Etapa 1: Carga y Estructura del Conocimiento

La Etapa 1 tiene como objetivo organizar y estructurar toda la información relevante del programa de Disciplina Socioemocional para que pueda ser utilizada eficazmente por el chatbot. Esta etapa es crítica, ya que de ella depende la calidad de las respuestas generadas por el modelo. A continuación, se detallan los pasos necesarios.

# 1. Recolección de materiales fuente

Recolectar todos los documentos proporcionados por la Secretaría de Educación, tales como:  
- Manuales del programa  
- Guías de implementación  
- Fichas temáticas y de actividad  
- Cursos y materiales audiovisuales (videos, infografías)  
- Protocolos de atención a riesgo y bienestar  
- Directorio de instituciones de apoyo

# 2. Conversión a formatos estructurados

Convertir los materiales a formatos estructurados y legibles por modelos de NLP, como:  
- Markdown (.md)  
- Texto plano (.txt)  
- HTML  
- PDF indexado (con OCR si es necesario)  
- CSV o JSON en caso de datos tabulares  
Esto se puede automatizar con herramientas como PyMuPDF, pdfminer, o herramientas de OCR (Tesseract) para imágenes.

# 3. Limpieza y segmentación del contenido

Una vez extraído el texto, se debe limpiar y segmentar en bloques semánticamente coherentes:  
- Párrafos bien formateados  
- Títulos y subtítulos jerarquizados  
- Metadatos útiles (tema, fuente, tipo de documento)  
- Identificación de contenido sensible o confidencial  
Este paso es clave para la indexación posterior en la base de conocimiento.

# 4. Embeddings y base de datos semántica

Indexar los bloques de contenido usando modelos de embeddings (ej. OpenAI embeddings, SBERT, BETO).  
Almacenar los vectores en una base de datos semántica como FAISS, Weaviate o Chroma.  
Esto permite que el chatbot recupere fragmentos relevantes ante cada consulta del usuario.

# 5. Validación de contenido cargado

Revisar junto con expertos en disciplina socioemocional que los contenidos estén completos, actualizados y correctamente interpretados.  
Este paso asegura que el chatbot opere bajo criterios éticos y pedagógicos sólidos.