9. (a) Explain how the pagerank algorithm can be implemented using the MapReduce paradigm.

(b) Enumerate the key steps of the algorithm to find the paths of length two on the web. Assume that your algorithm receives as input relation links consisting of URLs (a, b) such that there are is a link from a to b. Your solution must use natural join computation using MapReduce.

6 + 6 = 12

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) 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) Adusted Cosine similarity (d) None.

(ii) 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.

(iii) 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.

(iv) An e-commerce website has 22 million users and an inventory of 10,000 books. Then which is a more efficient form of collaborative filtering?
(a) user-user (b) item-item (c) user-item (d) item-user.

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

(vi) 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) 0.25 (d) 0.4.

(vii) 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) 3 (d) 4.

(viii) In a set S of 100 items of which 60 are relevant, a recommendation algorithm retrieves 50 items of which 20 are irrelevant. What are the precision and recall values?
(a) Precision = 0.6, Recall = 0.5 (b) Precision = 0.5, Recall = 0.5
(c) Precision = 0.5, Recall = 0.6 (d) Precision = 0.4, Recall = 0.6.

(ix) Which is the data warehousing component in the Hadoop ecosystem?
(a) Hive (b) Pig (c) Sqoop (d) Flume.

(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) 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

3. (a) Enumerate the key steps in the tag based collaborative filtering using social ranking.

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

6 + 6 = 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 an user will buy an item when all three input attributes are TRUE:

Attributes→ Users↓	Attr1	Attr2	Attr3	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) 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. State the functions of the following components of the Hadoop ecosystem:
(a) Apache Ambari
(b) Apache Zookeeper
(c) Apache Oozie
(d) Apache HBase.

3 + 3 + 3 + 3 = 12

7. (a) Explain the Apache Hadoop Filesystem.

(b) Mention and explain the design goals of HDFS.

6 + 6 = 12

Group – E

8. (a) What is the difficulty of implementing Djikstra's shortest path algorithm in MapReduce?

(b) Explain how the "Parallel Breadth-First Search" algorithm solves the single source shortest path problem using MapReduce.

(c) Given that the well known "Six degrees of separation" has been reduced to 3.57 for the users on Facebook, roughly how many MapReduce iterations required on the average on the Facebook graph by the "Parallel Breadth-First Search" algorithm?

4 + 4 + 4= 12