Register Number:……………..

KUMARAGURU
college of technology
character is life

B.E DEGREE EXAMINATIONS: NOV/DEC 2022

(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 analyzing 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. List the four types of Text Mining tasks. CO1 [K_1]
2. Identify the drawback of using conditional entropy in mining syntagmatic relations. CO1 [K_3]
3. What is the difference between LDA and PLSA? CO2 [K_2]
4. What do you infer from the estimated parameter values of PLSA? CO2 [K_2]
5. Define Agglomerative Clustering? CO3 [K_1]
6. Write the K-means clustering algorithm. CO3 [K_2]
7. Recall the two ways of defining the output sentiment class label? CO4 [K_2]
8. “The use of longer n-grams as features may cause overfitting”. Justify the sentence. CO4 [K_5]
9. State Apriori property. CO5 [K_1]
10. How to compute confidence for an association rule A⇒B? CO5 [K_1]

Answer any FIVE Questions:-
PART B (5 x 16 = 80 Marks)
(Answer not more than 400 words)

11. a) Illustrate with an example to explain the discovery of paradigmatic relations between the words? (8) CO1 [K_3]

b) Explain how Mutual Information can be used for syntagmatic relation mining with an example. (8) CO1 [K_3]

12. a) Compare unigram language model with two-component mixture model. (8) CO2 [K_2]

b) Explain how Expectation-Maximization (EM) algorithm can be used to compute the ML estimate of two-component mixture model. (8) CO2 [K_1]

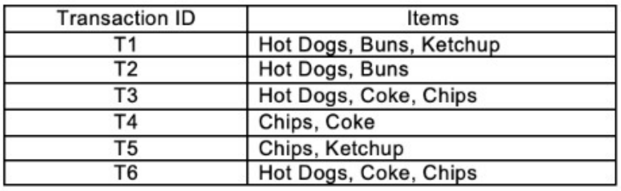
13. a) Construct with an example the working of K-Nearest Neighbor algorithm. (10) CO3 [K_2]

b) How do you evaluate the performance of Text Categorization algorithms? (6) CO3 [K_4]

14. a) Explain the two steps involved in Latent Aspect Rating Analysis with an example. (10) CO4 [K_3]

b) Show how a binary logistic regression can be used to solve multilevel rating prediction. (6) CO4 [K_3]

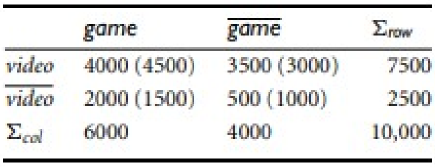
15. Find the frequent itemsets and strong association rules for the following transaction table using Apriori algorithm. Assume that minimum support threshold s=33.33% and minimum confidence threshold c=60%. (16) CO5 [K_3]



Transaction ID	Items
T1	Hot Dogs, Buns, Ketchup
T2	Hot Dogs, Buns
T3	Hot Dogs, Coke, Chips
T4	Chips, Coke
T5	Chips, Ketchup
T6	Hot Dogs, Coke, Chips

16. a) Write the FP-growth algorithm for mining frequent itemsets. (10) CO5 [K_2]

b) The contingency table below shows the observed and expected values (within parenthesis) of the transactions with respect to Game and Video purchases. Perform correlation analysis using Lift and χ2. (6) CO5 [K_3]



	game	$\overline{game}$	Σ_{row}
video	4000 (4500)	3500 (3000)	7500
$\overline{video}$	2000 (1500)	500 (1000)	2500
Σ_{col}	6000	4000	10,000

****