AYI NEDJIMI Consultants

Tokenization vs Embedding : Différenc

29 April 2026 • Mis à jour le 29 April 2026 • 48 min de lecture

Tokenization vs embedding expliqué : BPE, WordPiece, SentencePiece, tik
Cohere, BGE. Impact RAG, context window, pricing.

La tokenization et l'embedding constituent les deux piliers fondamentaux sur lesqu
artificielle moderne appliquee au langage naturel. Chaque mot que vous tapez dan
soumettez a un moteur de recherche semantique, chaque document que vous ind
inevitablement par ces deux etapes de transformation. Pourtant, la confusion entr
les praticiens de l'IA. La tokenization decoupe le texte brut en unites discretes que
transforme ces unites en vecteurs numeriques dans un espace multidimensionnel
similarite semantique. Comprendre cette distinction, ainsi que la relation profonde
indispensable pour optimiser les systemes de RAG (Retrieval-Augmented Generat
gerer efficacement les fenetres de contexte et construire des pipelines NLP perfo
algorithmes de tokenization (BPE, WordPiece, SentencePiece, Unigram), les mode
le comptage de tokens avec tiktoken, l'impact sur le RAG et la tarification, et les d
multilingue.

À RETENIR

A retenir : La tokenization est le decoupage du texte en unites (tokens). L'emb
tokens ou de sequences de tokens en vecteurs numeriques. La tokenization p
pipeline. Le choix du tokenizer affecte directement la qualite des embeddings

Qu'est-ce que la tokenization et pourquoi est-elle necessaire ?

Les modeles de langage, qu'ils soient des LLM generatifs ou des modeles d'embe
Un ordinateur ne comprend que des nombres. La tokenization est le processus qu
sequence d'entiers, chacun correspondant a une entree dans un vocabulaire pred
pipeline de traitement du langage naturel.

L'histoire de la tokenization en NLP a connu plusieurs approches successives. La t
le texte aux espaces et a la ponctuation. Simple mais problematique : le vocabulai
milliers de mots dans une langue comme le francais), les mots rares ou inconnus n
de OOV, Out-Of-Vocabulary), et les variations morphologiques (manger, mangeon
entites completement differentes.

La tokenization par caracteres (character-level) resout le probleme de vocabulaire
suffisent) mais perd toute notion de sens au niveau du token et produit des seque
pour les modeles Transformer dont la complexite est quadratique en fonction de la

La tokenization en sous-mots (subword tokenization) est le compromis qui a domi
Elle decoupe les mots frequents en tokens complets et les mots rares en sous-uni
exemple, "anticonstitutionnellement" pourrait etre decoupe en ["anti", "constitu", "
un vocabulaire de taille geree (typiquement 32 000 a 128 000 tokens), une couver
preservation partielle de l'information morphologique.