Shirpur Education Society's

R. C. Patel Institute of Technology, Shirpur

(An Autonomous Institute)

Course Structure and Syllabus

Final Year B. Tech

Computer Science and Engineering (Data Science)

With effect from Year 2023-24

Shahada Road, Near Nimzari Naka, Shirpur, Maharashtra 425405
Ph: 02563 259 802, Web: www.rcpit.ac.in

Machine Learning - IV(PCCS7010T)

Teaching Scheme	**Examination Scheme**
Lectures : 03 Hrs./week	Term Test : 15 Marks
Credits : 03	Teacher Assessment : 20 Marks
	End Sem Exam : 65 Marks
	Total Marks : 100 Marks

Prerequisite:Knowledge of

1. Basic Machine Learning
2. Database Management System

Course Objectives:

To teach advance concepts of data management and data analysis for Big Data.

CO	**Course Outcomes**	**Blooms Level**	**Blooms Description**
CO1	Evaluate the need of MapReduce framework.	L5	Evaluate
CO2	Apply appropriate method to handle big data.	L3	Apply
CO3	Apply suitable analysis method to draw conclusions from given big data.	L3	Apply

Course Contents

Unit-I — 04 Hrs.

Map-Reduce: The Map Tasks, Grouping by Key, The Reduce Tasks, Combiners, Details of Map-Reduce Execution, Coping with Node Failure; Algorithms using MapReduce: Matrix-Vector multiplication by MapReduce, Selection, Projection, Natural Join, Union, Intersection, Difference, Matrix Multiplication.

Unit-II — 08 Hrs.

Mining Data Stream: The Stream Model, Sampling Data in a Stream, Filtering Streams: The Bloom Filter; Counting distinct element in the Stream: The Count-Distinct Problem, The Flajolet-Martin Algorithm; Estimating Moments: The Alon-Matias-Szegedy Algorithm for Second Moments, Higher-Order Moments, Dealing with Infinite Streams; Counting ones in a window: The cost of exact count, The DGIM algorithm, Storage Requirement, Query Answering.

Unit-III — 08 Hrs.

Link Analysis: PageRank: Search Engine, Term Spam, PageRank, Structure of Web, Avoiding Dead Ends, Spider Traps and Taxation, Efficient Computing of PageRank: Transition Matrices, Iteration using MapReduce, Topic Sensitive PageRank: Biased Random Walk, Using Topic Sensitive PageRank, Inferring Topics from Words. Link Spam: Architecture, Analysis of Spam Farm, Combating Link Spam, Trust Rank, Spam Mass. Hubs and Authorities, HITs Algorithm.

Unit-IV — 06 Hrs.

Frequent Itemsets: The Market-Basket model: Association Rules, A-Priori, Representation of Market-Basket Data, Monotonicity of itemset, Handling Larger Datasets in Main Memory: The Multistage Algorithm, The Multihash Algorithm, Limited-Pass Algorithms: Randomized Algorithm, Avoiding error in sampling algorithms, Counting Frequent Itemsets in a Stream: Frequent Itemsets in a decaying window.

Unit-V — 06 Hrs.

Clustering: Clustering Strategies, The Curse of Dimensionality, Hierarchal Clustering in a Euclidean Space and Non-Euclidean Spaces, The CURE Algorithm: Initialization, Completion, Representing Clusters in a GRGPF Algorithm, Initializing Cluster Tree, Adding Points, Splitting and Merging Clusters. Clustering for Streams and Parallelism: The Stream-Computing Model, Initializing, Merging Buckets, Answering Queries.

Unit-VI 07 Hrs.

Social Network Analysis: Social Networks as Graphs, Clustering of Social Network Graphs: Distance Measure, Betweenness, The Girvan- Newman Algorithm, Betweenness to find communities, Direct discoveries of Communities: Finding Cliques, Complete Bipartite Graphs, Partition of Graphs: Normalized Cuts, Contents: Finding overlapping communities, Maximum Likelihood Estimation, The Affiliation Graph Model, SimRank: Random walkers on Social Media, Approximate SimRank, Counting Triangles, Neighborhood Properties of a graph.

Text Books:

1. Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman,"Mining of Massive Datasets", Stanford Press, 2020.
2. Donald Miner, Adam Shook, "MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems", OReilly, 2013.

Reference Books:

1. Suk-Man Ivy Tong, "Techniques in Data Stream Mining", Open Dissertation Press, 2017.
2. Leszek Ruthowski, Maciej Jawordki, Plotr Duda, "Stream Data Mining: Algorithms and Their Probabilistic Properties", Springer, 2019
3. A biefet, "Adaptive Stream Mining: Pattern Learning and Mining from Evolving Data Stream", IoS Press, 2010
4. Amy N Langville, Carl D. D. Meyer, "Google's PageRank and Beyond: The Science of Search Engine Rankings", Princeton University Press 2011.
5. Dr. Chandrashekhar Raghuvanshi, Dr. Hari Om Sharan, "Frequent Pattern Mining in Large Databases", AkiNik Publication, 2022.
6. Tanmoy Chakraborty, "Social Network Analysis", Wiley Publication, 2021.

Web Links:

1. Concept Drift: https://ebrary.net/199293/engineering/sampling_data_streams
2. Search Engine: https://moz.com/blog/search-engine-algorithm-basics

Evaluation Scheme:

Theory :

Continuous Assessment (A):

Subject teacher will declare Teacher Assessment criteria at the start of semester.

Continuous Assessment (B):

1. Two term tests of 15 marks each will be conducted during the semester.

2. Best of the marks scored in both the tests will be considered for final grading.

End Semester Examination (C):

1. Question paper based on the entire syllabus, summing up to 65 marks.

2. Total duration allotted for writing the paper is 3 hrs.

Machine Learning - IV Laboratory (PCCS7010L)

Practical Scheme
Practical : 02 Hrs./week
Credit : 01

Examination Scheme
Teacher Assessment : 25 Marks
End Sem Exam : 25 Marks
Total : 50 Marks

Course Objectives:

1. Understand MapReduce.
2. Perform Video Summarization and Community detection.

CO	Course Outcomes	Blooms Level	Blooms Description
CO1	Perform, Implement and Execute Matrix Multiplication, sorting using MapReduce.	L6	Create
CO2	Apply MapReduce for Market basket Analysis.	L3	Apply
CO3	Video Summarization using Cure Algorithm and Community detection using Girvan	L2	Understand
CO4	Perform Similarity analysis using SimRank.	L3	Apply

List of Laboratory Experiments

Suggested Experiments:

1. Execute Matrix Multiplication using MapReduce.
2. Perform Sorting using MapReduce.
3. Implement Bloom Filter using MapReduce.
4. Approximate the number of unique elements in a data stream or database in one pass using Flajolet-Martin Algorithm.
5. Compute stochastic matrix from a given graph, compute PageRank vector and return the results.
6. Identify which page belongs to Link farm in a given graph. Compute trustrank vector.
7. Perform Market-Basket analysis using MapReduce.
8. Video Summarization using Cure Algorithm.
9. Community detection using Girvan- Newman Algorithm.
10. Similarity analysis using SimRank.

Minimum eight experiments from the above suggested list or any other experiment based on syllabus will be included, which would help the learner to apply the concept learnt.

Evaluation Scheme:

Laboratory:

Continuous Assessment (A):

Laboratory work will be based on PCCS7010T with minimum 08 experiments to be incorporated. The distribution of marks for term work shall be as follows:

1. Performance in Experiments: 05 Marks
2. Journal Submission: 05 Marks
3. Viva-voce: 05 Marks
4. Subject Specific Lab Assignment/Case Study: 10 Marks

The final certification and acceptance of term work will be subject to satisfactory performance of laboratory work and upon fulfilling minimum passing criteria in the term work.

End Semester Examination (C):

Oral/ Practical examination will be based on the entire syllabus including, the practicals performed during laboratory sessions.