To,

IITD-AIA Foundation of Smart Manufacturing

Subject: **Weekly Progress Report for Week 2**

Dear sir,

Following is the required progress report to the best of my knowledge considering relevant topics to be covered.

What's happening this week:

- Deep Reinforcement Learning
- Deep Learning Limitations and new Frontiers
- Python Libraries like Textblob
- Text cleaning and Preprocessing techniques
- Data extraction techniques from website
- Normalization and Stemming
- Word Embedding or vectorization
- Word2Vec

My Understanding of INTP23-ML-01: Chatbot for FSM

Scope:

This project involves developing an interactive conversational interface that can assist users by providing instant responses to their queries, guiding them through the website, and resolving common issues. The chatbot should be equipped with natural language processing capabilities, allowing it to understand and respond to user input accurately. It can offer student support, retrieve information from the website, provide student with the information he wants, guide him to the website of fsm.

Solution:

This kinda chatbot can be made using natural language processing (NLP) where it involves designing and implementing an intelligent conversational bot that can understand and respond to user queries in a natural and helpful manner.

Approach:

It involves collecting all the data from the scratch of the website and then designing the conversational flow and user interface. NLP and Ml models will be applied to make the bot work with different languages.

Weekly Progress:

June 12:

Covered topics like Deep reinforcement learning.

- Reinforcement learning is a general-purpose framework for decision making which comes with states and action which turns in reasoning.
- Deep reinforcement learning defines the useful state space, action space, and reward by learning the data and gets us the insights from the this process.
- DRL framework consists of an agent, an environment, and a learning process. The agent interacts with the environment by observing the current state, taking actions, and receiving rewards. It then uses this information to update its policy or value function using techniques such as deep Q-networks (DQN).

June 13:

Learnt about limitations of deep learning and it's new frontiers. And also python libraries used for NLP.

- Universal Approximation theorem: A single layer is enough to make a abritary precision or and function. that layer can be exponential big.Deep learning limitations :
- Deep Learning Limitations:
    1. The system is very fond of data.
    2. Easily fool by adversarial examples(Forcefully given inputs).
    3. poor at predicting uncertainty(he is not sure about his own probability).
- Python Libraries:
    1. TextBlob: It provides a simple API for common NLP tasks such as tokenization, part-of-speech tagging, noun phrase extraction, sentiment analysis, and more. It is built on the top of NLTK library.
    2. Gensim: It is an open-source Python library for topic modeling and document similarity analysis. Its applications include like topic modeling, document similarity analysis, text classification, and recommendation systems.

June 14:

Studied some of the concepts of text cleaning and its pre-processing. And also done its practical implementation on a dataset.

- Studied text cleaning and pre-processing methods like tokenization , Noise Entities Removal, Data Visualization for Text Data, Parts of Speech (POS) Tagging.
- Tokenization are of 3 types White-space Tokenization, Regular Expression Tokenization, Word and sentence tokenization and this used to transform the data into smaller tokens for further processing.
- Noise Entities Removal are used to Remove unwanted Punctuation marks and stopped words.
- For visualizing the data we use techniques like frequency graphs,frequency distribution and word clouds.
- Parts of speech (POS) tagging is for identifying various kinds of speech and used in sematic analysi. POS tags are the basis of the lemmatization process for converting a word to its base form (lemma).
- This all the process is done on a red wine dataset which can be viewed on my github.

June 15:

Studied techniques of extracting the data from a website.

- Studied how to extract data from the website.
- Learned and implement web scrapping on a web page product.
- Studied how we can extract data from API in an web page.

June 16:

Studied Text cleaning techniques like Normalization, Stemming and Lemmatization with its practical implementation on red wine dataset. also chunking too.

- Normalization is the process of converting a token into its base form. It makes the text more uniform and easier to work with. Some of the normalization techniques are lower-casing, removing punctuation, removing special character, handling nos and lemmatization or stemming.
- Stemming refers to the process of removing the suffixes and prefixes of the word to obtain the root word.
- Lemmatization is an organized & step by step procedure of obtaining the root form of the word, as it makes use of vocabulary (dictionary importance of words) and morphological analysis (word structure and grammar relations).
- Difference between both of them is lemmatization converts the word to its meaningful base form, whereas stemming simply chops off the ends of a word using heuristics that often leading to incorrect meanings and spelling errors.
- In Chunking, we try to extract meaningful phrases from unstructured data in the form of text by tokenizing. It works on top of Part of Speech(POS) tagging.

June 17:

Studied Word embedding/ Vectorization, Term frequency-inverse document frequency with practical implementation.

- As it is difficult to process the raw form of the text so we require numerical numbers as inputs to perform any sort of task. For that we use Word Embedding or Vectorization which converts texts into number vectors and there may be different numerical representations of the same text.
- Vector representation of a one-hot encoded vector represents in the form of 1, and 0 where 1 stands for the position where the word exists and 0 everywhere else.
- In the N-Gram method, a document term matrix is generated, and each cell represents the count.
- Term frequency-inverse document frequency (TF-IDF) gives a measure that takes the importance of a word into consideration depending on how frequently it occurs in a document and a corpus. Here Term frequency denotes the frequency(TF) of a word in a document and Inverse document frequency (IDF)is the logarithmic ratio of no. of total documents to no. of a document with a particular word.

June 18:

Studied word2vec model.

- Word2Vec model is used for Word representations in Vector Space and is a neural network model that attempts to explain the word embeddings based on a text corpus.
- Word2Vec model composed of two preprocessing modules or techniques i.e., Continuous Bag of Words (CBOW),Skip-Gram.
- The aim of the CBOW model is to predict a target word in its neighborhood, using all words. To predict the target word, this model uses the sum of the background vectors.This model works in both the cases single context word and multiple context words.
- The skip-gram model is the exact opposite of the CBOW model. In this case, the target word is given as the input, the hidden layer remains the same, and the output layer of the neural network is replicated multiple times to accommodate the chosen number of context words.
- Skip-Gram model works well with a small amount of the training datasets, and can better represent rare words or phrases. However, the CBOW model is observed to train faster than Skip-Gram, and can better represent more frequent words which mean gives slightly better accuracy for the frequent words.
- Done sentiment analysis on twitter’s positive and negative comments.

GANTT CHART FOR INTP23-ML-01

Task ID	Task Name	Start date	End date	Duration
1	Deep Learning	01-06-23	01-06-23	1
2	Python Libraris	02-06-23	02-06-23	1
3	Linear modelling	03-06-23	03-06-23	1
4	Deep Neural Network	04-06-23	04-06-23	1
5	Problems in network	05-06-23	05-06-23	1
6	Prevention Techniques	06-06-23	06-06-23	1
7	Resoruces (Project)	07-06-23	07-06-23	1
8	NLP	08-06-23	08-06-23	1
9	NLTK,Scipy libraries	09-06-23	09-06-23	1
10	Models for DL	10-06-23	10-06-23	1
11	Gradient Descent	11-06-23	11-06-23	1
12	Deep Reinforcement learning	12-06-23	12-06-23	1
13	Deep Learning Limitations	13-06-23	13-06-23	1
14	Python Libraries	14-06-23	14-06-23	1
15	Data extraction	15-06-23	15-06-23	1
16	Text cleaning	16-06-23	16-06-23	1
17	Text cleaning	17-06-23	17-06-23	1
18	Word2Vec model	18-06-23	18-06-23	1
19	Task no 19	19-06-23	19-06-23	1
20	Task no 20	20-06-23	20-06-23	1
21	Task no 21	21-06-23	21-06-23	1
22	Task no 22	22-06-23	22-06-23	1
23	Task no 23	23-06-23	23-06-23	1
24	Task no 24	24-06-23	24-06-23	1
25	Task no 25	25-06-23	25-06-23	1
26	Task no 26	26-06-23	26-06-23	1
27	Task no 27	27-06-23	27-06-23	1
28	Task no 28	28-06-23	28-06-23	1
29	Task no 29	29-06-23	29-06-23	1
30	Task no 30	30-06-23	30-06-23	1
31	Task no 31	01-07-23	01-07-23	1
32	Task no 32	02-07-23	02-07-23	1
33	Task no 33	03-07-23	03-07-23	1
34	Task no 34	04-07-23	04-07-23	1
35	Task no 35	05-07-23	05-07-23	1
36	Task no 36	06-07-23	06-07-23	1
37	Task no 37	07-07-23	07-07-23	1
38	Task no 38	08-07-23	08-07-23	1
39	Task no 39	09-07-23	09-07-23	1
40	Task no 40	10-07-23	10-07-23	1
41	Task no 41	11-07-23	11-07-23	1
42	Task no 42	12-07-23	12-07-23	1
43	Task no 43	13-07-23	13-07-23	1
44	Task no 44	14-07-23	14-07-23	1
45	Task no 45	15-07-23	15-07-23	1
46	Task no 46	16-07-23	16-07-23	1
47	Task no 47	17-07-23	17-07-23	1
48	Task no 48	18-07-23	18-07-23	1
49	Task no 49	19-07-23	19-07-23	1
50	Task no 50	20-07-23	20-07-23	1