Actividad 2. Ingeniería de *prompts* con modelos en la nube a través de Mistral AI

**Objetivos**

En esta actividad vas a trabajar con el LLM (*large language model*) Mistral 7B, creado por la compañía **Mistral AI**. Un LLM es un modelo capaz de generar texto en lenguaje natural en respuesta a un *prompt*. El *notebook* que vas a utilizar fue presentado por primera vez en el *workshop* AIM219 en el principal evento de Amazon Web Services, re:Invent.

Por otro lado, vas a ejecutar este *notebook* en AWS SageMaker Studio Lab, el entorno de experimentación que proporciona Amazon para probar su plataforma de inteligencia artificial, SageMaker. Al utilizar un entorno de *notebooks* respaldado por la infraestructura de computación de Amazon, tendrás acceso a una capacidad de computación superior a la de un PC convencional.

Adicionalmente, vas a importar el modelo Mistral 7B desde Hugging Face, el mayor repositorio de modelos de IA en Internet.

El objetivo del *notebook* es ilustrar varias características del LLM y describir técnicas de *prompting*. La ingeniería de *prompts* es una actividad clave para sacar el máximo partido a estos modelos, y es una habilidad muy demandada en el ámbito de la IA generativa. En esta actividad vas a aprender las principales técnicas de *prompting,* experimentando con ellas en un entorno de computación en la nube.

**Pautas de elaboración**

Para la elaboración de esta actividad es necesario haber completado la actividad 1 de la asignatura, dado que se reutilizará el entorno de SageMaker Studio Lab y el repositorio de ejemplos clonado en el entorno de trabajo.

La realización de esta actividad comprende las siguientes actividades:

1. Inicio de sesión en SageMaker Studio Lab.
2. Análisis del código fuente del *notebook.*
3. Ejecución del *notebook.*
4. Cumplimentación de la batería de preguntas incluidas al final de esta sección.

A continuación, se detalla cada uno de estos pasos:

**Inicio de sesión en Sage Maker Studio Lab**

Para ello, utilizarás las credenciales creadas en la actividad 1. Tienes que iniciar una nueva sesión del *runtime,* haciendo clic en «Start runtime».

**Análisis del código fuente**

Abre el *notebook* titulado prompting-mistral7B.ipynb. Este *notebook* se encuentra en la siguiente ruta:

/studio-lab-examples/generative-ai/mistral

El *kernel* que debes utilizar para ejecutar este *notebook* es sagemaker-distribution:Python*.*

Lee atentamente las instrucciones del *notebook*.

**Ejecución del *notebook***

Ahora debes ejecutar de forma secuencial cada celda del *notebook*. Es importante que compruebes en cada paso que no se produce ningún error.

La primera celda instala varios paquetes con dependencias. Es posible que tengas que reiniciar el *kernel* después haberla ejecutada. De otro modo, pueden aparecer errores en las celdas sucesivas.

**Lista de preguntas**

Responde a cada pregunta que se formula a continuación. Cuando se pida una explicación, debes proporcionar al menos un párrafo en la respuesta.

1. **Sección *Libraries*.** ¿Cuál es la ventaja de utilizar el módulo ctransformers sobre modelos creados con el módulo transformers? ¿Por qué es necesaria esta ventaja especialmente en el caso de un LLM?
2. **Sección *Loading the model*.** AutoModelForCausalLM es una clase de Hugging Face que representa un modelo genérico. ¿Con qué dos métodos puede ser instanciada? ¿Cuál es la diferencia entre ambos?
3. **Sección *Let’s try a basic prompt*.** ¿Por qué se imprime la respuesta al prompt a través de un bucle for? ¿Para qué sirve el parámetro max\_new\_tokens? ¿Qué es un token?
4. **Sección *Instruction fine-tuning*.** ¿Por qué utilizamos la llamada a la función PromptTemplate.from\_template() en lugar de especificar el prompt directamente? ¿Qué significa [INST] [/INST] en este modelo?
5. **Sección *Few-shot prompts*.** ¿Cuál es la ventaja principal de la técnica de *few-shot prompts* respecto al proceso convencional de entrenamiento? ¿Con qué clase de Langchain podemos implementar el *few-shot prompting*?
6. **Sección *Augmenting LLM response with context*.** En el modelo Mistral 7B, ¿qué ocurre cuando la respuesta a una pregunta no se puede derivar del contexto proporcionado? Proporciona la evidencia en el *notebook*.
7. **Sección *Hallucinations*.** En términos generales, ¿qué es una *hallucination* en un LLM? ¿Qué dos métodos pueden usarse para evitarlas?
8. **Sección *Providing chat context*.** ¿Para qué se utiliza la variable resp en el *notebook*? ¿Es eficaz el uso de contexto en la llamada context\_template.format(context=resp, question=question)? ¿Por qué?
9. **Sección *More examples of LLM use cases*.** ¿Qué instrucciones se pueden proporcionar a Mistral 7B para evitar que se imprima información sensible de carácter personal en la respuesta? Extrae la respuesta del *notebook*.
10. ¿Qué significa el símbolo “<|im\_end|>”? ¿Para qué se utiliza en el código fuente del *notebook*?

**Extensión y formato**

El entregable de la actividad debe incluir cada una de las preguntas de la sección anterior con su correspondiente respuesta. Salvo en los casos en los que se solicite un dato puntual, la respuesta debe incluir al menos **un párrafo** con la explicación solicitada. Adicionalmente, se debe incluir como adjunto un *export* del *notebook* ejecutado, **en formato HTML**.

La extensión **máxima** del documento es de cuatro páginas.

Rúbrica

|  |  |  |  |
| --- | --- | --- | --- |
| Título de la actividad | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Criterio 1 | Entrega del *export* con todas las celdas del *notebook* ejecutadas de forma secuencial y sin errores en la salida. | 2 | 20 % |
| Criterio 2 | Entrega de una respuesta correcta a cada una de las 10 preguntas formuladas. | 8 | 80 % |
|  |  | **10** | **100 %** |