(c) Name the three modes in which Hadoop can be run. Explain briefly the utility of these modes and when to use them.

4 + 2 + (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 jth element is v_j. In the said context answer the following:-
    i. What would be the matrix vector product?
    ii. What would be the Map function?
    iii. What would be the Reduce function?

3 + 4.5 + 4.5 = 12

9. Work out all the necessary MR steps for the following problem.
Assume that you have five files, and each file contains two fields (a key and a value, in Hadoop terms) [CITY, TEMPERATURE] that represent a city and the corresponding temperature recorded in that city on some day. The data in each file looks like this:

Tokyo, 31
Barcelona, 35
New York, 32
Rome, 32
Tokyo, 30
Rome, 33
New York, 28

Out of all the data that have been collected, you need to find the maximum temperature for each city across all the data files (note that each file might have the same city represented multiple times).
[*Hint*: Using the MR framework, break this down into suitable mapper tasks to work on these files by going through the data and returning the maximum temperature for each city. For example, the results produced from one mapper task for the data above would look like this:
(Tokyo, 31) (Barcelona, 35) (New York, 32) (Rome, 33)]

12

INTELLIGENT WEB AND BIG DATA
(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) __________ 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) Work out the Jaccard Similarity Index and the Jaccard Distance between the two given sets A = {0,1,2,5,6} and B = {0,2,3,4,5,7,9}.
(a) 71% and 29% (b) 67% and 33%
(c) 33% and 67% (d) 29% and 71%.

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

(iv) What type of architecture is conceptually recommended for learning and embedding collective intelligence in your Web applications?
(a) Event-driven Service-oriented (b) Event-driven Synchronous
(c) Polling-based Service-oriented (d) Polling-based Synchronous.

(v) 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.

(vi) Work out the approximate retrieval time for a 1000-GB dataset distributed across a 1000-node cluster, assuming an average data transfer rate of 1 Gbps.
(a) 1000 ms (b) 500 ms
(c) 8000 ms (d) 4000 ms.

(vii) Give one example of 'Derived Intelligence'.
(a) Wikis (b) Text Mining
(c) Tagging (d) Rating.

(viii) Which of the following clustering requires merging approach ?
(a) Partitional (b) Hierarchical
(c) Naive Bayes (d) None of the above mentioned.

(ix) A _________ signal is sent from the Task-Tracker to the Job-Tracker every few minutes to check its status whether the node is dead or alive.
(a) SOS (b) May-Day
(c) health-check (d) heart-beat.

(x) 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 above mentioned.

Group – B

2. (a) Draw the architecture schematic for a typical 'Collective Intelligence' (CI) enabled *non-event-driven* Web-based application with *synchronous services*. What changes will be needed to make it an *event-driven* one?

(b) Name the three different classes of intelligence and provide one example each.

(c) Provide one example each for 'Structured Data', 'Unstructured Data' and 'Semi-structured Data'.

(d) "A typical Google Search is an example of a synchronous service" – comment on this statement, in the context of CI.

(2 + 2) + 3 + 3 + 2 = 12

3. (a) What are the different steps of text mining? Explain each of them.

(b) How does tagging work? What are the different types of tagging? Explain how intelligence is extracted from user tagging.

4 + (2 + 3 + 3) = 12

Group – C

4. (a) Compare and contrast the two approaches for 'Collaborative Filtering' – 'Memory-based' and 'Model-based'.

(b) Mention the three important properties of 'Mathematical Distance'.

(c) Work out the three pair-wise cosine similarities for three sample documents D1, D2, and D3 while doing content-based recommendation, considering the four most frequently occurring terms in each of them as follows: D1 = {corner-kick, goal, penalty, set-piece}; D2 = {corner-kick, goal, header, wing}; and D3 = {free-kick, handball, throw-in, wing}.

(d) Work out the following for the two given points (9, 6, 3) and (2, 4, 6) – 'Manhattan Distance', 'Euclidean Distance', and 'Cosine Similarity'.

3 + 3 + 3 + 3 = 12

5. (a) What is 'Clustering'? It is often said, "A general solution based on plain SQL queries is deficient and impractical for clustering." even though it is possible to retrieve records grouped together, using SQL's SELECT statement with ORDER BY clause, for a set of records in a database that contains book information. Then why do we say so? Explain with suitable example(s).

(b) Explain, in brief, any two different types of categorization for clustering algorithms.

(c) What is 'Classification'? Provide one real-life example.

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

3 + 4 + 3 + 2 = 12

Group - D

6. (a) Discuss in detail about Hadoop Distributed File System.

(b) What is set and map and explain their operation.

(c) Explain streaming mechanism with Hadoop.

4 + 4 + 4 = 12

7. (a) With respect to a typical Hadoop architecture, explain how distributed storage and distributed processing are taken care of.

(b) Explain, in brief, what HBase is.