BIG DATA AND WEB TECHNOLOGY
(CSEN 4182)

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) 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.

(ii) What do 'Content-based Techniques' build, and then associate similar items based on similarities between those?
(a) Vector Spaces (b) Term Vectors
(c) Decision Trees (d) Decision Matrices.

(iii) Hadoop is a framework that works with a variety of related tools. Common cohorts include:
(a) MapReduce, Hive and HBase
(b) MapReduce, MySQL and Google Apps
(c) MapReduce, Hummer and Iguana
(d) MapReduce, Heron and Trumpet.

(iv) 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.

(v) What is the process of adding freeform text, either words or small phrases, to items called?
(a) Tagging (b) Voting
(c) Blogging (d) Rating.

(vi) Work out the 'Cosine-based Similarity' between the two given points (6, 10, 4) and (4, 9, 6).
(a) 0.82 (b) 0.97 (c) 0.28 (d) 0.79.

(vii) Point out the wrong statement:
(a) k-means clustering is a method of vector quantization
(b) k-means clustering aims to partition n observations into k clusters
(c) k-nearest neighbor is same as k-means
(d) none of the mentioned.

(viii) Which type of clustering method does the 'Agglomerative Clustering' fall under?
(a) Hierarchical (b) Partitioned
(c) Probabilistic (d) None of the above.

(ix) Mention the default 'Block Size' as well as the default 'Replication Factor' for a multi-node single-master Apache Hadoop cluster.
(a) 128 MB and Two (b) 64 MB and Four
(c) 128 MB and Three (d) 64 MB and Three.

(x) Work out the approximate processing time for a 100-TB dataset distributed across a 2000-node cluster, assuming an average data scanning rate of 50 MB per second.
(a) 34 minutes (b) 17 minutes
(c) 23 hours (d) Can't do.

Group – B

2. (a) What are the ways of extracting information from external sites/blogs?

(b) How can metadata be developed from unstructured text?

(c) Explain in detail how a customer journey through a web page can help design a recommendation engine.

4 + 4 + 4 = 12

3. (a) Draw a generic data model for CI-enabling a web application.

(b) What do 'Term Frequency' (TF) and 'Inverse Document Frequency' (IDF) signify in text processing? Explain in brief.
For an 'Instagram-like' web-site, here is one data set about '*Users rating Photos*'.

Photo→	CR7	LM10	DM10
User↓ SOURAV-G	003	004	002
ABHISHEK-B	002	002	004
RANBIR-K	001	003	005

Apply the 'Cosine-based Similarity' computation to justify the following two statements:

(c) CR7 and LM10 seem to be very similar to each other.

(d) ABHISHEK-B and RANBIR-K seem to be very similar to each other.
Apply the 'Correlation-based Similarity' computation to justify the following two statements:

(e) CR7 and DM10 are strongly negatively correlated.

(f) ABHISHEK-B and RANBIR-K are highly correlated.

4 + 2 + 1 + 1 + 2 + 2 = 12

Group – C

4. (a) Web intelligence data are known to be dynamic, loosely structured and complex. Are standard clustering algorithms suitable for such data? Justify.

(b) Why do nature inspired clustering algorithms have an edge over traditional ones?

(c) Describe any nature inspired clustering algorithm of your choice.

4 + 3 + 5 = 12

5. Answer the following questions:

(i) What is 'Clustering'? Is it possible to do clustering using SQL's SELECT statement with ORDER BY clause for a set of records in a database that contains book information? If so, then why do we say that a general solution based on plain SQL queries is deficient and impractical? Furnish suitable example(s).

(ii) Explain, in brief, three different types of categorizations of clustering algorithms. What are these based on?

(iii) Highlight the strengths of 'ROCK'. Why is 'DBSCAN' primarily different from other algorithms that are based on the notion of links or the direct distance of the points from each other?

(iv) Briefly explain the two main issues with clustering in very large datasets.

(3 + 3 + 2 + 4) = 12

Group – D

6. (a) What are the 3 modes in which Hadoop can be run?

(b) Explain briefly the utility of these modes and when to use them.

(c) Explain the high level architecture of Hadoop.

3 + 4 + 5 = 12

7. (a) Explain streaming mechanism within hadoop.

(b) How does the data ingestion happen within Hadoop and how does MapReduce play a role?

(c) How can MapReduce patterns be used in Big Data Environments?

4 + 4 + 4 = 12

Group – E

8. Meri Chinipahari is the solution architect of 'Chehra', a 'Facebook-like' web-site that maintains a list of friends – note that friends are a bi-directional thing there; so if I'm your friend, you're mine.
Chehra has a lot of disk space and handles many requests every day. One common processing request is the typical "Akki and Sallu have 786 friends in common" feature. When you visit someone's profile, you see a list of friends that you have in common.
Meri has decided to pre-compute calculations, when it is possible, to reduce the processing time for such requests. Here is her rationale.
This list doesn't change frequently so it'd be wasteful to recalculate it every time you visited the profile. So, she has proposed using Map-Reduce technique so that Chehra can calculate everyone's common friends thrice a day and store those results. Later on, it's just a quick lookup. Since there is a lot of disk space, it's cheap.

Here are some useful assumptions.

(1) Assume that the friends data are stored as Person -> [List-of-Friends], and here is a sample list for five persons:
A -> B C D
B -> A C D E
C -> A B D E
D -> A B C E
E -> B C D

(2) Assume that each line in above list is an argument to a 'map' function, with the key being a friend along with the person, and the value being the list of friends, such that *map(A -> B C D)* will output the key-value pair as:
(A B) -> B C D
(A C) -> B C D
(A D) -> B C D

(3) Assume that each such key-value pair is passed as an argument to a 'reduce' function that simply intersects the lists of values, and outputs the same key with the result of the intersection; e.g.,

reduce ((A B) -> (A C D E) (B C D)) will output (A B) : (C D), meaning friends A and B have C and D as common friends.

Now considering when D visits B's profile, you work out all the necessary MapReduce steps, and enable Chehra to quickly look up (B D), and tell them that they have three friends in common, namely (A C E).

12

9. (a) What are the criteria of efficient graph algorithms?

   (b) Explain Breafth first search algorithm and how does it fit in MapReduce.

   (c) Explain page rank algorithm with an example.

4 + 4 + 4 = 12