Import re

Import string

Import nltk

From nltk.corpus import stopwords

From nltk.stem import WordNetLemmatizer

From nltk.tokenize import word\_tokenize

# Download NLTK data (if not already downloaded)

Nltk.download(‘stopwords’)

# Preprocessing function

Def clean\_text(text):

Text = text.lower()

Text = re.sub(f”[{re.escape(string.punctuation)}]”, ‘’, text)

Tokens = word\_tokenize(text)

Tokens = [word for word in tokens if word not in stopwords.words(‘english’)]

Lemmatizer = WordNetLemmatizer()

Tokens = [lemmatizer.lemmatize(word) for word in tokens]

Return ‘ ‘.join(tokens)

# Load and preprocess your dataset from a text file

Def load\_and\_preprocess\_data(file\_path):

With open(file\_path, ‘r’, encoding=’utf-8’) as file:

Lines = file.readlines()

Dialogue\_data = [line.strip().split(‘\t’) for line in lines]

Preprocessed\_data = [(clean\_text(pair[0]), clean\_text(pair[1])) for pair in dialogue\_data]

Return preprocessed\_data

# Specify your file path

Data\_file = r"C:\Users\Arunachala\Downloads\dialogs.txt"

# Load and preprocess the data

Dialogue\_data = load\_and\_preprocess\_data(data\_file)

# Print the first preprocessed dialog pair

Print(“First Preprocessed Dialog Pair:”, dialogue\_data[0])