Material descargable

Inteligencia Artificial para la Productividad

egg

Material complementario: El concepto de Token los Modelos de Lenguaje

El procesamiento por **tokens** es clave en el funcionamiento de los modelos de lenguaje como GPT-4. Un token es una unidad básica de texto que puede ser una palabra completa, una parte de una palabra, o incluso un signo de puntuación. Los modelos no procesan texto como un conjunto continuo de palabras, sino que dividen el texto en estos tokens, que son la base para generar respuestas.

Impacto en la Precisión

1. **Palabras largas o inusuales**: Cuando una palabra larga o menos común se divide en varios tokens, el modelo puede tener una representación más compleja de esa palabra. Si se pierden o malinterpretan fragmentos al tokenizarla, esto puede afectar la precisión de la respuesta.
   Ejemplo: La palabra "extraordinario" se procesa como un token en muchos casos, pero si se dividiera de manera incorrecta, el modelo podría perder el contexto o significado.
2. **Relación entre tokens**: Las relaciones entre los tokens determinan cómo el modelo comprende y genera texto. Si la tokenización divide erróneamente una palabra o frase clave, el modelo podría malinterpretar la intención original y producir una respuesta incorrecta o inexacta.

Malinterpretaciones y “Alucinaciones”

Las **alucinaciones** (respuestas generadas que no son precisas o tienen información fabricada) pueden surgir en parte debido a cómo el modelo procesa los tokens. Aquí algunas razones:

1. **Pérdida de contexto**: Si una secuencia de tokens no se procesa correctamente o si hay una división ineficiente del texto, el modelo puede perder el contexto global. Esto puede llevar a respuestas que parecen correctas pero que no tienen fundamento en la pregunta original.
2. **Fragmentación de información**: Cuando el texto se fragmenta en tokens, especialmente con términos técnicos o menos comunes, el modelo podría generar respuestas que combinan de manera incorrecta ideas o conceptos de diferentes tokens. Esto puede dar

lugar a "alucinaciones" donde el modelo construye algo que suena plausible pero que no tiene una base real.
Ejemplo: Si un término técnico se divide incorrectamente en varios tokens, el modelo puede generar información que no está vinculada a la realidad, mezclando conceptos erróneos.

3. **Sesgo en los tokens generados**: El modelo se entrena para predecir el siguiente token basado en los anteriores, lo que significa que la calidad de la tokenización afecta el flujo del texto generado. Si se produce una secuencia inesperada o mal dividida, el modelo puede seguir generando tokens que refuercen un error o malentendido, ampliando la posibilidad de generar alucinaciones.

Ejemplo de cómo las alucinaciones pueden surgir por tokenización:

Supongamos que el modelo recibe el término "criomicropropagación" (un término técnico relacionado con la biología). Si este término se tokeniza incorrectamente, el modelo podría interpretar y asociar solo parte de la palabra con otro concepto, lo que podría llevar a una alucinación como:

- Pregunta: ¿Qué es la criomicropropagación?
- Respuesta alucinada: *Es un proceso de congelación de microchips para mejorar su rendimiento* (aunque en realidad se refiere a la propagación de tejidos vegetales en condiciones controladas).

En resumen, la **tokenización** es crucial porque el modelo genera texto token por token. Si la división de los tokens no preserva adecuadamente el significado del texto original, se pueden generar malinterpretaciones, y el modelo puede generar respuestas con información incorrecta o inventada (alucinaciones).

¿Para qué sirve contar tokens?

El número de tokens es relevante principalmente por las siguientes razones:

1. **Límite de tokens**: Los modelos como GPT-4 tienen un límite máximo de tokens que pueden procesar en una sola consulta. Si superas ese límite, el modelo cortará parte de la entrada o la respuesta. Contar los tokens te permite asegurarte de que tu consulta no sea tan larga como para exceder el límite.
2. **Costos en modelos de pago**: En algunos servicios de IA, el costo de las consultas se basa en la cantidad de tokens que usas. Saber cuántos tokens tiene tu consulta te ayuda a optimizar el uso del modelo para ahorrar recursos.
3. **Procesamiento de texto largo**: Si trabajas con textos extensos, conocer el número de tokens te ayuda a dividir el contenido en partes más pequeñas, garantizando que el modelo procese toda la información correctamente.

¿Contar los tokens ayuda a evitar alucinaciones?

Las alucinaciones de los modelos no están directamente relacionadas con el número de tokens, pero contar los tokens sí puede **indirectamente** ayudar a reducir ciertos problemas de respuesta.

Las "alucinaciones" ocurren cuando el modelo genera información incorrecta o inventada, y esto puede suceder por diferentes motivos:

1. **Consultas demasiado largas**: Si envías consultas muy largas que están cerca del límite de tokens, el modelo podría tener que "resumir" la información para ajustarse a la cantidad permitida, lo que aumenta el riesgo de generar respuestas inexactas o incompletas.
2. **Falta de contexto**: Si una consulta excede el límite de tokens, el modelo puede cortar parte de la entrada, lo que provoca que pierda parte del contexto, llevando a respuestas incorrectas.
3. **Complejidad de la información**: Cuando un modelo debe lidiar con información compleja en pocos tokens, a veces intenta completar detalles de manera errónea, alucinando información para intentar "rellenar los huecos".

Entonces, ¿cómo evitar alucinaciones?

Para evitar alucinaciones, más allá de contar tokens, es importante que:

1. **Formules preguntas claras** y proporciones contexto suficiente, pero sin ser excesivamente verboso. Demasiado texto puede distraer al modelo.
2. **Verifiques la información generada por el modelo** cuando sea importante. Si el tema es crítico (por ejemplo, datos científicos, técnicos, etc.), vale la pena contrastar las respuestas del modelo con fuentes confiables.
3. **Utilices prompts más estructurados** para guiar al modelo hacia respuestas más precisas.

¿Qué hacer con el número de tokens?

1. **Si estás por encima del límite de tokens (para GPT-4, por ejemplo, hasta 8k o 32k tokens dependiendo de la versión), considera reducir la longitud de la entrada** o dividir el texto en partes más pequeñas.
2. **Si estás trabajando en tareas que involucran muchos textos o largas descripciones**, el conteo de tokens te ayudará a optimizar cómo organizas la información para mejorar la coherencia y precisión de las respuestas.

En resumen, contar los tokens es útil para evitar problemas como pérdida de información o costos innecesarios, pero no es la solución directa para evitar alucinaciones. La clave está en formular buenas consultas y ser crítico con las respuestas del modelo.