SRM VALLIAMMAI ENGINEERING COLLEGE

(An Autonomous Institution)

SRM Nagar, Kattankulathur – 603 203

DEPARTMENT OF INFORMATION TECHNOLOGY

QUESTION BANK

II SEMESTER (M. TECH-DATA SCIENCE)

DS3263 - BIG DATA MINING AND ANALYTICS

Regulation – 2023

Academic Year 2023-2024 (Even Semester)

Prepared by

Ms. S. PRIYA, Assistant Professor (O.G) / IT

SRM VALLIAMMAI ENGINEERING COLLEGE

(An Autonomous Institution)

SRM Nagar, Kattankulathur – 603 203.

DEPARTMENT OF INFORMATION TECHNOLOGY

M. TECH-DATA SCIENCE

QUESTION BANK

SUBJECT : DS3263 - BIG DATA MINING AND ANALYTICS

SEM / YEAR : II Sem / I Year

UNIT I - DATA MINING AND LARGE-SCALE FILES

SYLLABUS: Introduction to Statistical modeling – Machine Learning – Computational approaches to modeling – Summarization – Feature Extraction – Statistical Limits on Data Mining - Distributed File Systems – Map-reduce – Algorithms using Map Reduce – Efficiency of Cluster Computing Techniques.

PART-A

Q.No.	Question	Level	Competence
1	What is big data?	BTL1	Remember
2	Define Data Mining or Data Dredging.	BTL1	Remember
3	Give any 2 approaches to modelling data.	BTL2	Understand
4	Define Summarization	BTL1	Remember
5	What is Page Rank Idea?	BTL4	Analyze
6	Give any 2 important kinds of feature extraction from large-scale data.	BTL2	Understand
7	List out the Statistical Limits on Data Mining.	BTL1	Remember
8	Give an Example of Bonferroni's Principle.	BTL4	Analyze
9	Illustrate the concepts of Distributed File Systems.	BTL4	Analyze
10	If we had to abort and restart the computation every time one component failed, then the computation might never complete successfully. Give any 2 solutions to overcome this problem.	BTL4	Analyze
11	Define Distributed File System.	BTL3	Apply
12	What is Map-Reduce?	BTL4	Analyze
13	List out the Algorithms Using MapReduce.	BTL2	Understand
14	Write any 4 Major relational-algebra operations used in Big Data Analytics.	BTL2	Understand
15	Define Map Function?	BTL3	Apply
16	Define Reduce Function?	BTL3	Apply
17	Give any 2 Major characteristics with MapReduce systems?	BTL6	Create
18	What is Pregel and Giraph?	BTL5	Evaluate
19	What is Communication Cost Model?	BTL6	Create
20	What is Multi-way and Star Joins?	BTL5	Evaluate
21	How to managing computing node failures?	BTL3	Apply
22	What are Recursive Workflows?	BTL5	Evaluate

23	Define Reducers.	BTL2	Understand
24	What is Replication Rate and Reducer size?	BTL2	Understand

PART-B

Q.No.	Question	Marks	Level	Competence
1	A famous instance of clustering to solve a problem took place long ago in London, and it was done entirely without computers. The physician John Snow, dealing with a Cholera outbreak plotted the cases on a map of the city. Draw the Plotting cholera cases on a map of London and give proper Justification.	16	BTL2	Understanding
2	Explain Bonferroni's Principle.	16	BTL4	Analyze
3	Give suitable example how "bogus" data was handled by Bonferroni's Principle.	16	BTL4	Analyze
4	Explain the Following: i) Physical Organization of Compute Nodes ii) Large-Scale File-System Organization	 10 6	BTL1	Remember
5	A common sort of data-mining problem involves discovering unusual events hidden within massive amounts of data. Justify with suitable example.	16	BTL6	Create
6	How a program using MapReduce is executed. Justify.	16	BTL4	Analyze
7	Explain Multiway Joins in Cluster-Computing Environment	16	BTL1	Remember
8	Explain the Following using Map-Reduce Algorithm. i) Union, Intersection, and Difference by MapReduce ii) Computing Projections by MapReduce iii) Computing Selections by MapReduce	 6 5 5	BTL1	Remember
9	Explain Communication – Cost model for Cluster Computing.	16	BTL2	Understand
10	Explain Summarization with Suitable Examples.	16	BTL3	Apply
11	Explain the following using Map-Reduce Algorithms i) If the Vector v Cannot Fit in Main Memory ii) Matrix-Vector Multiplication by MapReduce	 10 6	BTL1	Remember
12	What are the basic Relational-Algebra Operations used in Map-Reduce Algorithms?	16	BTL5	Evaluate
13	Explain Recursive Extensions to MapReduce with suitable examples.	16	BTL1	Understand
14	Compare Multiway Joins and Star Joins.	16	BTL3	Apply
15	Give the efficiency of Cluster-Computing Algorithms.	16	BTL6	Create
16	Explain Statistical Limits on Data Mining with suitable examples.	16	BTL6	Create
17	Explain Algorithms Using MapReduce with suitable examples	16	BTL4	Evaluate

UNIT II - SIMILAR ITEMS

SYLLABUS: Nearest Neighbor Search – Shingling of Documents – Similarity preserving summaries – Locality sensitive hashing for documents – Distance Measures – Theory of Locality Sensitive Functions – LSH Families – Methods for High Degree of Similarities.

PART-A

Q.No.	Question	Level	Competence
1	Define Jaccard Similarity of Sets	BTL1	Remember
2	What is Similarity of Documents?	BTL3	Apply
3	What is Plagiarism?	BTL5	Evaluate
4	What are Mirror Pages?	BTL3	Apply
5	Define collaborative filtering.	BTL1	Remember
6	What is Shingling of Documents?	BTL4	Analyze
7	Define a k-shingle for a document.	BTL1	Remember
8	How to Hashing Shingles?	BTL5	Evaluate
9	Define Minhashing.	BTL4	Analyze
10	Define Distance Measures.	BTL2	Understand
11	Give the formula for n-dimensional Euclidean Distance.	BTL2	Understand
12	What is Manhattan distance?	BTL1	Remember
13	Define Cosine Distance.	BTL4	Analyze
14	What is Jaccard Distance?	BTL1	Remember
15	What is Hamming Distance?	BTL4	Analyze
16	Define Sketches.	BTL2	Understand
17	Give Applications of Locality-Sensitive Hashing.	BTL1	Remember
18	What are the two versions of fingerprint comparison?	BTL1	Remember
19	What is Prefix Indexing?	BTL2	Understand
20	Define Edit Distance.	BTL3	Apply
21	What is efficient Minhashing?	BTL2	Understand
22	What is Locality Sensitive hashing for Signatures?	BTL1	Remember
23	How to Generalize Locality-Sensitive Hashing?	BTL4	Analyze
24	Define Character Indexes and Position Indexes.	BTL2	Understand

PART-B

Q.No.	Question	Marks	Level	Competence
1	Sets of shingles are large. To overcome this, explain the Similarity-Preserving Summaries of Sets.	16	BTL4	Analyze
2	Explain the Following i) Jaccard Distance ii) Euclidean Distances	 6 10	BTL4	Analyze
3	Explain the steps involved in Analysis of the Banding Technique.	16	BTL3	Apply
4	What are Random Hyperplanes and the Cosine Distance? How it is related to LSH Families? Justify.	16	BTL6	Create
5	Explain the Following i) Matrix Representation of Sets ii) Minhashing iii) Minhash Signatures iv) Minhashing and Jaccard Similarity	 4 4 4 4	BTL2	Understand
6	Briefly explain the Theory of Locality-Sensitive Functions.	16	BTL2	Understand
7	i) Consider the two-dimensional Euclidean space (the customary plane) and the points (2, 7) and (6, 4). Compute L2-norm, L1-norm & L∞-norm ii) Let our two vectors be x = [1, 2, −1] and = [2, 1, 1]. Find the dot product of x.y. Calculate the Cosine angle between x and y (if x has L2-norm √ 6).	8 8	BTL3	Apply
8	How to Compute Minhash Signatures?	16	BTL1	Remember
9	What is general approach to LSH for Minhash Signatures?	16	BTL1	Remember
10	What is Prefix Indexing and Length-Based Filtering?	16	BTL1	Remember
11	Explain the Following i) Cosine Distance ii) Edit Distance	 8 8	BTL4	Analyze
12	Explain LSH Families for Other Distance Measures.	16	BTL2	Understand
13	Explain the Following i) Hamming Distance ii) Non-Euclidean Spaces	 8 8	BTL4	Analyze
14	Explain the Following i) Entity Resolution & Matching Fingerprints ii) Matching Newspaper Articles	 8 8	BTL5	Evaluate
15	Explain Shingling of Documents	16	BTL5	Evaluate
16	Explain Locality-Sensitive Hashing for Documents with suitable examples.	16	BTL6	Create
17	Give a brief account of Applications of Locality-Sensitive Hashing.	16	BTL1	Remember

UNIT III - MINING DATA STREAMS

SYLLABUS: Stream Data Model – Sampling Data in the Stream – Filtering Streams – Counting Distance Elements in a Stream – Estimating Moments – Counting Ones in Window – Decaying Windows.

PART-A

Q.No.	Question	Level	Competence
1	What is Stream Data Model?	BTL1	Remember
2	How to Sampling of Streams?	BTL1	Remember
3	What are Bloom Filters?	BTL1	Remember
4	What is Counting Distinct Elements?	BTL4	Analyze
5	What are Moments of Streams?	BTL2	Understand
6	How to Estimating Second Moments?	BTL2	Understand
7	How to Estimating Higher Moments?	BTL6	Create
8	How to Estimating the Number of 1's in a Window?	BTL5	Evaluate
9	How to Answering Queries About Numbers of 1's?	BTL5	Evaluate
10	How to get Closer Approximations to the Number of 1's?	BTL4	Analyze
11	Define Decaying Window.	BTL1	Remember
12	What is Filtering? Give an example.	BTL1	Remember
13	How to Maintaining Frequent Elements in an Exponentially Decaying Window?	BTL3	Apply
14	What are the basic Storage Requirements for the DGIM Algorithm?	BTL3	Apply
15	How to Maintaining the DGIM Conditions?	BTL2	Understand
16	Give the use of Flajolet-Martin Algorithm.	BTL2	Understand
17	Give the use of The Count-Distinct Problem	BTL2	Understand
18	What is General Sampling Problem?	BTL5	Evaluate
19	What are the issues present in Stream Processing?	BTL3	Apply
20	Give an example for Datar-Gionis-Indyk-Motwani Algorithm.	BTL4	Analyze
21	What is the use of Datar-Gionis-Indyk-Motwani Algorithm?	BTL2	Understand
22	Define Datar-Gionis-Indyk-Motwani Algorithm.	BTL1	Remember
23	Define Flajolet-Martin Algorithm.	BTL1	Remember
24	What is General Stream Sampling Problem?	BTL5	Evaluate

PART-B

Q.No.	Question	Marks	Level	Competence
1	With a suitable diagram Explain Data-Stream-Management System.	16	BTL4	Analyze
2	Explain Stream Data Model with suitable examples.	16	BTL6	Create
3	Explain Filtering streams with suitable example.	16	BTL2	Understand
4	What are Stream Queries present in Stream-Management System? Give Examples of Stream Sources	12 4	BTL2	Understand

5	Give the difference between Datar-Gionis-Indyk-Motwani Algorithm and Alon-Matias-Szegedy Algorithm	16	BTL3	Apply
6	Explain the Following i) General Sampling Problem ii) Filtering iii) The Bloom Filter iv) Analysis of Bloom Filtering	 4 4 4 4	BTL4	Analyze
7	Compare the features of The Count-Distinct Problem and The Flajolet-Martin Algorithm	16	BTL4	Analyze
8	What are the different ways to Sampling Data in a Stream?	16	BTL2	Understand
9	Explain The Count-Distinct Problem with suitable example.	16	BTL4	Analyze
10	Explain The Flajolet-Martin Algorithm with suitable example.	16	BTL4	Analyze
11	let us assume that a stream has a particular length n. Suppose we do not have enough space to count all the mi 's for all the elements of the stream. How to estimate the second moment of the stream using a limited amount of space using Alon-Matias-Szegedy Algorithm?	16	BTL5	Evaluate
12	Define the Following i) Decaying Window ii) Bloom Filter iii) Sampling Data in Stream iv) Higher-Order Moments	 4 4 4 4	BTL1	Remember
13	Suppose we have a window of length N on a binary stream. We want at all times to be able to answer queries of the form "how many 1's is there in the last k bits?" for any k ≤ N. Solve this problem use The Cost of Exact Counts to Counting Ones in a Window.	16	BTL5	Evaluate
14	Explain DGIM Conditions for Counting Ones in a Window	16	BTL3	Apply
15	Explain Datar-Gionis-Indyk-Motwani Algorithm with suitable examples.	16	BTL1	Remember
16	Explain The Alon-Matias-Szegedy Algorithm for estimating moments.	16	BTL2	Understand
17	What is Decaying Windows? What is the purpose for using Decaying Windows in Mining Data Streams?	16	BTL5	Evaluate

UNIT IV - LINK ANALYSIS AND FREQUENT ITEMSETS

SYLLABUS: Page Rank –Efficient Computation - Topic Sensitive Page Rank – Link Spam – Market Basket Model – A-priori algorithm – Handling Larger Datasets in Main Memory – Limited Pass Algorithm – Counting Frequent Item sets.

PART-A

Q.No.	Question	Level	Competence
1	Define the Term Spam.	BTL2	Understand
2	Give The Google Solution to Term Spam.	BTL3	Apply
3	Define PageRank.	BTL2	Understand
4	How to Computing PageRank on Strongly Connected Web Graphs?	BTL6	Create
5	What are Dead Ends?	BTL1	Remember
6	What are Spider Traps?	BTL4	Analyze

7	How to Representing Blocks of a Transition Matrix?	BTL4	Analyze
8	What is Topic-Sensitive PageRank?	BTL2	Understand
9	How to Creating Teleport Sets?	BTL6	Create
10	Define Link Spam.	BTL2	Understand
11	Define Trust Rank.	BTL5	Evaluate
12	Define Spam Mass.	BTL1	Remember
13	What is Market-Basket Data?	BTL1	Remember
14	What is The Pair-Counting Bottleneck?	BTL3	Apply
15	Explain Monotonicity of Frequent Item sets.	BTL2	Understand
16	Define A-Priori Algorithm.	BTL1	Remember
17	Define Toivonen's Algorithm.	BTL1	Remember
18	Define The SON Algorithm.	BTL1	Remember
19	Differentiate Multistage Algorithm and Multi hash Algorithm.	BTL5	Evaluate
20	What is PCY Algorithm?	BTL4	Analyze
21	What are Markov Processes?	BTL2	Understand
22	If the Distribution of the server approaches a limiting distribution v that satisfies v=Mv. It provided two conditions. What are they?	BTL5	Evaluate
23	How to avoid Dead Ends?	BTL2	Understand
24	What is Topic sensitive Page Rank?	BTL1	Remember

PART-B

Q.No.	Question	Marks	Level	Competence
1	Give the Efficient Computation of PageRank.	16	BTL2	Understand
2	Briefly Explain Topic-Sensitive PageRank with illustrations.	16	BTL1	Remember
3	Explain Architecture of a Spam Farm in Link spam.	16	BTL2	Understand
4	How to Analyse Spam Farm in Link spam?	16	BTL2	Understand
5	How to Formulate Hubbiness and Authority after PageRank was first implemented?	16	BTL4	Analyze
6	Explain the following using page rank i) Structure of the Web ii) Avoiding Dead Ends iii) Spider Traps and Taxation	 6 6 4	BTL4	Analyze
7	Explain the Following i) The Triangular-Matrix Method ii) The Triples Method iii) Monotonicity of Itemsets	 5 5 6	BTL1	Remember
8	How to Formalizing Hubbiness and Authority with suitable Link Matrix?	16	BTL6	Creating
9	Examine how spammers create link spam and explain several methods for decreasing the effectiveness of these spamming techniques, including TrustRank and measurement of spam mass.	16	BTL5	Evaluate
10	Explain A-Priori for All Frequent Item sets with suitable patterns.	16	BTL5	Evaluate

11	Explain Toivonen's Algorithm with suitable example.	16	BTL1	Remember
12	What are the Hybrid Methods present in Counting Frequent Items in a Stream	16	BTL4	Analyze
13	Explain The Algorithm of Park, Chen, and Yu with suitable example.	13	BTL3	Apply
14	Explain The Multistage Algorithm with suitable example.	13	BTL3	Apply
15	Explain The Multihash Algorithm with suitable example.	13	BTL3	Apply
16	Explain The Algorithm of Savasere, Omiecinski, and Navathe with suitable examples.	13	BTL3	Apply
17	Explain The SON Algorithm and MapReduce with suitable example.	13	BTL3	Apply

UNIT V CLUSTERING

SYLLABUS: Introduction to Clustering Techniques – Hierarchical Clustering –Algorithms – K-Means – CURE – Clustering in Non –– Euclidean Spaces – Streams and Parallelism – Case Study: Advertising on the Web – Recommendation Systems.

PART-A

Q.No.	Question	Level	Competence
1	Define Clustering	BTL1	Remember
2	Give the types of Clustering Algorithms.	BTL1	Remember
3	What is The Curse of Dimensionality?	BTL2	Understand
4	What is Clustroids?	BTL2	Understand
5	What are Centroids?	BTL1	Remember
6	How to Choose the Clustroid?	BTL5	Evaluate
7	What is Hierarchical Clustering?	BTL1	Remember
8	What is K-Means Algorithm?	BTL1	Remember
9	How to Initializing K-Means Algorithm?	BTL2	Understand
10	Define The BFR Algorithm.	BTL1	Remember
11	Define The CURE Algorithm.	BTL1	Remember
12	How to Representing Clusters in CURE?	BTL1	Remember
13	What is the full form of The GRGPF Algorithm and explain its purpose?	BTL4	Analyze
14	What are Clustering Streams?	BTL4	Analyze
15	What is Targeted Advertising?	BTL3	Apply
16	What is Bipartite Matching?	BTL4	Analyze
17	What is The Adwords Problem?	BTL2	Understand
18	Explain The Balance Algorithm with example.	BTL4	Analyze
19	What is Root-Mean-Square Error?	BTL5	Evaluate
20	What is The Netflix Challenge?	BTL4	Analyze
21	Define Hash Storage of Word Sets.	BTL1	Remember
22	How to matching Wordsets against Documents?	BTL4	Analyze
23	Give the competetive ratio of the Balance Algorithm.	BTL3	Apply

24	What is On and Off – Line Algorithms?	BTL2	Understand

PART-B

Q.No.	Question	Marks	Level	Competence
1	Explain the Following i) Points, Spaces, and Distances ii) Clustering Strategies iii) The Curse of Dimensionality	 5 5 6	BTL5	Evaluate
2	Will Hierarchical Clustering works in a Euclidean Space? Explain.	16	BTL6	Create
3	What are the Alternative Rules for Controlling Hierarchical Clustering?	16	BTL4	Analyze
4	Explain Hierarchical Clustering in Non-Euclidean Spaces.	16	BTL2	Understand
5	Explain K-means Algorithms with example.	16	BTL1	Remember
6	Explain The Algorithm of Bradley, Fayyad, and Reina with example.	16	BTL4	Analyze
7	Explain The CURE Algorithm.	16	BTL3	Apply
8	Explain GRGPF Algorithm in Clustering in Non-Euclidean Spaces.	16	BTL1	Remember
9	Explain A Stream-Clustering Algorithm with example.	16	BTL1	Remember
10	What are the Issues in On-Line Advertising?	16	BTL6	Create
11	Explain On-Line and Off-Line Algorithms.	16	BTL1	Remember
12	Explain Greedy Algorithms with example.	16	BTL2	Understand
13	Explain The Matching Problem of matching ads to search queries.	16	BTL4	Analyze
14	Explain the Following i) The Greedy Algorithm for Maximal Matching ii) The Greedy Approach to the Adwords Problem	 8 8	BTL3	Apply
15	The fundamental problem of search advertising is "adwords problem". Justify.	16	BTL5	Evaluate
16	Explain the steps involved in the Content-Based Recommendations.	16	BTL 6	Create
17	Explain the steps involved in Collaborative Filtering.	16	BTL 6	Create