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) Flickr is an online photo management and sharing site that allows users to tag. Which type of intelligence does it use?
(a) Explicit (b) Implicit (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) Large scale e-commerce sites usually implement
(a) User-based Recommendation (b) Item-based Recommendation
(c) Article-based Recommendation (d) None of the above

(iv) _____ is a common technique used to identify rule-like relationship patterns in large-scale sales transactions.
(a) Decision Tree (b) Association Rule Mining
(c) General Rule Mining (d) None of the above.

(v) The main goal of a content-based recommendation system is to _____.
(a) Recommend items similar to those liked by users with similar tastes
(b) Recommend items based on the user's own preferences for the item's content
(c) Recommend items based on the item's popularity in the system
(d) Recommend items based on the user's location and time of day

(vi) 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

(vii) NameNode serves as _____ and each DataNode serves as _____ .
(a) Master, Slave (b) Slave, Master
(c) Could be either master or slave (d) None of the above mentioned

(viii) In Hadoop, if there are 4 input splits of a file then how many Mappers will be running?
(a) 2 (b) 4 (c) 8 (d) 16

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

(x) _____ data is information that does not reside in a relational database but has some organizational properties that make it easier to analyse.
(a) Structured Data (b) Semi-structured Data
(c) Unstructured Data (d) All of the above.

Fill in the blanks with the correct word

(xi) _____ is an example of Derived Intelligence.

(xii) The full form of DIANA is _____.

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

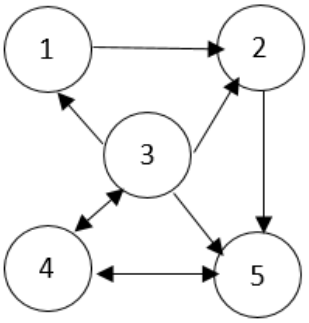
(xiv) During matrix-vector multiplication we can divide the matrix into vertical _____ of equal width.

(xv) _____ similarity measures the similarity between two sets by dividing the size of their intersection by the size of their union.

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) How do we compute correlation based similarity? *[(CO1)(Remember/LOCQ)]*

(b) Let there be three users A, B, and C. They have rated three articles named A1, A2, and A3 as shown in the following matrix:

	A1	**A2**	**A3**
A	1		
B		1	1
C	1		1

What are the related items based on the bookmarking patterns of the users? *[(CO6)(Apply/IOCQ)]*

(c) What can be the convergence criteria for the page rank algorithm?

[(CO6)(Analyse/IOCQ)]

3 + 6 + 3 = 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) State the Bayes Theorem. *[(CO4)(Remember/LOCQ)]*

(b) Discuss the Naïve Bayes Classifier. *[(CO4)(Understand/IOCQ)]*

(c) Consider the following dataset:

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Let there be a tuple X=(age=youth, income=medium, student=yes, credit_rating=fair). Using Naïve Bayes Classifier, classify this tuple X.

[(CO4)(Apply/HOCQ)]

3 + 4 + 5 = 12

Group - D

6. (a) Discuss the steps involved in writing a file in 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) Explain the workflow of a MapReduce job. *[(CO1)(CO3)(Understand/LOCQ)]*

(b) How does YARN handle resource allocation? *[(CO1)(CO3)(Understand/LOCQ)]*

(c) How do clients interact with HDFS and MapReduce components? *[(CO1)(Understand/IOCQ)]*

4 + 4 + 4 = 12

Group - E

8. (a) Consider a matrix M and 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? Suggest a solution to handle the problem mentioned in the above question. *[(CO4)(CO5)(Analyse/IOCQ)]*

6 + (3 + 3) = 12

9. (a) Discuss the Map and Reduce functions for computing Difference by MapReduce. *[(CO6)(Analyse/IOCQ)]*

(b) Use MapReduce algorithm to perform Intersection operation on the following tables. You do not need to write the algorithm. Just apply the algorithm on the data showing all the intermediate tables that get generated.

MAP WORKER 1

TABLE 1	
A	B
1	2
3	1

TABLE 2	
A	B
1	2
2	1

MAP WORKER 2

TABLE 1	
A	B
2	3
4	5

TABLE 2	
A	B
1	1
2	1

[(CO5)(CO6)(Apply/IOCQ)]

6 + 6 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	32.29	53.13	14.58