

Embeddings semánticos

Vectores que significan: la base de RAG y de la búsqueda corporativa.

CONCEPTOS CLAVE

- Un embedding es un vector denso de baja dimensión que codifica significado.
- Distancia canónica: $\cos(\vec{u}, \vec{v}) = \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\| \|\vec{v}\|} \in [-1, 1]$.
- Hitos: word2vec (estáticos) → **BERT** (contextuales) → Sentence-BERT (frases) → **E5/BGE** (retrieval SOTA).
- BERT desambigua: "banco.^{en} .^aprueba la hipoteca" tiene vector distinto a "banco del parque".

EJEMPLO A MANO

Distancia coseno a "banco financiero- (0,8, 0,5, 0,3):

Palabra	Vector 3D	Coseno
banco financiero	(0,8, 0,5, 0,3)	1,00
entidad bancaria	(0,7, 0,6, 0,4)	0,99
hipoteca	(0,6, 0,7, 0,4)	0,96
banco de parque	(-0,1, -0,4, 0,9)	0,29
árbol	(-0,5, -0,6, 0,7)	-0,12

La cercanía geométrica refleja cercanía semántica. Las primeras tres podrían recuperarse juntas para una query sobre productos bancarios.

CONCEPTO → APLICACIÓN CON EMBEDDINGS

Modelo	Caso	Salida
word2vec / GloVe	Buscador interno simple.	Palabra → vector estático (300d).
BERT base	Clasificación, NER.	Frase → N vectores (768d).
Sentence-BERT	Búsqueda semántica de párrafos.	Párrafo → 1 vector (768d).
E5/BGE large	RAG corporativo (Día 15).	Doc o pregunta → 1 vector (1024d).
CLIP	Búsqueda imagen-texto.	Imagen o texto → vector compartido.

Idea clave del Día 08. En RAG y búsqueda empresarial: distancia coseno + normalización a norma 1. Es la combinación canónica. Modelos E5 multilingual son SOTA y descargables gratis.

Para llevarte: lectura Reimers & Gurevych, *Sentence-BERT*; práctica: buscador semántico sobre 30 párrafos corporativos con sentence-transformers.