ARTIFICIAL INTELLIGENCE

CREATE A CHATBOT USING PYTHON

Name: Kalavathy K

Reg. No: 513521104019

Department: CSE

Year: III

NM ID: au513521104019

E-Mail: kaviyakalavathy@gmail.com

INTRODUCTION:

"Welcome to the world of Chatbot development with Python! In this project, we embark on a journey to create an intelligent and interactive chatbot powered by Python's Natural Language Processing (NLP) capabilities. Our goal is to design a chatbot that can engage in meaningful conversations, answer questions, and provide assistance across a range of domains. Through data acquisition, model training, and iterative improvement, we aim to craft a chatbot that not only understands user queries but also offers contextually relevant responses. Join us as we explore the fascinating realm of conversational AI and create a chatbot that's ready to assist and engage with users."

In the continuation of the previous documentation AI_Phase2, we will be looking forward on the development part of the project. In this phase, we will start building the model by loading the dataset and processing the data

DATASET LINK: https://www.kaggle.com/datasets/grafstor/simple-dialogs-for-chatbot

The following steps are followed in this development part.

1. EMPATHIZE

Understanding user needs is a critical aspect of designing and developing a chatbot that is effective and user-friendly. Developers should consider the following key user needs:

Clear Communication: Users need a chatbot that communicates clearly and in a language they understand. The chatbot should avoid jargon and complex terminology, providing responses that are easy to follow.

Efficiency: Users value efficiency in their interactions with a chatbot. They expect quick and accurate responses to their queries or requests. Developers should prioritize minimizing response times and streamlining interactions.

Relevance: Users want the chatbot to provide information or assistance that is relevant to their needs or context. Developers should ensure that the chatbot's responses and actions are tailored to individual users when possible.

Problem Solving: Users often turn to chatbots to solve problems or complete tasks. Developers should design the chatbot to be capable of addressing common user issues and helping users accomplish their goals.

Personalization: Personalization is increasingly important to users. Developers can enhance user experience by allowing chatbots to remember user preferences, past interactions, and other relevant information to deliver a more personalized experience.

User-Friendly Interface: The chatbot's user interface should be intuitive and user-friendly. Users need to be able to navigate the chatbot easily and understand how to use it without confusion.

Privacy and Security: Users are concerned about their data and privacy. Developers must implement robust security measures to protect user information and communicate transparently about data handling practices.

Availability: Users expect chatbots to be available when needed. Ensure the chatbot is accessible 24/7 if possible, or communicate its operating hours clearly to users.

Error Handling: Users appreciate a chatbot that can gracefully handle errors and unexpected inputs.

Feedback and Help: Users should have the ability to provide feedback on the chatbot's performance or request assistance when needed. Developers can

implement mechanisms for users to report issues or seek human assistance when the chatbot cannot provide a solution. Accessibility, Consistency, Multilingual Support, Integration, Simplicity, Empathy, Quality Content etc are some of the user needs in development of chatbot. By comprehensively understanding these user needs, developers can create chatbots that not only meet user expectations but also provide valuable and satisfying interactions. Regular user feedback and testing are essential to continuously refine the chatbot based on evolving user needs.

2.DEFINE:

"The problem to be solved is that users need an efficient and user-friendly way to schedule appointments with healthcare providers. Existing scheduling systems are often cumbersome, require phone calls, and are prone to errors, leading to patient frustration and missed appointments. The chatbot's objective is to streamline the appointment scheduling process, provide real-time availability information, and send timely reminders to both patients and healthcare providers, ultimately improving the patient experience and reducing appointment no-shows."This problem statement is clear and specific. It identifies the users' pain points (inefficient scheduling, missed appointments), the limitations of the current system, and the chatbot's role in addressing these issues (streamlining scheduling, real-time information, reminders).

3.Importing Libraries:

In this step we will import all the required libraries and packages from the pre-installed modules.

```
In [4]: import tensorflow as tf
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from tensorflow.keras.layers import TextVectorization
import re,string
from tensorflow.keras.layers import LSTM,Dense,Embedding,Dropout,LayerNormalization
from flask import Flask
```

4.Loading datasets:

In this step we will be loading the chatbots datasets acquired from Kaggle.

```
In [10]: df=pd.read_csv('dialogs.txt',sep='\t',names=['question','answer'])
           print(f'Dataframe size: {len(df)}')
           df.head()
           Dataframe size: 3725
Out[10]:
                                         question
                                                                               answer
             0
                             hi, how are you doing?
                                                             i'm fine, how about yourself?
              1
                        i'm fine. how about yourself?
                                                        i'm pretty good, thanks for asking.
             2
                   i'm pretty good. thanks for asking.
                                                      no problem, so how have you been?
              3 no problem, so how have you been?
                                                         i've been great, what about you?
                    i've been great, what about you? i've been good, i'm in school right now.
```

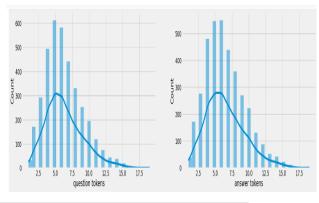
5. Ideate:

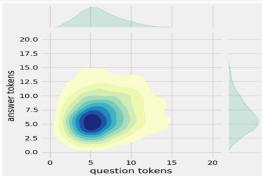
Data Preprocessing: (Data Visualization)

We will be pre-processing the data for the further development of model.

```
In [11]: df['question tokens']=df['question'].apply(lambda x:len(x.split()))
    df['answer tokens']=df['answer'].apply(lambda x:len(x.split()))
    plt.style.use('fivethirtyeight')
    fig,ax=plt.subplots(nrows=1,ncols=2,figsize=(20,5))
    sns.set_palette('Set2')
    sns.histplot(x=df['question tokens'],data=df,kde=True,ax=ax[0])
    sns.histplot(x=df['answer tokens'],data=df,kde=True,ax=ax[1])
    sns.jointplot(x='question tokens',y='answer tokens',data=df,kind='kde',fill=True,cmap='YlGnBu')
    plt.show()
```

O/P:





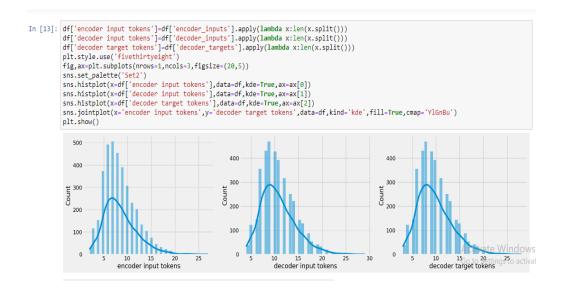
6.Text Cleaning:

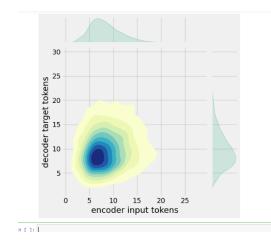
Text cleaning, also known as text preprocessing, is the process of preparing and standardizing textual data to make it suitable for analysis or natural language processing (NLP) tasks.

```
def clean_text(text):
    text=re.sub('-','
    text=re.sub('[.]'
                                                                        ,text.lower())
                                                                                 ',text)
',text)
',text)
',text)
',text)
           text=re.sub(
                                                                                 ',text)
                                                                                 ',text)
',text)
',text)
            text=re.sub(
text=re.sub(
text=re.sub(
text=re.sub(
text=re.sub(
text=re.sub(
                                                                                    ,text)
                                                                                 ',text)
                                                                                    ,text)
            text=re.sub(
text=re.sub(
text=re.sub(
                                                                                 ',text)
                                                                                  ',text)
',text)
             text=re.sub(
text=re.sub(
                                                                                     ,text)
                                                                                    ,text)
',text)
',text)
             text=re.sub(
text=re.sub(
             text=re.sub(
             text=re.sub('
return text
```

```
df.drop(columns=['answer tokens', 'question tokens'],axis=1,inplace=True)
df['encoder_inputs']=df['question'].apply(clean_text)
df['decoder_targets']=df['answer'].apply(clean_text)+' <end>'
df['decoder_inputs']='<start> '+df['answer'].apply(clean_text)+' <end>'
df.head(10)
```

decoder_inputs	decoder_targets	encoder_inputs	answer	question	
<start> i ' m fine . how about yourself ?</start>	i ' m fine . how about yourself ? <end></end>	hi , how are you doing ?	i'm fine. how about yourself?	hi, how are you doing?	0
<start> i ' m pretty good . thanks fo asking</start>	i ' m pretty good . thanks for asking .	i'm fine . how about yourself ?	i'm pretty good. thanks for asking.	i'm fine. how about yourself?	1
<start> no problem . so how have you been ?</start>	no problem . so how have you been ? <end></end>	i ' m pretty good . thanks for asking .	no problem. so how have you been?	i'm pretty good. thanks for asking.	2
<start> i ' ve been great . what abou you ?</start>	i've been great . what about you ? <end></end>	no problem . so how have you been ?	i've been great. what about you?	no problem. so how have you been?	3
<start> i ' ve been good . i ' m ir school ri</start>	i ' ve been good . i ' m in school right now	i ' ve been great . what about you ?	i've been good. i'm in school right now.	i've been great. what about you?	4
<start> what school do you go to ?</start>	what school do you go to ? <end></end>	i've been good . i'm in school right now .	what school do you go to?	i've been good. i'm in school right now.	5
<start> i go to pcc . <end></end></start>	i go to pcc . <end></end>	what school do you go to ?	i go to pcc.	what school do you go to?	6
<start> do you like it there ? <end></end></start>	do you like it there ? <end></end>	i go to pcc .	do you like it there?	i go to pcc.	7
<start> it 's okay . it 's a really big cam</start>	it's okay . it's a really big campus . $\mathrel{<_{\dots}}$	do you like it there ?	it's okay. it's a really big campus.	do you like it there?	8
<start> good luck with school . <end></end></start>	good luck with school . <end></end>	it's okay . it's a really big campus .	good luck with school.	it's okay. it's a really big campus.	9





After text preprocessing in the context of chatbot development, the text data is in a cleaner and more structured format, ready for further analysis, understanding, and use in natural language processing (NLP) tasks.

```
print(f"After preprocessing: {' '.join(df[df['encoder input tokens'].max()==df['encoder input tokens']]['encoder_inputs'].values.
print(f"Max encoder input length: {df['encoder input tokens'].max()}")
print(f"Max decoder input length: {df['decoder input tokens'].max()}")
print(f"Max decoder target length: {df['decoder target tokens'].max()}")
df.drop(columns=['question','answer','encoder input tokens','decoder input tokens','decoder target tokens'],axis=1,inplace=True)
     "max_sequence_length":30,
"learning_rate":0.008,
     "batch_size":149,
"lstm_cells":256,
      "embedding_dim":256,
      "buffer_size":10000
learning_rate=params['learning_rate']
batch_size=params['batch_size']
embedding_dim=params['embedding_dim']
lstm_cells=params['lstm_cells']
vocab_size=params['vocab_size']
buffer_size=params['buffer_size']
 max_sequence_length=params['max_sequence_length']
df.head(10)
 After preprocessing: for example , if your birth date is january 1 \, 2 \, , \, 1 \, 9 \, 8 \, 7 \, , write \, 0 \, 1 \, 1 \, 2 \, / \, 8 \, 7 \, .
 Max encoder input length: 27
 Max decoder input length: 29
Max decoder target length: 28
                             encoder_inputs
                                                                       decoder_targets
  0 \hspace{1cm} \text{hi , how are you doing ?} \hspace{1cm} \text{i'm fine . how about yourself? <end>} \hspace{1cm} \text{<start> i'm fine . how about yourself? <end>} \\
              i'm fine . how about yourself? i'm pretty good . thanks for asking . <end> <start> i'm pretty good . thanks for asking .
  2 i'm pretty good . thanks for asking . no problem . so how have you been ? <end> <start> no problem . so how have you been ? ...
        no problem . so how have you been ? i've been great . what about you ? <end> <start> i've been great . what about you ?
  4 i've been great . what about you ? i've been good . i'm in school right now ... <start> i've been good . i'm in school ri...
  5 i've been good . i'm in school right now . what school do you go to ? <end> <start> what school do you go to ? <end>
  i go to pcc .
                                                          do you like it there ? <end>
                                                                                                   <start> do vou like it there ? <end>
  8 do you like it there? it's okay . it's a really big campus . <... <start> it's okay . it's a really big cam...
        it's okay . it's a really big campus .
                                                         good luck with school . <end>
                                                                                               <start> good luck with school . <end>
```

7. Tokenization:

Tokenization is the process of splitting a text or a sequence of characters into individual units, known as tokens.

```
In [16]: def sequences2ids(sequence):
                        return vectorize_layer(sequence)
                  def ids2sequences(ids):
                        decode=
                        if type(ids)==int:
                              ids=[ids]
                        for id in ids:
    decode+=vectorize_layer.get_vocabulary()[id]+' '
                        return decode
                 x=sequences2ids(df['encoder_inputs'])
yd=sequences2ids(df['decoder_inputs'])
y=sequences2ids(df['decoder_targets'])
                 print(f'Question sentence: hi , how are you ?')
print(f'Question to tokens: {sequences2ids("hi , how are you ?")[:10]}')
print(f'Encoder input shape: {x.shape}')
print(f'Decoder input shape: {yd.shape}')
print(f'Decoder target shape: {y.shape}')
                  Question sentence: hi , how are you ?
                 Question to tokens: [1971 9
Encoder input shape: (3725, 30)
Decoder input shape: (3725, 30)
                                                                       45
                 Decoder target shape: (3725, 30)
       In [17]: print(f'Encoder input: {x[0][:12]} ...')
                   print(f'Decoder input: {yd[0][:12]} ...')
                   print(f'Decoder target: {y[0][:12]} ...')
                   Encoder input: [1971 9 45 24 8 194 7 0 0 0 0 Decoder input: [ 4 6 5 38 646 3 45 41 563 7 2 0] ... Decoder target: [ 6 5 38 646 3 45 41 563 7 2 0 0] ...
                                                                                                                       0] ...
 In [18]: data=tf.data.Dataset.from_tensor_slices((x,yd,y))
                data=data.shuffle(buffer_size)
                train_data=data.take(int(.9*len(data)))
                train_data=train_data.cache()
                train_data=train_data.shuffle(buffer_size)
                train_data=train_data.batch(batch_size)
train_data=train_data.prefetch(tf.data.AUTOTUNE)
                train_data_iterator=train_data.as_numpy_iterator()
                val_data=data.skip(int(.9*len(data))).take(int(.1*len(data)))
                val_data=val_data.batch(batch_size)
val_data=val_data.prefetch(tf.data.AUTOTUNE)
                _=train_data_iterator.next()
print(f'Number of train batches: {len(train_data)}')
print(f'Number of training data: {len(train_data)*batch_size}')
print(f'Number of validation batches: {len(val_data)}')
                print(f'Number of validation data: {len(val_data)*batch_size}')
                print(f'Encoder Input shape (with batches): {_[0].shape}')
print(f'Decoder Input shape (with batches): {_[1].shape}')
print(f'Target Output shape (with batches): {_[2].shape}')
O\P:
                         Number of train batches: 23
                        Number of training data: 3427
Number of validation batches: 3
                         Number of validation data: 447
                        Encoder Input shape (with batches): (149, 30)
Decoder Input shape (with batches): (149, 30)
                         Target Output shape (with batches): (149, 30)
```

8. Prototype:

Build encoder,

To build an encoder, you'll typically be working in the context of deep learning, and this encoder is often associated with autoencoders, recurrent neural networks (RNNs), or other neural network architectures.

```
In [19]: class Encoder(tf.keras.models.Model):
    def __init__(self,units,embedding_dim,vocab_size, "args, **kwargs) -> None:
        super().__init__("args, "*kwargs)
        self.units-units
        self.vocab_size-vocab_size
        self.embedding_dim=embedding_dim
        self.embedding_tembedding(
            vocab_size,
            embedding_tembedding(
            vocab_size,
            embedding_tembedding(
            vocab_size,
            embedding_tembedding',
            name='encoder_embedding',
            names'encoder_embedding',
            mask_zero-Irue,
            embedding_sinitializer=tf.keras.initializers.GlorotNormal()
        )
        self.normalize-LayerNormalization()
        self.lstm=LSTM(
            units,
            dropout=-4,
            return_sequences-Irue,
            name='encoder_lstm',
            kernel_initializer=tf.keras.initializers.GlorotNormal()
        )
}
```

```
def call(self,encoder_inputs):
    self.inputs-encoder_inputs
    x-self.embedding(encoder_inputs)
    x=self.normalize(x)
    x=bropout(.4)(x)
    encoder_outputs,encoder_state_h,encoder_state_c=self.lstm(x)
    self.outputs=lencoder_state_h,encoder_state_c]
    return encoder_state_h,encoder_state_c
encoder=Encoder(lstm_cells,embedding_dim,vocab_size,name='encoder')
encoder.call(_[0])
```

Build Decoder,

Build Training Model,

```
def train_step(self,batch):
    encoder_inputs,decoder_inputs,y-batch
    with tf.GradientTape() as tape:
        encoder_states=self.encoder(encoder_inputs,training=True)
        y_pred=self.decoder(decoder_inputs,encoder_states,training=True)
        loss=self.loss_fn(y,y_pred)
        acc=self.accuracy_fn(y,y_pred)
        variables=self.encoder.trainable_variables+self.decoder.trainable_variables
        gnads-tape.gradient(loss,variables)
        self.optimizer.apply_gradients(zip(gnads,variables))
        metrics={loss_incy_accuracy':acc}
        return metrics

def test_step(self,batch):
        encoder_inputs,decoder_inputs,y-batch
        encoder_inputs,decoder_inputs,encoder_states,training=True)
        y_pred=self.decoder(decoder_inputs,encoder_states,training=True)
        loss=self.loss_fn(y,y_pred)
        acc=self.accuracy_fn(y,y_pred)
        acc=self.accuracy_fn(y,y_pred)
        acc=self.accuracy_fn(y,y_pred)
        acc=self.accuracy_fn(y,y_pred)
        acc=self.accuracy_fn(y,y_pred)
        acc=self.accuracy_fn(y,y_pred)
        acc=self.accuracy_fn(y,y_pred)
        acciself.comenter(encoder,decoder,name='chatbot_trainer')
model.compile(
        loss=f.keras.losses.SparseCategoricalCrossentropy(),
        optimizer-tf.keras.optimizers.Adam(learning_rate=learning_rate),
        weighted_metrics=['loss', 'accuracy']
        So to Settings to a
model(:21)
```

```
Out[21]: <tf.Tensor: shape=(149, 30, 2443), dtype=float32, numpy=
array([[[9.39789286e-04, 5.81029046e-04, 3.49186215e-04, ...,
9.65607323e-05, 3.48445144e-03, 1.91266334e-03],
                           [1.59571457e-04, 7.51599000e-05, 4.00916470e-04, ..., 9.18773512e-05, 2.48697062e-04, 2.42815516e-03],
                            [1.79218347e-04, 1.28507416e-03, 6.64227584e-04, ...,
                             8.51919322e-05, 6.65480271e-04, 2.14136118e-04],
                           [1.00804398e-04, 2.44887429e-03, 2.69778934e-03, ..., 8.55395920e-05, 6.72195712e-03, 2.08562240e-04],
                            [1.00804398e-04, 2.44887429e-03, 2.69778934e-03, ...
                           8.55395920e-05, 6.72195712e-03, 2.08562240e-04], [1.00804398e-04, 2.44887429e-03, 2.69778934e-03, ...,
                             8.55395920e-05, 6.72195712e-03, 2.08562240e-04]],
                          [[1.48899248e-03, 1.11791574e-04, 1.69336592e-04, ...,
                           2.72725611e-05, 1.22156390e-03, 1.01141294e-03],
[7.83129595e-04, 2.81261309e-04, 5.02188013e-05, ...,
                             1.78441359e-03, 9.25112690e-05, 1.86895180e-04],
                           [1.20824232e-04, 5.75737329e-04, 4.34654357e-05, ..., 1.15328957e-03, 6.45271575e-05, 5.51116944e-04],
                            [3.07689370e-05, 6.83894672e-04, 1.33376860e-03, ....
                             1.96581837e-04, 1.09349994e-03, 6.80164376e-05],
                            [3.07689370e-05, 6.83894672e-04, 1.33376860e-03, ..., 1.96581837e-04, 1.09349994e-03, 6.80164376e-05],
                            [3.07689370e-05, 6.83894672e-04, 1.33376860e-03, ...,
                             1.96581837e-04, 1.09349994e-03, 6.80164376e-05]],
                            8.04460724e-05, 2.84239650e-03, 6.34894473e-03], [1.06537039e-03, 1.79962313e-04, 7.52430497e-05, ..., 3.37697216e-04, 8.45795556-04, 2.65551789e-04], [1.22962883e-04, 6.93331094e-05, 6.00464518e-05, ..., 3.87452892e-05, 9.59527737e-04, 1.02134619e-03],
                            [1.17018724e-04, 4.17125382e-04, 2.49348185e-03, ..., 8.36507315e-05, 8.44088104e-03, 9.44107742e-05], [1.17018724e-04, 4.17125382e-04, 2.49348185e-03, ..., 8.36507315e-05, 8.4368104e-03, 9.44107742e-05], [1.17018724e-04, 4.17125382e-04, 2.49348185e-03, ..., 8.36507315e-05, 8.44088104e-03, 9.44107742e-05]],
                           [[8.92622746e-04, 8.24281364e-04, 6.51491515e-04, ...,
                             2.69386510e-04, 3.47320316e-03, 2.96128052e-03],
                            [2.58860382e-04, 9.77864023e-04, 5.19359484e-04, ..., 4.62538999e-04, 1.51072303e-03, 1.34733960e-03],
                            [9.86914034e-04, 1.78585819e-04, 1.52172684e-03, ..., 4.31042863e-04, 2.96844955e-04, 6.79351564e-04],
                            [1.02885824e-04, 1.63706238e-04, 2.08607782e-03, ...,
                            4.12734407e-05, 2.38581165e-03, 1.92870823e-04], [1.02885824e-04, 1.63706238e-04, 2.08607782e-03, ...,
                            1.12734407e-05, 2.38581165e-03, 1.92870837e-04],

[1.02885824e-04, 1.63706238e-04, 2.08607782e-03, ...,

4.12734407e-05, 2.38581165e-03, 1.92870823e-04]],
                            [2.33066210e-04, 1.59448746e-03, 1.13014027e-03, ..., 2.15326436e-05, 2.81096133e-03, 1.51375134e-04], [2.33066210e-04, 1.59448746e-03, 1.13014027e-03, ...,
                              2.15326436e-05, 2.81096133e-03, 1.51375134e-04],
                             [2.33066210e-04, 1.59448746e-03, 1.13014027e-03, ...,
                              2.15326436e-05, 2.81096133e-03, 1.51375134e-04]],
                           [[1.10331795e-03, 5.95025660e-04, 3.74417345e-04, ...,
                              6.13579541e-05, 3.35158408e-03, 7.89283891e-04],
                             [1.57069473e-04, 4.53210610e-04, 1.73766151e-04, ...,
                             1.18845041e-04, 1.98629030e-04, 1.43995113e-03],
                            [7.53673958e-04, 3.93289747e-03, 1.64684563e-04, ...,
                              1.18359640e-04, 1.91663008e-03, 2.92276818e-04],
                             [1.15308554e-04, 5.29505778e-04, 1.03027304e-03, ...,
                              4.52578888e-06, 7.39003485e-03, 6.57932178e-05],
                             [1.15308561e-04, 5.29505836e-04, 1.03027222e-03, ...,
                              4.52579161e-06, 7.39003392e-03, 6.57932251e-05],
                            [1.15308561e-04, 5.29505836e-04, 1.03027222e-03,
```

4.52579161e-06, 7.39003392e-03, 6.57932251e-05]]], dtype=float32)>