UNIDAD DE INGENIERÍA EN COMPUTACIÓN
Centro Central
Inteligencia Artificial
Profesor: Kenneth Obando Rodríguez

TEC | Tecnológico de Costa Rica

Proyecto 2 - Investigación (15%) Fecha de Entrega: 9 de junio, 2024

Proyecto 2: Sistema RAG

1 Introducción

La inteligencia artificial ha revolucionado la manera en que interactuamos con la tecnología. En este proyecto, implementamos un chatbot utilizando avanzadas técnicas de procesamiento de lenguaje natural, específicamente mediante el uso de modelos generativos preentrenados y la técnica de RAG para proporcionar respuestas informadas y contextuales.

2 Objetivos

- Implementar un chatbot usando el modelo preentrenado llama3 con RAG.
- Realizar fine-tuning del modelo con un conjunto de documentos específicos para mejorar la precisión de las respuestas.
- Analizar y comparar el rendimiento del chatbot bajo diferentes configuraciones y prompts.

3 Descripción del Proyecto

Los estudiantes desarrollarán un chatbot utilizando el modelo llama3 o algún otro que tengan a bien, diseñado para responder preguntas y participar en diálogos basados en un conjunto de documentos específicos que servirán como contexto. Este proyecto se dividirá en varias fases:

Selección de la base de datos de documentos : Escoger una colección de documentos que servirá de base de conocimiento para el chatbot. Esta base puede consistir en artículos, libros, noticias, etc., que sean relevantes para un dominio específico de interés.

Implementación del modelo RAG : Configurar y adaptar el modelo llama3 para utilizarlo con la técnica de RAG, asegurándose de que pueda recuperar información de la base de datos seleccionada para generar respuestas.

Diseño de experimentos : Realizar pruebas con diversos prompts para evaluar cómo el chatbot responde a diferentes tipos de consultas y situaciones.

Análisis de resultados : Documentar y analizar los resultados obtenidos en los experimentos, evaluando la relevancia, coherencia, y precisión de las respuestas del chatbot. Además, deben investigar sobre las diferentes métricas que permitan comprobar la calidad del modelo.

Reporte final : Redactar un informe detallando el proceso de desarrollo, los experimentos realizados y los resultados obtenidos, incluyendo capturas de pantalla y ejemplos específicos.

4 Sección de Fine-Tuning del Modelo llama3

4.1 Objetivo del Fine-Tuning

Mejorar la capacidad del modelo llama3 para generar respuestas más precisas y contextualizadas, adaptándolo específicamente a las características y necesidades del conjunto de documentos seleccionado y los requisitos del chatbot.

4.2 Pasos para el Fine-Tuning

4.2.1 Preparación de los Datos

- **Selección de Datos:** Escoger un subconjunto representativo de documentos de la base de datos inicial que contenga variabilidad y riqueza en el lenguaje y temas relevantes.
- **Preprocesamiento:** Limpiar y preparar los textos para asegurar uniformidad en el formato, eliminando caracteres innecesarios o corrigiendo errores de codificación.

4.2.2 Creación del Dataset de Entrenamiento

- Dividir el conjunto de datos en entrenamiento, validación y prueba.
- Formatear los datos para el entrenamiento, asegurando que cada entrada esté adecuadamente emparejada con las respuestas esperadas o etiquetas correspondientes.

4.2.3 Configuración del Entorno de Entrenamiento

- Utilizar herramientas y librerías como Hugging Face Transformers y PyTorch.
- Establecer parámetros de entrenamiento, como la tasa de aprendizaje, tamaño de lote, número de épocas, entre otros.

4.2.4 Ejecución del Fine-Tuning

- Cargar el modelo preentrenado llama3.
- Aplicar el entrenamiento usando los datos preparados, ajustando el modelo a las particularidades del contexto definido por la base de datos.

- Monitorizar el desempeño del modelo durante el entrenamiento utilizando el conjunto de validación para ajustar los parámetros y evitar el sobreajuste.

4.2.5 Evaluación del Modelo Afinado

- Utilizar el conjunto de pruebas para evaluar la precisión, coherencia, y relevancia de las respuestas generadas por el modelo afinado.
- Comparar el rendimiento antes y después del afinamiento para medir las mejoras.

4.2.6 Integración con el Sistema de RAG

- Integrar el modelo afinado en el sistema de chatbot RAG.
- Realizar ajustes adicionales si es necesario para asegurar que la recuperación de información y la generación de respuestas funcionen de manera óptima.

4.3 Buenas Prácticas y Consideraciones

- **Iteración Continua:** El fine-tuning es un proceso iterativo. Es posible que se necesite ajustar los parámetros múltiples veces basándose en los resultados obtenidos.
- **Uso de Recursos:** Tener en cuenta los recursos computacionales disponibles, ya que el entrenamiento de modelos de lenguaje puede requerir una alta capacidad de procesamiento y memoria.
- **Ética y Bias:** Ser consciente de la posibilidad de introducir sesgos en el modelo durante el fine-tuning y tomar medidas para mitigar estos efectos.

5 Entregables

- Código fuente completo del chatbot, incluyendo scripts para la configuración del modelo RAG y la interfaz de usuario.
- Base de datos de documentos utilizada.
- Informe técnico que incluya descripción del proyecto, metodología, resultados de los experimentos, y análisis de los mismos, documente el proceso de fine-tuning, incluyendo detalles técnicos como configuraciones de entrenamiento, cambios realizados y evaluaciones de rendimiento.

6 Evaluación

Criterio	Excelente	Bueno	Aceptable	Insuficiente
Implementación técnica	Completa integración del modelo RAG con adaptaciones avanzadas y optimización del código.	Integración adecuada del modelo RAG con algunas adaptaciones.	Implementación básica del modelo RAG sin modificaciones significativas.	Implementación incompleta o incorrecta del modelo RAG.
Calidad de la base de datos	Base de datos extensa y altamente relevante al contexto del chatbot.	Base de datos adecuada con relevancia moderada.	Base de datos mínima o con poca relevancia.	Base de datos inadecuada o irrelevante.
Experimentación y análisis	Análisis exhaustivo y crítico con múltiples prompts y situaciones. Innovación en la aplicación y en los métodos de prueba.	Buen análisis con variedad de prompts, pero con alcance limitado en la experimentación.	Análisis básico con pocos prompts. Experimentación limitada.	Falta de análisis significativo. Experimentación insuficiente o inexistente.
Calidad del reporte	Reporte excepcionalmente bien escrito, organizado, con análisis detallado y ejemplos claros.	Reporte bien escrito y organizado con algunos análisis y ejemplos.	Reporte con la información necesaria pero organización o detalle deficientes.	Reporte incompleto, desorganizado o con múltiples errores.
Evaluación del Fine-Tuning	Fine-tuning que mejora significativamente la precisión y relevancia de las respuestas del chatbot.	Fine-tuning efectivo con mejoras notables en la respuesta del chatbot.	Fine-tuning realizado con mejoras mínimas en la respuesta del chatbot.	Fine-tuning ineficaz o no realizado.

Cuadro 1
Rúbrica de Evaluación del Proyecto de Chatbot con RAG y llama3