Register Number:……..

KUMARAGURU
college of technology
character is life

B.E/B.TECH DEGREE EXAMINATIONS: NOV/DEC 2023

(Regulation 2018)

Sixth Semester

INFORMATION SCIENCE AND ENGINEERING

U18ISE0015: Data Mining

COURSE OUTCOMES

CO1: Understand NLP techniques and text representation.

CO2: Understand mixture models and apply them for analysing topic from text.

CO3: Perform text clustering and categorization.

CO4: Analyze sentiment and mine opinion from text.

CO5: Understand pattern discovery concepts, approaches and evaluation measures.

Time: Three Hours **Maximum Marks: 100**

Answer all the Questions

PART A (10 x 2 = 20 Marks)

(Answer not more than 40 words)

1.	State the issues in Natural Language Processing	CO1	[K_2]
2.	What is the purpose of Entropy measure. Write the formula.	CO1	[K_2]
3.	Define unigram language model with an example.	CO2	[K_2]
4.	Find the value of P("Natural language processing"|θ) where the w and P(w|θ) are given in the following table.	CO2	[K_3]

w	P(w|θ)
Natural	0.1
Language	0.2
Learning	0.3
Processing	0.4

5.	Differentiate Generative and Discriminative Classifiers	CO3	[K_2]
6.	List the types of categorization	CO3	[K_2]
7.	Define CPLSA.	CO4	[K_2]
8.	Consider a scenario where we are in need to collect user responses from various geographic locations. what kind of technique should be used? Justify your answer.	CO4	[K_2]

9.	Describe the pros of FP growth algorithm			CO5	[K_2]
10.	T_id	Items bought		CO5	[K_3]
	10	Beer, Nuts, cheese			
	20	Beer, Coffee, cheese, Nuts			
	30	Beer, cheese, Eggs			
	40	Beer, Nuts, Eggs, Milk			
	50	Nuts, Coffee, cheese, Eggs, Milk			
	Given the transactions, mini-support (minsup) s = 50%, and minconf c = 50%, check whether the following rule is strong association rule or not.				
	{Beer, Nuts} ⇒ {cheese}				

Answer any FIVE Questions

PART B (5 x 16 = 80 Marks)

(Answer not more than 400 words)

11.	a)	Explain in detail about Paradigmatic and Syntagmatic relations for word predictions with examples	10	CO1	[K_2]
	b)	What is EOWC? How the similarity measure is calculated?	6	CO1	[K_2]
12.	a)	Write the formula for likelihood function for any two in generative probabilistic model with an example	8	CO2	[K_2]
	b)	How EM algorithm is compared with hill climbing and state the convergence of the same. State its merits and demerits.	8	CO2	[K_2]
13.		Cluster the following eight points (with (x, y) representing locations) into three clusters: A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9) Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2). The distance function between two points a = (x1, y1) & b = (x2, y2) is defined as $P(a, b) = \|x2 - x1\| + \|y2 - y1\|$ Use K-Means Algorithm to find the three cluster centers after the second iteration.	16	CO3	[K_3]

14.	a)	Assume that the likelihood function of PLSA has multiple local maxima and one global maximum. There exists an initial set of parameters for which PLSA will converge to the global maximum of the likelihood function - (True/False) Justify your answer and discuss about PLSA .	8	CO4	[K_4]
	b)	What is Text Categorisation. How sentiment classification output is processed in it?	8	CO4	[K_2]
15.		Define interestingness Measure. Discuss in detail about the types along with the purpose. Suppose University collected statistics on the number of students who take courses on Data Mining (DM) and Machine Learning (ML). Given the following 2 x 2 contingency table summarizing the statistics, calculate the $\chi 2$ score.	16	CO5	[K_3]

	DM	¬DM	Σrow
ML	700	300	1000
¬ML	500	1500	2000
Σcol	1200	1800	3000

16. A database has four transactions. Let min sup=60% and min conf =80%. 16 CO5 [K_3]

TID	Items bought (in the form of brand-item category)
T100	{King's-Crab, Sunset-Milk, Dairyland-Cheese, Best-Bread}
T200	{Best-Cheese, Dairyland-Milk, Goldenfarm-Apple, Tasty-Pie, Wonder-Bread}
T300	{Westcoast-Apple, Dairyland-Milk, Wonder-Bread, Tasty-Pie}
T400	{Wonder-Bread, Sunset-Milk, Dairyland-Cheese}

Write the Apriori algorithm along with its advantages and disadvantage. List the frequent k-itemset for the largest k, and all the strong association rules.

****