|  | **PES University, Bengaluru**  (Established under Karnataka Act No. 16 of 2013) | | **UE20CS933** |
| --- | --- | --- | --- |
| **March 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 is Generative AI? List any two concerns associated with it, and then suggest (any two) approaches to mitigate these concerns. (marks 1+ 2+2) | 5 |
| b) | What are Large Language Models (LLMs)? List any four limitations/drawbacks of LLMs (marks 1 +4) | 5 |
| c) | What is Attention in Transformer architecture? Using an example demonstrating how to compute Attention scores. (marks 2+5) | 7 |
| d) | What is Prompt Engineering. List any two prompting approaches. (marks 2 +1) | 3 |
|  | | | | |
| **SECTION B –40 MARKS** | | | | |
| 2 |  | Use the data.csv dataset as provided in the notebook as pandas DataFrame and process it as questioned below. |  |
| a) | Pre-Process the text feature as questioned below. (in the same sequence)   1. Remove the **accented** characters from text feature. (3 marks) 2. Remove **stopwords** from text feature. (3) 3. Remove **digits** from text feature. (3) 4. Remove **punctuations** from text feature. (3) 5. Eliminate **multiple spaces** from text feature. (3)   Note: Save this pre-processed text feature and use it as a feature for next questions. | 15 |
| b) | Find out the 5 most frequent words in the text corpus (from the preprocessed output of previous question 2.a) | 8 |
| c) | Vectorize the pre-processed text feature by building/training a Skip-Gram Word2Vec model. Use this Skip-Gram Word2Vec model to fetch the top 5 most similar word for the word 'food'. (marks 3+5) | 8 |
|  | d) | Vectorize the pre-processed text feature by building a CBOW Word2Vec model. Use the trained CBOW Word2Vec model to fetch the top 5 most similar word for the word 'food'. Is the output different than Skip-Gram’s output? (marks 3+5+1) | 9 |
|  | | | | |
| **SECTION C –40 MARKS** | | | | |
| 3 |  | The task specific pretrained transformers pipeline models is saved and provided. Use them to perform below Text processing tasks as questioned below. |  |
|  | a) | UsingSentence Classification - Sentiment Analysis model **classification\_pipeline\_model**, classify the sentence “Such a nice weather outside!” into positive/negative with a score. | 8 |
|  | b) | Using Named Entity Recognition model **ner\_pipeline\_model**, perform name-entity -recognition of sentence “'Hugging Face is a French company based in New-York.'” | 8 |
|  | c) | Using theQuestion Answering model **qa\_pipeline\_model**, provide the answer the **question** asked from the given **paragraph** (for question and paragraph refer notebook). | 8 |
|  | d) | Using Text Generation - Mask Filling model **tg\_pipeline\_model**,suggest the appropriate words for specified `**MISSING\_WORD\_Field**` in the given sentence. | 8 |
|  | e) | Using Summarization model **summarizer\_pipeline\_model**, provide summarization of the given **Long\_Tennis\_Article** as provided in notebook. | 8 |
|  | | | | |