Graphic Era HILL UNIVERSITY
Established by an Act of the State Legislature of Uttarakhand (Adhiniyam Sankhya 12 of 2011)
University under section 2(f) of UGC Act, 1956

End Term (Odd) Semester Examination December 2024

Roll no.....

Name of the Course and semester: B.tech-CSE(Specialization) Sem V
Name of the Paper: Natural Language Processing
Paper Code: TCS519
Time: 3 hour Maximum Marks: 100

Note:
(i) All the questions are compulsory.
(ii) Answer any two sub questions from a, b and c in each main question.
(iii) Total marks for each question is 20 (twenty).
(iv) Each sub-question carries 10 marks.

Q1. (2X10=20 Marks) (CO1)
a. Compare the performance of different morphological analysis models on various languages.
b. Explain the concept of topic modeling and its application in document analysis.
c. Describe the hierarchical structure of documents and the importance of identifying different levels of structure.

Q2. (2X10=20 Marks) (CO2)
a. Why is NLP considered a challenging field? Discuss the various sources of ambiguity in natural language
b. Define Chunking and its role in syntactic analysis. How does it help in understanding sentence structure?
c. List and explain the core tasks of NLP. Provide examples of real-world applications for each task.

Q3. (2X10=20 Marks) (CO3)
a. What is a regular expression, and why is it useful in text processing?
b. Explain the differences between Beautiful Soup and Scrapy for web scraping.
c. What are the legal and ethical considerations when performing web scraping for NLP projects?

Q4. (2X10=20 Marks) (CO4)
a Describe the difference between top-down and bottom-up parsing approaches.
b Evaluate the differences between manually annotated treebanks and automatically generated ones
c. Analyze the trade-offs between greedy and exhaustive parsing strategies

Q5. (2X10=20 Marks) (CO5)
a. Explain how SpaCy identifies named entities and assigns POS tags to words in a sentence.
b. Evaluate the performance of a baseline POS tagger using accuracy and F1-score.
c. Compare the effectiveness of term frequency (TF) and TF-IDF for text classification tasks.