PROJECT PHASE I

DATE	26-09-2023
NAAN MUTHALVAN TEAM MEMBERS ID	au513121104005 au513121104023 au513121104024 au5131211040708
PROJECT NAME	CREATE A CHATBOT USING PYTHON

TABLE OF CONTENTS:

Loading and Preprocessing the Dataset

INTRODUCTION:

This document outlines the critical process of loading and preprocessing a dataset, a foundational step in building a chatbot. Proper data preparation ensures that the chatbot model can learn and generate coherent responses.

DATA SOURCE:

The dataset is sourced from a CSV file containing conversations between users and a chatbot. It includes both user messages and corresponding chatbot responses.

STEPS FOR LOADING AND PREPROCESSING THE DATASET

Step 1: Loading the Dataset

We start by using the pandas library to load the dataset. Removing duplicate conversations is essential to maintain data consistency and integrity.

Code Snippet:

```
import pandas as pd

# Load the dataset

df = pd.read_csv("chatbot_dataset.csv")

 # Remove duplicate conversations

df = df.drop_duplicates()
```

Step 2: Text Preprocessing

I)Special Character and Digit Removal:

We remove special characters and digits from the text as they don't contribute to the conversation's meaning.

Code Snippet:

```
import re

 # Remove special characters and digits

df['user_message'] = df['user_message'].apply(lambda x: re.sub(r'[^a-zA-Z\s]', '', x))
```

II)Text Lowercasing:

To ensure uniformity, we convert all text to lowercase.

Code Snippet:

```
# Convert text to lowercase
df['user_message'] = df['user_message'].str.lower()
```

III)Sentence Tokenization:

The NLTK library is employed to break the text into sentences, phrases, or responses.

Code Snippet:

```
python
import nltk
nltk.download('punkt')
 # Tokenize sentences
df['user_message'] = df['user_message'].apply(nltk.sent_tokenize)
```

IV)Lemmatization:

Words within the sentences are lemmatized, reducing them to their base forms for dimensionality reduction and simplified text analysis.

Code Snippet:

```
from nltk.stem import WordNetLemmatizer
nltk.download('wordnet')



# Lemmatize words
lemmatizer = WordNetLemmatizer()
```

```
df['user_message'] = df['user_message'].apply(lambda x: [lemmatizer.lemmatize(word) for word in x])
```

Step 3: Data Split

The dataset is organized into input (user messages) and output (bot responses) pairs. This structure prepares the data for training a machine learning model.

```
# Split the dataset into input and output pairs

input_data = df['user_message']

output_data = df['bot_response']
```

PREPROCESSED DATA

The dataset is now ready for use in training a chatbot model, with clean and structured data facilitating meaningful interactions.

CONCLUSION:

In conclusion, this document has detailed the essential steps in loading and preprocessing the dataset, laying a strong foundation for chatbot development. These steps encompassed dataset loading, the removal of duplicate conversations, comprehensive text preprocessing including special character and digit removal, text lowercasing, sentence tokenization, and lemmatization. The data was thoughtfully organized into input and output pairs, priming it for machine learning model training. The preprocessed dataset now stands as a clean, structured, and coherent source of information, poised to empower the subsequent phases of the chatbot development process, including model selection, training, and performance evaluation, to ultimately achieve a chatbot capable of engaging in meaningful and contextually relevant conversations with users.