Roll No: | | | | | | | | | | | |

PRESIDENCY UNIVERSITY

GAIN MORE KNOWLEDGE
REACH GREATER HEIGHTS

Presidency University Act, 2013 of the Karnataka Act No. 41 of 2013 | Established under Section 2(f) of UGC Act, 1956
Approved by AICTE, New Delhi | Approved By BCI
Bengaluru

Mid Term Examinations - October 2026

Date: 05/10/2026 **Time:** 11:45 AM - 01:15 PM

Course Code: CSE3504 **Course Name:** Computational Linguistics and Natural Language Pro

Semester: Seventh Semester **Max. Marks:** 50 **Weightage:** 50%

CO - Levels	CO1	CO2
Marks	50	40

PART A: Answer Following Questions. 10 M

Q.N.	Questions	M	CO	BT
1	Given a tagset of |T| tags, and assuming a uniform distribution, state the value of the emission / lexical probability of each tag.	2	CO1	BT1
2	State the formula of tanh(x)	2	CO1	BT1
3	Given that there are k common n-grams between 2 strings (where k > 0), state the **maximum number** of (n+1)-grams between the 2 strings.	2	CO1	BT1
4	State the full form of HISK.	2	CO1	BT1
5	The decision boundary of the sigmoid activation function takes place at z = 0. State the value of sigmoid(0).	2	CO1	BT1

PART B : Answer Any 1 Following Questions. 10 M

Q.N.	Questions	M	CO	BT
6	A limerick is a poem which has 5 lines, with a rhyming scheme of AABBA (1st, 2nd and 5th line rhyme, as does the 3rd and 4th line rhyme separately). Prof. SAM decides to create a ***limerick classifier***, where, given a set of 5 lines, he has to decide whether or not they form a limerick. To do this, he creates a ***static 26-dimension vector representation*** for ***the last word on each line***, such that each dimension stores the ***number of times that the character appears in the word***. For example, the vector for the word ***sam*** is given as: Vector[“sam”] = [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0] He then goes on to calculate the ***cosine similarities*** between each pair of ***last words*** for each pair of lines. He concludes that the set of lines constitute a limerick if the ***cosine similarities*** of the ***last words*** of the lines (1,2), (1,5), (2,5), and (3,4) are the four highest cosine similarities of the 10 possible cosine similarity pairs. With that rule, **predict** whether or not the following set of 5 lines constitutes a limerick (convert all to lowercase while finding the vectors): There was an old man from **Peru**, Who dreamt he was eating his **shoe**, In the middle of the **night**, He woke up with a **fright**, To see it was definitely **true**!	10	CO2	BT3
7	Consider the similar situation as stated in the previous question, except this time, use **Normalized HISK** (up to character **trigrams only**), to evaluate the similarity between the words and **predict** whether the poem (in the previous question) is a limerick or not.	10	CO2	BT3

PART C: Answer Any 1 Following Questions. 10 M				
Q.N.	**Questions**	**M**	**CO**	**BT**
8	Consider the following **training data** for a **Naive Bayes classifier**: 1. Just plain boring - NEGATIVE 2. Entirely predictable and lacks energy - NEGATIVE 3. No surprises and very few laughs - NEGATIVE 4. Very powerful - POSITIVE 5. The most fun film of the summer - POSITIVE Now, **classify** the sentence "The most boring film of summer" as either POSITIVE or NEGATIVE using Bag-of-Words representations of the documents and the Naive Bayes classifier.	10	CO1	BT2
9	Consider a 2 input feed forward neural network. **Express** the Boolean circuit it represents in the following situations: a). W1 = m, W2 = n, B = 1. m = n, m > 0, y = 1 if Z > 0 b). W1 = m, W2 = n, B = -(m+n). m = n, m > 0, y = 1 if Z >= 0 NOTE: You must construct the truth tables as well for this!	10	CO1	BT2

PART D: Answer Any 1 Following Questions. 10 M				
Q.N.	**Questions**	**M**	**CO**	**BT**
10	Recall the definition of a confusion matrix. C[i][j] = the number of times an instance was classified as class i, when in fact, it belonged to class j. In other words, the rows represent the predictions of the classifier, and the columns represent the actual classes. Based on this, consider the following confusion matrix: **Actual C1 Actual C2 Actual C3** **Predicted C1** a b c **Predicted C2** d e f **Predicted C3** g h i Consider that all the numbers from a to i are non-negative integers. Based on that, **express formulae** for the **accuracy** of the classifier, as well as the **precision**, **recall**, and **F1 Score** for **all 3 classes**. Your final answers should be **in terms of the variables** from **a to i**. NOTE: Try to simplify the answer as much as possible.	10	CO2	BT2
11	Consider the following situation. **term DF(t) IDF(t)** T1 a 0 T2 1 b Given **N** documents in the corpus, **express** the values of **a** and **b** in terms of N. Use logarithm to Base 10 for the calculations. Once done, **compute** the values of a and b if N = 10000.	10	CO2	BT2

PART E: Answer Any 1 Following Questions. 10 M				
Q.N.	**Questions**	**M**	**CO**	**BT**

<table>
<tr>
<td>12</td>
<td>Associate the terms in column 1 with those of column 2.
<table>
<tr><th>Column 1</th><th>Column 2</th></tr>
<tr><td>1. Normalized HISK</td><td>A. Range is between -1 and 1</td></tr>
<tr><td>2. Accuracy</td><td>B. Big and Large</td></tr>
<tr><td>3. Synonym</td><td>C. Calculated for the classifier, as a whole</td></tr>
<tr><td>4. Antonym</td><td>D. Dog and Cat</td></tr>
<tr><td>5. Similarity</td><td>E. Matrix of size T * T</td></tr>
<tr><td>6. Association</td><td>F. Linear function</td></tr>
<tr><td>7. Transition Probability</td><td>G. Good and Bad</td></tr>
<tr><td>8. Tanh(x)</td><td>H. Range is between 0 and 1.</td></tr>
<tr><td>9. NOR gate</td><td>I. Non-Linear Function</td></tr>
<tr><td>10. XOR gate</td><td>J. Pen and Paper</td></tr>
</table>
NOTE: For your answer, you just need to write the number and letter. For example, if you are associating "Normalized HISK" and "Range is between -1 and 1", simply write 1A.</td>
<td>10</td>
<td>CO1</td>
<td>BT2</td>
</tr>
<tr>
<td>13</td>
<td>Compute the edit distance between the following pair of words "natural" and "language". Show the working by constructing the dynamic programming tables. Use a substitution cost of 1.</td>
<td>10</td>
<td>CO1</td>
<td>BT2</td>
</tr>
</table>