CSE 3152: Mining of Massive Datasets

Programme: B.Tech. (CSE)	Year: 3	Semester: VI
Course: Program Elective	Credits: 3	Hours: 40

Course Context and Overview:

Big data is transforming the world. But at the same time, it is so large that it does not fit in main memory. Thus it requires a different way to handle this massive data and mine it to get useful patterns out of it. This course will take an algorithmic view of applying algorithms to massive data rather than using data to train a machine learning model.

Prerequisite Courses:

Design and Analysis of Algorithms, Introduction to Data Science

Course Outcomes (COs):

On completion of this course, the students will have the ability to:
CO1 Understand massive data from different domains such as market basket, social media, web etc.
CO2 Apply MapReduce framework for creating parallel algorithms that succeed on very large amounts of data
CO3 Analyze algorithms applicable for massive data and results obtained from such implementation

Course Topics:

Contents	Lecture Hours
UNIT 1 **Introduction**	2
Data Mining, Statistical Modeling, Machine Learning, Computational Approaches to Modeling, Statistical Limits on Data Mining	

UNIT 2 **MapReduce**	5
Distributed File Systems, MapReduce Paradigm, Extensions to MapReduce, Complexity Theory for MapReduce	

UNIT 3 **Similar Items**	8
Near Neighbor Search, Shingling of Documents, Similarity-Preserving Summaries of Sets, Locality-Sensitive Hashing, Distance Measures, Methods for High Degrees of Similarity	
UNIT 4 **Mining Data Streams**	5
The Stream Data Model, Sampling, Filtering, Distinct Elements, Estimating Moments, Counting Ones in a Window, Decaying Windows	
UNIT 5 **Frequent Itemsets**	7
Market-Basket Model, A-Priori Algorithm, Multistage/Multihash/Limited-Pass Algorithms, Frequent Items in Streams	
UNIT 6 **Advertising on the Web**	4
Issues in On-Line Advertising, On-Line Algorithms, The Matching Problem, The Adwords Problem, Adwords Implementation	
UNIT 7 **Mining Social-Network Graphs**	4
Social-Networks as Graphs, Discovery of communities, Partitioning of Graphs, Finding Overlapping Communities	
UNIT 8 **Large-Scale Machine Learning**	5

The Machine Learning Model, Perceptrons, Support Vector Machines, Learning from Nearest Neighbors, Comparison of Learning Methods	

Textbook references:

Text Books:

· J. Leskovec, A. Rajaraman, J.D. Ullman: "Mining of Massive Datasets," Cambridge University Press, 2nd Edition, 2014

Reference books:

· P. Tan, M. Steinbach, V. Kumar: "Introduction to Data Mining," Addison-Wesley, 2006 · M.J. Zaki & W. Meira Jr.: "Data Mining and Analysis-Fundamental Concepts and Algorithms," Cambridge University Press, 2014

Evaluation Methods:

Component	***Weightage (%)***
Continuous evaluation (Quiz, Assignment)	35
Mid term	25
End term	40

Prepared By: Subrat K Dash

Last Update: 30th Nov 2018