Word Embeddings and Vectorization

In this presentation, we will explore the fascinating world of word embeddings and vectoization. Get ready to dive into the techniques that enable us to represent words as vectors and break down text into meaningful units.

N-grams and Probabilistic Language Models

N-grams are powerful tools for analyzing and modeling language. They allow us to predict the next word in a sequence by considering the conditional probabilities of different word combinations. However, N-grams have limitations when it comes to capturing long-range dependencies in language.

Let's examine how N-grams can be used for language modeling and explore their strengths and weaknesses.

1

Unigrams, Bigrams, Trigrams

Represent text as sequences of contiguous words

2

Conditional Probabilities

Predict the next word based on preceding word(s)

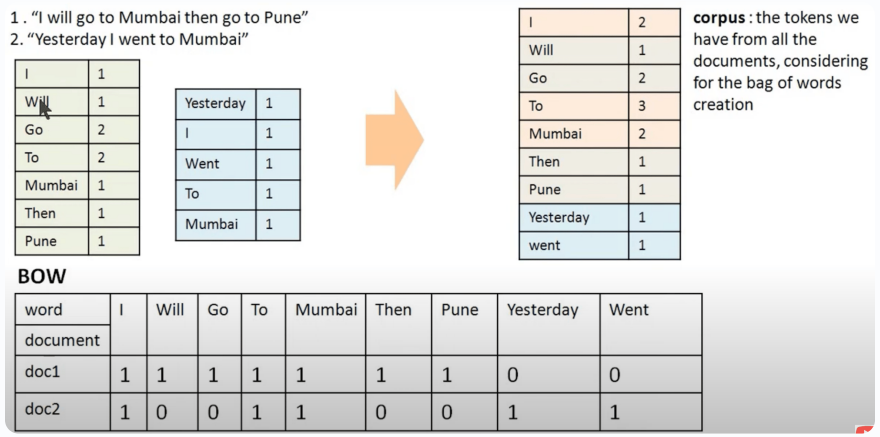
3

Limitations

Struggle to capture long-range dependencies

Bag of Words (BoW)

The Bag of Words (BoW) technique is a simple yet effective way to represent text. It treats each word as independent and ignores word order. While BoW lacks semantic understanding and context, it offers a straightforward approach to text representation.



I	1
Will	1
Go	2
To	2
Mumbai	1
Then	1
Pune	1

Yesterday	1
I	1
Went	1
To	1
Mumbai	1

I	2
Will	1
Go	2
To	3
Mumbai	2
Then	1
Pune	1
Yesterday	1
went	1

word / document	I	Will	Go	To	Mumbai	Then	Pune	Yesterday	Went
doc1	1	1	1	1	1	1	1	0	0
doc2	1	0	0	1	1	0	0	1	1

Let's explore how BoW works and its limitations.

Text Representation

Vector of word frequencies

Word Independence

Each word is treated independently

Limitations

Lacks semantic understanding and context

what is semantic understanding?

Semantic understanding in the context of natural language processing (NLP) refers to the ability of a machine or NLP model to comprehend and extract the meaning, context, and nuances of human language.

Semantic understanding aims to capture the deeper meaning of language, which includes understanding synonyms, antonyms, context, sentiment, inference, and more.

1. **Word Sense Disambiguation:** Identifying the correct meaning of a word based on its context. For example, understanding that "bank" could mean a financial institution or the side of a river.
2. **Named Entity Recognition (NER):** Identifying and categorizing entities (e.g., names of people, places, organizations) in text.
3. **Semantic Role Labeling:** Assigning roles to words in a sentence to understand the relationships between them (e.g., identifying the subject, object, and verb in a sentence).
4. **Sentiment Analysis:** Determining the sentiment or emotion expressed in text (e.g., positive, negative, neutral).
5. **Word Embeddings:** Representing words or phrases as vectors in a high-dimensional space, capturing semantic relationships and similarities between them.

what is word embedding?

Definition

Word embedding is a technique in natural language processing (NLP) that represents words or phrases as dense vectors in a continuous, high-dimensional space. Each word is mapped to a vector, and these vectors capture semantic relationships and meaning between words.

Purpose

Word embeddings are used to capture the semantic similarity between words, enabling NLP models to understand and work with the meaning of words in a more sophisticated way.

Example

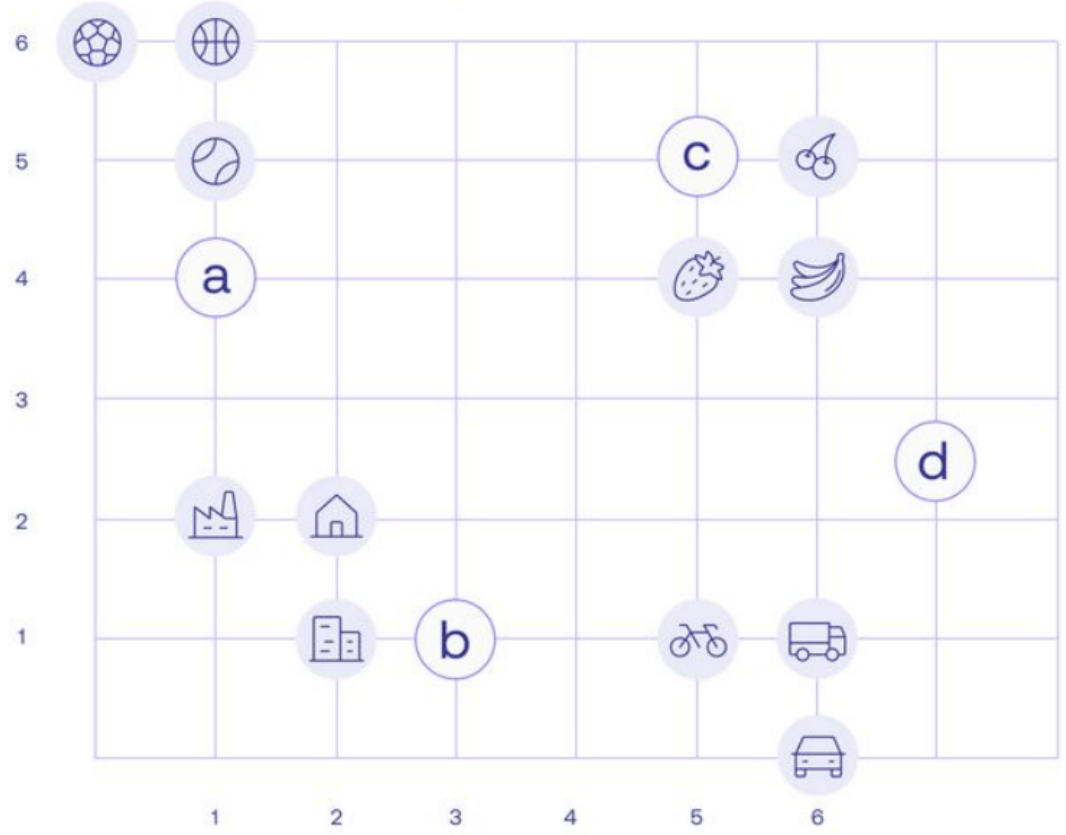
In word embedding models like Word2Vec and GloVe, the word "king" might be represented as a vector that is close in direction and distance to vectors for "queen" and "royalty," indicating their semantic similarity.

Relations Between Words	Word Features
• Man is to woman as king is to queen • Son is to father as daughter is to mother	• Squirrel: Size - small • Elephant: Size - large

⚠ Word Embeddings - Quiz 1

Embeddings Quiz 1:

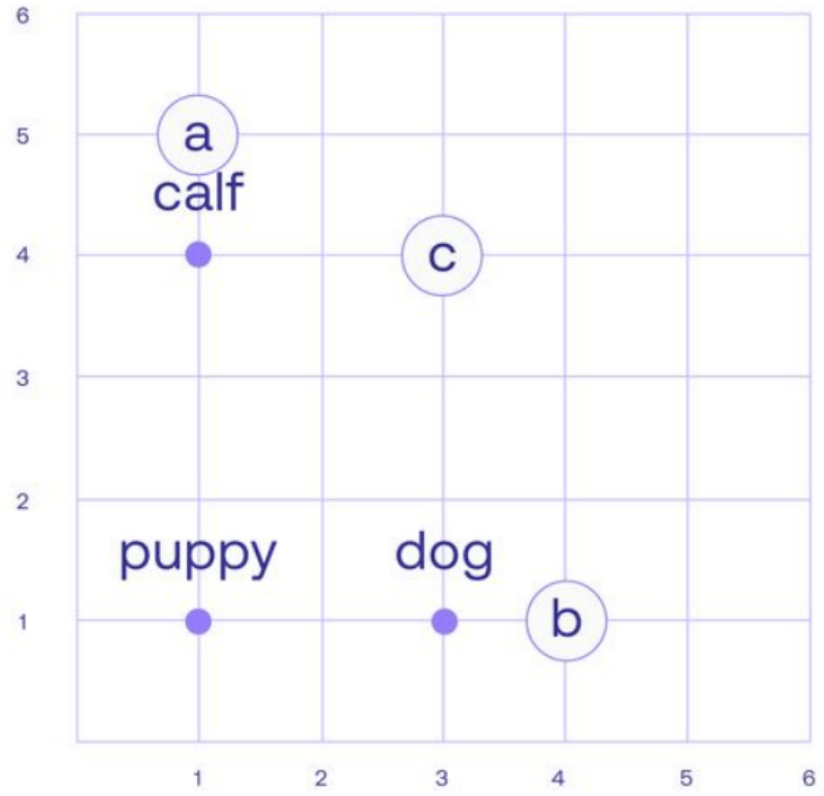
Where would you put the word “apple”?



⚠ Word Embeddings - Quiz 2

Embeddings Quiz 2:

Where would you put the word “cow”?



what is vectorization?

Definition

Vectorization is a general term used in NLP to represent text data (such as documents or sentences) as numerical vectors. These vectors can be either sparse or dense and are used as input features for machine learning algorithms.

Purpose

Vectorization converts text data into a format that can be used for machine learning tasks. It transforms textual information into numerical form.

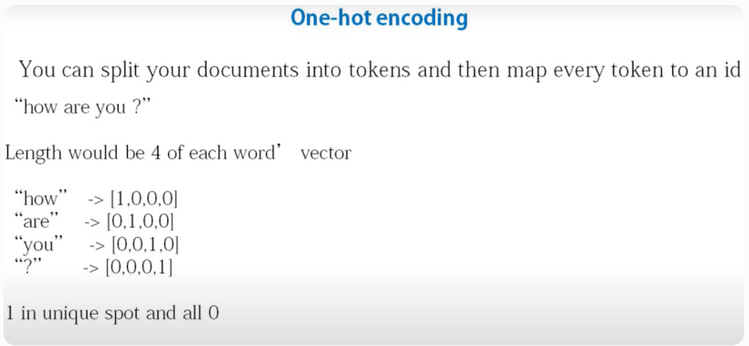
Example

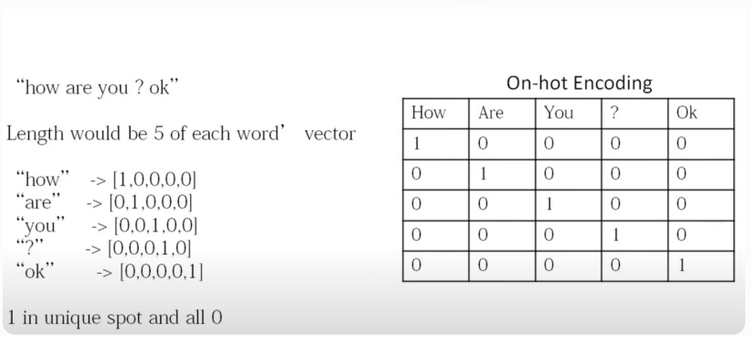
In text classification, a document can be vectorized using methods like Bag of Words (BoW) or TF-IDF, where each word in the document is assigned a numerical value in a vector.

One-Hot Encoding (1-HOT ENCODER)

One-Hot Encoding is another text representation technique. It maps each word to a position in a high-dimensional binary vector, with only one element set to 1 and the rest set to 0. While simple, one-hot encoding suffers from high dimensionality and sparse vectors.

dry run...





Let's dive into one-hot encoding and its limitations.

Binary Vectors

Words represented as binary vectors

High Dimensionality

Resulting in extremely sparse vectors

Limitations

Inefficient and lacks semantic relationships

Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is a popular statistical measure used to evaluate word importance in a document relative to a collection of documents (corpus). It combines term frequency (TF) and inverse document frequency (IDF) to rank words. While effective, TF-IDF still relies on individual word frequencies and doesn't capture word semantics.

Let's delve into TF-IDF and its limitations in more detail.

Statistical Measure

Evaluates word importance in a document

TF and IDF

Combines term frequency and inverse document frequency

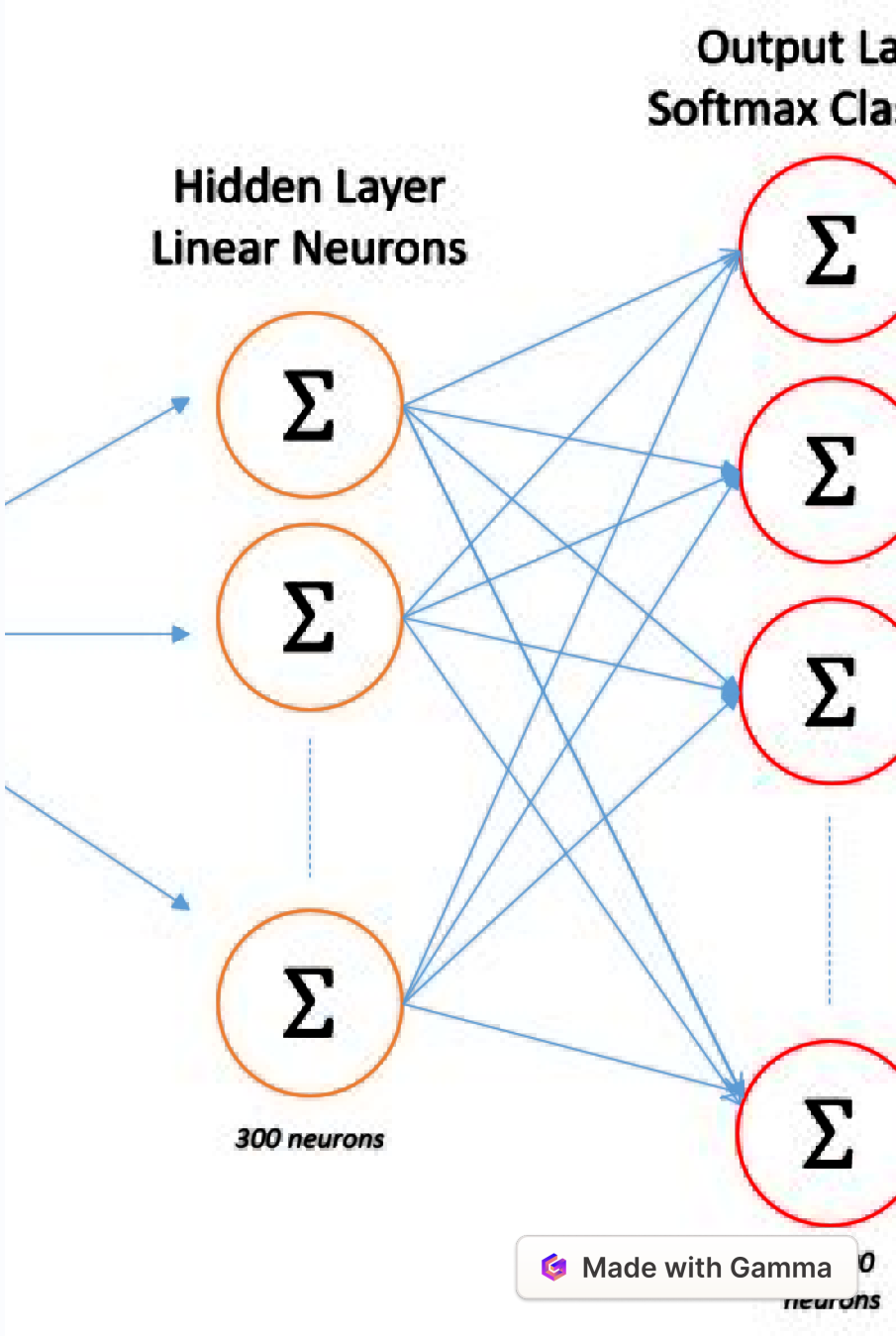
Limitations

Doesn't capture word semantics, relies on frequencies

Word2Vec

Enter the revolutionary technique of Word2Vec! Introduced in 2013 by Mikolov et al., Word2Vec uses neural networks to learn distributed representations of words based on their contexts. It solves the limitations of earlier techniques by capturing semantic relationships, context, and word similarity in a continuous vector space.

Let's explore the power of Word2Vec and its advancements in NLP tasks.



Summary and Beyond

We have journeyed through various text representation techniques, from N-grams to Word2Vec. Each technique has its own strengths and limitations. As a curious college student, you now have a solid foundation in understanding the world of word embeddings and tokenization.

So, are you ready to explore even more exciting NLP concepts? Let's dive deeper into the world of natural language processing and unlock endless possibilities!