BIRLA INSTITUTE OF TECHNOLOGY, MESRA, RANCHI
(END SEMESTER EXAMINATION)

CLASS: MTECH **SEMESTER : III**
BRANCH: AIML **SESSION : MO/2023**

SUBJECT: CS633 NATURAL LANGUAGE PROCESSING

TIME: 3 Hours **FULL MARKS: 50**

INSTRUCTIONS:
1. The question paper contains 5 questions each of 10 marks and total 50 marks.
2. Attempt all questions.
3. The missing data, if any, may be assumed suitably.
***4*. Before attempting the question paper, be sure that you have got the correct question paper.**
5. Tables/Data hand book/Graph paper etc. to be supplied to the candidates in the examination hall.

CO BL

Q.1(a) Outline the four main phases of Natural Language Processing. Briefly describe each phase, highlighting the key tasks involved in lexical analysis, syntactic analysis, semantic analysis, and pragmatic analysis. [5] 1,1,1

Q.1(b) [5] 1,2,1
(i) Discuss the concept of normalization in text preprocessing. Provide two examples of normalization techniques and explain how they contribute to improving the quality of textual data.
(ii) Compare and contrast lemmatization and stemming. Using a sample word, show the results of both lemmatization and stemming. Discuss when it might be preferable to use one technique over the other.

Q.2(a) Consider a corpus of text: "The cat sat on the mat." [5] 2,3,2
(i) Calculate the frequencies of each unique bi-gram in the given corpus. Provide the count for each bi-gram.
(ii) Using the bi-gram frequencies obtained in question 1, calculate the probabilities of each bi-gram. Express the probabilities in decimal form.
(iii) Apply add-one (Laplace) smoothing to the bi-gram probabilities calculated in question 2. Recalculate the probabilities with smoothing and provide the updated values.

Q.2(b) Define Neural Language Models and highlight their advantages over traditional N-gram models. How do neural language models handle long-range dependencies in language sequences? [5] 2,2,2

Q.3(a) Explain the concepts of Term Frequency-Inverse Document Frequency (TF-IDF) and Pointwise Mutual Information (PMI) in the context of word vectors. How do these measures contribute to capturing semantic relationships between words? [5] 3,3,3

Q.3(b) Provide an overview of Word2Vec. How does Word2Vec generate vector representations for words, and what are the key ideas behind its architecture? [5] 3,2,3

Q.4(a) Explain how Hidden Markov Models (HMMs) are used for parts-of-speech tagging. Discuss the key components of an HMM-based POS tagging system and the underlying principles. [5] 4,3,4

Q.4(b) Define Named Entity Recognition (NER) and explain its importance in natural language processing. How does NER differ from traditional parts-of-speech tagging, and what are its key challenges? [5] 4,3,4

Q.5(a) Discuss the working of CKY parsing algorithm. Explain how it processes a sentence to generate a parse table and identify possible parse trees. Provide a step-by-step explanation using a simple example. [5] 5,3,5

Q.5(b) Choose a specific text processing task (e.g., information extraction, question answering) and discuss how parsing techniques, especially syntactic parsing, can be applied to enhance the task's performance. Provide real-world examples and describe the impact of parsing on the chosen application. [5] 5,4,5

:::::22/11/2023 E:::::