B.TECH/CSE/7TH SEM/CSEN 4165/2017

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) What type of architecture is recommended for learning and embedding intelligence in your Web applications?
(a) Event-driven SOA (b) Event-driven Synchronous
(c) Polling-based SOA (d) Polling-based Synchronous.

(ii) __________ can best be described as a programming model used to develop Hadoop-based applications that can process massive amounts of data.
(a) MapReduce (b) Mahout
(c) Oozie (d) All of the mentioned.

(iii) What is the typical processing type for 'Search Services'?
(a) Asynchronous (b) Synchronous
(c) Both (d) None

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

(v) 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 Mentioned.

(vi) Which of the following is required by K-means clustering?
(a) Defined distance metric
(b) Number of clusters
(c) Initial guess as to cluster centroids
(d) All of the Mentioned.

(vii) Let the amount of time spent by nine readers on a news article be [5, 47, 50, 55, 47, 54, 100, 45, 50] seconds. Considering a validation window of two times the standard deviation, work out the outlier values in first iteration.
(a) 5 and 100 both (b) 5 only
(c) 100 only (d) none.

(viii) Which of the following combination is incorrect?
(a) Continuous – euclidean distance
(b) Continuous – correlation similarity
(c) Binary – manhattan distance
(d) None of the Mentioned.

(ix) Which of the following clustering requires merging approach?
(a) Partitional (b) Hierarchical
(c) Naive Bayes (d) None of the Mentioned.

(x) Work out the approximate processing time for a 100-TB dataset distributed across a 2000-node cluster, assuming an average data scanning rate of 50 MB per second.
(a) 34 minutes (b) 17 minutes
(c) 23 hours (d) Can't do.

Group – B

2. (a) "Collective Intelligence (CI) is the core component of Web 2.0." – explain in brief with help of any one suitable example from real-life.

(b) Mention one real-life example each of 'Synchronous Service' and 'Asynchronous Service' in CI. What type of service is a typical Google Search? Explain in brief.

(c) Mention one real-life example each of web-sites exploiting 'Explicit Intelligence', 'Implicit Intelligence', and 'Derived Intelligence'.

(d) Mention one real-life example each for the following types of metadata attributes – Numeric, Nominal Ordinal, and Nominal Categorical.

(e) Compare and contrast the two types of collaborative filtering approaches – 'Memory-based' and 'Model-based'.

2 + 3 + 1.5 + 1.5 + 4 = 12

3. (a) What are the different steps of text mining? Explain each of them.

(b) How does tagging work? What are the different types of tagging? Explain how intelligence is extracted from user tagging.

4 + (2 + 3 + 3) = 12

Group – C

4. Answer the following questions in the context of Recommendations:
   (i) What is a 'Recommendation Engine' (RE)? Name the two basic types of RE.
   (ii) Mention four important properties of 'Mathematical Distance'. What is 'Jaccard Similarity'?
   (iii) Mention the major advantage as well as the main limitation of 'Collaborative Filtering' (CF) type recommendations based on user similarity. Mention some other type of CF to tackle such limitation.
   (iv) Work out the pair-wise 'Cosine Similarities' for three sample documents Doc1, Doc2, and Doc3 while doing 'Content-based Recommendation', considering the four most frequently occurring terms in each of them as follows: Doc1 = {Google, shares, advertisement, president}; Doc2 = {Google, advertisement, stock, expansion}; and Doc3 = {NVidia, stock, semiconductor, graphics}.

(2 + 3 + 3 + 4) = 12

5. (a) What are the properties of distance measure?
   (b) What are the different types of similarity measure?
   (c) Describe any one of the email categorization algorithms and uses the same.

3 + 4 + 5 = 12

Group – D

6. (i) What is Hadoop? What are its two essential components? Mention any two of its optional components.
   (ii) Highlight, and explain in brief, any three of the major design considerations for Hadoop based on corresponding assumptions.
   (iii) In light of a typical Hadoop architecture, explain how distributed storage as well as distributed processing is taken care of.
   (iv) Explain, in brief, how Map-Reduce (MR) works for counting occurrences of distinct words in a given set of documents.

(4 × 3) = 12

7. Answer *all* the following questions, in a concise manner, with respect to Hadoop and/or Map-Reduce:
   (i) What is a 'Cluster'?
   (ii) What is a 'Rack'?

   (iii) What is a 'Namespace'?
   (iv) What is a 'Single-Node Cluster' (also known as 'Pseudo-Distributed Cluster')?
   (v) What is a Check-point'?
   (vi) What is an 'Edit-log'?

(6 × 2) = 12

Group – E

8. (a) How do you use graph data in MapReduce?
   (b) How will you invert a graph in MapReduce?
   (c) State the advantages and disadvantages of adjacency matrices.

4 + 4 + 4 = 12

9. Suppose we have an n × n matrix M whose element in row i and column j will be denoted m_{ij}. Suppose we also have a vector v of length n, whose jth element is v_j.

Assumptions:
1) Let us first assume that n is large, but not so large that vector v cannot fit in main memory. The matrix M and the vector v each will be stored in a file of the HDFS.
2) We assume that the row-column coordinates of each matrix element will be discoverable, either from its position in the file, or because it is stored with explicit coordinates, as a triplet (i, j, m_{ij}).
3) We also assume the position of element v_j in the vector v will be discoverable in the analogous way.

Questions:
(i) Describe, step-by-step, a Map-Reduce-based approach for this matrix-vector multiplication.
(ii) Explain what kind of typical problems can arise to slow down the computation in case the the vector v is so large that it does not fit in its entirety in main memory, this violating Assumption #1 above.
(iii) Suggest some solution (*other than using more powerful computing resources*) to handle problems mentioned in (ii) above, and its impact on the approach mentioned in (i) above.

(6 + 2 + 4) = 12