Sistema RAG para la AI ACT

Retrieval-Augmented Generation

CONTENIDO

INTRODUCCIÓN 2

¿QUÉ ES UN RAG Y PORQUÉ ES NECESARIO? 2

JUSTIFICACIÓN DEL USO DE TECNOLOGÍAS 3

ARQUITECTURA DEL SISTEMA 4

FUNCIONALIDAD DEL SISTEMA RAG 5

19 Diciembre 2024

Introducción

El desarrollo de un sistema RAG proporciona una solución robusta y escalable para la consulta de la AI ACT Europea (inglés o español). Esta arquitectura no solo garantiza un acceso rápido y preciso a la información, sino que también ofrece la capacidad de generar respuestas contextualizadas y coherentes, adaptándose a las necesidades específicas del usuario y del dominio. El despliegue de la infraestructura del RAG se ha realizado sobre AWS que nos provee una alta disponibilidad, seguridad y escalabilidad del sistema.

En este documento se detallan las especificaciones y el desarrollo de un sistema de Recuperación y Generación de Respuestas (RAG, por sus siglas en inglés) diseñado para consultar la AI ACT Europea. El sistema ha sido desarrollado utilizando la tecnología Langchain, y se han empleado diversas tecnologías avanzadas como los embeddings de Hugging Face con el modelo "all-mpnet-base-v2", el modelo GPT-4 de OpenAI, y la base de datos vectorial Pinecone. A continuación, se desglosarán los componentes del sistema, su justificación y la arquitectura propuesta para su implementación en AWS.

¿Qué es un RAG y por qué es necesario?

Un sistema de Recuperación y Generación de Respuestas (RAG) combina técnicas de recuperación de información con modelos de generación de texto para proporcionar respuestas precisas y de relevancia contextual a consultas de usuario. Este tipo de sistema es especialmente útil para la consulta de bases de datos de conocimiento, como la AI ACT Europea, por varias razones:

1. **Acceso Rápido y Preciso a la Información**: Un RAG puede buscar rápidamente en grandes volúmenes de datos y extraer la información más relevante, ahorrando tiempo y esfuerzo en comparación con la búsqueda manual.
2. **Generación de Respuestas Contextualizadas**: Utilizando modelos de lenguaje avanzados, un RAG no solo recupera información, sino que también puede generar respuestas que integren varios fragmentos de datos de manera coherente.
3. **Adaptabilidad y Escalabilidad**: Estos sistemas pueden adaptarse fácilmente a diferentes dominios y escalar para manejar grandes cantidades de consultas simultáneas, lo cual es crucial en aplicaciones a nivel europeo.

Justificación del uso de tecnologías

Modelo de Hugging Face: "all-mpnet-base-v2"

El modelo "all-mpnet-base-v2" de Hugging Face se ha seleccionado por su alta precisión en la creación de embeddings de texto. Este modelo es conocido por:

- **Alta Calidad de Embeddings**: Ofrece representaciones vectoriales densas y semánticamente significativas de textos, lo cual es esencial para la recuperación eficiente de información.
- **Eficiencia Computacional**: Está optimizado para ser utilizado en aplicaciones de alta demanda, proporcionando un buen equilibrio entre precisión y rendimiento.
- **Flexibilidad**: Puede ser fácilmente integrado en pipelines de procesamiento de lenguaje natural, como los proporcionados por Langchain.

GPT-4 de OpenAI:

GPT-4 se ha elegido el modelo de generación de texto debido a sus capacidades avanzadas en:

- **Comprensión del Lenguaje Natural**: GPT-4 puede entender y generar texto de manera coherente y contextualmente relevante, lo que es crucial para proporcionar respuestas precisas y útiles.
- **Versatilidad**: Es capaz de manejar una amplia variedad de tareas de procesamiento de lenguaje, desde la generación de respuestas hasta la síntesis de información.
- **Soporte Continuo**: OpenAI proporciona actualizaciones y soporte continuo, asegurando que el modelo se mantenga en la vanguardia de la tecnología de IA.

Pinecone

Pinecone se ha seleccionado como la base de datos vectorial para almacenar los embeddings debido a:

- **Optimización para Datos Vectoriales**: Pinecone está diseñada específicamente para manejar grandes volúmenes de datos vectoriales de manera eficiente.
- **Escalabilidad**: Puede escalar horizontalmente para manejar crecientes volúmenes de datos y consultas.

- **Integración Sencilla**: Ofrece una integración fácil con frameworks de IA y machine learning, permitiendo una implementación fluida en el pipeline de Langchain.

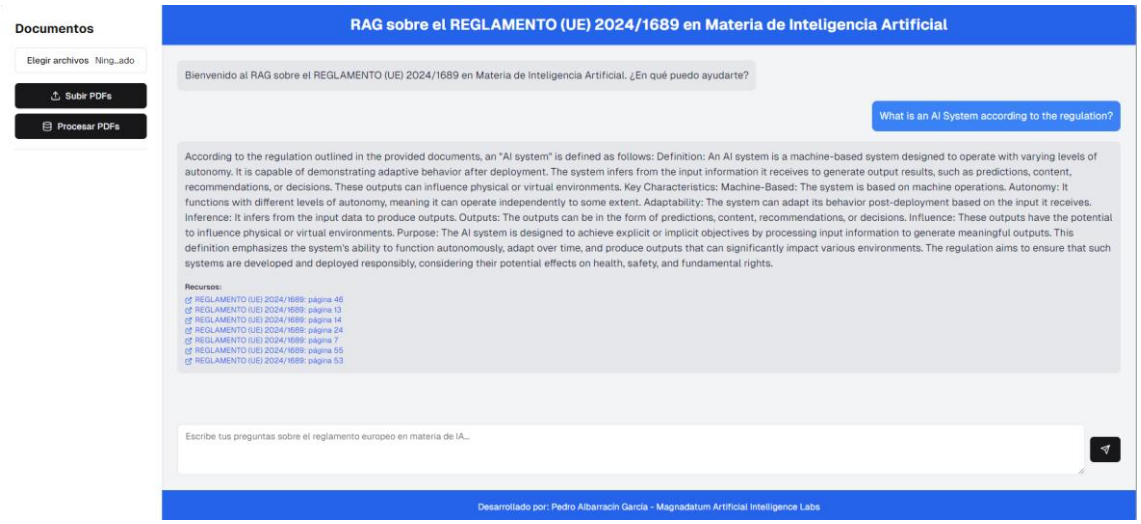
Arquitectura del Sistema

La arquitectura del sistema RAG para consultar la AI ACT Europea está diseñada para ser implementada sobre AWS, aprovechando sus servicios escalables y seguros. A continuación, se describe la arquitectura propuesta:

1. **Capa de Ingesta de Datos**:
    - **AWS S3**: Para el almacenamiento de datos sin procesar y documentos de referencia.
    - **AWS Lambda**: Para procesar y transformar los datos antes de la indexación.
2. **Capa de Procesamiento de Embeddings**:
    - **EC2 Instances**: Máquinas virtuales configuradas para ejecutar el modelo "all-mpnet-base-v2" de Hugging Face, generando embeddings a partir de los datos.
    - **Pinecone**: Almacenamiento de los embeddings generados, facilitando la recuperación rápida.
3. **Capa de Generación y Respuesta**:
    - **GPT-4 de OpenAI**: Acceso al modelo GPT-4 para la generación de respuestas basadas en la información recuperada.
    - **AWS API Gateway y Lambda**: Para gestionar las solicitudes de los usuarios y coordinar las llamadas a los modelos de Hugging Face y GPT-4.
4. **Capa de Frontend y Usuario**:
    - **AWS Amplify**: Para el desarrollo y despliegue de la interfaz de usuario, proporcionando una experiencia interactiva y receptiva.
5. **Capa de Seguridad y Monitoreo**:
    - **AWS IAM**: Gestión de identidades y acceso para asegurar los recursos.

AWS CloudWatch: Monitorización y logging para mantener la operatividad y detectar posibles problemas.

Funcionalidad del Sistema RAG



Este RAG (sistema de Recuperación Aumentada por Generación) está diseñado para facilitar la consulta y comprensión del **Reglamento Europeo de Inteligencia Artificial (AI Act)**, proporcionando una herramienta interactiva y eficiente para acceder a información relevante del reglamento y otros documentos relacionados.

Funcionalidad principal:

1. **Consulta Inteligente del AI Act**:

   El sistema permite a los usuarios realizar preguntas específicas sobre el contenido del reglamento, ofreciendo respuestas detalladas y precisas. Estas respuestas se generan mediante una combinación de procesamiento lingüístico avanzado y la extracción de contenido clave del documento.
2. **Referencias a Documentación Original**:

   Cada respuesta incluye enlaces directos a las secciones relevantes del reglamento u otros documentos. Esto permite a los usuarios verificar la información proporcionada o explorar el contexto completo de la respuesta.

3. **Subida e Indexación de Documentos**:

   Además del AI Act, el sistema permite a los usuarios subir sus propios documentos en formato PDF. Estos documentos se indexan automáticamente para ser integrados en el sistema, ampliando el rango de consultas posibles y personalizando el contenido accesible.
4. **Procesamiento de Archivos PDF**:

   Los usuarios pueden procesar múltiples documentos cargados para que sean analizados en conjunto, lo que permite realizar preguntas complejas que abarquen información de varias fuentes.
5. **Interfaz Amigable y Orientada al Usuario**:

   La interfaz está diseñada para facilitar su uso, con opciones claras como "**Subir PDFs**" y "**Procesar PDFs**", así como un campo para introducir preguntas. El sistema genera respuestas completas que no solo resuelven dudas puntuales, sino que también ayudan a entender conceptos clave.

Beneficios:

- **Ahorro de tiempo**: Los usuarios no necesitan leer el reglamento completo; pueden obtener información específica al instante.
- **Fiabilidad**: Al enlazar directamente con la documentación original, asegura que la información es siempre verificable.
- **Adaptabilidad**: La capacidad de integrar otros documentos permite personalizar el uso del sistema para adaptarlo a necesidades concretas.
- **Accesibilidad**: El diseño del sistema está orientado a facilitar su uso incluso para usuarios sin experiencia técnica o jurídica.

Además, este sistema RAG no solo se limita a responder las preguntas introducidas por el usuario, sino que utiliza técnicas avanzadas para enriquecer el contexto y ofrecer respuestas más precisas y relevantes. A continuación, se explica este proceso en detalle:

Generación Interna de Preguntas para Contextualización

Cuando el usuario introduce una pregunta en el sistema, el modelo realiza un procesamiento adicional que incluye los siguientes pasos:

1. **Análisis de la Pregunta del Usuario**:

   El sistema analiza la pregunta inicial para identificar las palabras clave, el tema central y el posible ámbito de la consulta. Esto incluye detectar si la pregunta es genérica, específica o ambigua.

2. **Generación de Preguntas Internas**:

   Basándose en la pregunta del usuario, el sistema fórmula automáticamente **tres preguntas adicionales** con el objetivo de:

   - **Ampliar el contexto:** Explorar aspectos relacionados que puedan ser relevantes para la respuesta.
   - **Especificar términos:** Desambiguar conceptos en caso de que la pregunta sea vaga o imprecisa.
   - **Cubrir posibles lagunas:** Asegurar que no se omita información importante relacionada con el tema.

Por ejemplo, si el usuario pregunta *"¿Cómo afecta el AI Act a los sistemas autónomos?"*, el modelo podría formular internamente preguntas como:

   - *¿Qué define el AI Act como sistema autónomo?*
   - *¿Qué obligaciones específicas impone el AI Act sobre estos sistemas?*
   - *¿Qué excepciones o flexibilidades ofrece el reglamento para los sistemas autónomos?*

3. **Recopilación de Respuestas Internas**:

   Cada una de estas preguntas internas se consulta dentro del contenido indexado (documentación cargada y reglamento). Esto genera múltiples fragmentos de información que son integrados en un único bloque de respuesta.

4. **Generación de la Respuesta Final**:

   La respuesta final presentada al usuario combina:
   - La información derivada directamente de la pregunta inicial.
   - El contexto adicional obtenido a partir de las preguntas internas.
   - Referencias específicas a las secciones del reglamento o documentos indexados que sustentan la respuesta.

Beneficios de este enfoque:

- **Mayor Precisión**: Al explorar múltiples dimensiones del tema, se garantiza que la respuesta sea lo más completa y relevante posible.
- **Desambiguación Automática**: Si la pregunta original es imprecisa, las preguntas internas ayudan al modelo a interpretar mejor la intención del usuario.
- **Enriquecimiento del Contexto**: El usuario no solo obtiene una respuesta directa, sino también información complementaria que puede ampliar su entendimiento.

Ejemplo Práctico del Proceso:

Supongamos que el usuario pregunta: *"¿Qué responsabilidades tienen los proveedores según el AI Act?"*, el modelo internamente generará preguntas tales como:

- *¿Cuáles son las definiciones clave de proveedor según el reglamento?*
- *¿Qué responsabilidades específicas se aplican a los proveedores durante el ciclo de vida del sistema de IA?*
- *¿Qué sanciones se aplican si no se cumplen estas responsabilidades?*

De esta manera, el sistema no solo responde a la pregunta del usuario, sino que garantiza que la respuesta incluya detalles críticos y sea respaldada por las secciones relevantes del reglamento.

Este enfoque innovador asegura que el usuario obtenga información de alta calidad, optimizando tanto la precisión como la utilidad de las respuestas.

En resumen, este RAG actúa como un asistente virtual experto en el AI Act europeo, ofreciendo una forma interactiva, eficiente y verificable de acceder a la normativa y otra documentación complementaria. Es una herramienta ideal para profesionales legales, técnicos y gestores interesados en entender y aplicar la normativa europea de inteligencia artificial.