Roll No.												

PRESIDENCY UNIVERSITY

BENGALURU

End - Term Examinations – MAY 2025

Date: 21-05-2025 **Time:** 01:00 pm – 04:00 pm

School: SOIS	**Program:** MCA	
Course Code: CSA4040	**Course Name:** Natural Language Processing	
Semester: IV	**Max Marks**: 100	**Weightage**: 50%

CO - Levels	**CO1**	**CO2**	**CO3**	**CO4**	**CO5**
Marks	**20**	**20**	**30**	**30**	

Instructions:

(i) Read all questions carefully and answer accordingly.

(ii) Do not write anything on the question paper other than roll number.

Part A

Answer ALL the Questions. Each question carries 2marks. **10Q x 2M=20M**

1.	Name the algorithm that is used to generate the best possible state sequence, given the observation sequence.	**2 Marks**	**L1**	**CO3**
2.	State true or false. The XOR gate is a linear separable function.	**2 Marks**	**L1**	**CO3**
3.	Name the activation function that takes a vector as input and returns a probability distribution.	**2 Marks**	**L1**	**CO3**
4.	State any multilingual pre-trained language model that consists of only Indian languages (and English).	**2 Marks**	**L1**	**CO3**
5.	State GPE in the context of named entity recognition.	**2 Marks**	**L1**	**CO3**
6.	Write the output of the following program: from nltk.stem import PorterStemmer stemmer = PorterStemmer() print(stemmer.stem(“dancing”))	**2 Marks**	**L1**	**CO4**
7.	Name the NLP task which uses the NLTK resource “averaged_perceptron_tagger_eng”	**2 Marks**	**L1**	**CO4**

8.	Recall the function in NLTK, which takes a text as input and splits the text into sentences, returning a list of sentences as output.	**2 Marks**	**L1**	**CO4**
9.	State the Python library which is used to parse CSV files.	**2 Marks**	**L1**	**CO4**
10.	Fill in the blanks for the below code to load the Spacy model for different NLP tasks: import spacy nlp = spacy.____("en_core_web_sm")	**2 Marks**	**L1**	**CO4**

Part B

Answer the Questions. **Total Marks 80M**

11. Associate the entries in column A with those of columns B and C. **20 Marks** **L2** **CO1**

Column A	**Column B**	**Column C**
A. Sentiment Analysis	**F.** Syntactic Grammars	**K.** 1954
B. Part-of-Speech Tagging	**G.** Document Classification	**L.** Colourless Green Ideas Sleep Furiously
C. Noam Chomsky	**H.** Machine Translation	**M.** Turing Test
D. Alan Turing	**I.** Word Classification	**N.** Penn Treebank
E. Georgetown Experiment	**J.** Imitation Game	**O.** Binary Polarity (Eg. Positive / Negative)

NOTE: For your answers, you ONLY NEED TO WRITE the letters (Eg. AFK). No need to write all 3 entities in your group.

Or

12. Classify each of the following sentences as either positive, negative, neutral, or sarcastic: **20 Marks** **L2** **CO1**

A. Huggingface Transformers was recognized as the Best System at EMNLP 2020.
B. Detecting Sarcasm is very easy ;)
C. The rain in Spain stays mainly in the plain.
D. The major flaw of *Dragon Ball Z* is the extremely repetitive fighting that goes on for close to 300 episodes.
E. Donald Trump, who is a stable genius, is the greatest President of the United States in its history.
F. The plot of the movie is extremely unpredictable.
G. The steering of the car is extremely unpredictable.
H. The battery of the mobile phone gives a good backup of 2 days.
I. The battery of the mobile phone gives a bad backup of 2 hours.

	J. The battery of the mobile phone gives an awesome backup of 2 hours.			

13.	For the following data, draw the trellis, and predict the part-of-speech tags for the sentence "Count the vote". **Emission Probability:** **Transition Probability:** Mention the Viterbi probabilities and back-pointers for each node, as well as the part-of-speech tags.	**20 Marks**	**L3**	**CO3**

Emission Probability:

	DT	**NN**	**VB**
The	0.2	0	0
Fans	0	0.1	0.2
Watch	0	0.3	0.15
Race	0	0.1	0.3
Count	0	0.2	0.2
Vote	0	0.1	0.2
Stop	0	0.2	0.2

Transition Probability:

	DT	**NN**	**VB**
$(START)	0.8	0.1	0.1
DT	0	0.9	0.1
NN	0.1	0.4	0.5
VB	0.5	0.4	0.1

Or

14.	For the data given in **Question 13**, draw the trellis, and predict the part-of-speech tags for the sentence "Stop the count". Mention the Viterbi probabilities and back-pointers for each node.	**20 Marks**	**L3**	**CO3**

15.	Solve the below problem(s) using the **glove-wiki-gigaword-50** model from the gensim library: a. Write a function to find the similarity score between 2 words. Call the function to find the similarity between the words "dog" and "cat".	**20 Marks**	**L3**	**CO2**

	b. Write a function to complete the analogy of wordA:wordB=wordC:? Your function must return the missing word. Call the function to complete the analogy "man:woman=king:?"			
	Or			
16.	Solve the problem of language identification by writing a function in Python to find out the language of a given text. Your function should convert the text into a bag-of-words representation and the Multinomial Naïve Bayes classifier.	**20 Marks**	**L3**	**CO2**

17.	Find the sentiment scores of a text using VADER and Use that information to classify whether the text is positive or negative. Test the code by calling the function for sentiment scores prediction. Call the function for the input "I love the new design of the website."	**20 Marks**	**L3**	**CO4**
	Or			
18.	Use a Hidden Markov Model with the Forward Algorithm using the following parameters: • States: ['U1', 'U2', 'U3'] • Observations: ['R', 'G', 'B'] (Red, Green, Blue) • Observation Sequence: ['R', 'R', 'G', 'G', 'B'] • Initial Probabilities: [0.4, 0.3, 0.3] • Transition Probabilities: [[0.1, 0.4, 0.5], [0.6, 0.2, 0.2], [0.3, 0.4, 0.3]] • Observation Probabilities: [[0.3, 0.5, 0.2], [0.1, 0.4, 0.5], [0.6, 0.1, 0.3]]	**20 Marks**	**L3**	**CO4**