| | |
|---|---|
| SASTRA DEEMED TO BE UNIVERSITY (U/S 3 OF THE UGC ACT, 1956) THINK MERIT \| THINK TRANSPARENCY \| THINK SASTRA | **School of Computing**<br>**Course Code: CSE405R01**<br>**Course Name: Natural Language Processing**<br>**Duration: 3 hrs** **Max Marks: 50** |

1. Perform the following task using NLTK for the given corpus: tokenize, remove stop words, stemming. Generate bigrams for the consecutive tokens and find whether the bigrams form an interesting collocation are not using t test and chi square text for a critical value 1.962. **File name "India.txt".**

2. Apply forward and backward procedure to find the probability of the week's sequence "ABAABAC". Always the first day of the week starts with a neutral mood.