Transformers

Trabajo Práctico 2

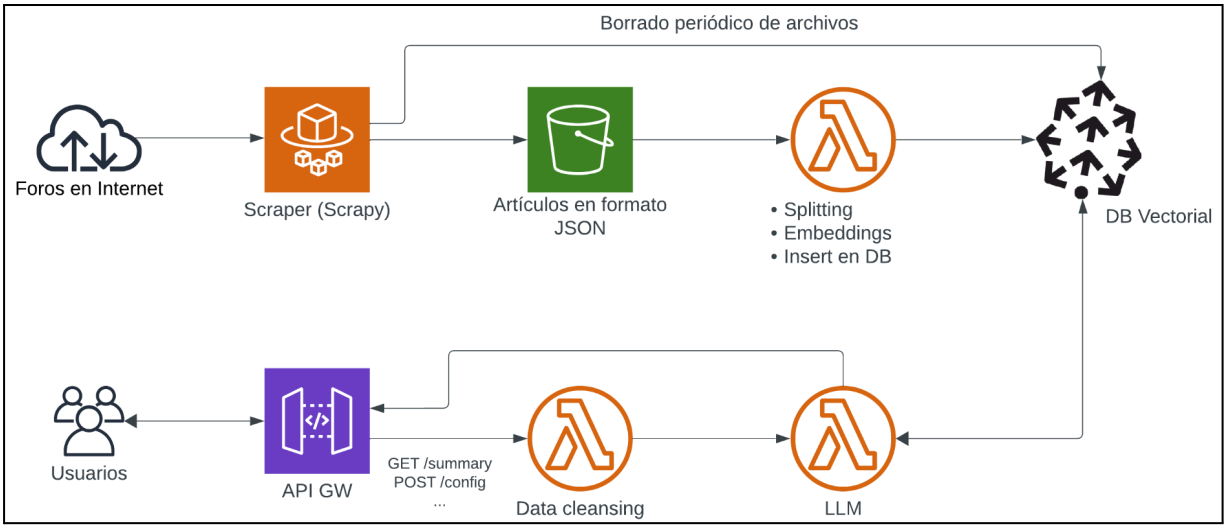
Temas Avanzados en Deep Learning

Grupo 4:

- Nicolás Birsa
- Salvador Castagnino
- Valentín Ye Li

Puesta en producción

Pensando en una arquitectura de tipo serverless usando los servicios de AWS, nuestra arquitectura en producción podría ser como la siguiente:



Donde se tiene como consideraciones:

- El uso de un Fargate para el scraper, debido a la posibilidad de necesitar manejar grandes cantidades de memoria para los documentos, lo que hace ineficiente el uso de una lambda para esto.
- Un cron job dentro del fargate para borrar periódicamente archivos viejos de la db vectorial, a fines de que los resúmenes sean siempre de noticias recientes.

Consideraciones para Responsible AI y Safety

A fines de cumplir con las recomendaciones provistas por organismos internacionales como la UNESCO, nuestro modelo debería tener las siguientes características:

- A la hora de dar un resumen, listar los distintos artículos usados como base, a fines de que los usuarios puedan saber claramente qué se usó como base para la respuesta provista.
- En casos de encontrar algún exploit en el modelo para producir respuestas maliciosas o imparciales, implementar más controles en la etapa de data cleansing. Así el modelo logra una mayor robustez.
- Indicarle al usuario en todo momento el modelo LLM base usado para darle las respuestas para tener una mayor transparencia.
- No pedirle datos al usuario más allá de los necesarios para la operación básica del modelo, que son sus tópicos y autores de interés.