AUTUMN END SEMESTER EXAMINATION-2023
5th Semester B.Tech (DE-I)

NATURAL LANGUAGE PROCESSING

IT 3035

(For 2022 (L.E), 2021 & Previous Admitted Batches)

Time: 3 Hours   Full Marks: 50

Answer any SIX questions.
Question paper consists of four SECTIONS i.e. A, B, C and D.
Section A is compulsory.
Attempt minimum one question each from Sections B, C, D.
The figures in the margin indicate full marks.
Candidates are required to give their answers in their own words as far as practicable and all parts of a question should be answered at one place only.

SECTION-A

1.   Answer the following questions.   [1 × 10]

(a)   Define a HMM.

(b)   What is the necessity of lemmatization?

(c)   Write the phrase structure of the following sentences using Brown tags.

        a. The homework is very easy.

        b. Go and sleep on your bed.

(d)   Differentiate between attachment and coordination ambiguity with examples.

(e)   Why accuracy is not a good measure in natural language processing?

(f)   Define prefixes, infixes, suffixes and circumfixes with appropriate examples.

(g) Kullback-Leibler divergence do not satisfy the triangle inequality. State True or False. Justify your answer.

(h) What is POS tagging?

(i) What does recall refer to in classification?

(j) Define derivation and compounding in natural language processing with suitable examples.

SECTION-B

2. (a) Consider the following corpus $C3$ of four sentences. [4]
   - \<s\> *monkies love banana*\</s\>
   - \<s\> *one monkey is eating apple*\</s\>
   - \<s\> *three monkies are eating jackfruit*k \</s\>
   - \<s\> *one is silent*\</s\>

   Assume a bi-gram language model.

   Calculate P(\<s\> *silent monkey is eating apple*\</s\>)

   (b) Differentiate between BoW and tf-idf model with suitable example. [4]

3. Compute the edit distance between the words "INDORE" and "INDOOR" using dynamic programming. [8]

SECTION-C

4.

\<S\>	Bapu	krishu	will	see	cricket	\<E\>
	N	N	M	V	N	
\<S\>	Bapu	will	watch	Rishu	\<E\>	
	N	M	V	N		
\<S\>	will	Rishu	play	Football	\<E\>	
	M	N	V	N		
\<S\>	Rishu	Will	play	cricket	\<E\>	
	N	M	V	N		

[8]

Using Viterbi algorithm find out the appropriate POS tagging for the test data "Bapu watch krishu".

5.

No.	Height	Tone	Hair	Species
1	5	sharp	Small	Girl
2	6	rude	Long	Boy
3	6	rude	Small	Boy
4	5	sweet	Small	Girl
5	6	sharp	Long	Boy
6	4	soft	Long	Boy
7	4	soft	Small	Girl
8	5	sweet	Long	Boy

[8]

If a person has, height=4, Tone=rude, Hair=Small, find out the probable gender of the person?

6. Given the following data: [8]

<s> I love pizza</s>

<s> I like pizza</s>

<s>pizza I do like </s>

<s> do I taste pizza</s>

<s>tasty pizza is yummy</s>

Assume that we use a bigram language model based on the above data.

Which of the following sentences is better, i.e., gets a higher probability with this model?

(1) <s>pizza is tasty</s>

(2) <s>pizza is yummy</s>

SECTION-D

7. (a) Describe classic NLP pipeline. Describe lexical ambiguity with example. [4]

(b) Simplify the following grammer: [4]

S ----> aA | aBB

A ----> aaA | λ

B ----> bB | bbC

C ----> B

8. Write short notes: [3+3+2]

(a) Maximum likelihood estimation

(b) Perplexity

(c) IBM model-1

*****