Informe sobre el Chatbot del Río Paraná

Objetivo

El objetivo del proyecto es crear un chatbot capaz de responder de manera precisa y útil preguntas relacionadas con el Río Paraná, incluyendo consultas sobre alturas del río, distancias entre ciudades a lo largo del río, y otros aspectos relevantes. Para lograr esto, se utilizan diferentes fuentes de datos y técnicas de procesamiento de lenguaje natural.

Desarrollo

1. Instalaciones e importaciones necesarias

Se instalan y se importan las bibliotecas y paquetes necesarios para el desarrollo del chatbot. Esto incluye bibliotecas para procesamiento de texto, manipulación de datos, aprendizaje automático, graficación y acceso a recursos externos como bases de datos y modelos de lenguaje.

2. Base de Datos Vectorial (ChromaDB)

Se utiliza ChromaDB para almacenar y consultar información sobre documentos relacionados con el Río Paraná. Estos documentos fueron obtenidos de fuentes públicas de internet en formato PDF, por lo tanto se necesita procesarlos y limpiarlos para una correcta utilización. Se implementa una función para agregar documentos a la base de datos y se realizan consultas para recuperar información relevante.

3. Embeddings

Se emplean modelos de embeddings para representar textos de manera vectorial. Se utilizan modelos pre-entrenados para calcular embeddings de textos relacionados con alturas del río, distancias entre ciudades y otros temas relevantes.

4. Base de Datos Tabulares

Se recopilan y procesan datos tabulares relacionados con alturas del río en diferentes ciudades a lo largo del Río Paraná. Esta información fue obtenida de fuentes oficiales. Se procesan los archivos obtenidos para unificar y ordenar los datos en un sólo dataframe que luego será convertido a cadena de texto.

5. Base de Datos de Grafos

Se construye un grafo dirigido para representar relaciones entre ciudades a lo largo del Río Paraná, incluyendo información demográfica y distancias entre ciudades. Se visualizan dichos grafos para corroborar su correcta creación y luego se convierte la información a cadena de texto.

6. Clasificador

Se entrena un clasificador para predecir la categoría de las preguntas y seleccionar la fuente de datos más adecuada para proporcionar respuestas. Se utilizan modelos de aprendizaje automático (Multinomial Naive Bayes) para clasificar las consultas en categorías como alturas del río, distancias entre ciudades y otros temas. Luego se entrena el modelo para comprobar su rendimiento mediante métricas como ‘accuracy’.

7. Retriever

Se implementa un mecanismo de recuperación para obtener información relevante en función de la pregunta formulada y de la clasificación previamente realizada. Se realizan consultas sobre las bases de datos disponibles y se seleccionan los documentos o datos más relevantes para responder la pregunta.

8. RAG (Retrieve and Generate)

Se utiliza un modelo de generación de texto (Hugging Face) para generar respuestas basadas en el contexto recuperado y la pregunta formulada. Se implementa un sistema de diálogo que utiliza la información contextual para generar respuestas coherentes y relevantes.

Conclusiones

El chatbot del Río Paraná es una herramienta útil para obtener información rápida y precisa sobre diversos aspectos relacionados con el río. Su capacidad para recuperar y generar respuestas basadas en diferentes fuentes de datos lo hace versátil y efectivo en la atención de consultas relacionadas con el tema.

Futuras Mejoras

Para mejorar el chatbot del Río Paraná, se podrían considerar las siguientes mejoras:

- Ampliar y mejorar las fuentes de datos disponibles.
- Refinar los modelos de procesamiento de lenguaje natural para mejorar la precisión de las respuestas.
- Implementar un sistema de retroalimentación para mejorar continuamente la calidad de las respuestas.
- Incorporar capacidades de aprendizaje automático para adaptarse y mejorar con el tiempo.

Referencias

Se han utilizado diversas bibliotecas y recursos externos en el desarrollo del chatbot del Río Paraná, incluyendo bibliotecas de Python como NLTK, spaCy y scikit-learn, así como modelos de lenguaje pre-entrenados como Sentence Transformers y modelos de generación de texto de Hugging Face.