Tokenizers in NLP

What are they, and why are they used?

What are Tokenizers?

A tokenizer is a program that breaks up text into smaller pieces or **tokens**. There are many different tokenizers, but the word and character tokenizers are the most common**.** There is also another type that is **sub-word (n-gram) tokenizers.**

Word Tokenizers

Tokenization is a process that occurs when there are white spaces in between letters. Tokenizers divide up text into individual tokens.

1. The first step is called **tokenization**;
2. The second step is called **lemmatization**;
3. The third step is called **stemming**;
4. The fourth and final step is called **inflection**.

Lemmatization happens when we replace all of the different forms of a word with its root form. For example, in English we often use “read,” “reads,” and “read,” but in some languages like Spanish and French, each of these forms will be replaced with their root form: “leer.” The process is known as lemmatization.

The third step is called **stemming**, where we **replace a word with** parts of words that aren’t actual letters, such as the suffixes “-tion,”

“-sion,” and “-ment.**its root form or stem.** For example, the most common way to stem words in English is by removing “-ing” and replacing it with “.” There are also more complicated ways of stemming that involve removing parts of words that aren’t actual letters, such as the suffixes “-tion,” “-sion,” and “-ment.”

Finally, there’s **inflectional morphology, which includes certain grammatical rules that usually involve making a word plural, or changing it from a verb to a noun.** For example, if you add ‘er’ to the word ‘design’, it becomes ‘designer’ but the former is a verb whereas the latter becomes a noun.

Character Tokenizers

Character tokenizers break up text into individual characters. **They are often used to categorize language more granularly than word tokenizers.** For example, a sentence written in French would be broken apart with a character tokenizer, and each of the letters could be categorized as being an uppercase letter, a lowercase letter, a number, or a symbol. Character tokenizers are also used for languages like Chinese and Japanese that don’t use the same alphabet as English (called “Romanization”), as well as languages like Arabic that need to be fully understood before they can be translated into another language.

Tokenizers are used in a variety of ways today, some of which include:

TOOLS USED FOR TOKENIZATION

1. NLTK or Natural Language Toolkit:

   One of the most popular applications of NLTK is tokenization or the process of breaking text into smaller pieces. By separating the text into individual tokens, NLTK can help with a variety of tasks, such as classification, information extraction, machine translation, and text-to-speech applications.

2. Stanford CoreNLP:
3. **Chainer NLP Tokenizer:**

   Chainer NLP Token is an open-source neural network framework for Natural Language Processing (NLP). Chainer NLP Tokenizer is a tokenizer that uses a neural network to separate sentences into words and classify the word types.

   It can be used as Tokenizer in most NLP frameworks such as spaCy; however, it is difficult to use only with external APIs because it requires another “token” and “word” data structure along with sentence data.

   So, the next time you’re trying to learn a foreign language or just want to know what your phone is saying, remember you have a tokenizer right at your fingertips!