INTELLIGENT WEB AND BIG DATA
(CSEN 4126)

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

(ii) What could be the three pillars of a Video streaming recommendation engine like Netflix?
(a) History of films and TV Series, History of User on Netflix, Taggers who tag content
(b) History of films and TV Series, History of User on Netflix, Machine Learning Algorithm
(c) History of User on Netflix, History of films and TV Series, Taggers who tag content
(d) History of User on Netflix, Taggers who tag content, Machine Learning Algorithm.

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

(iv) In Big data environments, velocity refers —
(a) data can arrive at fast speed
(b) enormous data sets can accumulate within very short periods of time
(c) velocity of data translates into the amount of time it takes for the data to be processed
(d) all of the mentioned above.

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

(vi) Self-organizing maps are the best example of which of the following?
(a) Unsupervised learning (b) Supervised learning
(c) Reinforcement learning (d) Missing data imputation.

(vii) Which of the following can be considered as the main source of unstructured data?
(a) Twitter (b) Facebook
(c) Webpages (d) All of the mentioned above.

(viii) Apache Hadoop _____ provides a persistent data structure for binary key-value pairs.
(a) GetFile (b) SequenceFile (c) Putfile (d) all of the mentioned

(ix) _____ is a platform for developing data flows for the extraction, transformation and loading (ETL) of huge data sets, as well as for data analysis.
(a) Spark (b) HBase (c) Hive (d) Pig

(x) _____ class allows you to specify the InputFormat and Mapper to use on a per-path basis.
(a) MultipleOutputs (b) MultipleInputs
(c) SingleInputs (d) None of the mentioned

Group- B

2. (a) What do you mean by Web Intelligence? How can we create web intelligent document and queries? Give suitable examples. [(CO4)(Remember/LOCQ)]
(b) How can metadata be developed from unstructured text? [(CO2)(Understand/LOCQ)]
(c) Explain in detail how a customer journey through a web page can help design a recommendation engine. [(CO1)(Analyze/IOCQ)]
(2 + 4) + 3 + 3 = 12

3. (a) What is the job of a ranking algorithm? Briefly explain how the Page Rank algorithm works. [(CO3)(Analyse/LOCQ)]
(b) How does tagging work? What are the different types of tagging? [(CO3)(Understand/LOCQ)]
(c) What are the different steps of information extraction? Explain in brief. [(CO4)(Analyse/IOCQ)]
(3 + 3) + 3 + 3 = 12

Group - C

4. (a) What is the need for classification? Give a brief overview of classifiers. [(CO3)(Understand/LOCQ)]
(b) Explain with example Supervised and Unsupervised Classification. [(CO3)(Analyze/IOCQ)]
(c) Researchers are studying biodiversity in two rainforests. They catalog specimens from six different species, A, B, C, D, E and F. Two species are shared between the two rainforests. What is Jaccard coefficient? [(CO3)(Analyze/HOCQ)]
(2 + 3) + 4 + 3 = 12

5. (a) What is 'Recommendation Engine' (RE)? Name the two basic types of RE. [(CO4)(Understand/LOCQ)]
(b) Consider a set of 12 points {(185, 72), (170, 56), (168, 60), (179, 68), (182, 72), (188, 77), (180, 71), (180, 70), (183, 84), (180, 88), (180, 67), (177, 76)} to which K-means clustering is applied for k = 2. If (185, 72) and (170, 56) are the initial cluster seed

points for clusters A and B, respectively. How many points are there in the two clusters? [(CO2,CO3)(Analyze/LOCQ)]

4 + 8 = 12

Group – D

6. (a) What are the three main components of Hadoop? [(CO2)(Remember/LOCQ)]
   (b) What are the most common input formats defined in Hadoop? [(CO2)(Remember/LOCQ)]
   (c) Explain the Pseudo-distributed Mode in Hadoop. [(CO2)(Understand/IOCQ)]
   (d) What is Hadoop streaming? [(CO2)(Remember/LOCQ)]

3 + 3 + 3 + 3 = 12

7. (a) What is HBase? "HBase is a data model designed to produce quick random access to huge amounts of structured data". Do you agree with this statement? Justify. [(CO6)(Analyse/HOCQ)]
   (b) Differentiate between HBase and HDFS and explain the meaning of horizontally scalability characteristics of HBase. How does Storage Mechanisms work in HBase? [(CO4)(Understand/HOCQ)]

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

Group - E

8. (a) What are the 3 phases in which MapReduce algorithms execute? Explain each phase. [(CO5)(Remember/LOCQ)]
   (b) Big data graphs are typically sparse - Explain why. What is the best representation of such types of graphs? [(CO5)(Understand/IOCQ)]

(2 + 6) + (2 + 2) = 12

9. (a) How will you use graph in Map Reduce? [(CO4)(Remember/LOCQ)]
   (b) How will you compute Relational Algebra projections by Map Reduce? [(CO3)(Understand/LOCQ)]
   (c) State the advantages and disadvantages of adjacency matrices. [(CO4)(Analyse/IOCQ)]

4 + 4 + 4 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	65.63	18.75	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