

VISUALISATION STEPS

STEPS INVOLVED IN BUILDING CHAT BOT WITH PYTHON AND NLTK:

- ✓ READING STEP CORPUS.
- ✓ PRE-PROCESSING.
- ✓ TOKENIZATION, STEMMING, AND LEMMATIZATION.
- ✓ BAG OF WORDS.
- ✓ ONE HOT ENCODING.

1. READING STEP CORPUS:

The first step in building a chat bot is to gather a text corpus, which is a collection of text data. This corpus serves as the basis for the chat bot's language understanding. We can obtain text data from various sources, such as websites, books, or existing datasets.

2. PRE-PROCESSING:

Pre-processing involves cleaning and preparing the text data. This includes:

- Removing stop words (common words like “and”, “the”, “in” that add little meaning).
- Converting text to lowercase to ensure uniformity.
- Handling special characters, punctuation, and other noise in the text.

3. TOKENISATION, STEMMING AND LEMMATIZATION:

Tokenisation: Splitting the text into individual words or tokens.

Stemming: Reducing words to their root form.

Lemmatisation: It reduce their words to their base or dictionary form.

4. BAG OF WORDS(BOW):

- The bag of words representation converts text data into a numerical format.
- It creates a dictionary of unique words in the corpus and counts the frequency of each word in a document.
- The result is a sparse matrix where rows represent words and cell values represent word frequencies.

5. ONE-HOT ENDING:

One-hot encoding is used to represent categorical data, such as words in the Bag of words model, as binary vectors. Each word in the vocabulary is assigned a unique index, and the one-hot vector for a word is a binary vector with a 1 at the index of the word and 0s elsewhere.