DATA SCIENCE WITH PYTHON : TOKENIZATION & LEMMATIZATION #1856

I AM DEEPTHI M.

SERIAL NUMBER:198

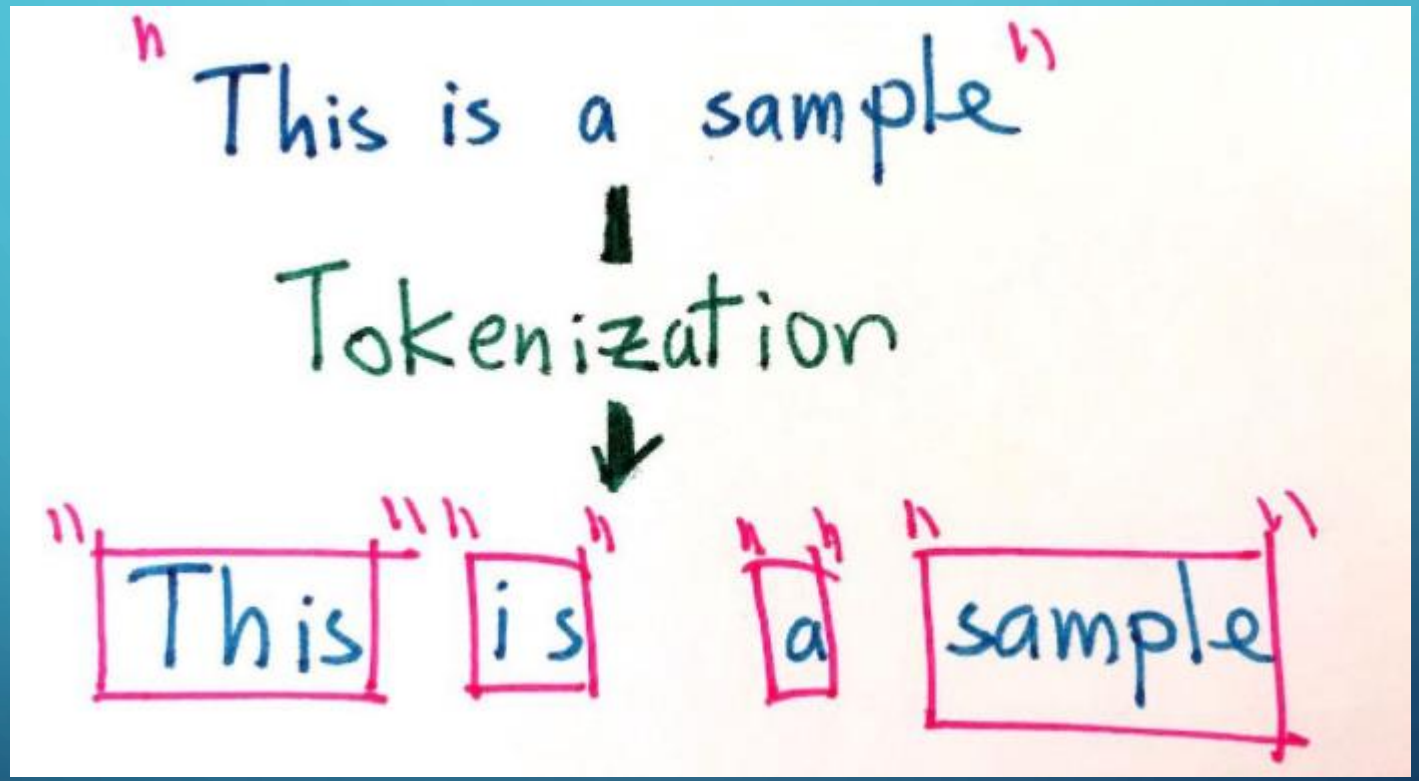
BATCH-5

Tokenization

It is a common task in NLP.

It is the building blocks of NLP.

It is the way of separating a text into smaller units called “tokens”.



Types of Tokenization:

Tweet Tokenizer:

- Specifically designed for tokenizing tweets.

MWE tokenizer:

- Multi-Word Expression.
- Certain group of multiple words are treated as one entity during tokenization.

Regular Expression tokenizer

- Developed using regular expression.
- Sentence are split based on occurrence of a pattern.

Whitespace Tokenizer:

- Splits a string whenever a space, tab, or newline character is present.

Word Punkt Tokenizer:

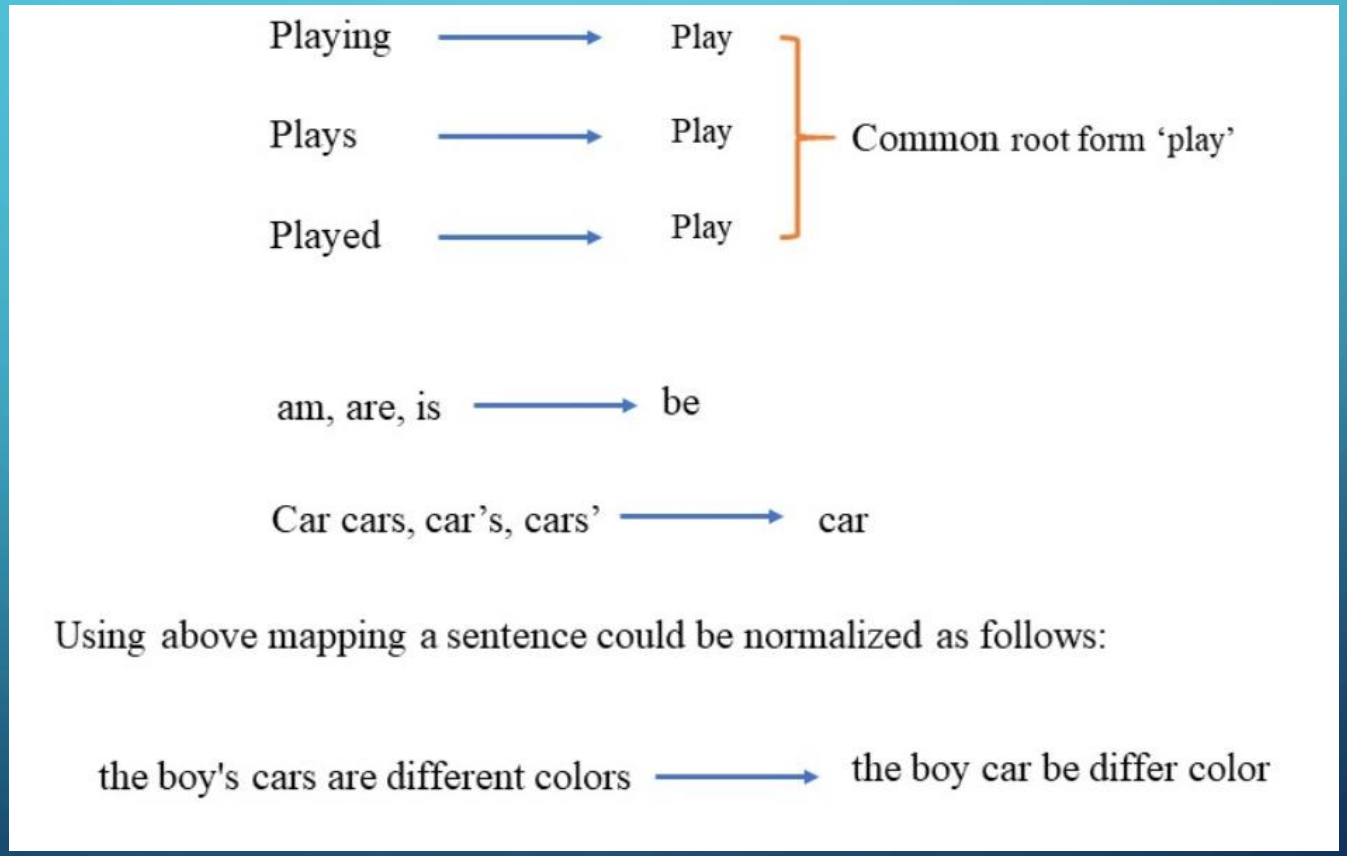
- Splits text into a list of characters and digits.

Why tokenization? Or advantages.

- We can easily apply the NLP models.
- Easy to evaluate the text.
- Easy to create word clouds.
- Easy to implement pre-processing techniques like Lemmatization, stemming, stop word removal.

Lemmatization:

- It derives the root word in a text.



Why Lemmatization? Or Advantages

- Helps to reduce the length of words.
- Increases the model performs.
- Decreases the training time.
- Improves the accuracy.