|  | **PES University, Bengaluru**  (Established under Karnataka Act No. 16 of 2013) | | **UE20CS933** |
| --- | --- | --- | --- |
| **October 2024: END SEMESTER ASSESSMENT (ESA)**  **M TECH DATA SCIENCE AND MACHINE LEARNING\_ SEMESTER II**  **UE20CS933 - NATURAL LANGUAGE PROCESSING** | | | |
| Time: 3 Hrs | | Answer All Questions | Max Marks: 100 |

| **INSTRUCTIONS** | | | | |
| --- | --- | --- | --- | --- |
| * All questions are compulsory. * Section A should be handwritten in the answer script provided. * Section B and C are coding questions which have to be answered in the system. | | | | |
|  | | | | |
| **SECTION A – 20 MARKS** | | | | |
| 1 | a) | What do you mean by NLP? list 4 real-world applications of NLP. (Marks- 3 + 2) | 5 |
| b) | What are Large Language Models (LLMs)? List any four limitations/drawbacks of LLMs (Marks 1 + 4) | 5 |
| c) | Discuss RNN cell and its drawback. How LSTM overcomes RNN drawbacks? (Marks 3+2) | 5 |
| d) | Explain Named Entity Recognition with an example. | 5 |
| **SECTION B –40 MARKS** | | | | |
| **2** |  | Given the dataset reviews.csv, perform the following preprocessing steps: | **20** |
| a) | Load the dataset and display the first 5 rows. | 5 |
| b) | Remove any duplicate reviews. | 5 |
| c) | Clean the text by removing punctuation, converting to lowercase, and removing stopwords. | 10 |
| **3** | a) | Convert the cleaned text into word embeddings using TF-IDF. | 10 |
| b) | Display the shape of the resulting TF-IDF matrix. | 2 |
| c) | Get the vocabulary of TF-IDF. | 8 |
|  | | | | |
| **SECTION C –40 MARKS** | | | | |
| **4** |  | **Model building using Naive Bayes** | **20** |
| a) | Split the dataset into training and testing sets. | 5 |
| b) | Train a Naive Bayes classifier on the training set. | 10 |
| c) | Evaluate the model on the testing set and display the accuracy. | 5 |
| **5** |  | **Model Building using LSTM** | **20** |
| a) | Convert the cleaned text into sequences using Tokenizer. | 5 |
| b) | Build and compile an LSTM model for sentiment analysis. | 10 |
| c) | Train the model and evaluate its performance on the testing set. | 5 |
|  | | | | |