Roll No											

PRESIDENCY UNIVERSITY
BENGALURU

SCHOOL OF ENGINEERING
MID TERM EXAMINATION - OCT 2023

Semester : Semester VII - 2020
Course Code : CSE3134
Course Name : Sem VII - CSE3134 - Text Mining and Text Analytics
Program : B. TECH

Date : 31-OCT-2023
Time : 2:00PM - 3:30PM
Max Marks : 60
Weightage : 30%

Instructions:
(i) Read all questions carefully and answer accordingly.
(ii) Question paper consists of 3 parts.
(iii) Scientific and non-programmable calculator are permitted.
(iv) Do not write any information on the question paper other than Roll Number.

PART A

ANSWER ALL THE QUESTIONS **(5 X 2 = 10M)**

1. List the advantages of turning text data into **high-quality information** or **actionable knowledge?**
(CO1) [Knowledge]
2. Label the words in the following statement with the proper part-of-speech tag from the Universal Tagset.
A dog is chasing a boy on the playground
(CO1) [Knowledge]
3. Expand EOWC.
(CO1) [Knowledge]
4. List the issues with "Term as Topic".
(CO2) [Knowledge]
5. Define Topic.
(CO2) [Knowledge]

PART B

ANSWER ALL THE QUESTIONS **(4 X 5 = 20M)**

6. My _____ eats _____ on Saturday
His _____ eats _____ on Tuesday

Explain how the intitution helps in fill in the blanks above using paradigmatic and syntagmatic word association rules.
(CO1) [Comprehension]

7. Apply the lexical analysis, syntatic analysis and semantic analysis on the given statement.

A dog is chasing a boy on the playground.

(CO1) [Comprehension]

8. The anonymous nature of the internet and the many communication features operated through it contribute to the increased risk of internet-based crimes. Today, text mining is making cybercrime prevention easier for enterprise organizations as well as law enforcement by establishing more context around the intelligence they are being fed. This enables them to pinpoint real threats and limit the number of false positives created by keywords taken out of context.
Explain the kind of crimes are these and how Probabilistic Topic Models can help us to prevent them

(CO2) [Comprehension]

9. Suppose a fisherman has three baskets; one basket contains 9 pieces of Tuna and 1 Salmon, the other contains 9 pieces of Salmons & 1 Tuna while the last basket contains 5 pieces of Salmons & 5 Tunas.

Now, if the fisherman just randomly placed his hands in the first basket to pick a fish, knowing that there are 9 Tunas and only one Salmon means there is a higher probability that he will pick a Tuna and since there is a higher probability of picking up a Tuna it would not be surprising if he does. In contrast, if the fisherman picked up the Salmon from the basket we would be relatively surprised.
The second basket has a lot more Salmon than Tunas and because there is now a higher probability of picking up a Salmon we would not be very surprised if it happened and because there is a relatively low probability of picking the Tuna, it would be relatively surprising if he does.
The third basket has an equal number of Tunas and Salmons thus regardless of what fish he picks up we would be equally surprised. Combined, these baskets tell us that **Surprise** is in some way **inversely** related to **probability**. In other words, when the probability of picking up a Salmon is low as it was in the first basket, the surprise is high and when the probability of picking up a Tuna is high, the surprise is low.

Explain how surprise is related to entropy.

(CO2) [Comprehension]

PART C

ANSWER ALL THE QUESTIONS **(2 X 15 = 30M)**

10. a) Define TF
b) Define IDF
c) When a **100-word document** contains the term “cat” **12 times**, Calculate the TF of **cat.**
d) Let’s say the size of the corpus is 10,000,000 million documents. If we assume there are 0.3 million documents that contain the term “cat”, then the IDF (i.e. log {DF}) is given by the total number of documents (10,000,000) divided by the number of documents containing the term “cat” (300,000). Calculate the IDF of **cat.**
e) Calculate the TF-IDF score of word **cat** in the given corpus.

(CO1) [Application]

11. Explain the Mutual Information with an example.

(CO2) [Application]