BENNETT UNIVERSITY
THE TIMES GROUP

Enrolment No: E22CSEU0827 Name of Student: MADHAV GUPTA

Department/ School: ~~E22CSEU0~~ SCSET

END-TERM EXAMINATION, ODD SEMESTER DECEMBER 2024

COURSE CODE: CSET346 **MAX. DURATION: 2 HRS**

COURSE NAME: Natural Language Processing

PROGRAM: B.Tech **TOTAL MARKS: 40**

Mapping of Questions to Course and Program Outcomes										
Q.No.	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5
CO	1	1	2,3	1,2	2,3	1	2,3	1,3	1,2	2,3
PO	1	1,2	2,3	3, 4	2,3	2,3	1,2	4,5	1 ,2	3,5
BTL	1	2	1	2	3	1	3	2	3	4

GENERAL INSTRUCTIONS: -

1. Do not write anything on the question paper except **name, enrolment number** and **department/school.**
2. Carrying mobile phones, smartwatches and any other non-permissible materials in the examination hall is an act of **UFM.**

COURSE INSTRUCTIONS:

a) All the answers should be written with proper index number.

b) In case of incorporating any diagram, it should be neat and clear.

c) All questions are compulsory to answer.

SECTION A

[5Q × 3 Marks = 15 Marks]

A1) a) Explain the different components of NLP?

b) Describe the disadvantages of N-gram model? **[2 + 1 = 3 Marks]**

A2) Explain Vanishing and Exploding Gradients with proper example? **[3 Marks]**

A3)a) Illustrate the disadvantages of context free language modelling techniques in NLP?

b) Write the Markov assumption? **[2 + 1 = 3 Marks]**

A4) a) Describe Perplexity along with the formula? **[1 + 2 = 3 Marks]**

b) Explain the key components of Hidden Markov Models?

A5) Define smoothing? How does smoothing deals with zero probability? **[2 + 1 = 3 Marks]**

SECTION B

[5Q × 5 Marks = 25 Marks]

B1) Consider the following training data [5×1 =5 Marks]

<S>I am Mike</S>

<S>Mike I am</S>

<S>Mike I like</S>

<S>Mike I do like </S>

<S>Do I like Mike</S>

Assume that we use a bigram language model based on above data. What does the most probable next word predict by the model of the following data?

1. <S>Mike _____
2. <S>Mike I do_____
3. <S>Mike I am _____
4. <S>Do I like_____
5. <S>I_____

B2) a) Explain the Multinomial Naïve Bayes algorithm and write the mathematical expression for text classification?

b) Consider the following dataset of Documents classification problem, where 4 documents (Doc1, Doc2, Doc3, and Doc4) and their corresponding classes (A and B) are there. Predict the class of Doc5 using Multinomial Naïve Bayes Classifier? [2+3 = 5 Marks]

Doc Number	Word contains	Class
Doc1	Apple, Mango, Apple	A
Doc2	Apple, Apple, Banana	A
Doc3	Apple, Cherry	A
Doc4	Lemon, Orange, Apple	B
Doc5	Apple, Apple, Apple, Lemon, Orange	?

B3) a) Is stemming technique improve effectiveness in the text processing tasks? Justify your answer?

b) Explain the different types of Tokenization in NLP with proper example? [2 + 2 + 1 = 5 Marks]

c) Why is regular expression necessary in text processing? Justify your answer?

B4) a) Explain the components of GRU? [3 + 2 = 5 Marks]

b) Explain the BERT model in NLP. How does BERT differ from traditional language models?

B5) A weather prediction system uses a Hidden Markov Model (HMM) to forecast the weather for the next few days. There are three possible hidden states: **Sunny (S), Cloudy (C),** and **Rainy (R).** The observable outputs are two weather conditions: **Umbrella (U)** and **Raincoat (RC).**

The model has the following parameters:

Transition Probabilities: [arrow (→) represents the transition from one state to another state]

$P(S \rightarrow S) = 0.5$ $P(S \rightarrow C) = 0.3$ $P(S \rightarrow R) = ?$

$P(C \rightarrow S) = 0.4$ $P(C \rightarrow C) = ?$ $P(C \rightarrow R) = 0.4$

$P(R \rightarrow S) = 0.3$ $P(R \rightarrow C) = 0.3$ $P(R \rightarrow R) = ?$

Emission Probabilities:

$P(U \mid S) = 0.7$ $P(U \mid C) = 0.3$ $P(U \mid R) = 0.1$

$P(RC \mid S) = 0.3$ $P(RC \mid C) = 0.7$ $P(RC \mid R) = 0.9$

Initial State Probabilities:

$P(S) = 0.5$ $P(C) = 0.3$ $P(R) = 0.2$

a) Draw the state transition diagram for the given data? **[1 + 2 + 2 = 5 Marks]**

b) Given that the weather on day 1 is sunny, what is the probability that the weather for the next 5 days will be "Sunny – Sunny - Cloudy- Cloudy - Rainy?

c) Determine the probability of hidden sequence (Sunny – Sunny - Cloudy - Cloudy - Rainy) that led to the observed sequence: (Umbrella – Umbrella – Raincoat– Umbrella – Raincoat) ?