CREATE A CHATBOT IN PYTHON

TEAM MEMBERS

1.SUYAMBULAKSHMI V
2.SHARINITHA R
3.AYESHASITHIKA S
4.PIRITHIBA M

Phase 3 – Document submission

OBJECTIVE:

- The Objective of this project is to create a high-quality support to users, ensuring a positive user experience and customer satisfaction chatbot in Python that provides exceptional customer service, answering user queries on a website or application.

PREPROCESSING:

- Here preprocessing includes,
    1. Data cleaning
    2. Removing stop-words
    3. Stemming and lemmatization

DATA CLEANING:

- By using pandas library “Dataframe” ,We can organize the given dataset.

PROGRAM:

These are the needed libraries needed to do preprocessing process which is

1. Checking Null data.
2. Checking Invalid data.
3. Removing stop – words.
4. Stemming.
5. Lemmatizing.

```
import re #regular expression to extract data
import nltk
import pandas as pd
import numpy as np

import spacy #spacy have some simple algorithm to do tokenization of
the given dataset
from spacy.lang.en.stop_words import STOP_WORDS #english stopwords are
extracted
len(STOP_WORDS) #To know about how many stopwords are available
nlp = spacy.load("en_core_web_sm") #using english language as the
preferred language

df = pd.read_csv("dialogs.txt", sep='\t', header=None, names=
["column1", "column2"])

df
```

OUTPUT:

this is the output of the above program,

	column1	column2
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?
4	i've been great. what about you?	i've been good. i'm in school right now.
...	...	...
3720	that's a good question. maybe it's not old age.	are you right-handed?
3721	are you right-handed?	yes. all my life.
3722	yes. all my life.	you're wearing out your right hand. stop using...
3723	you're wearing out your right hand. stop using...	but i do all my writing with my right hand.
3724	but i do all my writing with my right hand.	start typing instead. that way your left hand ...

3725 rows × 2 columns

By using this code we can find out there is no null data.

CHECKING INVALID DATA AND NULL DATA:

```
invalid_data = df[df.isnull().any(axis=1)]
null_data = df[df.isnull().all(axis=1)]
```

REMOVING INVALID DATA AND NULL DATA:

If any null data and invalid data is there, then by using this code we can remove that.

```
df = df.drop(invalid_data.index)
df = df.drop(null_data.index)
```

and the removed invalid and null datas are omitted and the cleaned data is stored in “cleaned_dailogs.txt” by changing the current directry.

```
df.to_csv("cleaned_dailogs.txt", sep="\t", index = False)
```

REMOVING STOP-WORDS:

```
from string import punctuation
df = pd.read_csv("cleaned_dailogs.txt", sep='\t', header=None,
names=["column1", "column2"])

# Get the text from the DataFrame by header name
text_by_header_name = df["column1"].to_string() +
df["column2"].to_string()


stopwords = set(STOP_WORDS)

def removing_stopwords(text):
  doc = text.split()

  no_stop_words =[word for word in doc if word not in STOP_WORDS and
word not in punctuation]
  return " ".join(no_stop_words)

df["column1"] = df["column1"].apply(removing_stopwords)
df["column2"] = df["column2"].apply(removing_stopwords)


display(df["column1"])
display(df["column2"])
```

By using this **removing_stopwords** function, we can able to remove stop words.

Finally we have succeccfully removed all the stop words from the dataset,

OUTPUT:

this is the output of the above program,

```
0                                          column1
1                                column1_lemmatized
2                                        hi, doing?
3                               i'm fine. yourself?
4                    i'm pretty good. thanks asking.
                              ...
3720    that's good question. maybe it's old age.
3721                                    right-handed?
3722                                       yes. life.
3723          you're wearing right hand. stop much.
3724                               writing right hand.
Name: column1, Length: 3725, dtype: object
0                                                 column2
1                                       column2_lemmatized
2                                    i'm fine. yourself?
3                         i'm pretty good. thanks asking.
4                                          problem. been?
                                  ...
3720                                     right-handed?
3721                                        yes. life.
3722               you're wearing right hand. stop much.
3723                                    writing right hand.
3724    start typing instead. way left hand half work.
Name: column2, Length: 3725, dtype: object
```

We have just executed to see the results to verify the given process is done or not.

STEMMING AND LEMMATIZATION:

For stemming, this is the code.

```
stemmer = nltk.PorterStemmer()

df["column1_stemmed"]=df["column1"].apply(lambda x: stemmer.stem(x))
df["column2_stemmed"]=df["column2"].apply(lambda x: stemmer.stem(x))
```

Forlemmatization, this is the code.

```
lemmatizer = nltk.WordNetLemmatizer()


df["column1_lemmatized"] = df["column1"].apply(lambda x:
lemmatizer.lemmatize(x))
```

```
df["column2_lemmatized"] = df["column2"].apply(lambda x:
lemmatizer.lemmatize(x))
df
```

OUTPUT:

this is the output of the above program,

	column1	column2	column1_stemmed	column2_stemmed	column1_lemmatized	column2_lemmatized
0	hi, doing?	i'm fine. yourself?	hi, doing?	i'm fine. yourself?	hi, doing?	i'm fine. yourself?
1	i'm fine. yourself?	i'm pretty good. thanks asking.	i'm fine. yourself?	i'm pretty good. thanks asking.	i'm fine. yourself?	i'm pretty good. thanks asking.
2	i'm pretty good. thanks asking.	problem. been?	i'm pretty good. thanks asking.	problem. been?	i'm pretty good. thanks asking.	problem. been?
3	problem. been?	i've great. you?	problem. been?	i've great. you?	problem. been?	i've great. you?
4	i've great. you?	i've good. i'm school right now.	i've great. you?	i've good. i'm school right now.	i've great. you?	i've good. i'm school right now.
...	...	...	...	...	...	...
3720	that's good question. maybe it's old age.	right-handed?	that's good question. maybe it's old age.	right-handed?	that's good question. maybe it's old age.	right-handed?
3721	right-handed?	yes. life.	right-handed?	yes. life.	right-handed?	yes. life.
3722	yes. life.	you're wearing right hand. stop much.	yes. life.	you're wearing right hand. stop much.	yes. life.	you're wearing right hand. stop much.
3723	you're wearing right hand. stop much.	writing right hand.	you're wearing right hand. stop much.	writing right hand.	you're wearing right hand. stop much.	writing right hand.
3724	writing right hand.	start typing instead. way left hand half work.	writing right hand.	start typing instead. way left hand half work.	writing right hand.	start typing instead. way left hand half work.

3725 rows × 6 columns

CONCLUSION:

These are the basic preprocessing steps that we have followed for this project for creating chatbot using python (NLP).