**Development Part - 1**

**Group Members:**

**Prabha.K,**

**GowthamRaj. J,**

**Dileep. B,**

**Mathesh. S,**

**Gokul. M.**

Email ID: industryevoluator@gmail.com

Faculty Mentor:ezhimathia@jkkn.ac.in

**AI-ChatBot**

Phase - 3

**AI Chatbot Project - Data Processing, Cleaning, and EDA**

**Project Overview:**

Project Name: AI Chatbot using Python with NLP (Pretrained Model)

Dataset: dialogs.txt

**Data Loading**

Load your dataset text file into Python using a suitable library (e.g., pandas, numpy, openai, or custom code). Ensure that the dataset is accessible for data processing.

```python

import pandas as pd

# Load the dataset

dataset = pd.read\_csv("your\_dataset.txt")

```

**Data Cleaning**

Describe the steps you took to clean the dataset. Include explanations for data cleaning procedures, such as:

* Removing duplicates
* Handling missing values
* Handling outliers
* Standardizing text data

```python

# Example data cleaning (modify according to your dataset)

dataset = dataset.drop\_duplicates()

dataset = dataset.dropna()

dataset['text\_column'] = dataset['text\_column'].str.lower()

```

**Data Preprocessing**

Explain the preprocessing steps you applied to make the dataset suitable for NLP tasks, such as:

* Tokenization
* Stopword removal
* Lemmatization or stemming
* Feature extraction (e.g., TF-IDF)

```python

# Example data preprocessing (modify according to your dataset)

from nltk.corpus import stopwords

from nltk.tokenize import word\_tokenize

nltk.download('stopwords')

nltk.download('punkt')

stop\_words = set(stopwords.words('english'))

def preprocess\_text(text):

words = word\_tokenize(text)

words = [word for word in words if word.isalnum() and word.lower() not in stop\_words]

return ' '.join(words)

dataset['text\_column'] = dataset['text\_column'].apply(preprocess\_text)

```

**Exploratory Data Analysis (EDA)**

Perform EDA to understand the dataset better. Include details such as:

* The number of samples in the dataset
* Data statistics (mean, median, min, max, etc.)
* Visualizations (e.g., histograms, word clouds, bar charts) to understand the distribution of data

```python

# Example EDA (modify according to your dataset)

import matplotlib.pyplot as plt

num\_samples = len(dataset)

average\_text\_length = dataset['text\_column'].apply(lambda x: len(x.split())).mean()

# Visualization

plt.figure(figsize=(8, 5))

plt.bar(['Num Samples', 'Average Text Length'], [num\_samples, average\_text\_length], color=['blue', 'green'])

plt.title('Data Exploration')

plt.ylabel('Count / Length')

plt.show()

```

**Conclusion**

Summarize the key findings from the data processing, cleaning, and EDA steps. Explain how these steps have prepared the dataset for your chatbot project.