Roll No.												

GAIN MORE KNOWLEDGE
REACH GREATER HEIGHTS

PRESIDENCY UNIVERSITY

BENGALURU

End - Term Examinations – MAY 2025

Date: 20-05-2025 **Time:** 09:30 am – 12:30 pm

School: SOCSE	**Program:** B. Tech- CAI/COM/CSE/CSG	
Course Code: CSE3188	**Course Name:** Natural Language Processing	
Semester: VI	**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 a pre-trained language model which has approximately 12 million parameters and supports a dozen Indian languages.	**2 Marks**	**L1**	**CO3**
2.	Suppose 2 annotators annotate a lexical resource which features **ordered** classes. Name an inter-annotator agreement measure which is suitable for this task.	**2 Marks**	**L1**	**CO3**
3.	State true or false. We do not use the probability of the word sequence (P(W)), in our forward probability calculations in HMM, because P(W) = 1, irrespective of the sequence of words.	**2 Marks**	**L1**	**CO3**
4.	Consider that we have a set of |T| tags. If we assume a **uniform distribution** for the tags, state the emission probability for an unknown word, given a tag.	**2 Marks**	**L1**	**CO3**
5.	PoS tagging is a sequence labeling task, which uses the Viterbi algorithm. Name the algorithm paradigm of the Viterbi algorithm.	**2 Marks**	**L1**	**CO3**
6.	**LAB Question**. NLTK's translate.bleu_score library has 2 different functions to evaluate the BLEU score. Name both of them.	**2 Marks**	**L1**	**CO4**
7.	**LAB Question**. Name the package which we use to download and use for GloVe word vectors.	**2 Marks**	**L1**	**CO4**
8.	**LAB Question**. Name the resource which we use for PoS tagging in NLTK.	**2 Marks**	**L1**	**CO4**
9.	**LAB Question**. Name the Python library which contains functions for evaluation of text classification outputs.	**2 Marks**	**L1**	**CO4**

10.	**LAB Question**. Name the Python library which uses “en_core_web_sm” as a resource.	2 Marks	L1	CO4

Part B

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

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

Column A	**Column B**
A. Speech	**K.** Co-reference Resolution
B. Lexical Analysis	**L.** Word Boundary Detection
C. Syntactic Analysis	**M.** Cat & Dog
D. Pragmatics and Discourse	**N.** Part-of-Speech Tagging
E. Synonymy	**O.** Pen & Paper
F. Similarity	**P.** Natural Language Ambiguity
G. Association	**Q.** Good & Bad
H. Antonymy	**R.** Huge & Large
I. Ordered Classes	**S.** Cohen’s Unweighted Kappa
J. Unordered Classes	**T.** Cohen’s Weighted Kappa

NOTE: For your answers, you **ONLY NEED TO WRITE** the letters (Eg. AK). No need to write the entries.

Or

12. Consider a situation where we have a set of ***k*** classes (numbered from C_1 to C_k). We have 2 annotators, A1 and A2. A1 is a diligent annotator who assigns every instance to the proper class. Hence, for each C_i, A1 assigns N_i instances. A2, on the other hand, is a lazy annotator, who assigns every instance to the same class (without loss of generality, let us say that A2 assigns all N instances to the class C_1). Verify the equality between the observation and expectation matrix and, based on that result, show that the Kappa value is 0, irrespective of the Kappa used. — **20 Marks** — **L2** — **CO1**

13. Tag the following text: "The races watch the fans" using the Viterbi Algorithm. Assume that you have only 3 tags – DT, VB, and NN. You can use the following tables: — **20 Marks** — **L3** — **CO3**

Emission	**DT**	**NN**	**VB**
The	0.2	0	0
Fans	0	0.1	0.2
Watch	0	0.3	0.15
Races	0	0.1	0.3

Transition	**DT**	**NN**	**VB**	**^ (END)**
$ (START)	0.8	0.2	0	0
DT	0	0.9	0.1	0
NN	0	0.5	0.5	1
VB	0.5	0.5	0	1

	Draw the trellis. For each **NON-ZERO Emission Probability** node, calculate the Viterbi Probabilities, as well as the back-pointers. Then, you should tag the sentence.			
	Or			
14.	Tag the following text: "The watch races the fans" using the Viterbi Algorithm. The tagset, probabilities, etc. are the same as given in **Question 13**. Draw the trellis. For each **NON-ZERO Emission Probability** node, calculate the Viterbi Probabilities, as well as the back-pointers. Then, you should tag the sentence.	**20 Marks**	**L3**	**CO3**

15.	**LAB Question.** ***Deutschlandanglization*** (yes, this is a made-up word) is when we transcribe English in such a way that **ALL nouns** (not just proper nouns) start with an uppercase letter. A ***Deutshdlandanglizer*** is a system that ***Deutschladanglizes*** the text. Explain how we build a ***Deutschdlandanglizer***, where we capitalize only nouns using a Hidden Markov Model, **WITHOUT** using a part-of-speech tagger, or a part-of-speech tagged corpus. You are given only a ***Deutschdlandanglized*** corpus (without explicit states written). For this question, you must list out (a) the different states, (b) the list of observations, and (c) how you calculate the different probabilities. Use your ***Deutschdlandanglizer*** to then ***Deutschdlandanglize*** the following texts: **a.** many hands make light work : using essay traits to automatically score essays **b.** a survey on using gaze behaviour for natural language processing **c.** happy are those who grade without seeing : a multitask learning approach to grade essays using gaze behaviour d. eyes are the windows to the soul : predicting the rating of text quality using gaze behaviour e. ASAP++ : enriching the ASAP automated essay grading dataset with essay attribute scores	**20 Marks**	**L3**	**CO2**
	Or			
16.	**LAB Question.** One of the ways in which we evaluate the quality of a machine translation system is by using the BLEU score. However, that relies on having reference data. In the absence of reference data (but assuming that we have a system trained from source to target and target to source language), write a function to evaluate both systems. This is called **Round-Trip Translation**. Assume that you have the following functions: 1. **translateSentence**(sentence, source, target), which translates the sentence from the source language to the target language. 2. **getBLEU**(candidate, reference, weights), which returns the BLEU score between the candidate and reference sentences. Weights is a 4-tuple which is the set of weights to be given to the unigram, bigram, trigram, and 4-gram precisions. Write a function to perform Round-Trip Translation. Then, calculate the values returned by: a) getBLEU("Stuff my heart agricultural to know he wants things to keep in mind", "Things my heart used to know things it yearns to remember", (0.5, 0.5, 0, 0)) b) getBLEU("Stuff my heart agricultural to know he wants things to keep in mind", "Things my heart used to know things it yearns to remember ", (0.4, 0.6, 0, 0)) Consider that the texts are case-insensitive.	**20 Marks**	**L3**	**CO2**

17. **LAB Question.** Professor SAM wants to create a document corpus. So, he takes 20 unlabeled documents, and asks 2 annotators - PCM & PGM - to label them as either COMEDY or TRAGEDY. Here is the result of the classifications by the 2 annotators:

Document	**PCM Label**	**PGM Label**
D01	Tragedy	Comedy
D02	Comedy	Comedy
D03	Tragedy	Comedy
D04	Comedy	Comedy
D05	Comedy	Comedy
D06	Comedy	Comedy
D07	Tragedy	Comedy
D08	Tragedy	Comedy
D09	Tragedy	Tragedy
D10	Tragedy	Tragedy
D11	Tragedy	Comedy
D12	Comedy	Comedy
D13	Comedy	Comedy
D14	Tragedy	Comedy
D15	Comedy	Comedy
D16	Tragedy	Tragedy
D17	Tragedy	Tragedy
D18	Tragedy	Tragedy
D19	Tragedy	Tragedy
D20	Tragedy	Comedy

Write a program to calculate the unweighted, linear weighted and the quadratic weighted Kappas. Show that all 3 Kappas are equal (irrespective of what ratings the annotators gave) and find out the value of the Kappa in the above example.

20 Marks **L3** **CO4**

Or

18. **LAB Question.** Two annotators (A1 and A2) are using the following tagset:

NN = Noun, VB = Verb, JJ = Adjective, RB = Adverb, FW = Function word, and PM is a punctuation mark.

They are annotating the following text: "You enter a very dark room , and sitting there in the gloom , is Dracula , so how do you say goodbye ?"

A1 = FW FW VB FW JJ NN PM FW FW FW VB FW PM VB FW NN FW VB FW VB FW FW NN PM

A2 = FW VB FW RB JJ NN PM FW VB FW FW FW NN PM VB NN PM FW FW VB FW VB NN PM

Write a function, which takes the 2 annotators' annotations as input and calculates the ***appropriate*** Kappa. Also, calculate the ***appropriate*** Kappa between the 2 annotators.

20 Marks **L3** **CO4**