|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  Roll No |  |  |  |  |  |  |  |  |  |  |  |

PRESIDENCY UNIVERSITY BENGALURU

 SCHOOL OF ENGINEERING

 MAKEUP EXAMINATION - JULY 2024

|  |  |
| --- | --- |
| **Semester :6** | **Date :** **03-07-2024** |
| **Course Code :** **CSE 243** | **Time :** **09.30am to 12.30am** |
| **Course Name :** **Natural Language Processing** | **Max Marks :100** |
| **Program :** **B.TECH (2019 Batch)** | **Weightage :50%** |

**Instructions:**

1. *Read all questions carefully and answer accordingly.*
2. *Question paper consists of 3 parts.*
3. *Scientific and non-programmable calculator are permitted.*
4. *Do not write any information on the question paper other than Roll Number.*

|  |
| --- |
| **PART A** |
|  **ANSWER ANY 4 QUESTIONS 4Q X 5M=20M** |
| 1 | Explain the stages of NLP in detail. | (CO 1) | [Knowledge] |
|  |
| 2 | Explain the TF-IDF algorithm in NLP. Find TF-IDF vectors use log (with base 10) for calculating IDF. D1=The best team plays the finals D2=India won a medal in the finals. | (CO2) | [Comprehension] |
|  |
| 3 | Discuss the common similarity measures used in Natural Language Processing (NLP). Explain Cosine similarity in detail. | (CO1) | [Knowledge] |
|  |
| 4 | An AI model when evaluated shows 0.92% accuracy. The value of Precision and Recall are 0.75 % each. What will be the value of F1 Score? | (CO2) | [Application] |

|  |
| --- |
| **PART B** |
|  **ANSWER ANY 5 QUESTIONS 5Q X 10M=50M** |
| 5 | Consider the problem of title casing. Title casing is where we capitalize the first letter of some parts of speech, while other words start with lower case. Explain, using a HMM, how we will perform title casing, given that we have a very large list of titles, but no part-of-speech tagger, or a part-of-speech tagged corpus. You will need to explain the states, as well as how you calculate the initial, transition and emission probabilities. | (CO2) | [Knowledge] |
|  |
| 6 | Describe the application of the Histogram Intersection String Kernel in Natural Language Processing (NLP). | (CO2) | [Knowledge] |
|  |
| 7 | BIOSE is another variant of the BIO tags, where we have the following expansions:* B = Beginning of the Named Entity span.
* I = Inside the Named Entity span.
* O = Outside the Named Entity span.
* S = Single word Named Entity
* E = Ending of the Named Entiity span

 For example, the named entity "United Arab Emirates" will be tagged as "B-LOCATION I-LOCATION ELOCATION" (NOTE: B and E take precedence over I). Similarly, the name "Mausam" will be tagged as "S-PERSON". Now, consider the following sentences:* European authorities fined [Google] a record $5.1 billion on Wednesday for abusing its power in the mobile phone market and ordered the company to alter its practices.
* [Barry Schwartz] entered the classroom and asked questions to the students about human nature and thinking skills.
* [Barcelona] is the capital of [Catalunya] in [Spain].
* [Narendra Modi] is the Prime Minister of [India].
* [Jimmy Doolittle] led a famous raid on [Tokyo] during World War II.

 For each of words in the [spans], assign the appropriate BIOSE tag. Assume that the only NER classes are PERSON, LOCATION, and ORGANIZATION. | (CO3) | [Comprehension] |
|  |
| 8 | Assume that we are using a small, 26-dimension vector to represent our words, such that each dimension represents the count of the character (from a to z) of our words. Eg. "sandeep" = [1, 0, 0, 1, 2, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]. For each word pair, compute the dot product and cosine similarity. * word1 = sitting, word2 = kitten
* word1 = donkey, word2 = money
* word1 = grain, word2 = grail
* word1 = table, word2 = stall
* word1 = hello, word2 = helm
 | (CO3) | [Comprehension] |
|  |
| 9 | Consider the following documents (Yes, each bullet point is a document):* Principles of Artificial Intelligence
* Artificial Intelligence for Gaming
* Artificial Intelligence and Machine Learning
* Artificial Intelligence for Game Development

 Assume only the following terms:* Principles
* Artificial
* Intelligence
* Gaming
* Machine
* Learning
* Game
* Development

 Write down the raw counts matrix, and generate the TF-IDF matrix, whose elements are weighted by the product of the TF and the IDF. Consider that the logarithm we are using is in base 10. | (CO2) | [Comprehension] |

|  |
| --- |
| **PART C** |
|  **ANSWER ANY 2 QUESTIONS 2Q X 15M=30M** |
| 10 | Consider the problem of title casing. Title casing is where we capitalize the first letter of some parts of speech, while other words start with lower case. Explain, using a HMM, how we will perform title casing, given that we have a very large list of titles, but no part-of-speech tagger, or a part-of-speech tagged corpus. You will need to explain the states, as well as how you calculate the initial, transition and emission probabilities. | (CO2) | [Comprehension] |
|  |
|  |
| 11 | Consider the following probabilistic context free grammar: (CO4) [Comprehension] Starting from the non-terminal symbol S, derive a parse tree, such that the probability of the parse tree is for the sentence "children hear songs with friends". Draw the parse tree from the derivations. | (CO4) | [Comprehension] |
|  |
|  |