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) Which of the following is correct?
(a) Dangling page is connected to all other pages
(b) Dangling page increases the page rank of other pages
(c) Dangling page has no contribution in page rank
(d) both (a) and (b).

(ii) How is the dice coefficient (DSC) related to the Jaccard distance (JD)?
(a) DSC=JD (b) 2 DSC=JD
(c) DSC=3 JD (d) DSC=2 JD.

(iii) If you consider that rating (one, two, three, four, and five) of items is an attribute for item classification, then what kind of attribute is this rating?
(a) Ordinal (b) Numerical
(c) Categorical (d) None of these.

(iv) Which statement is true?
(a) Support vector machine gives optimal linear decision boundary
(b) Perceptron model gives optimal linear decision boundary
(c) Bayes classifier gives linear decision boundary
(d) Logistic regression does not introduce non-linearity in decision boundary.

(v) What is the default block size in a typical multi-node single-master Hadoop cluster?
(a) 64 MB (b) 64 KB
(c) 128 MB (d) 128 KB.

(vi) What is the default replication factor in a typical multi-node single-master Hadoop cluster?
(a)2 (b) 3
(c) 4 (d) 5.

(vii) In HDFS, by default what is the frequency of heartbeat sent by the DataNode to the NameNode?
(a) 5 seconds (b) 3 seconds
(c) 4 seconds (d) 6 seconds.

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

(ix) Which component of the Hadoop ecosystem is a non-relational distributed database?
(a) Hive (b) Pig
(c) Oozie (d) HBase.

(x) In Hadoop, the wrapper class for Float is
(a) FWritable (b) FloatWritable
(c) FlWritable (d) FloatWrite.

Fill in the blanks with the correct word

(xi) When the DataNode starts up, it scans through its local file system and sends the list of hosted data blocks to the NameNode. This list of hosted data blocks is known as _____.

(xii) Classification task is _____ prediction whereas the regression task is _____ prediction.

(xiii) In a fully connected neural network (FCNN), the weights are updated during _____.

(xiv) The daemons associated with the MapReduce phase are _____ and TaskTracker.

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

Group - B

2. (a) Why is text indexing required? *[(CO1)(Remember/LOCQ]*

(b) Compute page rank of all the pages/nodes in the following diagram (Fig.1) using a random surfer model. Consider the damping factor (d) is 0.75, initial rank of each of all pages is 1 and check up to iteration 2. *[(CO1)(Understand/IOCQ)]*

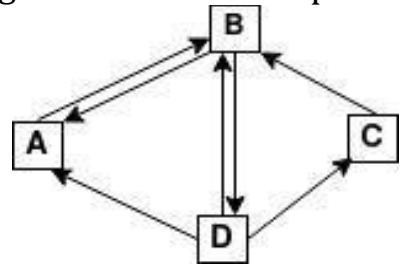

Fig.1

(c) What are the basic differences in features of Web 1.0 and Web 3.0? *[(CO1)(Remember/LOCQ)]*

2 + 6 + 4 = 12

3. (a) Find the Euclidean norm of the given vector $\boldsymbol{x} = [3\ 7\ 1]^T$. What does an outer product of two vectors generate? Show with an example. *[(CO3)(Analyse/HOCQ)]*

(b) Three users (A, B, C) have bookmarked three articles (A1, A2, A3) as shown in the following matrix (Fig.2). Find the item-item similarity matrix (here an article is an item). If a user buys the article A2, which related article will you recommend that user to buy? *[(CO1)(Apply/IOCQ)]*

	A	B	C
A1	1		1
A2		1	
A3		1	1

Fig.2: Bookmarking matrix

7 + 5 = 12

Group - C

4. (a) What is the assumption in the Bayes Classifier? How is the training of a Bayes Classifier done? *[(CO5)(Understand/LOCQ)]*

(b) Write any one classification model for e-mail classification in detail. *[(CO6)(Apply/IOCQ)]*

(c) What do you mean by the sensitivity of a classifier? *[(CO5)(Remember/LOCQ)]*

6 + 4 + 2 = 12

5. (a) "k-means algorithm converges in a finite number of steps"- Justify your answer. Does centroid initialization affects the k-means algorithm? *[(CO5)(Analyse/IOCQ)]*

(b) Given a dataset {0, 2, 4, 6, 24, 26}, initialize the k-means clustering algorithm with 2 cluster centres c1=3 and c2=4. What are the values of c1 and c2 after one iteration of k-means? What are the values of c1 and c2 after the second iteration of k-means? *[(CO5)(Apply/IOCQ)]*

6 + (3 + 3) = 12

Group - D

6. (a) Discuss the steps involved in writing a file in the Hadoop Distributed File System (HDFS). *[(CO5)(Understand/LOCQ)]*

(b) What is Secondary NameNode and what is its function? *[(CO1)(Remember/LOCQ)]*

(c) Why is the NameNode considered as the single point of failure in Hadoop 1.0? How is this problem overcome in Hadoop 2.0? *[(CO3)(Analyse/IOCQ)]*

5 + 4 + 3 = 12

7. (a) What is MapReduce? Discuss the steps involved in solving the word count problem with MapReduce. *[(CO2,CO6)(Apply/IOCQ)]*

(b) What is YARN? Discuss in detail. *[(CO1,CO3)(Understand/LOCQ)]*

(c) Describe the HBase component of the Hadoop ecosystem. *[(CO5)(Remember/LOCQ)]*

(2 + 4) + 3 + 3 = 12

Group - E

8. (a) Consider a matrix M and a vector **v**. Describe a MapReduce-based approach to solve the matrix-vector multiplication. Clearly mention the Map and the Reduce functions involved in the solution. *[(CO4,CO5)(Apply/HOCQ)]*

   (b) What problem will arise if the vector **v** is so large that it does not fit entirely in the main memory? *[(CO4,CO5)(Analyse/HOCQ)]*

   (c) Suggest a solution to handle the problem mentioned in the above question Q 8. (b). *[(CO4,CO5)(Analyse/IOCQ)]*

6 + 2 + 4 = 12

9. (a) How can we compute Natural Join by MapReduce? Explain in detail. *[(CO6)(Apply/IOCQ)]*

   (b) Discuss the Map and Reduce functions for computing the following relational algebra operations with MapReduce:
   (i) Grouping and Aggregation
   (ii) Selection. *[(CO6)(Understand/LOCQ)]*

4 + (4 × 2) = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	38.54	45.83	15.63

Course Outcome (CO):

After the completion of the course students will be able to

1. Learn the basics of web Intelligence and Big data.
2. Acquire fundamental enabling techniques and scalable algorithms like Hadoop, Map Reduce etc in big data analytics.
3. Understand the key issues in big data management and its associated applications in intelligent business and scientific computing.
4. Interpret business models and scientific computing paradigms.
5. Understand and practice big data analytics.
6. Apply the knowledge of Big Data and web intelligence on industry applications.

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