INTELLIGENT WEB AND BIG DATA
(CSEN 4126)

Time Allotted : 2½ hrs **Full Marks : 60**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 4 (four) from Group B to E, taking one from each group.

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

Group – A

1. Answer any twelve: **12 × 1 = 12**

Choose the correct alternative for the following

(i) Blogs and wikis are following kind of intelligence
(a) Implicit (b) Explicit (c) Derived (d) None of the above

(ii) Give two examples of 'Implicit Intelligence'
(a) Searching and Recommending (b) Rating and Voting
(c) Bookmarking and Tagging (d) Blogs and Wikis

(iii) Attributes having enumerated values with no ordering are
(a) Categorical (b) Ordinal (c) Numerical (d) None of the above

(iv) The basic idea of this system is that if users shared same interest in the past, then they will also have similar tastes in the future. This is known as
(a) Collaborative Recommendation (b) Content-based Recommendation
(c) Both (a) and (b) (d) None of the above

(v) Which Clustering technique requires a merging approach?
(a) Partitioning (b) Hierarchical (c) Naïve Bayes (d) None of the above.

(vi) _____ is a Qualitative data.
(a) Nominal Data (b) Discrete Data
(c) Continuous Data (d) None of the above

(vii) In Hadoop, the InputFormat class for reading in sequence files is called as
(a) SequenceFileInputFormat (b) SequenceFSInputFormat
(c) SequenceHDFSInputFormat (d) FSequenceInputFormat

(viii) What is the data warehousing component of the Hadoop ecosystem that can perform reading, writing, and management of large data sets in a distributed environment using SQL-like interface?
(a) MapReduce (b) Hive (c) Pig (d) HBase

(ix) Which of the following is not a relational algebra operation?
(a) Union (b) Intersection (c) Difference (d) Extraction.

(x) The process of partitioning a relation R into tuples based on the values of some attribute or set of attributes is known as

(a) Natural Join (b) Grouping and Aggregation
(c) Union (d) Intersection.

Fill in the blanks with the correct word

(xi) _____ is the process of grouping a set of data objects into multiple groups or clusters.

(xii) The full form of AGNES is _____.

(xiii) _____ is a common technique used to identify rule-like relationship patterns in large-scale sales transactions.

(xiv) The Secondary NameNode is also called as _____.

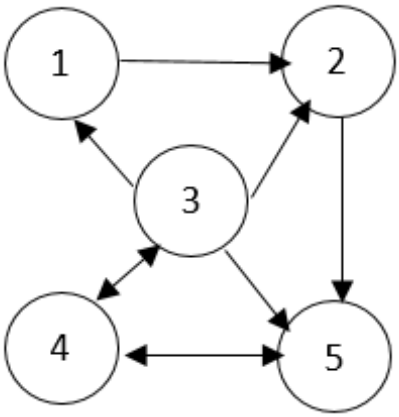
(xv) The full form of HDFS is _____.

Group - B

2. (a) What is Web 3.0? Differentiate between Web 1.0, Web 2.0, and Web 3.0.

[(CO1)(Understand/LOCQ)]

(b) Consider the following network of five web pages:



Considering the basic page rank algorithm, compute the page rank of the five pages after Iteration 0, Iteration 1, and Iteration 2. *[(CO6)(Analyse/IOCQ)]*

(2 + 3) + 7 = 12

3. (a) What are the different forms of user interactions? Explain with examples.

[(CO1)(Understand/LOCQ)]

(b) Let there be two users named John and Jane. They have tagged three articles A1, A2, and A3 as follows: John has tagged A1 with the tags apple, fruit, and banana. John has tagged A2 with the tags orange, mango, and fruit. Jane has tagged A3 with the tags cherry, orange, and fruit. The vocabulary consists of six tags: apple, fruit, banana, orange, mango, and cherry. Based on all the above information, find how much related are the three articles A1, A2, and A3 to one another.

[(CO6)(Analyse/HOCQ)]

6 + 6 = 12

Group - C

4. (a) What do you mean by supervised and unsupervised learning? *[(CO1)(Remember/LOCQ)]*

   (b) Given a dataset of the following points:
   A(2, 10), B(2, 5), C(8, 4), D(5, 8), E(7, 5), F(6, 4), G(1, 2), H(4, 9)
   Initialize k-means clustering algorithm with 3 cluster centers c1=A(2, 10), c2=D(5, 8), c3=G(1, 2). Consider Euclidean distance as the metric. What are the values of c1, c2, and c3 after one iteration of k-means? What are the values of c1, c2, and c3 after the second iteration of k-means? *[(CO5)(Apply/IOCQ)]*

   (c) Which method is more robust----K-means or K-medoids? Justify. *[(CO5)(Analyse/HOCQ)]*

3 + 6 + 3 = 12

5. (a) Describe any algorithm that can be used for categorization of emails. *[(CO6)(Analyse/LOCQ)]*

   (b) Explain how Laplacian correction can be used to avoid computing probability values of zero in a Naïve Bayes Classifier. *[(CO4)(Analyse/HOCQ)]*

   (c) What do you mean by sensitivity of a classifier? *[(CO4)(Understand/LOCQ)]*

5 + 4 + 3 = 12

Group - D

6. (a) Describe the following components of HDFS architecture:
   (i) NameNode
   (ii) DataNode
   (iii) Secondary NameNode. *[(CO1)(Remember/LOCQ)]*

   (b) Discuss the concept of HDFS federation. *[(CO2,CO3)(Understand/IOCQ)]*

(3 + 3 + 3) + 3 = 12

7. (a) Discuss the role of JobTracker and TaskTracker in a Hadoop cluster. *[(CO1)(Remember/LOCQ)]*

   (b) What are benefits of having multiple NameNodes in Hadoop? *[(CO1)(Understand/HOCQ)]*

   (c) What are the different OutputFormat classes in MapReduce? *[(CO2)(Remember/LOCQ)]*

5 + 4 + 3 = 12

Group - E

8. (a) Discuss the Map and Reduce functions for computing Selections by MapReduce. *[(CO6)(Understand/IOCQ)]*

   (b) Implement Matrix Multiplication as the cascade of two MapReduce operations. Explain the Map and Reduce functions for each of the operations. *[(CO4)(CO5)(Apply/HOCQ)]*

4 + (4 + 4) = 12

9. Discuss the Map and Reduce functions for computing the following problems with MapReduce:
   (i) Selection
   (ii) Projection
   (iii) Difference.

[(CO6)(Understand/LOCQ)]

(4 + 4 + 4) = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	53.13	20.83	26.04