INSTITUTE OF AERONAUTICAL ENGINEERING
(Autonomous)
Dundigal - 500 043, Hyderabad, Telangana

COURSE CONTENT

MINING MASSIVE DATASETS								
II Semester: CSE								
Course Code	**Category**	**Hours / Week**			**Credits**	**Maximum Marks**		
		L	**T**	**P**	**C**	**CIA**	**SEE**	**Total**
BCSD20	**Elective**	3	-	-	3	40	60	100
Contact Classes:48	**Total Tutorials: Nil**	**Total Practical Classes: Nil**				**Total Classes: 48**		
Prerequisite: Data Science, Data Mining								

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 sets, clustering, issues in online advertising, recommendation systems and mining social network graphs.

II. COURSE OBJECTIVES:
The students will try to learn:

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 in detail.

III.COURSE OUTCOMES:
After successful completion of the course, students will be able to:

CO 1 Apply MapReduce for massive data analysis
CO 2 Develop and implement algorithms for massive data sets and methodologies in the context of data mining.
CO 3 Understand the algorithms for extracting models and information from large datasets
CO 4 Understand the model of recommendation systems and its applications
CO 5 Gain experience in matching various algorithms for particular classes of problems.

IV. COURSE CONTENT:
MODULE-I: DATA MINING (10)
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.

MODULE-II: SIMILARITY SEARCH (10)
Finding Similar Items-Applications of Near-Neighbor Search, Shingling of Documents, Similarity Preserving Summaries of Sets, and Distance Measures.
Streaming Data: Mining Data Streams-The Stream Data Model, Sampling Data in a Stream, Filtering Streams.

MODULE-III: LINK ANALYSIS (09)
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..

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 (10)
Mining Social-Network Graphs - Social Networks as Graphs, Clustering of Social-Network Graphs, Partitioning of Graphs, Sim rank, Counting Triangles.

V. TEXT BOOKS:

1. Jure Leskovec, Anand Rajaraman, Jeff Ullman, Mining of Massive Datasets, 3rd edition, 2010.

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.

VII. WEB REFERENCES:

1. http://www.mmds.org/
2. https://web.stanford.edu/class/cs246/

VIII E-Text Books:

1. http://infolab.stanford.edu/~ullman/mmds/book.pdf
2. https://www.e-booksdirectory.com/details.php?ebook=5300
3. https://www.kdnuggets.com/2010/12/book-mining-massive-datasets.html