INTELLIGENT WEB AND BIG DATA
(CSEN 4182)

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) In the PageRank algorithm, if the hyperlink matrix stores probabilities on outgoing links in the columns and those on incoming links along the rows
(a) the rows add up to 1
(b) the columns add up to 1
(c) both the rows and columns add up to 1
(d) neither the rows nor the columns add up to 1.

(ii) Which of the following is used to primarily tackle the problem of dead ends in the web graph?
(a) Stochasticity adjustment
(b) Primitivity adjustment
(c) Both by stochasticity adjustment and primitivity adjustment
(d) Neither by primitivity adjustment nor by stochasticity adjustment

(iii) With a matrix of user ratings for items, if we represent the items along the rows and users along the columns, subtract row averages from each entry and compute dot products to find item-item similarity, we are computing
(a) Cosine similarity (b) Pearson's similarity
(c) Adjusted Cosine similarity (d) None.

(iv) 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.

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

(vi) Mention the default block size as well as the default replication factor for a typical multi-node single-master Apache Hadoop cluster.

(a) 128 MB and two (b) 128 MB and three
(c) 64 MB and three (d) 64 MB and four.

(vii) In a sentiment analysis task to label movie reviews as pos or neg, if for review S,P(pos) = 0.5, P(S|pos) = 0.9, P(S|neg) = 0.3, compute P(pos|S).
(a) 0.33 (b) 0.45 (c) 0.5 (d) 0.75.

(viii) Consider the directed graph G = (V,E), V = {y, a, m}. In which of the following cases will the problem of spider trap occur at node m?
(A) E = {(y,y), (y, a), (a, y), (a, m), (m, m)}.
(B) E = {(y,y), (y, a), (a, y), (a, m)}.
(a) A only (b) B only (c) Both A and B (d) Neither A nor B

(ix) Suppose Book1 is described by keywords {a, b, c, d} and Book2 by keywords {c, d, e, f, g, h}. The dice coefficient between Book1 and Book2 calculated based on this information is
(a) 0.5 (b) 0.2 (c) 0.4 (d) 0.25.

(x) There are 10 items and 10 users. User A rates the first 9 items 1,2,3,3,2,1,1,2,2. The 10th item is rated by the other 9 users as 3,2,1,3,2,1,3,3,1. Using the RF-Rec Predictors, the predicted rating of item 10 by user 1 is
(a) 1 (b) 2 (c) 4 (d) 3.

Group – B

2. (a) What are stochasticity and primitivity adjustments in web search? Which of them is used to primarily tackle the problem of spider traps in the web graph? In the pagerank algorithm with primitivity and stochasticity adjustment, if a node has 5 outlinks and the damping factor is 0.8, then with what probability is each such outlink assumed to be visited?

(b) Assume that we run the Topic Specific Pagerank algorithm for targeted search on the graph G= (V, E) with V= {A, B, C, D} and E = {(A, B), (A, C), (A, D), (B, A), (B ,D), (C, A), (D, C), (D, B)} and initial pagerank vector (A, B, C, D) = (1 0 0 0), β = 0.8 and the topic specific set = {A}. What is the pagerank vector after 1 iteration? Show all steps.

6 + 6 = 12

3. (a) What is 'Tagging' and how do tags help? Name the different types of tags, with one suitable example for each. What is a 'Tag Cloud'?

(b) Draw a data persistence model for tagging done by users for an online news portal (like say yahoo.com).

(c) Farmer-A has tagged Article-1 for tags {apple, banana, fruit}, and has tagged Article-2 for tags {fruit, mango, orange}; Farmer-B has tagged Article-3 for tags {cherry, fruit, orange}. Derive the vocabulary and use that to create a blank tabular format for collection of raw data about tagging these fruit-related articles by different farmers.

(1 + 3 + 1) + 4 + 3 = 12

Group – C

4. (a) Consider a set of 6 points {(0.5, 0.5), (1.5, 1.5), (0.86, 0.99), (0.1, 1), (0.2, 0.9), (1.7, 1.1)} to which k-means clustering is applied for k = 2. If (0.5, 0.5) and (1.5, 1.5) are the initial cluster seed points for clusters A and B respectively, how many points are there in the two clusters after the first round of allocation?

   (b) Consider the following table and using Bayes' classifier predict whether a user will buy an item when all three input attributes are TRUE:

Attributes→ User↓	Attribute 1	Attribute 2	Attribute 3	Buy?
A	F	T	T	F
B	T	F	T	F
C	T	T	F	T
D	T	T	F	F
E	T	T	T	T

6 + 6 = 12

5. (a) Consider the following user-item ratings matrix.

Items / **Users**	**Item1**	**Item2**	**Item3**
A	3	4	2
B	2	2	4
C	1	3	5

Compute the Cosine similarity between users A and B. Show all the steps.

   (b) Compute the Pearson similarity between users B and C. Show all the steps.

   (c) Compute the Adjusted Cosine similarity between item1 and item2. Show all steps.

4 + 4 + 4 = 12

Group – D

6. State the functions of the following components of the Hadoop ecosystem:
   (i) Ambari (ii) Zookeeper (iii) YARN (iv) Hive.

(3 + 3 + 3 + 3) = 12

7. (a) What are the three modes in which Hadoop can be run?

   (b) Explain briefly the utility of any two of these modes and when to use them.

   (c) Explain, with the help of a suitable schematic diagram, the high-level architecture of Hadoop.

3 + 4 + 5 = 12

Group – E

8. 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 j^{th} 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 the 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 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

9. Indicate the Map and Reduce functions in each of the following MapReduce computations:
   (i) Computing Selection and Projection.
   (ii) Computing Grouping and Aggregation
   (iii) Computing Natural Join.

(4 + 4 + 4) = 12

Department & Section	Submission link:
ECE and AEIE	https://classroom.google.com/c/MTQxODcyMTk3NDAw/a/Mjc0ODQ1ODMxODcx/details