CHATBOT IN PYTHON

Phase 3 : Development Part 1

Preprocessing the dataset

Datasets :

So far, this tutorial has focused on loading data off disk. You can also find a dataset to use by exploring the large catalog of easy-to-download datasets at TensorFlow Datasets.

As you have previously loaded the Flowers dataset off disk, let's now import it with TensorFlow Datasets.

Configure dataset for performance :

To train a model with this dataset you will want the data:

- To be well shuffled.
- To be batched.
- Batches to be available as soon as possible.

Get the Dataset :

To create a machine learning model, the first thing we required is a dataset as a machine learning model completely works on data. The collected data for a particular problem in a proper format is known as the **dataset**.

Importing the Datasets :

Now we need to import the datasets which we have collected for our machine learning project. But before importing a dataset, we need to set the current directory as a working directory. To set a working directory in Spyder IDE, we need to follow the below steps:

1. Save your Python file in the directory which contains dataset.
2. Go to File explorer option in Spyder IDE, and select the required directory.
3. Click on F5 button or run option to execute the file.

Summary:

In this article, you will learn about data preprocessing in Machine Learning: 7 easy steps to follow.

1. Acquire the dataset
2. Import all the crucial libraries
3. Import the dataset

4. Identifying and handling the missing values
5. Encoding the categorical data
6. Splitting the dataset
7. Feature scaling

Acquire the dataset :

Acquiring the dataset is the first step in data preprocessing in machine learning. To build and develop Machine Learning models, you must first acquire the relevant dataset. This dataset will be comprised of data gathered from multiple and disparate sources which are then combined in a proper format to form a dataset. Dataset formats differ according to use cases. For instance, a business dataset will be entirely different from a medical dataset. While a business dataset will contain relevant industry and business data, a medical dataset will include healthcare-related data.