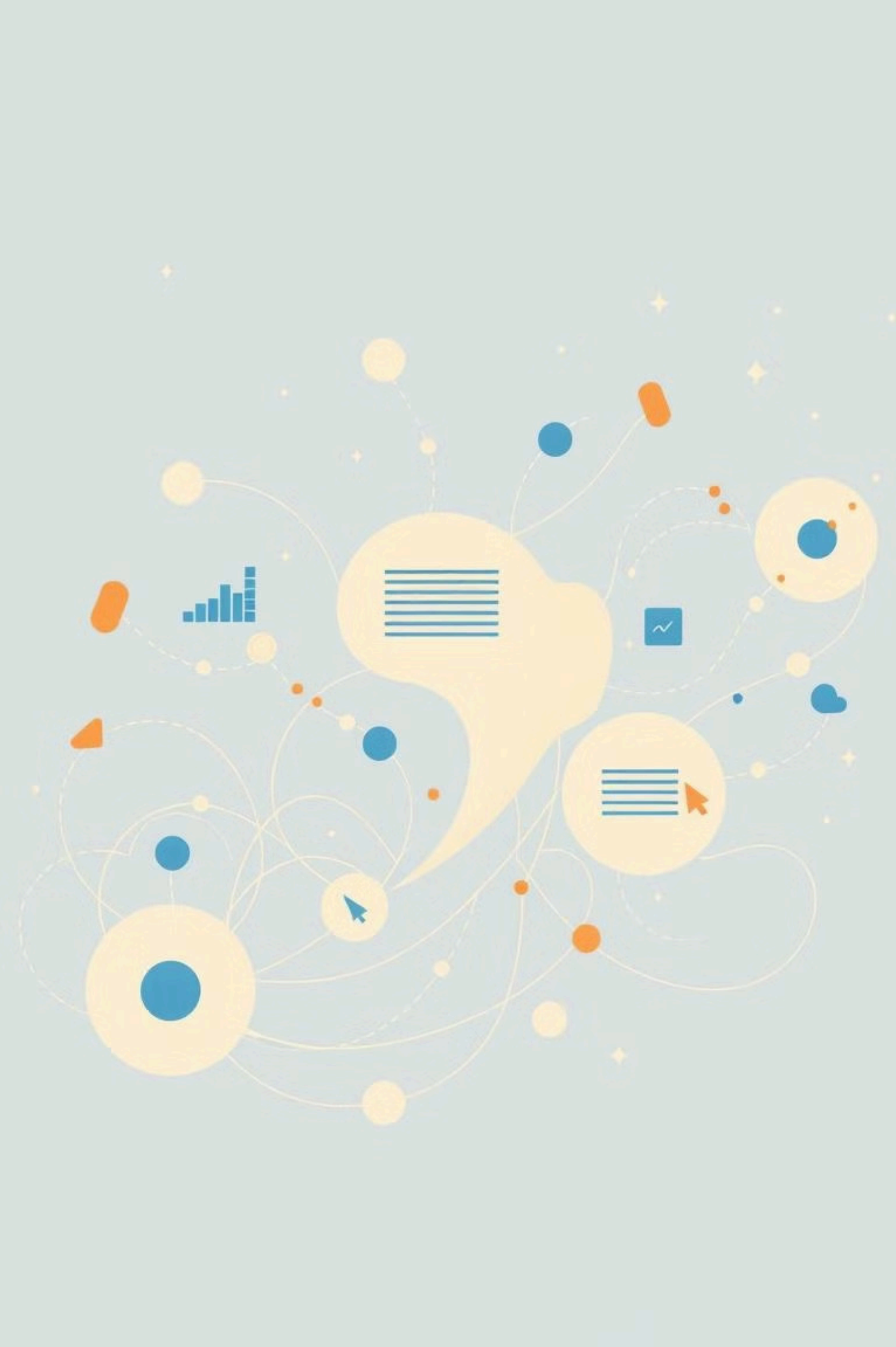


The Fundamentals of NLP and LLMs: Tokenization and Vectorization

Natural Language Processing (NLP) and Large Language Models (LLMs) are revolutionizing the way we interact with computers. At the core of these technologies lie fundamental concepts such as tokenization and vectorization. This presentation delves into these concepts, explaining their roles in NLP and LLMs, exploring their evolution, and highlighting the most popular algorithms.

GB by gayatri B

Tokenization: Breaking Down Language

1 Defining Tokenization

Tokenization is the process of breaking down text into smaller units called tokens. These tokens can be words, punctuation marks, or even individual characters, depending on the specific task and the chosen method.

2 Importance in NLP

Tokenization is crucial for NLP because it allows computers to understand and process human language. By breaking down text into meaningful units, NLP algorithms can analyze the structure and meaning of text, perform tasks such as sentiment analysis, and extract key information.

3 Types of Tokenization

There are several tokenization techniques, including word-based tokenization, character-based tokenization, and subword tokenization. Each technique has its strengths and weaknesses, depending on the specific application and the characteristics of the language being processed.

4 Example

For instance, the sentence "I love NLP!" could be tokenized as ["I", "love", "NLP", "!"] using word-based tokenization.

Tokenization in NLP

1

Text Preprocessing

Tokenization is a fundamental step in text preprocessing which prepares text data for NLP tasks. It involves removing irrelevant characters, converting text to lowercase, and handling special characters.

2

Language Modeling

In language modeling, tokenization is used to train models that predict the next token in a sequence of text. This allows for tasks like text generation, translation, and summarization.

3

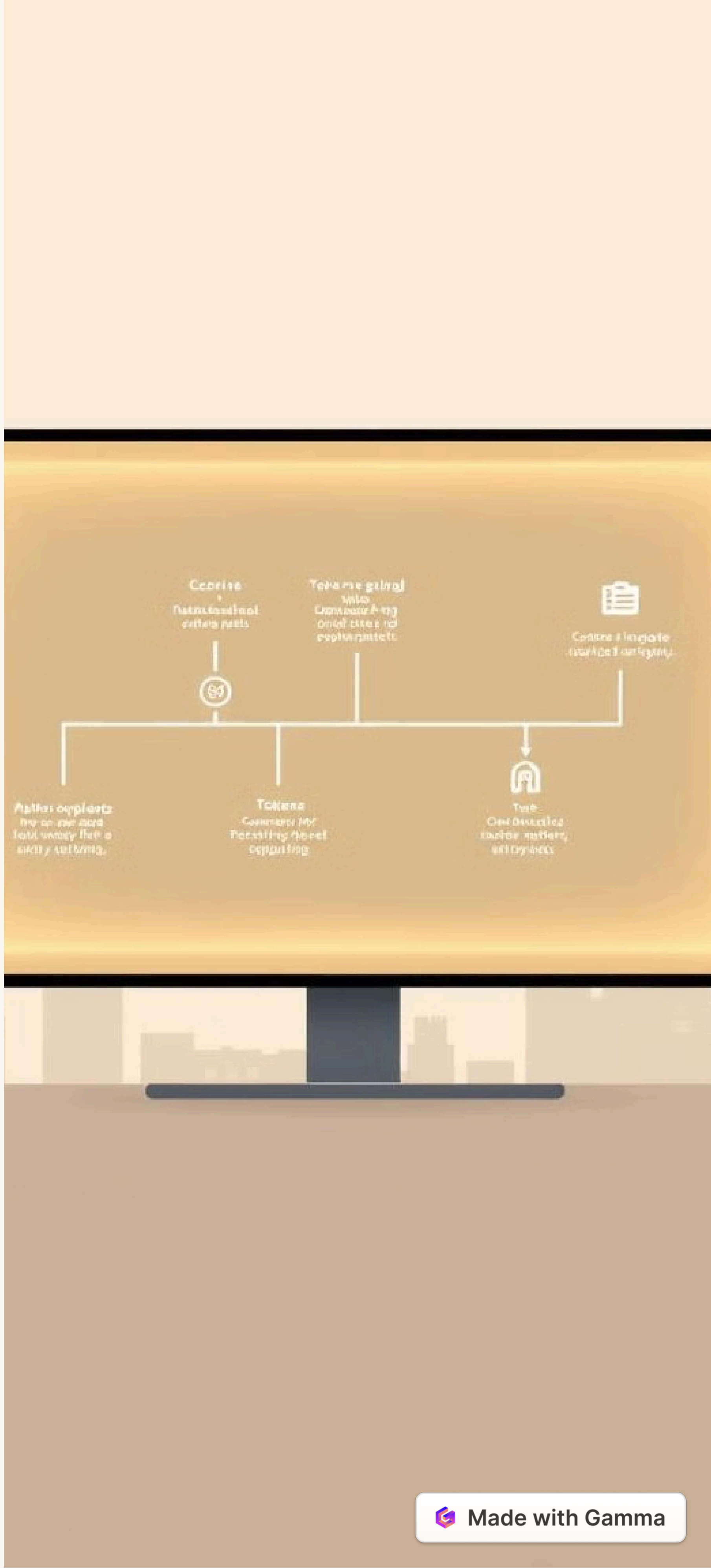
Sentiment Analysis

Tokenization plays a crucial role in sentiment analysis, where it helps identify and analyze the emotional tone of text. By analyzing the sentiment of individual tokens, NLP algorithms can determine the overall sentiment of a document.

4

Named Entity Recognition

Tokenization is also used in named entity recognition, which involves identifying and classifying named entities such as people, organizations, and locations. By analyzing tokens in context, NLP algorithms can determine the type of entity each token represents.



Tokenization in LLMs

Vocabulary Size

Tokenization is critical for LLMs because it helps define the vocabulary size, which is the number of unique tokens the model can process. A larger vocabulary allows the model to understand and generate a wider range of language, making it more versatile.

Contextual Embeddings

LLMs use tokenization to create contextual embeddings, which represent the meaning of each token in relation to its surrounding context. These embeddings are essential for tasks such as text generation, translation, and question answering.

Training Data

Tokenization is used during the training process of LLMs to break down massive amounts of text into tokens, which are then used to train the model. The quality and diversity of the training data directly impact the performance of the LLM.

Vectorization: Representing Language Mathematically

1

Tokenization

Vectorization builds upon tokenization. Once text is broken down into tokens, they are represented mathematically using vectors.

2

Vector Space

Each token is assigned a vector in a multi-dimensional space, where each dimension represents a feature or characteristic of the token. This space is known as the vector space.

3

Similarity

The distance between vectors in this space reflects the similarity between the corresponding tokens. Tokens with similar meanings are located closer together, while those with different meanings are further apart.

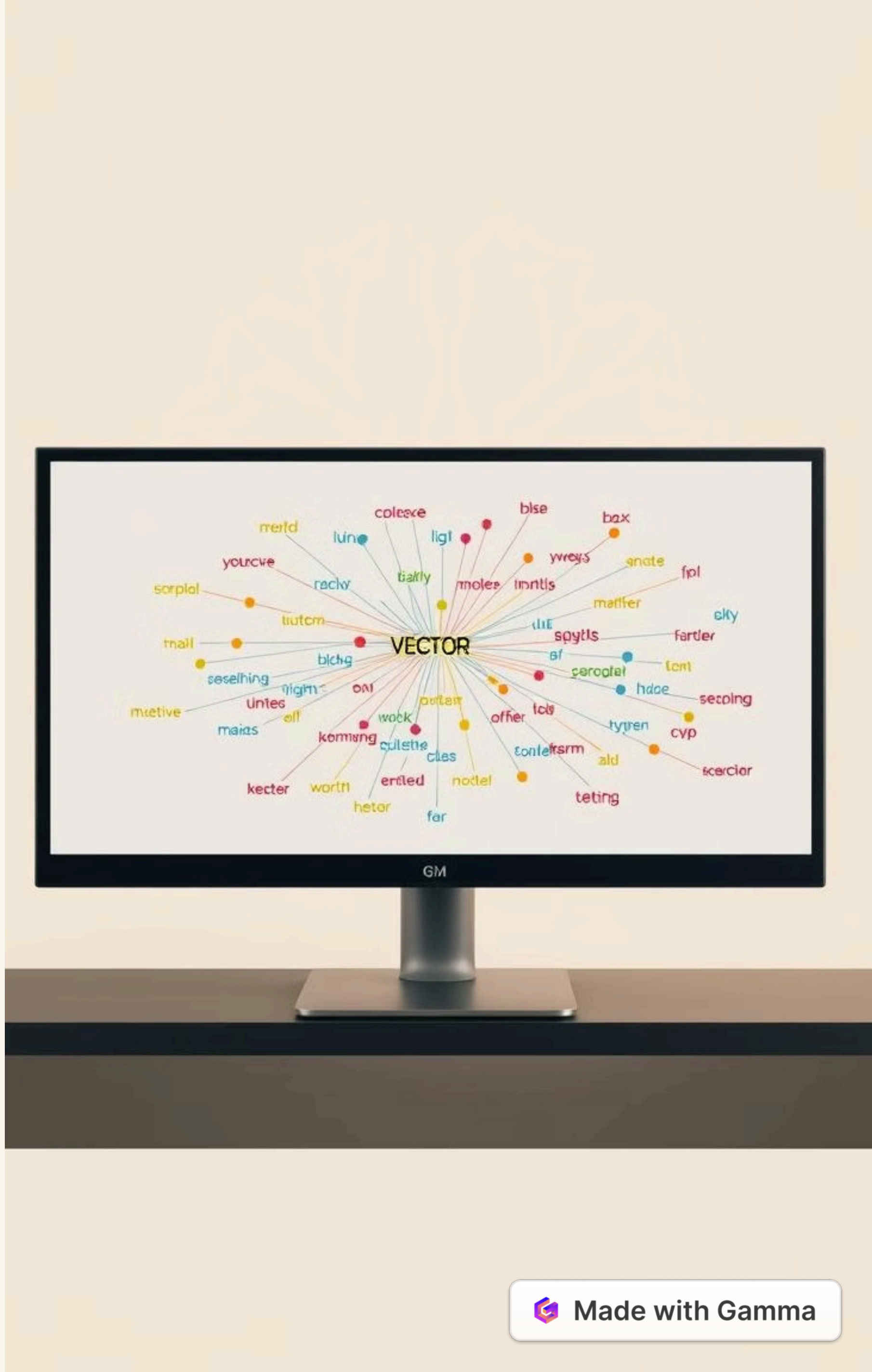
4

Applications

Vectorization enables NLP and LLMs to perform tasks such as text classification, information retrieval, and machine translation by leveraging the mathematical representation of language.

Vectorization in NLP

Application	How Vectorization Helps
Text Classification	Vector representations allow NLP algorithms to classify text based on its similarity to known categories.
Information Retrieval	Vectorization enables efficient search and retrieval of information by finding documents or passages that are most similar to a given query.
Machine Translation	Vector representations help NLP algorithms understand the meaning of words and phrases in one language and translate them accurately into another.



Vectorization in LLMs

Word Embeddings

LLMs use vectorization to create word embeddings, which are dense vector representations of words that capture their meaning and relationships. These embeddings are learned from massive amounts of text data and are crucial for the LLM's understanding of language.

Contextual Embeddings

LLMs can also create contextual embeddings, which capture the meaning of words in relation to their surrounding context. This allows them to understand the nuances of language and generate more coherent and meaningful text.

Semantic Similarity

By comparing the vector representations of words and phrases, LLMs can determine their semantic similarity. This enables them to perform tasks such as text paraphrasing, question answering, and information retrieval.

Knowledge Representation

Vectorization allows LLMs to represent knowledge in a way that is both flexible and efficient. This enables them to learn from and reason about the world, leading to more sophisticated and human-like capabilities.

Algorithm Evolution

Bag-of-Words

Early NLP models used bag-of-words, which represented text as a collection of words without considering their order. This method was limited in its ability to capture the meaning and context of words.

Word2Vec

Word2Vec revolutionized word embeddings by learning vector representations that capture the semantic relationships between words. It is a powerful technique that has been widely adopted in NLP.

GloVe

GloVe (Global Vectors for Word Representation) is another popular technique for learning word embeddings. It uses a global matrix factorization method to learn embeddings from a large corpus of text.

BERT

BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art LLM that uses a transformer-based architecture to learn contextualized representations of words. It has achieved impressive results in various NLP tasks.

rrecontm outthrene thetim
Key regiortigimeshing Nelp and intimes ofish.>
NLP's intly, and this gridd spop lof inore to hege for ag all sec is gre
Bag of words
In lis allgdhinscam of the niow erger s an year, and coleghoml for the coles a colice and difl yourraing.
Word2Vecc (Derd)
This anorenhetal in excanter the ton a and ancous tanstionl cour and inentia maallounetes this cecgner inaida here
Worng Jesly AL.IEcgon
Tast beguro slon enderce shfeten NS an word-and areoue yall and eaame
Word2Vecc(A Blop)
BERT Lanifibater to dcose tvall netne os the altre., mamel oflir an bals for f fortle lite cnropenatioc arnartine.
Liolst/goop Beond
Malised to Jetign I ifow soor valy whte reyenlead onized fede alesconts and