**Tokenization** is used in natural language processing to split paragraphs and sentences into smaller units.

The first step of the NLP process is gathering the data (a sentence) and breaking it into understandable parts (words).

Here’s an **example** of a string of data:

“What restaurants are nearby?“

In order for this sentence to be understood by a machine, **tokenization is performed on the string to break it into individual parts. With tokenization**, we’d get something like this:

‘what’ ‘restaurants’ ‘are’ ‘nearby’

A large **challenge** is being able to segment words when spaces or punctuation marks don’t define the boundaries of the word. This is especially common for symbol-based languages like Chinese, Japanese, Korean, and Thai.

**Kinds of Tokenization:**

**Word Tokenization:**

**Word Tokenization**: This is the most basic form of tokenization, where the text is divided into individual words. For example, the sentence "I love cats" would be tokenized into the following tokens: ["I", "love", "cats"].

Word tokenization’s accuracy is based on the **vocabulary** it is trained with.

it’s **difficult** for word tokenization to separate unknown words or Out Of Vocabulary (OOV) words.

**Character Tokenization**: In this type of tokenization, the text is divided into individual characters. For example, the word "tokenization" would be tokenized into the following tokens: ["t", "o", "k", "e", "n", "i", "z", "a", "t", "i", "o", "n"].

**Sentence Tokenization**: In this type of tokenization, the text is divided into sentences. The exact definition of a sentence may vary depending on the language and specific requirements. For example, the text "I love cats. They are adorable." would be tokenized into the following tokens: ["I love cats.", "They are adorable."

**Lemmatization** is a technique used in language processing to convert different forms of a word into their base form. The base form is called a lemma, and it represents the word's most simple and canonical form.

For **Example**, let's consider the word "running." In lemmatization, the lemma of "running" would be "run" because "run" is the base form of the word.

The goal of lemmatization is to simplify words and reduce their variations.

**Stemming** is another technique used in Natural Language Processing (NLP) to reduce words to their base or root form, called a stem. It aims to remove prefixes and suffixes from words to obtain a simplified version of the word.

In simple words, stemming trims down words to their core form by removing common word endings. For example, the words "running," "runs," and "ran" would all be stemmed to the root form "run."

Lemmatization:

Lemmatization aims to convert words to their base or canonical form, known as a lemma.

It considers the context, grammar, and part of speech of words to determine their base form accurately.

Lemmatization produces valid words that exist in the language and have proper linguistic normalization.

It uses dictionaries, linguistic rules, or machine learning algorithms to generate lemmas.

Example: The word "running" would be lemmatized to "run" because "run" is its base form.

Stemming:

Stemming aims to reduce words to their root form, called a stem, by removing prefixes and suffixes.

It applies a set of predefined rules to truncate words, without considering linguistic context or part of speech.

Stemming can result in stems that are not actual words.

It is a simpler and faster process compared to lemmatization.

Example: The words "running," "runs," and "ran" would all be stemmed to the root form "run."