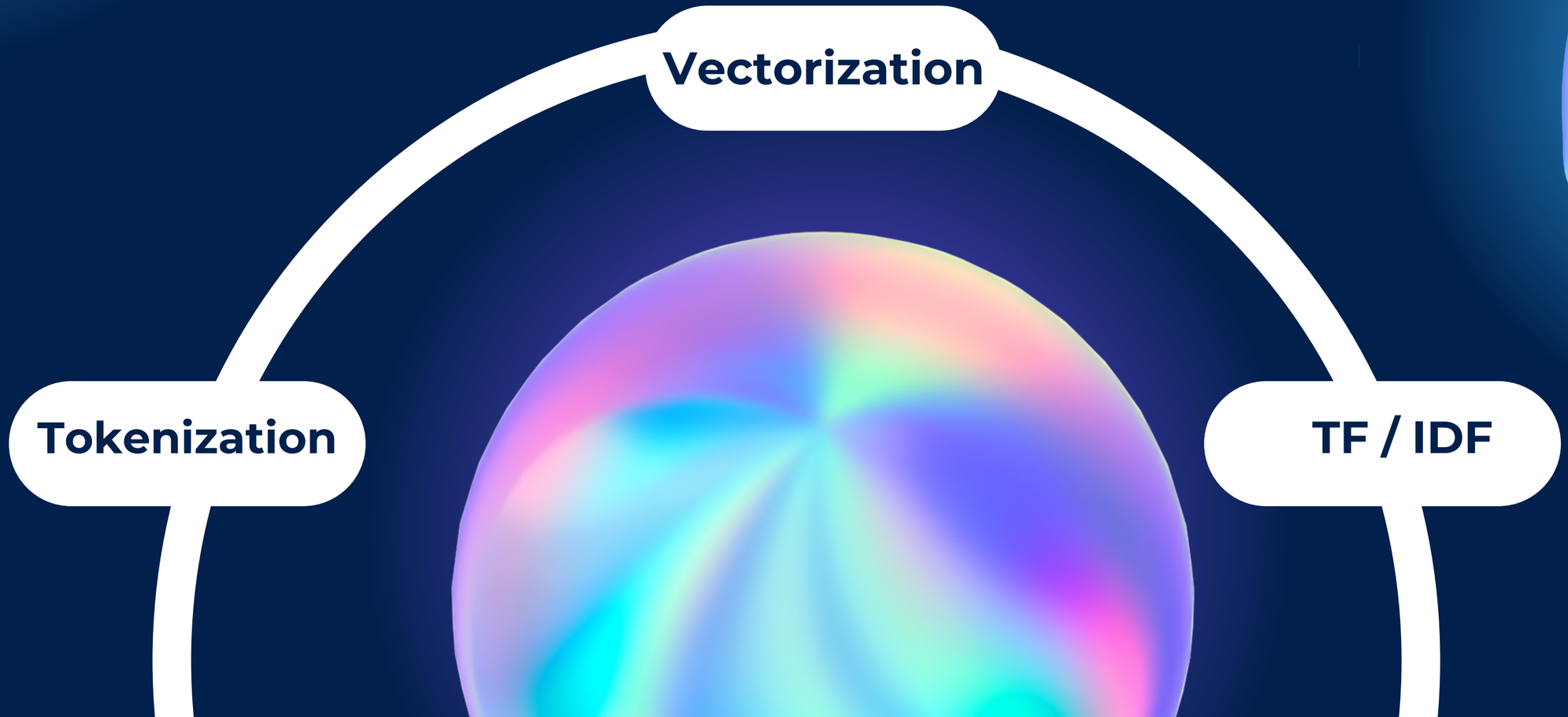
Presentation On



Janki Gajjar

Bhargav Joshi

Aarti Vadukar

Utsav Raychura

Aashtha Gondaliya

Khushali Mesvaniya

Tokenization

- Tokenization, in the realm of Natural Language Processing (NLP) and machine learning, refers to the process of converting a sequence of text into smaller parts, known as tokens.

- These tokens can be as small as characters or as long as words.

- The primary reason this process matters is that it helps machines understand human language by breaking it down into bite-sized pieces, which are easier to analyze.

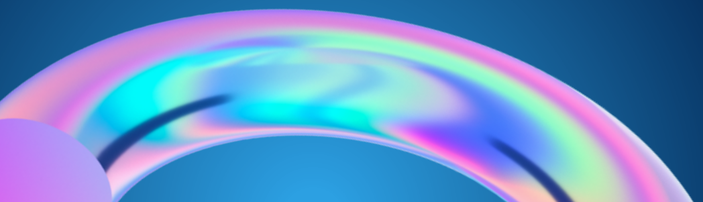
Type Of Tokenization :

1. Word tokenization
- This method breaks text down into individual words.

2. Character tokenization
- The text is segmented into individual characters.

3. Sentence tokenization
- The sentence tokenization is to identify and extract each separate sentence from a given paragraph or document.

Vectorization



- Vectorization is the process of converting text data into numerical vectors.

- Imagine you have a collection of documents. Each document is represented as a vector, where each dimension of the vector corresponds to a unique word in the entire collection.

- If a word appears in a document, its corresponding dimension in the vector gets a value indicating its frequency or presence.

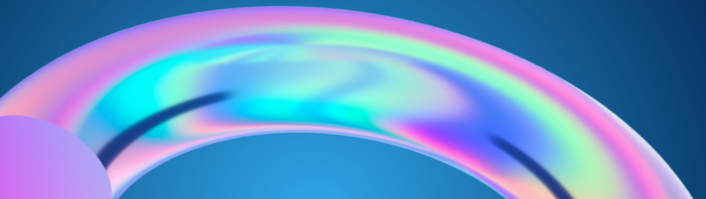
- This allows us to perform mathematical operations on text data and enables machine learning algorithms to work with textual information.

- Vectorization in NLP is important because many machine learning algorithms and statistical models require numerical input.

- By converting textual data into numerical representations, vectorization enables the application of these models and algorithms to NLP tasks.

TF / IDF

Term Frequency-Inverse Document Frequency



- TF-IDF stands for Term Frequency Inverse Document Frequency of records.

- TF-IDF is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents.

- It consists of two parts:
  1. Term Frequency (TF)
  2. Inverse Document Frequency (IDF).

- Term Frequency (TF) measures how frequently a term appears in a document. The more often it appears, the higher its TF value.

- Inverse Document Frequency (IDF) measures how unique or rare a term is across all documents in the collection. Rare terms are weighted higher.

- TF-IDF is calculated by multiplying TF with IDF. This results in higher scores for terms that are frequent in the document but rare in the entire collection, thus capturing their significance.

Vectorization v/s TF/IDF

Consider a corpus consisting of three documents:

Document 1: "The cat sat on the mat."
Document 2: "The dog played in the yard."
Document 3: "The cat and the dog are friends."

Vectorization:

	the	cat	sat	on	mat	dog	played	in	yard	and	are	friends
Doc 1	1	1	1	1	1	0	0	0	0	0	0	0
Doc 2	1	0	0	0	0	1	1	1	1	0	0	0
Doc 3	2	1	0	0	0	1	0	0	0	1	1	1

- Each row represents a document, and each column represents a word in the vocabulary.
- The numbers represent the frequency of each word in the corresponding document.

TF / IDF :

$$\text{TF (Term Frequency)} = \frac{\text{(Number of times term appears in a document)}}{\text{(Total number of terms in the document)}}$$

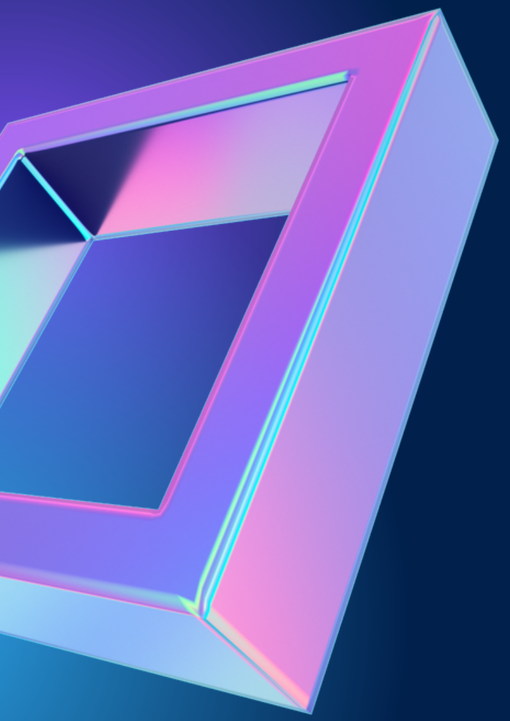
$$\text{IDF (Inverse Document Frequency)} = \frac{\text{(Total number of documents)}}{\text{Number of documents containing the term)}}$$

For the word "cat":

TF in Document 1 = 1/6

TF in Document 3 = 1/7

IDF = log(3/2)



Thank You