Course Code : CST 409–1 **CXDW/RW – 18 / 5586**

Eighth Semester B. E. (Computer Science and Engineering) Examination

Elective IV

WEB INTELLIGENCE AND BIG DATA

Time : 3 Hours] [Max. Marks : 60

Instructions to Candidates :—

(1) All questions carry marks as indicated against them.

(2) Assume suitable data and illustrate answers with neat sketches wherever necessary.

1. (a) Illustrate Locality Sensitive hashing and its use in finger print matching. 4(CO 1)

(b) Evaluate and plot the S - curve for S = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 having r = 5 and b = 40. Also find the threshold and prove its correctness. 3(CO 1)

(c) Discuss how text indexing is done on web along with an example. Find the complexity of index creation. 3(CO 1)

OR

(d) What are the different Vs that characterize Big Data ? Is it necessary that all the characteristics must be satisfied to categories any problem into Big Data. 3(CO 1)

2. (a) Consider the following TF :—

WORD	TF-D1	TF-D2	TF-D3	Doc freq of word (N_w)
Car	30	6	25	68068
Auto	5	66	0	60540
Insurance	5	66	0	1900
Best	15	0	18	24235

Consider the total number of documents as 90000

Find the following :—

(1) Inverse Document Frequency (IDF)

(2) TF-IDF

(3) Find the normalised TF

(4) Compute TF-IDF using normalized TF

(5) Compare and comment on the results obtained using TF and Normalized TF. 7(CO 3)

(b) Can machine learning be used to find whether a person intends to shop or surf ? How can conditional probability be used to predict an event of buy if the following search keywords are given-Red, Flower, Gift and Cheap, derive the equation of probability of a buy=yes /no. 3(CO 4)

OR

(c) What is the significance of finding Precision and Recall for a classifier ? Also state their meaning and use. 3(CO 4)

3. (a) Consider a company collects 5 readings of temperature in a month. They have data for 4 months with temperature ranging from 10 to 50. Find the mean temperature recorded for a month using Map-Reduce technique. 5(CO 2)

OR

(b) Create an example document in MongoDB for an e-commerce website having reviews and ratings about two to three products. Write a query to find the posts that have ratings greater than 4. 5(CO 2)

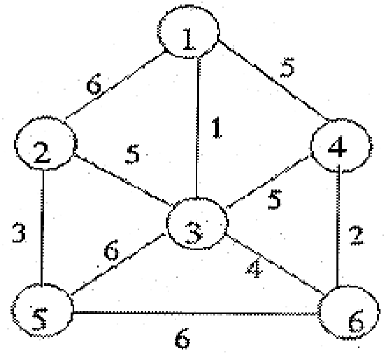
(c) Explain Big table and how sharding works in Big table. 5(CO 2)

4. (a) What is the problem of long-tail ? Explain the three techniques to deal with long-tail. 4(CO 1,CO 3)

(b) Apply clustering on graph of web pages to determine the websites of similar domain, by using the following approaches :—

(1) Delete branches with maximum weight.

(2) Deleting the inconsistent edges.



6(CO 3)

OR

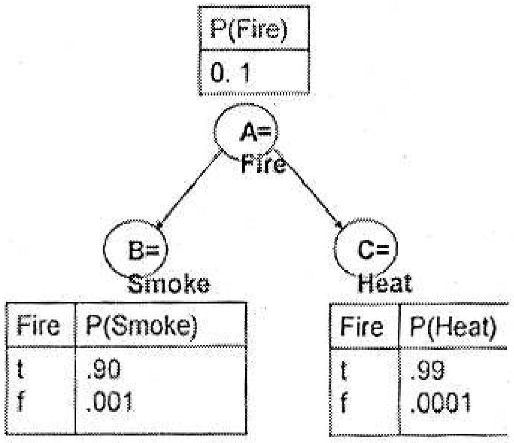
(c) Explain the Aprori algorithm. Also draw the frequent item set tree if there are 5 items in the transaction. Justify the need of pruning. 6(CO 3)

5. (a) Explain Predicate logic and its application on web data. 5(CO 3)

(b) Apply Marginalization to find :

P(Smoke = T) and P(Fire = T Smoke = T).

Also find P(Fire = T | Smoke = T).



P(Fire)
0. 1

Fire	P(Smoke)
t	.90
f	.001

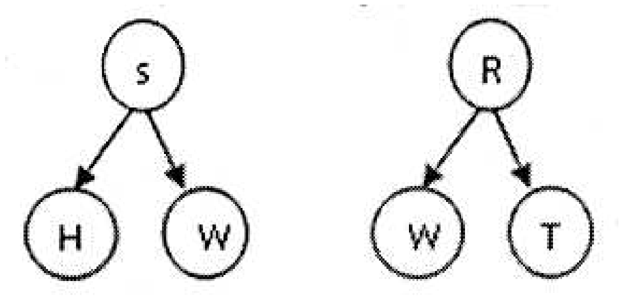
Fire	P(Heat)
t	.99
f	.0001

5(CO 4)

OR

(c) Multiple Na¨¡ve Bayes classifiers are combined to form a Bayesian network. Justify and explain.

Combine the networks given below to form Bayesian network and show how will P (H, W, T, S, R) be evaluated.



5(CO 4)

6. (a) Justify the need of prediction.

Explain linear predication and how linear prediction will work to predict the next in the sequence of 10 numbers as given below :

-0.14, 0.24, 0.71, 0.87, 0.83, 0.78, 0.89, 0.82, 0.66, 0.29

Consider order 5 model for illusration. 5(CO 1)

OR

(b) Design a blackboard architecture for self driving car system. 5(CO 1)

(c) Design and organize the Spare Distributed Memory (SDM) for the input data 1001101110.

Show the complete sparse memory with the output from memory. (Radius = 3) Reference Address is 0101101110 and 5 Address of the memory : 0001001010, 0001101100, 000000100, 0111101110, 0000011100. 5(CO 1,CO 3)