SPECIAL SUPPLE B.TECH/CSE/7TH SEM/CSEN 4165/2018

WEB INTELLIGENCE AND BIG DATA
(CSEN 4165)

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 type of clustering method does the agglomerative clustering fall under?
(a) Hierarchical (b) Partitioned
(c) Probabilistic (d) None of the above.

(ii) Hadoop is a framework that works with a variety of related set of tools; name one such set.
(a) Map-Reduce, Heron, Trumpet
(b) Map-Reduce, Hive, HBase
(c) Map-Reduce, MySQL, Google Apps
(d) Map-Reduce, Hummer, Iguana.

(iii) 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.

(iv) Work out the approximate processing time for a 500-TB dataset distributed across a 2000-node cluster, assuming an average data scanning rate of 100 MB per second.
(a) 63 minutes (b) 42 minutes
(c) 21 hours (d) Can't do.

(v) Which of the following can best be described as a programming model used to develop Hadoop-based applications that can process massive amounts of data?
(a) Map-Reduce (b) Mahout
(c) Oozie (d) All of the above-mentioned.

(vi) The daemons associated with the Map-Reduce phase are _____ and task-tracker.
(a) Map-tracker (b) Job-tracker
(c) Reduce-tracker (d) All of the above-mentioned.

(vii) Which of the following statements is incorrect?
(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 the same as k-means.
(d) None of the above-mentioned.

(viii) Which are two examples of implicit intelligence?
(a) Searching and Recommending (b) Rating and Voting
(c) Bookmarking and Tagging (d) Blogs and Wikis.

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

(x) 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 above-mentioned.

Group – B

2. (a) Is a typical Google Search an example of a 'Synchronous Service' or an 'Asynchronous Service' in the context of Collective Intelligence (CI)? Explain in brief.

(b) Provide <u>one</u> real-life example <u>each</u> for: (i) Web-sites exploiting 'Explicit Intelligence', 'Implicit Intelligence', and 'Derived Intelligence', and (ii) following types of metadata attributes – Numeric, Nominal Ordinal, and Nominal Categorical.

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

2 + (3 + 3) + 4 = 12

3. (a) What is 'Content' in the context of Collective Intelligence (CI)? Provide <u>one</u> suitable example.

(b) Draw a typical data persistence model for reviews done by users about restaurants for a Web-site like Zomato.

(c) Draw a typical data persistence schema either for a Blogging System or for a Wiki System.

3 + 4 + 5 = 12

Group – C

4. (a) What is a 'Recommendation Engine' (RE)? Name the two basic types of RE.

(b) What is 'Jaccard Similarity'?

(c) Explain, in brief, how Collaborative Filtering (CF) based on User Similarity and Collaborative Filtering (CF) based on Item Similarity work for the RE of (say) an online music store.

(d) Why is CF with Item Similarity preferred for a very popular e-commerce site like Amazon?

(e) Prove that the 'Cosine-based Similarity' between the two given points (6, 9, 4) and (4, 10, 6) is 97%.

(1 + 2) + 2 + 3 + 2 + 2 = 12

5. (a) What is 'Clustering'? Provide one real-life example.

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

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

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

3 + 3 + 4 + 2 = 12

Group – D

6. (a) What is Hadoop? With respect to a typical Hadoop architecture, explain how distributed storage as well as distributed processing are taken care of.

(b) Highlight, and explain in brief, any two of the major design considerations for Hadoop, based on corresponding assumptions.

8 + 4 = 12

7. (a) What are the three modes in which Hadoop can be run?

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

   (c) Explain, with the help of a suitable schematic diagram, the high level architecture of Hadoop.

3 + 4 + 5 = 12

Group – E

8. Explain, step-by-step and with help of a suitable example, how Map-Reduce (MR) works for counting occurrences of distinct words in a given set of documents.

12

9. Work out all the necessary Map-Reduce (MR) steps for the following problem.
   Assume that you have five files, and each file contains two columns (a key and a value, in Hadoop terms) that represent a city and the corresponding temperature recorded in that city on some day. In this example, CITY is the key and TEMPERATURE is the value.
   The data in each file looks somewhat like this:
   Trivandrum, 31
   Bangalore, 25
   New Delhi, 32
   Ranchi, 32
   Trivandrum, 30
   Ranchi, 33
   New Delhi, 28
   Out of all the data that have been collected, you have to find the maximum temperature for each city across all of the data files (note that each file might have the same city represented multiple times).
   [*Hint*: Using the Map-Reduce 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: (Trivandrum, 31) (Bangalore, 25) (New Delhi, 32) (Ranchi, 33)].

12