmic techniques that are used for large datasets. III This course will cover stream processing algorithms for data streams that arrive constantly, page ranking algorithms for web search, and online advertisement systems that are studied indetail. <p>III. COURSE OUTCOMES: After successful completion of the course, students should be able to:</p> <table border="1" style="width: 100%;"> <tbody> <tr> <td>CO 1</td> <td>Handle massive data using Map Reduce</td> <td>Understand</td> </tr> <tr> <td>CO 2</td> <td>Develop and implement algorithms for massive data sets and methodologies in the context of data mining.</td> <td>Apply</td> </tr> <tr> <td>CO 3</td> <td>Understand the algorithms for extracting models and information from large datasets</td> <td>Understand</td> </tr> <tr> <td>CO 4</td> <td>Develop recommendation systems</td> <td>Apply</td> </tr> <tr> <td>CO 5</td> <td>Gain experience in matching various algorithms for particular classes of problems.</td> <td>Apply</td> </tr> </tbody> </table> <p>IV. SYLLABUS:</p> <p>MODULE-I: DATA MINING (09) Introduction-Definition of Data Mining-Statistical Limits on Data Mining. Map Reduce and the New Software Stack-Distributed File Systems, Map Reduce, Algorithms Using Map Reduce.</p> <p>MODULE-II: SIMILARITY SEARCH (8) Finding Similar Items-Applications of Near-Neighbor Search, Shingling of Documents, Similarity-Preserving Summaries of Sets, Distance Measures. Streaming Data: Mining Data Streams-The Stream Data Model, Sampling Data in a Stream, Filtering Streams.</p> <p>MODULE-III: LINK ANALYSIS (10) Page Rank, Efficient Computation of Page Rank, Link Spam. Frequent Item sets-Handling Larger Datasets in Main Memory, Limited-Pass Algorithms, Counting Frequent Items in a Stream. Clustering-The CURE Algorithm, Clustering in Non-Euclidean Spaces, Clustering for Streams and Parallelism.</p>									CO 1	Handle massive data using Map Reduce	Understand	CO 2	Develop and implement algorithms for massive data sets and methodologies in the context of data mining.	Apply	CO 3	Understand the algorithms for extracting models and information from large datasets	Understand	CO 4	Develop recommendation systems	Apply	CO 5	Gain experience in matching various algorithms for particular classes of problems.	Apply
CO 1	Handle massive data using Map Reduce	Understand																					
CO 2	Develop and implement algorithms for massive data sets and methodologies in the context of data mining.	Apply																					
CO 3	Understand the algorithms for extracting models and information from large datasets	Understand																					
CO 4	Develop recommendation systems	Apply																					
CO 5	Gain experience in matching various algorithms for particular classes of problems.	Apply																					

MODULE-IV: ADVERTISING ON THE WEB (09)

Issues in On-Line Advertising, On-Line Algorithms, The Matching Problem, The Adwords Problem, Adwords Implementation.

Recommendation Systems-A Model for Recommendation Systems, Content-Based Recommendations, Collaborative Filtering, Dimensionality Reduction, The NetFlix Challenge.

MODULE-V: MINING SOCIAL-NETWORK GRAPHS: (09)

Mining Social-Network Graphs - Social Networks as Graphs, Clustering of Social-Network Graphs,

Partitioning of Graphs, Simrank, Counting Triangles.

V. TEXT BOOKS:

1. Jure Leskovec, Anand Rajaraman, Jeff Ullman, Mining of Massive Datasets, 3rd Edition.

VI. REFERENCE BOOKS:

1. Jiawei Han & Micheline Kamber, Data Mining – Concepts and Techniques 3rd Edition Elsevier.
2. Margaret H Dunham, Data Mining Introductory and Advanced topics, PEA.
3. Ian H. Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann