3 * 4 = 12

9. (a) Explain how Local Aggregation in MapReduce increases efficiency. Explain the role of combiners in the same.

   (b) What is distributed Cache in MapReduce Framework? Explain.

(4 + 4) + 4 = 12

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
<u>any 5 (five)</u> from Group B to E, taking <u>at least one</u> 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) 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 above.

(ii) Movie1 receives a rating of 1 each from users A and C and Movie2 receives a rating of 1 each from users B and C. Then the cosine similarity between Movie1 and Movie2 is
(a) 1 (b) 0.5 (c) $\sqrt{2}$ (d) $1/\sqrt{2}$.

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

(iv) Assume a dataset for users with 2 input attributes X and Y and an output attribute Z. Using the Naive Bayes approach, what is P(Z=T|X=T,Y=F)/P(Z=F|X=T,Y=F) if
P(X=T|Z=T)= 1/2 P(Y=F|Z=T)= 1/2 P(Z=T)= ½ and
P(X=T|Z=F)= 2/3 P(Y=F|Z=F)= 1/3 P(Z=F)=1/2
(a) 8/9 (b) 9/8 (c) 3/8 (d)1/72.

(v) Which of the following is NOT a hierarchical agglomerative clustering algorithm?
(a) K-means (b) ROCK (c) MST (d) single link.

(vi) If H is the web or hyperlink matrix, then the pagerank vector
(a) is an eigenvector of H with eigenvalue 0.5
(b) is an eigenvector of H with eigenvalue 1
(c) is not an eigenvector of H
(d) not an eigenvector.

(vii) For a binary classification problem, if TP denotes the number of true positives and FP denotes the number of false positives, then TP/(TP+FP) denotes
(a) Precision (b) Recall (c) Accuracy (d) None of the above.

(viii) If instead of computing cosine similarities between all pairs of users, we partition the users into k roughly equal sized clusters and compute pairwise similarities within each cluster, the time for the correlation computation roughly
(a) increases by O(k) (b) decreases by O(k)
(c) increases by O($\sqrt{k}$) (d) decreases by O(klog k).

(ix) Mention the default 'Block Size' as well as the default 'Replication Factor' for a multi-node single-master Apache Hadoop cluster.
(a)128 MB and Two (b) 64 MB and Four
(c)128 MB and Three (d) 64 MB and Three.

(x) Which of the following are techniques to meet HDFS design goals?
(a) Simplified coherency model
(b) Data replication
(c) Move computation close to the data
(d) All of the above

Group – B

2. (a) Consider the following frequency of tags used in 3 articles.

Tags→ Articles	apple	fruit	banana	orange	mango	cherry
Article1	4	8	6	3		
Article2		5		8	5	
Article3	1	4		3		10

Also consider the following frequency of tags used by 2 users.

Tags→ Users	apple	fruit	banana	orange	mango	cherry
A	1	2	1	1	1	
B		1		1		1

Compute the cosine similarity between the articles. Also find cosine similarity between the users.

(b) Find the cosine similarity between the users and the articles. Then find the two most relevant articles for user A and the most relevant user for Article 1.

6 +6 = 12

3. (a) Explain the steps of document retrieval.

(b) What is the Page rank vector?

(c) Describe the web surfing model.

5 + 3 + 4 = 12

Group – C

4. (a) Consider the adjacency matrix A with users along rows and items along columns where an entry corresponds to an edge in the bipartite graph in the activation spreading method of Huang et al. The entries are 0 otherwise. Let T be the transpose of A, and P*Q represent the result of multiplying matrices P and Q in the usual way. A path of length k exists between an user i and an item j iff the (i,j)th entry in the matrix A*T*A*T*A is positive. Showing all steps, compute the value of k.

(b) In k-means clustering what is k? How is the value of k chosen? What is the convergence criteria? Enumerate the key steps of the algorithm.

6 + 6 = 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. (a) Explain the pig architecture with a neat diagram.

(b) Write at least two differences between pig and hive.

8 + 4 = 12

7. (a) What are the common types of failures in HDFS and how are these handled?

(b) What are the benefits of multiple namenodes and namespaces in Hadoop?

6 + 6 = 12

Group – E

8. Design a MapReduce algorithms to take a very large file of integers and produce as output:
(i) The largest integer.
(ii) The average of all the integers.
(iii) The same set of integers, but with each integer appearing only once.
(iv) The count of the number of distinct integers in the input.