Manual de Usuario: Sistema de Recuperación Aumentada (RAG) con Texto a Voz (TTS)

Bienvenido al sistema RAG, una herramienta poderosa para realizar búsquedas semánticas y resúmenes sobre su propia documentación. Esta guía le ayudará a configurar sus archivos y a utilizar todas las funciones de la aplicación.

1. Preparación y Carga de Documentos (El Proceso de Indexación)

Para que el sistema RAG pueda responder preguntas sobre sus documentos, estos deben ser leídos, divididos y convertidos en un formato numérico (vectores). Este proceso se llama **Indexación**.

1.1. Estructura de Carpetas

La aplicación está diseñada para buscar documentos en una carpeta específica y categorizarlos automáticamente por **Tema**.

1. **Directorio Principal:** Localice la carpeta raíz del proyecto.
2. **Carpeta de Datos (./libros/):** Todos sus documentos PDF deben colocarse dentro del directorio llamado libros/.
3. **Temas (Categorización):** Para filtrar su búsqueda por temas, **debe crear subcarpetas** dentro de libros/ con el nombre del tema que desea usar como filtro.

Ejemplo de Estructura Correcta:

```
.
├── libros/
│   ├── Finanzas_Personales/   <-- Este es un Tema/Filtro
│   │   └── guia_inversion.pdf
│   └── Historia_Latinoamericana/ <-- Este es otro Tema/Filtro
│       └── ensayo_independencia.pdf
└── ...
```

Importante: La aplicación solo acepta archivos en formato **PDF**.

1.2. Ejecución del Script de Indexación

Una vez que haya organizado sus archivos PDF en la carpeta libros/, debe ejecutar el script de

indexación:

1. Abra la terminal o línea de comandos en la carpeta raíz del proyecto.
2. Ejecute el siguiente comando de Python:
   python index_data.py

3. **Proceso:** El script:
   - Detectará los nuevos archivos PDF.
   - Procesará cada archivo bajo las diferentes estrategias de segmentación (chunking).
   - Creará o actualizará las bases de datos vectoriales (chroma_db_rag_...).
   - Registrará los archivos procesados en el archivo processed_files.txt.

Debe ejecutar este script cada vez que añada, elimine o modifique documentos en la carpeta ./libros/.

2. Guía de Uso de la Aplicación (Interfaz Streamlit)

La aplicación se utiliza a través de una interfaz de chat en Streamlit.

2.1. Panel Lateral de Configuración (Sidebar)

Utilice la barra lateral izquierda para configurar su sesión de consulta.

1. **Selección de Estrategia de Segmentación:**
   - **¿Qué es?:** Define cómo se dividió el texto del documento para la búsqueda.
   - **Recomendación:** La opción **Recursive (Recomendada)** suele ofrecer los mejores resultados de calidad, ya que intenta preservar el contexto lógico de los párrafos.
   - *Nota: Cambiar esta opción recarga la base de datos vectorial completa.*
2. **Filtrar documentos por tema:**
   - **¿Qué es?:** Permite restringir la búsqueda solo a los archivos contenidos dentro de la subcarpeta seleccionada en ./libros/.
   - **Opción TODOS:** Busca en la totalidad de los documentos indexados, sin restricción de tema.
3. **Seleccionar Modo de Operación:**
   - **RAG (Búsqueda Documental):** Es el modo principal. Utiliza su pregunta para buscar fragmentos de texto relevantes en la base de datos (Chroma) y los pasa al LLM (Mistral) para que genere una respuesta basada en sus documentos.
   - **Resumen (Summary):** El LLM genera un resumen del texto que usted ingresa en la caja de chat, sin consultar los documentos indexados.
4. **Diagnóstico del Sistema:**
   - Revise el **Estado de la Base de Datos (Chroma)** y el **Estado de conexión LLM** para asegurarse de que ambos muestren ✅ **Listo**. Si hay un error, el chat no funcionará.

2.2. Interacción en el Chat

1. **Formular la Pregunta:** Escriba su consulta en la caja de texto inferior y presione Enter.

    - **Modo RAG:** Pregunte como si estuviera hablando con un experto sobre sus documentos (ej: "¿Cuáles son los tres pasos clave para invertir, según el documento 'guia_inversion.pdf'?").
    - **Modo Resumen:** Pegue el texto que desea resumir.
2. **Visualización de la Respuesta:** La respuesta del LLM aparecerá en el chat.
    - **Fuentes:** Si está en **Modo RAG**, la respuesta incluirá una caja de información con las **Fuentes encontradas**, mostrando el nombre de los archivos de donde se extrajo la información.

3. Uso de la Función Text-to-Speech (TTS)

La aplicación incluye un botón para escuchar las respuestas generadas por el LLM.

1. **Botón de Reproducción:** Justo debajo de cada respuesta del asistente, encontrará un botón con un icono de altavoz.
2. **Primer Clic:** Presione el botón. Aparecerá un mensaje de **Cargando...** mientras el sistema envía el texto de la respuesta al servicio **Gemini TTS API**.
3. **Reproducción:** Una vez que el audio se genera, comenzará a reproducirse automáticamente a través de los altavoces de su dispositivo.
4. **Segundo Clic (Detener):** Si presiona el botón mientras el audio está en reproducción o cargándose, la reproducción se detendrá.

> **Requisito:** Para que el TTS funcione, el sistema requiere que la **API Key de Google** esté configurada correctamente en el entorno de ejecución. Si el botón no aparece, la clave no está disponible.