Ayudantía 12

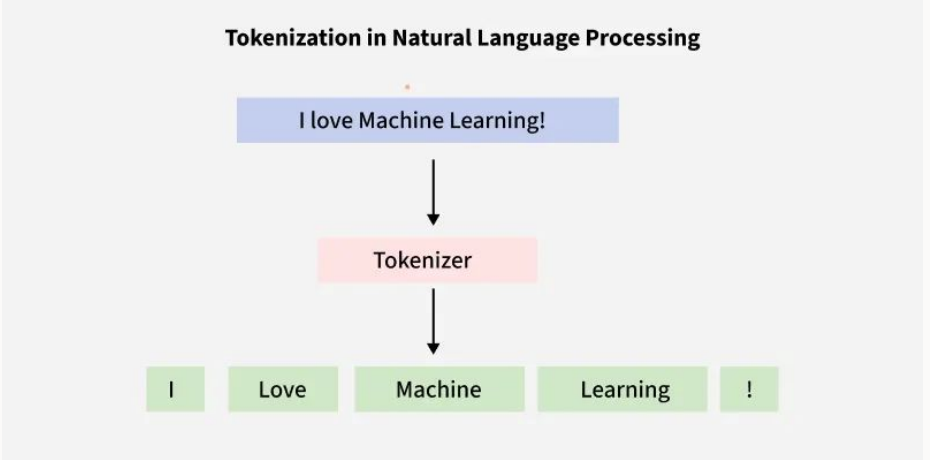
Tokenización

Por Carlos Olguin y Fernando Arévalo

21 de Noviembre 2025

Tokenizar?

- Buscaremos tomar datos en formato texto asociado a un lenguaje (NLP).
- El tokenizar nos dará una herramienta para poder



Fuente : https://www.geeksforgeeks.org/nlp/nlp-how-tokenizing-text-sentence-words-works/

Algunas definiciones

Definición

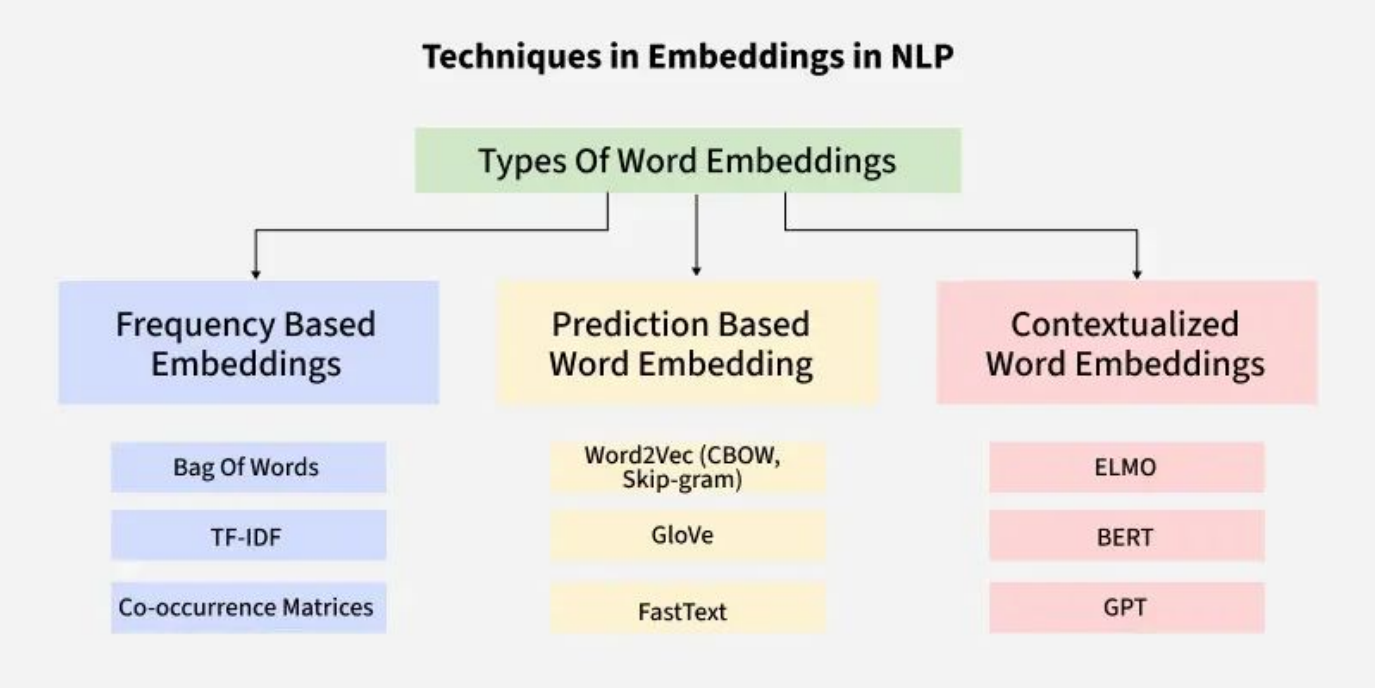
Tokenización Es el proceso de dividir un texto (corpus) en unidades más pequeñas llamadas "tokens" (palabras, caracteres o subpalabras). Es el primer paso para convertir texto en números.

Algunas definiciones

Definición

Embedding La representación vectorial (numérica) de esos tokens donde palabras similares tienen números cercanos

De esta manera



Fuente: https://www.geeksforgeeks.org/nlp/word-embeddings-in-nlp/

Motivación

- Los computadores no entienden texto, solo entienden números.
- Para procesar lenguaje natural (**NLP**!), necesitamos traducir nuestro lenguaje humano a un formato matemático que la red neuronal pueda procesar (Matrices/Tensores).
- Esto será ocupado en nuestro modelos de manera efectiva.
- Cada posible tokenización tiene posibles modificaciones, cambios y ajustes posibles! (Rellenar caracteres, ocupar la frecuencia, tamaño de ventanas, etc).

Tipos de Tokenización

Tokenización por Palabras (Word Tokenization)

La separación del corpus se hace palabra a palabra.

Ventajas:

- Es fácil de entender e implementar.
- Funciona bien con idiomas con separación clara entre palabras como el español y el inglés.

Desventajas:

- Tiene el problema Out Of Vocabulary
- Vocabulario más extenso implica un modelo más pesado. (¡el inglés tiene más de 500.000 palabras!)

Tipos de Tokenización

Tipos de Tokenización

Tokenización por Caracteres (Char-based)

La separación del corpus se hace por caracteres.

Ventajas:

- No existe el problema Out Of Vocabulary.
- Funciona bien para tareas como corrección de ortografía.
- El vocabulario es más ligero.

Desventajas:

- En lenguajes como el español y el inglés, cada token tiene muy poco significado por sí mismo.
- El modelo que usemos tendrá que procesar una cantidad muy grande de tokens.

Tipos de Tokenización

Tipos de Tokenización

Tokenización por Subpalabras (Subword/BPE)

Los tokens son subdivisiones de las palabras del corpus.

Ventajas:

- Reduce el problema Out Of Vocabulary.
- Permite representar palabras más raras o complejas.
- Funciona especialmente bien en lenguas aglutinantes

Desventajas:

- Implementación más compleja.

Tipos de Tokenización

En Resumen...

- Para poder utilizar texto en una red neuronal, se tiene que convertir a vectores de números que lo representen, es decir, crear **embeddings**.
- Para hacerlo, se debe dividir el texto en **tokens** por medio de la **tokenización**.
- Existen distintos tipos de tokenización que separan el texto de formas diferentes:
  - Por palabras
  - Por caracteres
  - Por subpalabras
  - Y varios más...

Vamos a Colab