Diplomado en Ciencia de Datos e Inteligencia Artificial

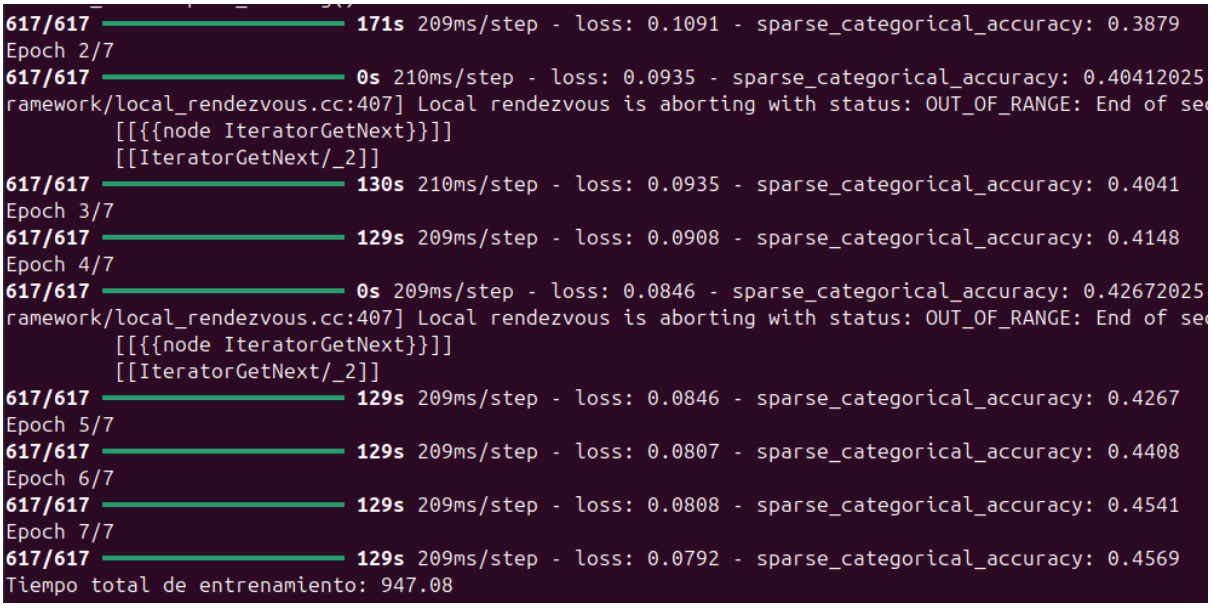
Proyecto: Chatbot CSE

Alumno: Mauricio Miranda Baron

Se decidió crear un chatbot para la coordinación de servicios escolares de la Universidad Autónoma Metropolitana Unidad Iztapalapa encargada de realizar distintos tipos de trámites, proporcionar información relevante sobre la licenciatura, maestría y posgrado. Anteriormente se contaba con un chatbot que actualmente no está en funcionamiento, el chatbot está basado en reglas y fue creado con aproximadamente 500 preguntas con archivos y documentación que fue otorgada.
El objetivo era realizar un procedimiento de fine tuning con el cual se buscaba darle contexto acerca de los servicios que proporciona la coordinación y preguntas frecuentes. Por el costo computacional que llevaba realizar fine tuning completo se decidió implementar una técnica llamada LoRA o Low Rank Adaptation que permite ajustar un modelo de lenguaje sin tener que hacerlo de forma completa, computacionalmente es mucho más eficiente y mejor en casos donde se cuenta con pocos datos.
Utilice como el ejemplo de keras de generative deep learning llamado “GPT2 Text Generation With KerasHub” y de cómo aplicar la tecnica LoRA el ejemplo de keras llamado “Parameter-efficient fine-tuning of GPT-2 with LoRA”,.
Decidí utilizar el modelo gemma3, creado por google, de los modelos pre entrenados que ofrece keras. Específicamente el modelo *gemma3_instruct_1b,* este modelo es multilenguaje y es especializado para tareas como un chatbot. Los datos que recibe este modelo por su preprocesador son del tipo pregunta y respuesta.

Toda la información fue recabada mediante la página oficial de la coordinación (https://www.cseuami.org/), de la cual se obtuvieron preguntas, respuestas y algunas otras se crearon con la información encontrada, dando un total de 617 preguntas. Para obtener la información principalmente se tenía contemplado hacer web scraping con selenium pero por la estructura de la página web (dinámica) fue más difícil hacerlo así y se optó por hacerlo de forma manual.

Para el entrenamiento me apoye de este ejemplo proporcionado por google para el entrenamiento de su modelo gemma en el cual muestran como hacer fine tuning con la tecnica LoRA https://ai.google.dev/gemma/docs/core/lora_tuning?hl=es-419.
Se pudo entrenar el modelo en una GPU RTX A5000 con 22GB de VRAM, aunque también era posible entrenarlo con google colab en la versión gratuita pero tardando más en el entrenamiento usando la GPU T4.



Como resultados se tuvo que el modelo no genera buenas respuestas muy probablemente por el número de datos con el que se cuenta. Una alternativa es utilizar un técnica llamada RAG o Retrieval-Augmented Generation en donde se combina el uso de un LLM y una base de datos. Esta base de datos incluirá toda la información de los servicios que proporciona la coordinación así como algunos documentos importantes, proporcionando aún más información al sistema que solo preguntas y respuestas. De esta manera el modelo podría crear mejor información.

Al finalizar el entrenamiento la idea igual era crear una api para poder probar el modelo e integrarlo de forma fácil con cualquier sistema que lo necesitara. La API se creó con FastAPI un framework para crear apis de forma fácil con python.