Final Assessment Test – November 2018

VIT
Vellore Institute of Technology

Course: **CSE3024** - **Web Mining**
Class NBR(s): **5997 / 5998 / 5999 / 6000 / 6001 / 6005**
Time: **Three Hours**

Slot: **F2+TF2**
Max. Marks: **100**

Answer any TEN Questions
(10 X 10 = 100 Marks)

1. a) With a neat diagram, explain the architecture of IR models. [5]

   b) Briefly explain the different classes of web security with possible solutions. [5]

2. Find out the document order (ranking) with respect to the query "T4 and T5", considering the following table. Compare Cosine similarity and Euclidean Distance way to find the most relevant Document.

	T1	T2	T3	T4	T5
D1	0	0	0	66	0
D2	0	0	0	22	36
D3	0	70	0	21	0
D4	0	0	15	40	22
D5	30	0	0	30	31
D6	0	0	0	0	54

3. Create a decision tree for the data values given in the table using Entropy (Information Gain) based attribute finding.

Refund	Marital Status	Taxable Income	Cheat
Yes	Single	125K	No
No	Married	100K	No
No	Single	70K	No
Yes	Married	120K	No
No	Divorced	95K	Yes
No	Married	60K	No
Yes	Divorced	220K	No
No	Single	85K	Yes
No	Married	75K	No
No	Single	90K	Yes

4. Alice wants to buy a iPhone 8. But she don't know if it is a great phone to buy or not. She searches the customer review for iPhone 8 and found the following review.

Doc	Words	Class
1	Don't buy	Negative
2	Phone got hanged	Negative
3	Battery drains fast	Negative
4	Durable phone	Positive
5	Great camera	Positive

She needs to know if the review containing "Great phone buy it", words is positive or negative. Help Alice to classify her sentiment using Naïve Bayes classifier.

Scanned by CamScanner

5. Apply the hierarchical clustering on the following TF-IDF matrix and draw the dendrogram. A, B, C, D, E and F are the documents. T1 and T2 are the terms. Using Single linkage and complete linkage method.

	T1	T2
A	0.40	0.53
B	0.22	0.38
C	0.35	0.32
D	0.26	0.19
E	0.08	0.41
F	0.45	0.30

6. → Compute the hub and authority score for the following graph using HITS algorithm.
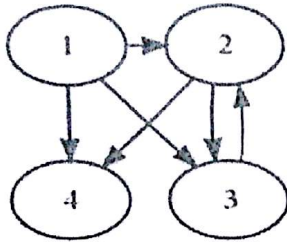
7. → Considering the following graph, compute the Closeness Centrality.
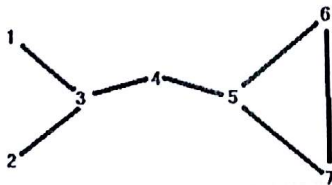
8. → Briefly explain latent semantic indexing process by applying Singular Value Decomposition approach with suitable example.

9. Using Expectation Maximization (Probability) algorithm compute mean, standard deviation and probability sampling, identify the data value 0.49 belong which cluster.

A	0
B	0.961
A	0
A	0
B	0.780
A	0
B	0
A	0
B	0.980
A	0
B	0.135
A	0.490
B	0.928
B	0
B	0.658
A	0
A	0
A	0.387
A	0.570
B	0

10. a) Encode the following decimal values using Elias Gamma, Elias Delta and Golomb (b=8) **[5]** Encoding techniques. Decimals to encode (show three encodes for every decimal separately): 11 and 19.

 [5]

 b) Compute the proximity centrality for the following graph
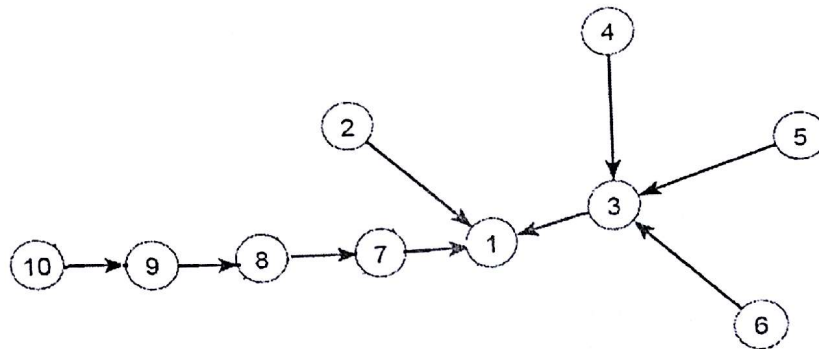
11. a) Explain the working mechanism of focused crawler and context-focused crawler. **[5]**

 [5]

 b) Consider the following database having five transactions.

TID	ITEMS BOUGHT
T1001	{Mango, Onion, ~~Nut~~, Kite, Eggs, Yo-yo}
T1002	{~~Rice~~, Onion, ~~Nut~~, Kite, Eggs, Yo-yo}
T1003	{Mango, ~~Apple~~, Kite, Eggs}
T1004	{Mango, ~~Umbrella~~, ~~Corn~~, Kite, Yo-yo}
T1005	{~~Corn~~, Onion, Kite, ~~Ice-cream~~, Eggs}

Let min_sup = 60% and min_conf = 77%. Find all frequent itemsets using Apriori algorithm. List all of the *strong* association rules with support and confidence.

⇔⇔⇔

Scanned by CamScanner