|  |
| --- |
|  |
| **Implementación de un Chatbot de Servicios Legales en el área de Medio Ambiente en Ecuador**   * Tania Gualli * Víctor Montaluisa |
| Quito, 06 de marzo de 2025 |

Tabla de contenido

[1 DEFINICIÓN DEL PROBLEMA 1](#_Toc192120654)

[2 RECOLECCIÓN DE DATOS 1](#_Toc192120655)

[3 PREPROCESAMIENTO DE DATOS 2](#_Toc192120656)

[4 ANÁLISIS EXPLORATORIO DE DATOS 2](#_Toc192120657)

[5 SELECCIÓN Y ENTRENAMIENTO DEL MODELO 2](#_Toc192120658)

[6 EVALUACIÓN DEL MODELO 3](#_Toc192120659)

[7 OPTIMIZACIÓN Y AJUSTE 3](#_Toc192120660)

[8 DESPLIEGUE Y PUESTA EN PRODUCCIÓN 3](#_Toc192120661)

[9 MONITOREO Y MANTENIMIENTO 3](#_Toc192120662)

[10 CONCLUSIONES Y RECOMENDACIONES 4](#_Toc192120663)

# 

# DEFINICIÓN DEL PROBLEMA

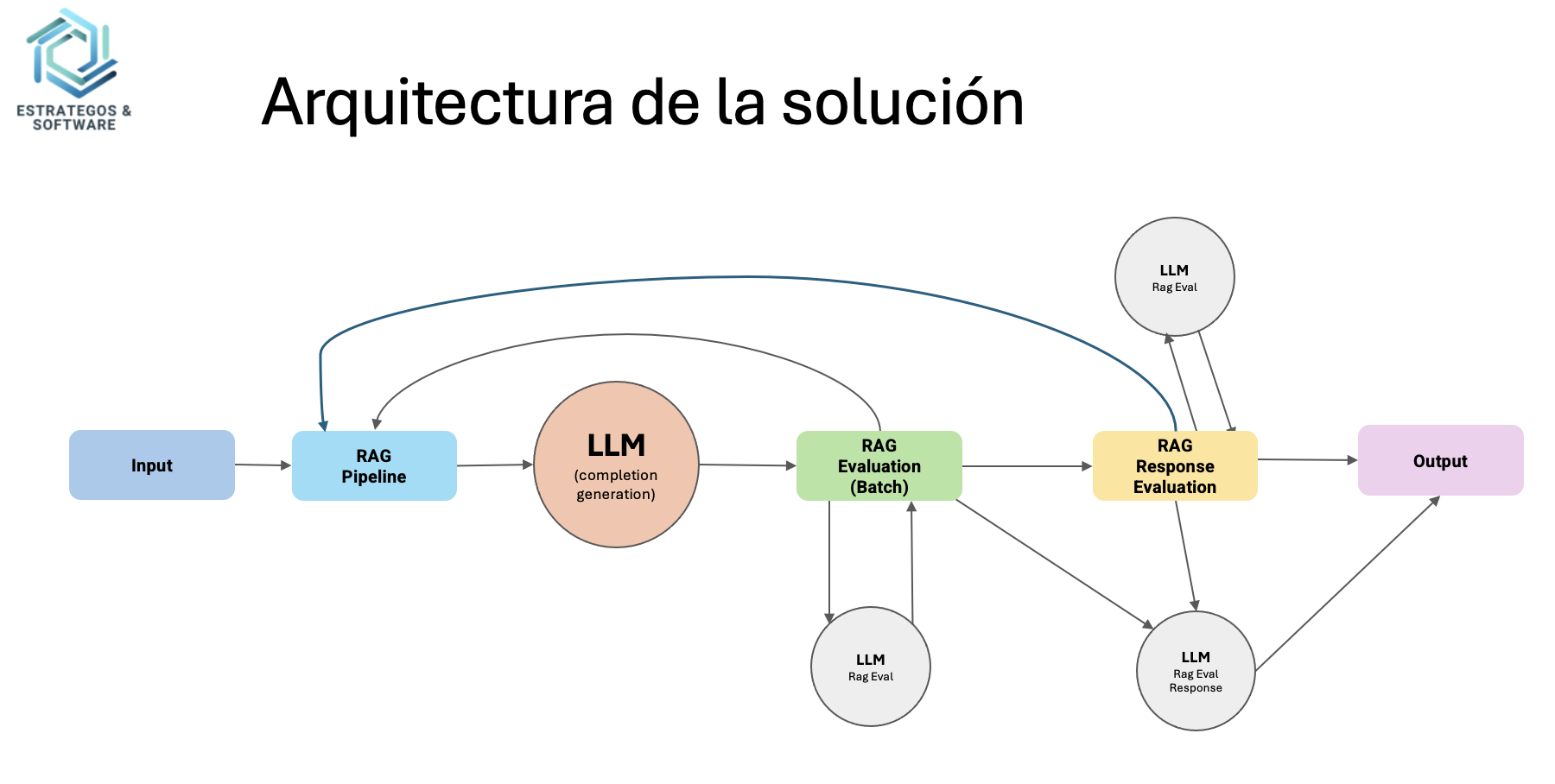
El acceso a información sobre normativas ambientales en Ecuador puede ser complicado para ciudadanos y empresas debido a la dispersión y volumen de los documentos legales. Se ha desarrollado un **chatbot de servicios legales especializado en medio ambiente** que responde a preguntas frecuentes basadas en documentos legales indexados en una base de datos vectorial.

Se plantearon los siguientes objetivos:

* Responder consultas sobre normativas ambientales de forma rápida y precisa.
* Integrar tecnologías de búsqueda semántica e inteligencia artificial.
* Estar disponible en un sitio web para facilitar el acceso a la información.

# ARQUITECTURA DE LA SOLUCIÓN

De manera general la Arquitectura de la solución presentada es la siguiente:



## flujo de trabajo estándar de RAG

El flujo de trabajo estándar de RAG consta de **cuatro etapas principales**:

### ****Preprocesamiento y Almacenamiento de Documentos (Indexación)****

* Se recopilan documentos o información de referencia, en este caso documentos legales en formato pdf.
* Los documentos se dividen en fragmentos manejables.
* Se generan representaciones numéricas (**embeddings**) para cada fragmento usando un modelo de lenguaje como hiiamsid/sentence\_similarity\_spanish\_es.
* Los embeddings se almacenan en una **base de datos vectorial** como **FAISS, ChromaDB o Weaviate** para permitir búsquedas rápidas basadas en similitud semántica.

### ****Recuperación de Información (Retrieval)****

* Cuando un usuario hace una consulta, se genera su embedding.
* La base de datos vectorial busca los fragmentos más similares a la consulta.
* Se devuelve el contexto más relevante encontrado en los documentos indexados.

### ****Generación de Respuesta (Augmented Generation)****

* Se combina la consulta del usuario con los fragmentos recuperados.
* Se envía esta información como entrada a un modelo de lenguaje avanzado (como GPT-4, Llama 2, Mistral, etc.).
* El modelo genera una respuesta basada en el contexto relevante.

### ****Evaluación y Mejora del Modelo****

* Se miden métricas como precisión, coherencia y relevancia.
* Se ajustan los parámetros del retriever (por ejemplo, número de documentos recuperados k).
* Se optimizan los embeddings o se refinan los documentos indexados.

## ****Evaluación de RAG en Batch (RAG Batch Evaluation)****

Evaluar un sistema RAG en batch significa probarlo en un conjunto grande de consultas en lugar de una por una. Esto permite analizar su rendimiento de manera sistemática y mejorar sus resultados.

### ****¿Cómo se realiza una evaluación en batch?****

1. **Definir un conjunto de pruebas**
   * Se prepara una lista de consultas típicas que los usuarios podrían hacer.
   * Se definen respuestas esperadas, basadas en documentos reales.
2. **Ejecutar las consultas en el sistema RAG**
   * Se procesan todas las consultas de forma automatizada.
   * Se guardan las respuestas generadas por el modelo.
3. **Comparar respuestas generadas con respuestas esperadas**
   * Se usan métricas de evaluación como:
     + **Exactitud (Accuracy):** ¿El modelo responde correctamente?
     + **Recall@K:** ¿La respuesta relevante está en los k fragmentos recuperados?
     + **BLEU/ROUGE:** ¿La respuesta generada es similar a una respuesta esperada?
4. **Analizar los resultados y hacer ajustes**
   * Si las respuestas son irrelevantes, se pueden mejorar los embeddings o la base de datos vectorial.
   * Si las respuestas generadas son incorrectas, se pueden ajustar los parámetros del modelo generativo.

# RECOLECCIÓN DE DATOS

Los datos utilizados provienen de documentos legales oficiales en formato PDF, incluyendo:

* Leyes ambientales publicadas en el Registro Oficial de Ecuador.
* Normativas municipales y provinciales relacionadas con gestión ambiental.
* Regulaciones de residuos y recursos naturales emitidas por el Ministerio del Ambiente.

# PRE-PROCESAMIENTO DE DATOS

El procesamiento de los documentos legales se realizó en el script 01\_generacion\_indices.py e incluyó:

* **Extracción de texto**: Uso de PyMuPDF para extraer texto de PDFs editables y pytesseract para OCR en documentos escaneados.
* **Limpieza de texto**: Eliminación de caracteres especiales, normalización y lematización con spaCy.
* **Segmentación en fragmentos**: División en párrafos procesables.
* **Generación de embeddings**: Uso del modelo hiiamsid/sentence\_similarity\_spanish\_es para representar el texto en vectores numéricos.
* **Almacenamiento** en FAISS y ChromaDB para recuperación eficiente.

Posteriormente para la generación de preguntas y respuestas se realizó la carga de datos en Langchain, con la base de datos ChromaDB que contiene los documentos procesados en la fase anterior.

# ANÁLISIS EXPLORATORIO DE DATOS

Se analizaron los documentos para:

* Evaluar la calidad y relevancia del contenido extraído.
* Identificar términos clave y sinónimos en legislación ambiental.
* Determinar estructuras de texto que facilitaran la segmentación y búsqueda.

# SELECCIÓN Y ENTRENAMIENTO DEL MODELO

El chatbot utiliza una combinación de búsqueda semántica y generación de respuestas:

* **Embeddings** generados con hiiamsid/sentence\_similarity\_spanish\_es.
* **FAISS** para consultas rápidas basadas en similitud de texto.
* **ChromaDB** para recuperación semántica optimizada.
* **Modelo de lenguaje llmware/dragon-deci-7b-v0** para respuestas generadas con contexto legal relevante. Se realizó la creación de preguntas y respuestas con esta LLM que está afinada especialmente para tareas de RAG.

# EVALUACIÓN DEL MODELO

Para la evaluación del modelo de preguntas y respuestas se realizó:

* Creación de datos sintéticos para la evaluación RAG.
* Uso de Ragas para una evaluación completa del rendimiento del modelo RAG a través de varias métricas.

En 03\_q&a\_evaluacion.py, se implementaron pruebas para medir:

* **Precisión de respuestas** comparadas con documentos legales originales.
* **Relevancia de documentos recuperados** mediante métricas de similitud.
* **Tiempo de respuesta** en consultas reales.

# OPTIMIZACIÓN Y AJUSTE

Se realizaron mejoras en el rendimiento del sistema:

* Ajuste de embeddings **para mejorar la similitud de búsqueda.**
* Filtrado de textos irrelevantes **para evitar ruido en los resultados.**
* Organización de metadatos **para mejorar la precisión de búsqueda.**

# DESPLIEGUE Y PUESTA EN PRODUCCIÓN

El chatbot se desplegó usando **FastAPI** (02\_consulta\_indice.py):

* **Endpoints principales:**
  + /buscar/faiss: Búsqueda en FAISS.
  + /buscar/chromadb: Búsqueda semántica en ChromaDB.
* **Integración con un sitio web** para consultas en línea.

# MONITOREO Y MANTENIMIENTO

Se implementó un sistema de monitoreo con logs para:

* Registrar consultas y mejorar la base de conocimiento.
* Actualizar documentos legales en la base de datos.
* Optimizar respuestas mediante almacenamiento en caché.

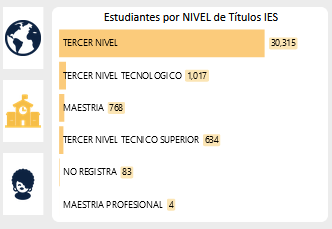


Figura 1: Nueva implementación en visualización-títulos

# CONCLUSIONES Y RECOMENDACIONES

El desarrollo del chatbot de servicios legales en medio ambiente ha demostrado ser una solución eficaz para mejorar el acceso a la información normativa en Ecuador. Su integración con tecnologías de inteligencia artificial y bases de datos vectoriales ha permitido ofrecer respuestas rápidas y precisas a consultas legales.

Entre los principales logros se encuentran:

* **Optimización de la búsqueda semántica**, reduciendo el tiempo de respuesta y aumentando la precisión de los resultados.
* **Accesibilidad mejorada**, al permitir a ciudadanos y empresas obtener información legal sin necesidad de asesoramiento especializado.
* **Arquitectura escalable**, que permite la integración de nuevas fuentes de datos y la actualización continua del contenido legal.

Para mejorar el sistema en el futuro, se recomienda:

* **Ampliar la base de documentos indexados**, incorporando regulaciones más específicas y actualizaciones periódicas.
* **Integrar modelos más avanzados de NLP**, como GPT-4 o mistral, para mejorar la comprensión del lenguaje natural.
* **Optimizar la eficiencia del almacenamiento en caché** mediante Redis u otras tecnologías de almacenamiento rápido.
* **Implementar una interfaz más intuitiva** para mejorar la experiencia del usuario y facilitar la navegación en la base de conocimientos.

En conclusión, el chatbot representa un avance significativo en la digitalización del acceso a normativas ambientales en Ecuador y sienta las bases para futuras implementaciones en otras áreas del derecho.