B.TECH/AEIE/ECE/7TH SEM/CSEN 4126/2021

INTELLIGENT WEB AND BIG DATA
(CSEN 4126)

Time Allotted : 3 hrs **Full Marks : 70**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 5 (five) from Group B to E, taking at least one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A
(Multiple Choice Type Questions)

1. Choose the correct alternative for the following: **10 × 1 = 10**

(i) _________ can best be described as a programming model used to develop Hadoop-based applications that can process massive amounts of data.

(a) MapReduce (b) Mahout
(c) Oozie (d) All of the mentioned.

(ii) Which of the followings is used by Google to determine the importance of a particular page?

(a) Singular Value Decomposition (b) Page Rank
(c) Fast Map (d) All of the mentioned.

(iii) Point out the correct statement:

(a) The choice of an appropriate metric will influence the shape of the clusters
(b) Hierarchical clustering is also called HCA
(c) In general, the merges and splits are determined in a greedy manner
(d) All of the Mentioned.

(iv) Which of the following is finally produced by Hierarchical Clustering?

(a) Final estimate of cluster centroids
(b) Tree showing how close things are to each other
(c) Assignment of each point to clusters
(d) All of the Mentioned.

(v) For an ecommerce website, which of the following is explicit data?

(a) Order history (b) Page views
(c) Cart events (d) Product feedback.

(vi) With a matrix of user ratings for items, if we represent the items along the rows and users along the columns, subtract row averages from each entry and compute dot products to find item-item similarity, we are computing

(a) Cosine similarity (b) Pearson's similarity
(c) Adjusted Cosine similarity (d) None.

(vii) Which of the following combination is incorrect?
(a) Continuous – euclidean distance (b) Continuous – correlation similarity
(c) Binary – manhattan distance (d) None of the Mentioned.

(viii) Which of the following is required by K-means clustering?
(a) Defined distance metric (b) Number of clusters
(c) Initial guess as to cluster centroids (d) All of the Mentioned.

(ix) The daemons associated with the MapReduce phase are _______ and task-trackers.
(a) job-tracker (b) map-tracker
(c) reduce-tracker (d) all of the mentioned .

(x) Mention the default block size as well as the default replication factor for a typical multi-node single-master Apache Hadoop cluster.
(a) 128 MB and two (b) 128 MB and three
(c) 64 MB and three (d) 64 MB and four.

Group – B

2. (a) What are the ways of extracting information from external sites/blogs? [(CO1)(Remember/LOCQ)]
(b) How can metadata be developed from unstructured text? [(CO1) (Understand/LOCQ)]
(c) Explain in detail how a customer journey through a web page can help design a recommendation engine. [(CO2) (Analyze/IOCQ)]

4 + 4 + 4 = 12

3. Write short notes on the followings:
(i) Tag cloud
(ii) Page Rank algorithm
(iii) Hierarchical clustering. [(CO2) (Understand/IOCQ)]

(4 + 4 + 4) = 12

Group – C

4. (a) What are the properties of distance measure? [(CO3) (Understand/LOCQ)]
(b) What are the different types of similarity measure? [(CO2) (Understand/ LOCQ)]
(c) Describe any one of the email categorization algorithms.[(CO2) (Analyze/IOCQ)]

4 + 4 + 4 = 12

5. (a) Consider a set of 6 points {(0.5, 0.5), (1.5, 1.5), (0.86, 0.99), (0.1, 1), (0.2, 0.9), (1.7, 1.1)} to which k-means clustering is applied for k = 2. If (0.5, 0.5) and (1.5, 1.5) are the initial cluster seed points for clusters A and B respectively, how many points are there in the two clusters after the first round of allocation? [(CO2) (Evaluate/HOCQ)]
(b) How do you choose optimal number of clusters while applying K-means algorithm? [(CO3) (Analyze/HOCQ)]

8 + 4 = 12

Group – D

6. (a) What are the 3 modes in which Hadoop can be run? [(CO5)(Understand/LOCQ)]
   (b) Explain briefly the utility of these modes and when to use them. [(CO4) (Understand/IOCQ)]
   (c) Explain the high-level architecture of Hadoop. [(CO5)(Analyze/IOCQ)]

3 + 4 + 5 = 12

7. (a) Explain streaming mechanism within Hadoop. [(CO3) (Understand /IOCQ)]
   (b) State the functions of the following components of the Hadoop ecosystem:
   (i) Zookeeper (ii) YARN (iii) Hive. [(CO4) (Understand/IOCQ)]

3 + (3 × 3) = 12

Group – E

8. Suppose we have an n × n matrix M whose element in row i and column j will be denoted m_{ij}. Suppose we also have a vector **v** of length n, whose j-th element is $v_{j.}$

<u>Assumptions:</u>

- Let us first assume that n is large, but not so large that vector v cannot fit in main memory. The matrix M and the vector v each will be stored in a file of the HDFS.
- We assume that the row-column coordinates of each matrix element will be discoverable, either from its position in the file, or because it is stored with explicit coordinates, as a triplet (i, j, m_{ij}).
- We also assume the position of the element v_j in the vector v will be discoverable in the analogous way.

Answer the **following questions** based on the above mentioned assumptions:

(a) Describe, step-by-step, a Map-Reduce-based approach for solving matrix-vector multiplication. [(CO4)(CO5) (Evaluate/HOCQ)]

(b) Briefly discuss about the typical problems which may arise, in case the vector **v** is so large that it does not fit entirely in the main memory, violating the first assumption. [(CO4) (CO5)(Evaluate/HOCQ)]

(c) Suggest a solution (other than using more powerful computing resources) to handle the problem mentioned in the above **question Q8(b)**, and also discuss its impact on the map reduce approach for matrix-vector multiplication problem. [(CO4) (CO5) (Evaluate/HOCQ)]

(6 + 2 + 4) = 12

9. (a) Explain Breadth first search algorithm and how does it fit in MapReduce. [(CO5)(CO6)(Apply/IOCQ)]
   (b) Indicate the Map and Reduce functions in MapReduce based computation of the following two problems: [(CO5)(CO6) (Apply/IOCQ)]
   (i) Computing Grouping and Aggregation
   (ii) Computing Natural Join.

4 + (2 × 4) = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	19.8%	55.2%	25%

Course Outcome (CO):

After the completion of the course students will be able to

1. Understand the basic concepts related to Web Intelligence and Big Data.
2. Explain the terms data mining, neural networks, support vector machine, text analytics, text mining, web mining etc.
3. Learn how to use and deploy various web/social/mobile analytics platforms.
4. Understand the importance of Web intelligent as the art of customizing items in response to the needs of the users.
5. Learn the concepts of Hadoop and MapReduce.
6. Apply big data technologies in business intelligence using geospatial data, location-based analytics, social networking, reality mining, and cloud computing

*LOCQ: Lower Order Cognitive Question; IOCQ: Intermediate Order Cognitive Question; HOCQ: Higher Order Cognitive Question

Department & Section	**Submission link:**
AEIE, ECE	https://classroom.google.com/c/NDA1MzQ5NzQ4MDQz/a/NDQzODYwMzUyMzcz/details