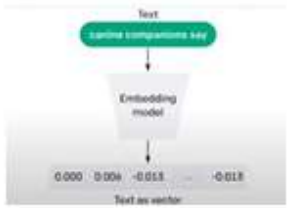
Archivos necesarios
(PDF, DOCX,TXT,ETC)

Procesamiento
(transformación a vectores)

Base de datos
(vectorDB)



Procesamiento
(transformación a vectores)

consulta

(Pregunta del usuario)

respuesta

LLM

Motor de GPT

Explicación

1. Archivos necesarios (PDF, DOCX, TXT, etc.)

•**Entrada**: El flujo comienza con la carga de archivos como PDFs, DOCX o archivos de texto (TXT) que contienen la información que se quiere procesar.

•**Contenido**: Estos documentos contienen información textual relevante que se quiere utilizar para consultas o análisis futuros en este caso el Reglamento Académico, Calendario Académico con fechas muy importantes para ti, Código de Ética, Política Integral de la Ley 21369 y el Reglamento de convivencia y responsabilidad disciplinaria. Esto también se podría usar con cualquier docuento.

2. Procesamiento (transformación a vectores)

•**Conversión**: Los archivos cargados se procesan y convierten en **vectores**. El proceso de vectorización implica tomar el contenido textual y transformarlo en una representación numérica que capta el significado de las palabras o frases en un espacio vectorial.

•**Embeddings**: Los embeddings de texto son generados por un modelo de embeddings como el que se mostró en la imagen anterior, donde las frases o palabras son representadas por secuencias de números.

3. Base de datos (VectorDB)

•**Almacenamiento**: Los vectores resultantes de la etapa de procesamiento se almacenan en una base de datos especializada llamada **VectorDB**. Esta base de datos está optimizada para almacenar y consultar información en forma de vectores, permitiendo una recuperación eficiente basada en similitudes.

•**Acceso rápido**: Gracias a que los textos están almacenados como vectores, se puede hacer búsquedas rápidas por similitud entre vectores, lo que permite encontrar información relacionada de manera eficiente.

4. Consulta

- **Petición**: Un usuario o sistema hace una consulta, que puede ser una pregunta o una búsqueda de información relacionada.
- **Vectorización de la consulta**: La consulta también se transforma en un vector mediante un proceso similar al aplicado a los documentos originales. Esto garantiza que la consulta esté en el mismo espacio vectorial que los documentos almacenados.

5. Procesamiento (transformación a vectores)

- **Comparación**: La consulta vectorizada se compara con los vectores almacenados en la base de datos (VectorDB). La base de datos busca los vectores más cercanos a la consulta basándose en alguna métrica de similitud, como la distancia coseno o euclidiana.
- **Recuperación**: Se seleccionan los documentos o fragmentos más relevantes basados en esta similitud.

6. Respuesta

- **Resultados**: Los resultados de la búsqueda o consulta son presentados en forma de texto o información relevante.

7. LLM (Motor de GPT)

- **Generación de respuesta**: Si se usa un modelo de lenguaje grande (LLM), como GPT, este puede generar una respuesta basada en la información recuperada. El LLM puede usar su capacidad de generar texto para crear respuestas más detalladas o naturales, mejorando la comprensión de los resultados.
- **8. Entrega de la respuesta**
- **Salida**: Finalmente, se genera la respuesta completa y optimizada, ya sea simplemente recuperando documentos o incluso generando texto con un modelo LLM que refuerza la calidad y naturalidad de la respuesta.