Laboratorio – Large Language Models
CC3092 - Deep Learning y Sistemas Inteligentes

De manera individual,

Utilice el código compartido en la actividad para correr el modelo TinyLlama de manera local. (Recomendación: Si por cualquier motiva encuentra problemas corriendo el código en su computadora, utilice un Notebook de Google Colab, el código fue probado en este ambiente).

Practica:

1. Haga inferencia del prompt definido en el script utilizando el modelo: *TinyLlama/TinyLlama_v1.1*
2. Repita el proceso con el modelo: *Doctor-Shotgun/TinyLlama-1.1B-32k-Instruct*
3. Repita el proceso con algún modelo distinto de su elección. Solo tenga cuidado al momento de seleccionar el modelo por temas de tamaño o recursos que pueda necesitar.
4. Ejecute el mismo prompt en ChatGPT.
5. Compare las respuestas de cada modelo. ¿En qué se parecen? ¿En qué son distintas?

Teoría:

Utilice las páginas de todos modelos en HuggingFace.com para compararlos y explicar las diferencias en rendimiento. Enfóquese en los siguientes puntos:

a. Tiempo necesario para que el modelo responda.
b. Recursos (memoria, procesamiento, tiempo, parámetros) consumidos por cada modelo.
c. Calidad de la respuesta.