Embeddings: La Representación Vectorial del Significado

¿Qué son los Embeddings?

Un **embedding** es una representación numérica (vector) de una pieza de información, como texto, imágenes o sonido. En el contexto de los LLMs, los embeddings convierten palabras, frases o documentos enteros en vectores de números de punto flotante.

El objetivo es que las palabras con significados similares tengan representaciones vectoriales cercanas en un espacio multidimensional.

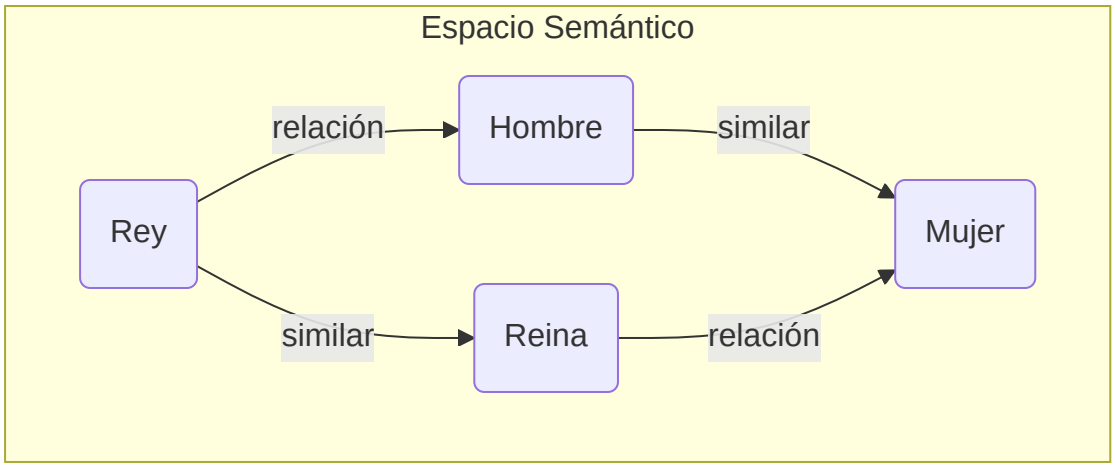
¿Por qué son importantes?

Los modelos de machine learning, incluidos los LLMs, no pueden procesar texto en su forma original. Necesitan números.

- **Capturan el Significado Semántico**: Los embeddings capturan las relaciones contextuales y semánticas entre las palabras.
- **Reducción de la Dimensionalidad**: Convierten datos de alta dimensionalidad (como el vocabulario) en un espacio de menor dimensionalidad.
- **Permiten Cálculos**: Una vez que el texto se convierte en vectores, podemos realizar operaciones matemáticas sobre ellos para encontrar similitudes, diferencias, etc.

Visualizando el Espacio de Embeddings

Imagina un espacio donde las palabras se organizan según su significado. Palabras como "rey" y "reina" estarían cerca, al igual que "caminar" y "correr".



La famosa relación `vector('Rey') - vector('Hombre') + vector('Mujer')` resulta en un vector muy cercano a `vector('Reina')`.

Proceso de Creación de Embeddings

Los embeddings se aprenden durante el proceso de entrenamiento de un modelo de lenguaje.

1. **Tokenización**: El texto se divide en tokens.
2. **Inicialización**: A cada token se le asigna un vector aleatorio.
3. **Entrenamiento**: El modelo ajusta estos vectores a medida que aprende a predecir palabras en su contexto. Las palabras que aparecen en contextos similares terminan con vectores similares.

Aplicaciones de los Embeddings

- **Búsqueda Semántica**: Encontrar documentos o frases que son conceptualmente similares, no solo que coinciden en palabras clave.
- **Clasificación de Texto**: Categorizar textos según su contenido.
- **Agrupamiento (Clustering)**: Agrupar documentos similares.
- **Sistemas de Recomendación**: Recomendar ítems (películas, productos) a los usuarios.
- **Base para RAG**: Son el pilar fundamental del componente de recuperación en los sistemas RAG.