Costos

GEMINI

El costo de la API de Gemini se basa principalmente en la cantidad de texto que procesas, medido en "tokens". Un token es aproximadamente equivalente a una palabra para el texto en inglés, pero puede variar para otros idiomas. Para entender el costo por registro MARCXML, necesitamos estimar cuántos tokens contendrá cada registro una vez que lo envíes a la API.

Aquí te presento los pasos y consideraciones para hacer esta estimación:

1. Analiza la estructura de tus registros MARCXML:

- **Tamaño promedio de los registros:** ¿Qué tan extensos suelen ser tus registros MARC21? ¿Tienen muchos campos y subcampos?
- **Cantidad de texto en los campos:** Identifica los campos que contienen principalmente texto (por ejemplo, títulos, notas, encabezamientos de materia, etc.) y estima la cantidad promedio de texto en ellos.
- **Elementos XML:** Si bien los elementos XML en sí mismos no se consideran texto para el conteo de tokens de la API, una estructura XML muy repetitiva podría influir indirectamente en la cantidad de información textual que envías.

2. Realiza pruebas con algunos registros de muestra:

La forma más precisa de estimar el costo es tomar una muestra representativa de tus registros MARCXML y realizar algunas pruebas:

- **Convierte a texto plano:** Una forma de aproximar la cantidad de tokens es convertir tus registros MARCXML a texto plano. Puedes hacer esto extrayendo el contenido textual de los diferentes campos MARC. Herramientas de procesamiento XML o incluso scripts sencillos pueden ayudarte con esta tarea.
- **Cuenta las palabras:** Una vez que tengas la representación en texto plano de tus registros de muestra, puedes usar un contador de palabras para tener una idea aproximada de la cantidad de tokens por registro. Recuerda que la conversión de tokens no es exactamente palabra por palabra, pero te dará una buena base.

3. Considera el modelo de Gemini que utilizarás:

El costo por cada 1,000 tokens varía según el modelo de Gemini que elijas. Los modelos más capaces suelen tener un costo más elevado. Consulta la documentación oficial de la API de Gemini para conocer los precios actuales de los diferentes modelos.

4. Calcula el costo por registro:

Una vez que tengas una estimación del número de tokens por registro (basada en tus pruebas) y conozcas el costo por 1,000 tokens del modelo que planeas usar, puedes calcular el costo por registro de la siguiente manera:

```
Costo por registro = (Número de tokens por registro / 1000) * Precio por 1,000 tokens
```

Ejemplo ilustrativo:

Supongamos que, después de analizar tus registros de muestra y convertirlos a texto plano, encuentras que un registro promedio contiene alrededor de 500 tokens. Y supongamos que el modelo de Gemini que planeas usar tiene un precio de $0.001 por cada 1,000 tokens (este es un precio hipotético, verifica los precios reales en la documentación de Google AI).

Entonces, el costo por registro sería:

```
Costo por registro = (500 / 1000) * $0.001 = $0.0005
```

En este ejemplo hipotético, cada registro te costaría $0.0005.

Consideraciones adicionales:

- **Prompting:** La forma en que formules tus prompts a la API de Gemini también puede influir en el número de tokens procesados. Un prompt más extenso resultará en un mayor costo.
- **Respuestas de la API:** Ten en cuenta que la respuesta de la API también generará tokens y tendrá un costo asociado. La longitud de la respuesta dependerá de la tarea específica que le pidas a Gemini (por ejemplo, si solo pides una confirmación o si solicitas información adicional).
- **Volumen total:** Multiplica el costo por registro por la cantidad total de registros que planeas catalogar para obtener una estimación del costo total del proyecto.
- **Monitoreo:** Es importante monitorear tu uso de la API a través de la consola de Google Cloud para tener un control preciso de los costos y ajustar tu estrategia si es necesario.

Recomendaciones:

- **Comienza con un lote pequeño de registros de prueba:** Realiza pruebas con un conjunto reducido de tus registros para obtener una estimación más precisa del costo por registro antes de procesar todo el catálogo.
- **Optimiza tus prompts:** Intenta formular tus prompts de la manera más eficiente posible para minimizar la cantidad de tokens enviados.
- **Considera diferentes modelos de Gemini:** Evalúa si un modelo menos costoso podría ser suficiente para tu tarea de catalogación, especialmente si no requieres la máxima capacidad del modelo más avanzado.
