**Phase3**

**DEVELOPMENT-PART I**

**Project Name : CREATE A CHATBOT IN PYTHON**

**Name: SRISHIVANTH R F**

**Introduction:**

Chatbot often powered by sophisticated language models like GPT, have gained popularity in natural language processing tasks. These models can generate human-like responses based on input text and are trained on vast amounts of diverse textual data.

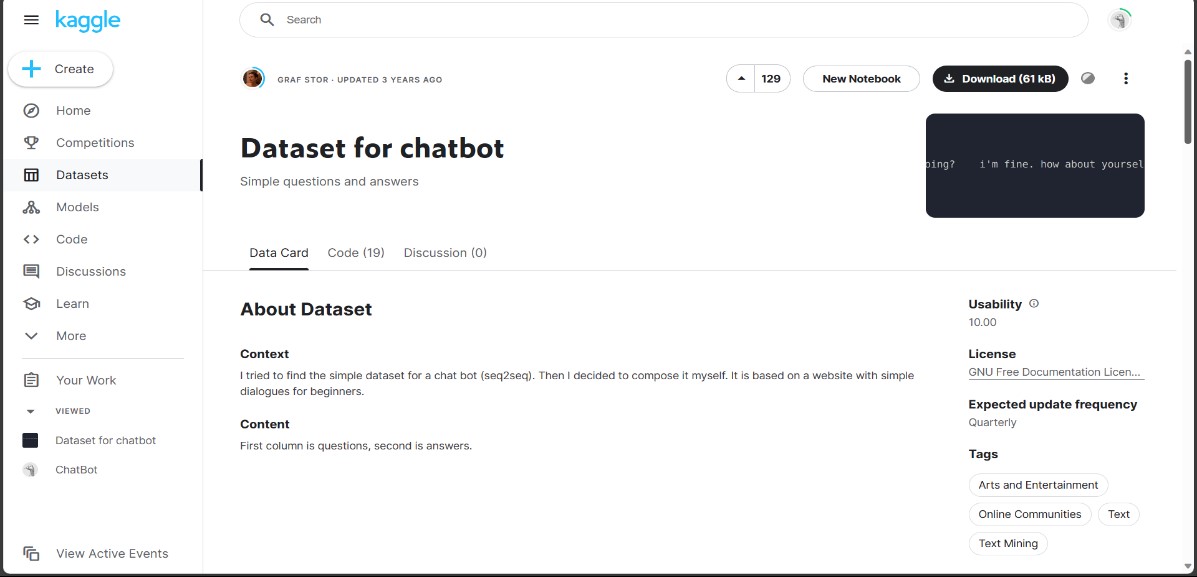
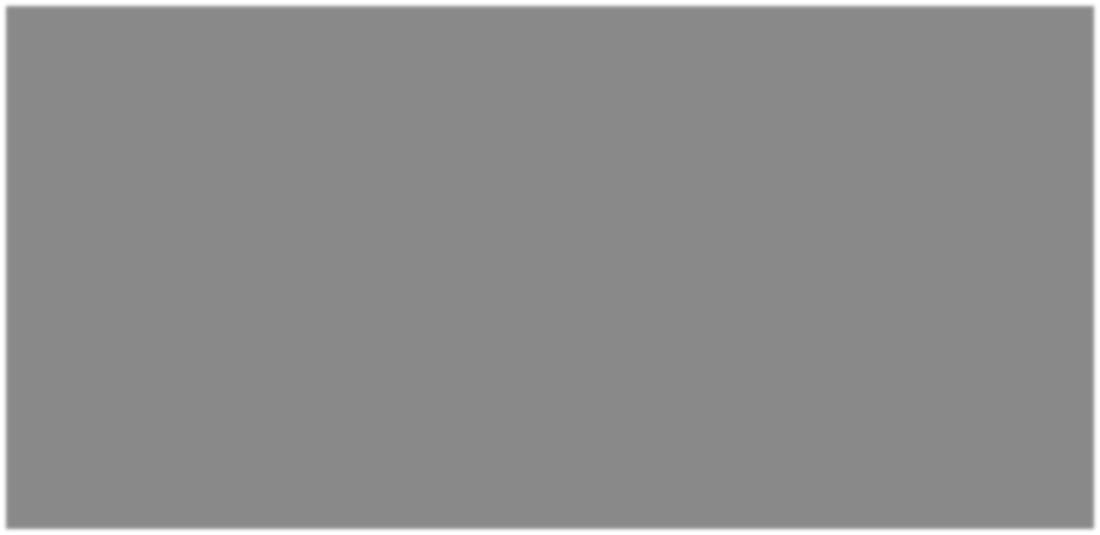
To effectively train such models, a crucial step is the preprocessing of the dataset. In this Python script, we outline a systematic approach to prepare a dataset for training a chatbot.

**Set up a development environment:**

To get started, we need to set up our development environment.we are using

Jupyter Notebook (Python3).

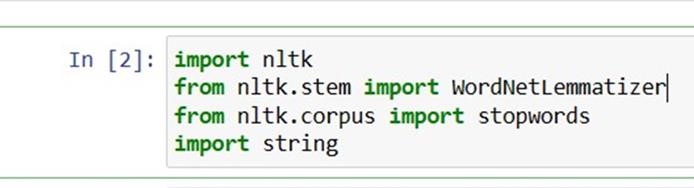
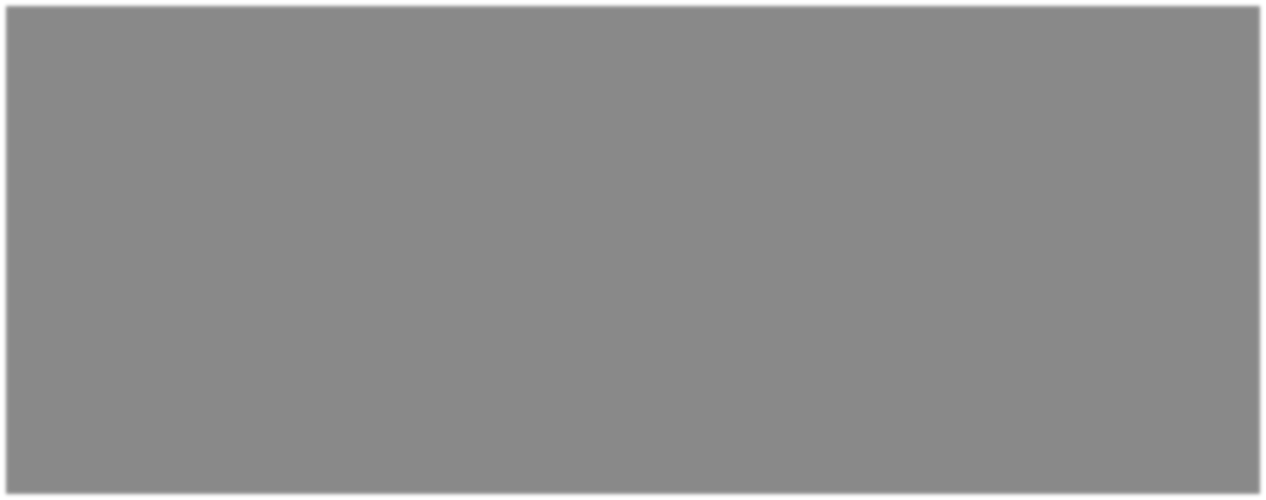
**Download corresponding dataset:**



**Data preprocessing:**

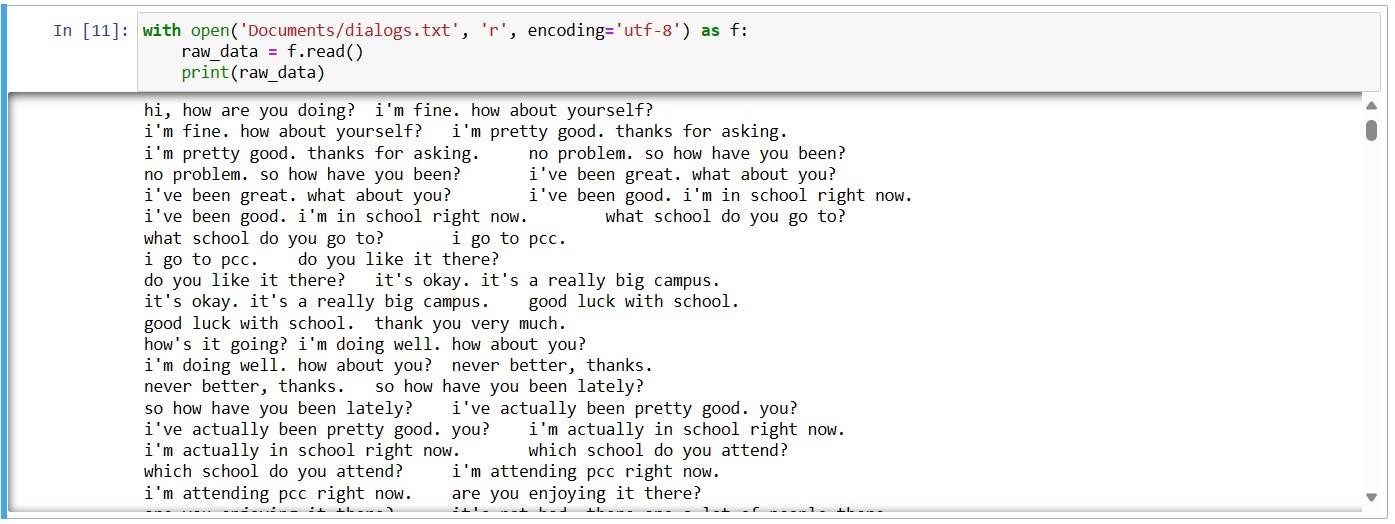
Data preprocessing is a crucial step in the data mining and data analysis process that involves transforming raw data into a format that can be understood and analyzed by computers and machine learning algorithms.

**Importing Libraries:**

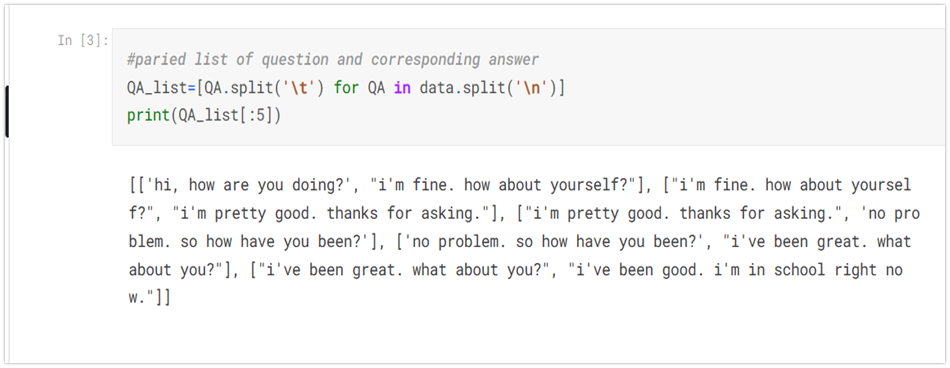


**Statement segmentation:**

Reading the dataset



**Question and Answers:**

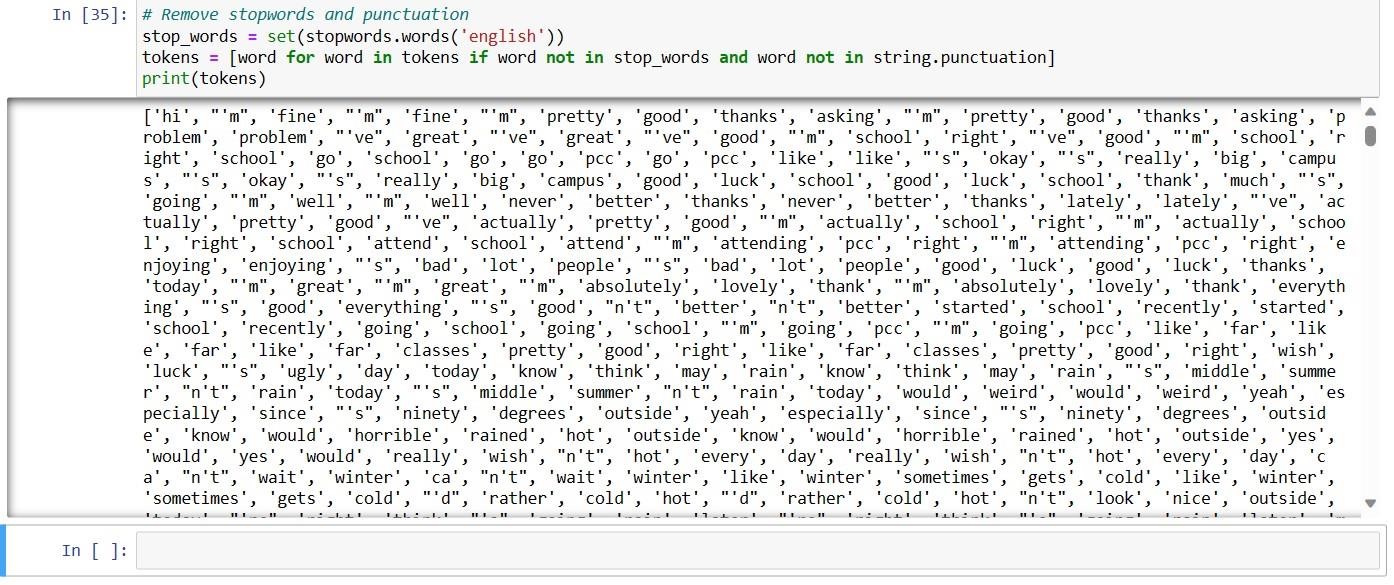


**Normalization:**

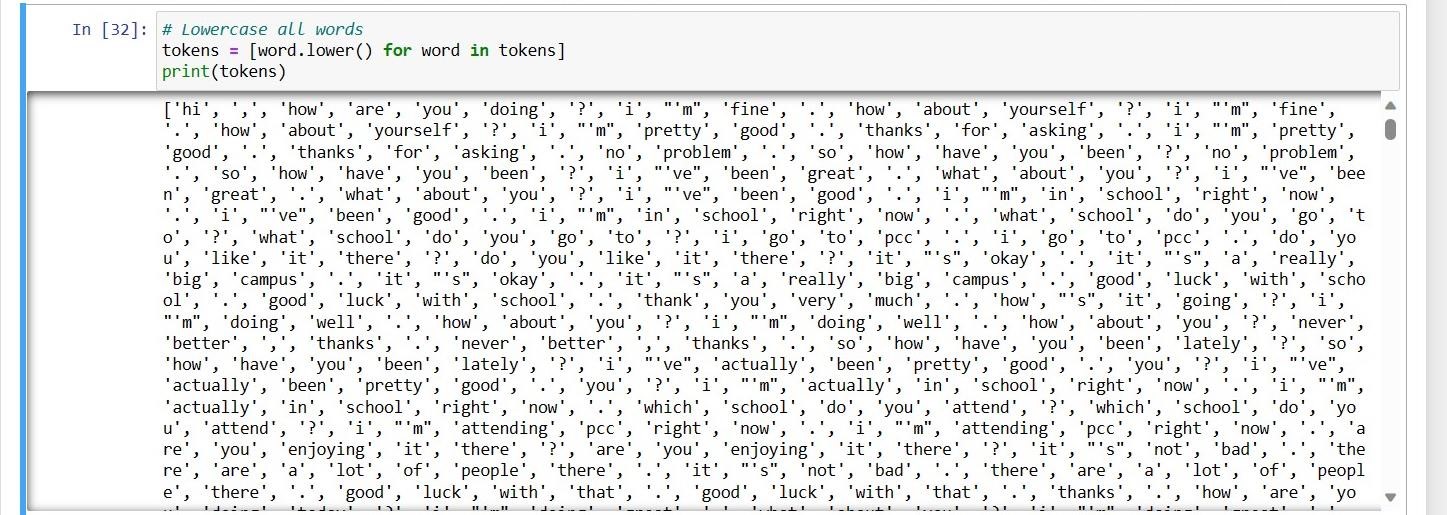
The goal of normalizing text is to *group related tokens together*, where tokens are usually the words in the text.

Depending on the text you are working with and the type of analysis you are doing, you might not need all of the normalization techniques in this post.

**Removing Punctuations:**



**Converting Uppercase to Lowercase:**

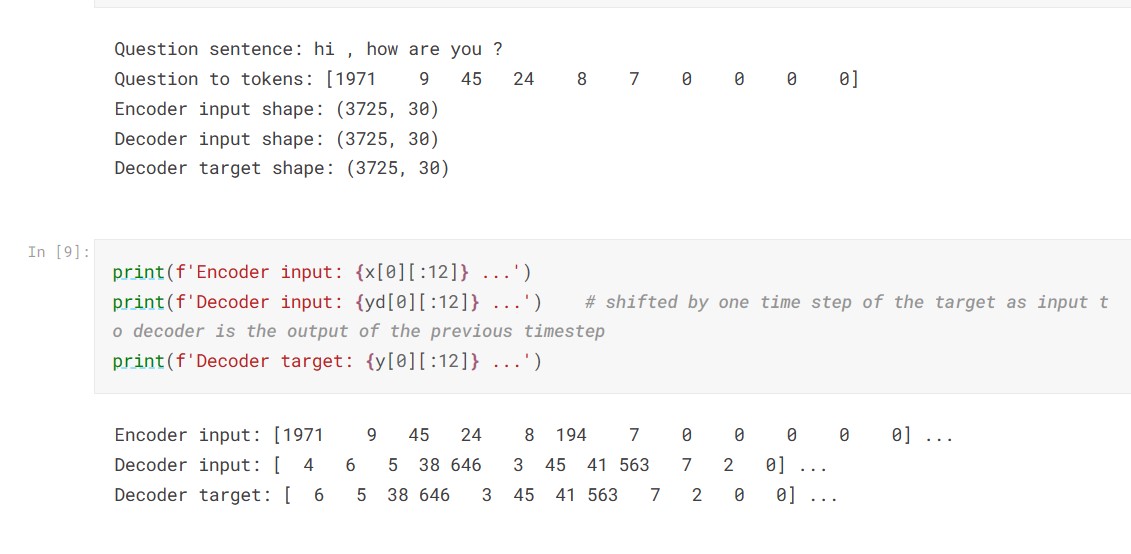


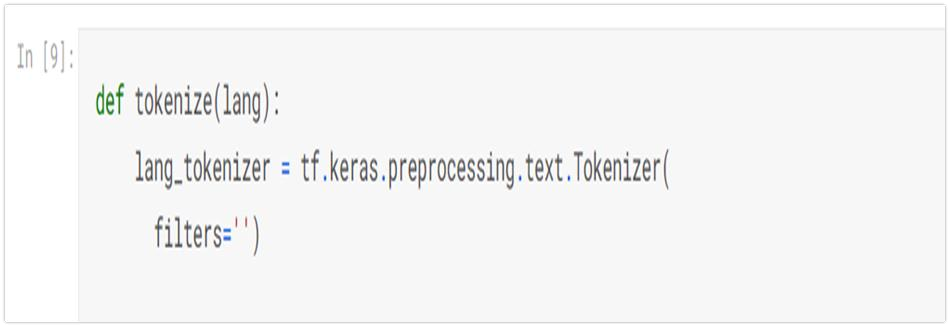
**Tokeniz**[**ation:**](https://en.wikipedia.org/wiki/Data_element)

Tokenization, when applied to data security, is the process of substituting a sensitive data element with a non-sensitive equivalent, referred to as a token, that has no intrinsic or exploitable meaning or value.



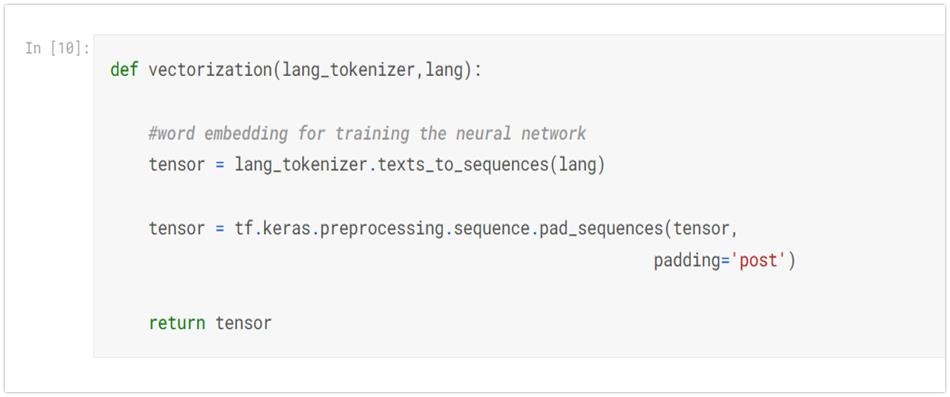




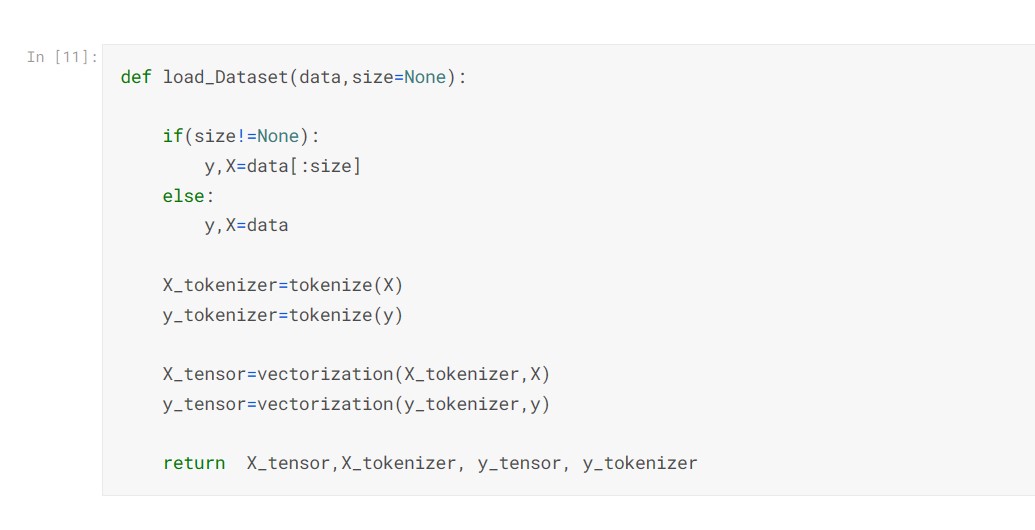
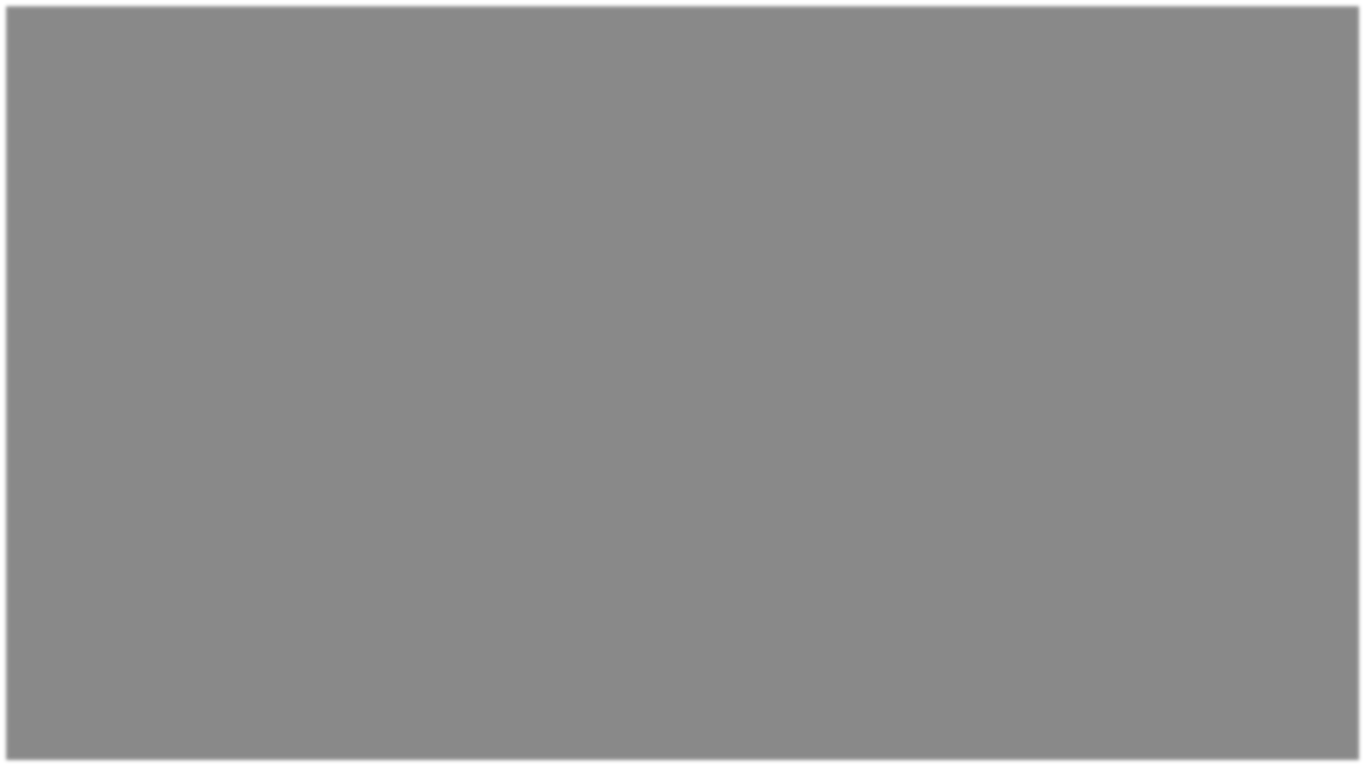


**Word Embedding**:

It is an approach for representing words and documents. Word Embedding or Word Vector is a numeric vector input that represents a word in a lowerdimensional space. It allows words with similar meanings to have a similar representation. They can also approximate meaning. A word vector with 50 values can represent 50 unique features.



**Loading dataset** :





**Text Cleaning:**



**Output:**



**Conclusion:**

Preprocessing is an essential step in chatbot development. It involves cleaning and normalizing the data, removing irrelevant information, and tokenizing the text. Preprocessing helps optimize inputs and outputs for better results . Once data is collected for training a chatbot, it’s important to pre-process it to ensure it’s clean and ready for use. And creation of virtual environment is an most important part of building chatbot.