

Transformación de texto en vectores para analizar la similitud semántica

□ ¿Qué es esto de «embedding»?

Una **incrustación (embedding)** es una forma de convertir texto en números, o más concretamente, en vectores. Esto permite comparar textos por **similitud semántica**. Por ejemplo, «gato» y «felino» son parecidos en significado, y sus vectores estarán cerca entre sí.

□ Ejemplo paso a paso con datos

Supón que estás trabajando con una tabla de datos que tiene una columna llamada `summary`, que contiene resúmenes de artículos:

id	summary
1	El cambio climático afecta al planeta.
2	Las emisiones de CO ₂ causan calentamiento.
3	Messi marca un gol en el último minuto.

□ Paso 3: Realizas una transformación

1. **Realizas una transformación** (puede estar en una plataforma como OpenAI, Databricks, etc.).
2. **Le dices que tome la columna `summary`** para convertir el texto en vectores.

3. **Seleccionas el modelo de embedding**, esto es un modelo entrenado para crear esas representaciones vectoriales.

□ ¿Qué ocurre internamente?

Cada resumen se convierte en un vector. Algo así (simplificado):

id	summary	incrustar
1	El cambio climático afecta al planeta.	[0.67, 0.12, 0.45...]
2	Las emisiones de CO ₂ causan calentamiento.	[0.66, 0.10, 0.47...]
3	Messi marca un gol en el último minuto.	[0.13, 0.80, 0.24...]

Los resúmenes 1 y 2 tienen vectores parecidos porque hablan de temas similares (medioambiente, cambio climático). El resumen 3, que habla de fútbol, tiene un vector muy distinto.

□ ¿Para qué sirve esto?

Después puedes usar estos vectores para:

- **Buscar similitudes:** ¿qué resúmenes hablan de lo mismo?
- **Agrupar por temas automáticamente (clustering).**
- **Buscar el resumen más parecido a uno nuevo.**
- **Clasificar textos en categorías temáticas.**