RAG (Retrieval Augmented Generation) Aplicado à Ciência de Dados

Olá! Bem-vindo(a) à nossa newsletter semanal sobre Ciência de Dados!

Nesta edição, vamos explorar o conceito de **RAG (Retrieval Augmented Generation)** e como essa tecnologia pode transformar a forma como manipulamos dados e tomamos decisões nos negócios.

Vamos começar pelo começo: **O que é RAG?** RAG (Geração Aumentada por Recuperação) é uma técnica de inteligência artificial que combina o poder de recuperação de informações relevantes em grandes bases de dados com a geração de texto. Basicamente, ele funciona trazendo o melhor de dois mundos: a **precisão** de dados históricos e contextuais e a **criatividade** de modelos de linguagem natural, principalmente LLM ou Modelos de Linguagem Grandes como o GPT. Isso permite que sistemas RAG façam recomendações ou gerem respostas muito mais contextualizadas e detalhadas do que um modelo isolado de IA.

Por que o RAG é importante para a Ciência de Dados e Negócios? A principal vantagem do RAG é que ele oferece **respostas mais precisas** e **contextualizadas** ao extrair informações relevantes de fontes específicas e usá-las para melhorar a geração de texto. Na área de **ciência de dados**, onde o volume de informações pode ser vasto e complexo, o RAG permite extrair insights que vão além do que um modelo tradicional de IA poderia alcançar sozinho. Para as **empresas**, isso representa um salto na qualidade das tomadas de decisão, já que a capacidade de recuperar informações com precisão permite que dados históricos e atuais sejam utilizados de maneira estratégica.

Embora o RAG ofereça muitas vantagens, ele também apresenta alguns desafios:

- **Curadoria de dados**: É essencial que as bases de dados sejam precisas e atualizadas, o que pode exigir grande esforço de manutenção.
- **Treinamento de modelos**: Integrar recuperação de informações com modelos de linguagem natural requer uma combinação de técnicas de IA avançadas.
- **Escalabilidade**: Em grandes volumes de dados, manter a performance do sistema RAG pode se tornar complexo.

Algumas das principais aplicações do RAG incluem: **Automação de Atendimento ao Cliente**: Fornece respostas rápidas e contextualizadas para dúvidas de clientes, baseado em informações recuperadas de bases de conhecimento internas; **Geração de Documentos Jurídicos**: Automação na criação de contratos, baseando-se em exemplos anteriores e gerando novos documentos personalizados; **Análise de Dados**: Extração de insights valiosos ao combinar dados históricos com novas tendências, proporcionando uma visão completa para tomada de decisões; entre outras.

Passo a Passo de Implementação de RAG

Vamos pensar em um exemplo prático: você é um pequeno empresário do ramo imobiliário e quer automatizar o processo de criação de contratos de locação ou compra e venda. O RAG pode ajudar a gerar contratos automaticamente, buscando informações anteriores (cláusulas padrão, termos preferenciais etc.) e adaptando-as a novos clientes. Aqui está um passo a passo simples para implementar o RAG:

1. **Criação de uma base de dados de contratos antigos**: Colete todos os contratos anteriores e armazene-os em um formato acessível (JSON, CSV, banco de dados). Para isso, você pode

utilizar o SQLite ou o PostgreSQL, que são bancos de dados gratuitos e de código aberto, ideais para armazenar e consultar dados de forma eficiente.

2. **Desenvolvimento ou integração com um modelo de linguagem natural**: Utilize um modelo de IA como o GPT, que será responsável por gerar novas versões dos contratos. Existem vários modelos de código aberto e gratuitos, como o Llama, Falcon, Mistral, Phi, Gemma entre outros. Essas ferramentas oferecem modelos de linguagem poderosos, acessíveis e escaláveis para microempresas.

3. **Configuração da camada de recuperação de dados**: Configure um sistema de busca que permita recuperar partes relevantes dos contratos antigos, como cláusulas padrões, exigências de clientes, ou modelos específicos. Você pode utilizar ferramentas de busca de código aberto, como o Elasticsearch ou o Whoosh. Esses sistemas permitem indexar documentos e realizar consultas de maneira eficiente.

4. **Combinação das informações**: Integre o modelo de IA com a camada de recuperação, permitindo que ele gere um novo contrato com base nas informações recuperadas, ajustando os detalhes ao cliente atual. Para a integração, você pode utilizar bibliotecas de código aberto como Haystack, que facilita a implementação de sistemas RAG ao integrar recuperação de dados e geração de texto com IA.

5. **Automatização do processo**: Crie um sistema que permita preencher automaticamente as informações básicas de um novo contrato e permita ao modelo gerar a versão final personalizada. Ferramentas como Flask ou FastAPI podem ser utilizadas para construir interfaces simples de aplicação web que conectam as etapas de recuperação e geração de dados, permitindo ao usuário final interagir com o sistema.

6. **Validação e Ajustes**: Sempre inclua uma etapa de revisão e validação humana antes de finalizar o documento. Para facilitar essa etapa, ferramentas de automação de fluxo de trabalho, como Camunda ou Apache Airflow, podem ser utilizadas para gerenciar as etapas de revisão e aprovação dos contratos, garantindo que nada seja publicado sem a devida supervisão.

O **RAG** está transformando a forma como trabalhamos com dados e informações, permitindo maior precisão e contexto nas respostas geradas. Se você trabalha com grandes volumes de informações ou precisa automatizar a geração de documentos e relatórios complexos, essa pode ser uma solução extremamente poderosa.

Esperamos que você tenha achado esta edição útil. Na próxima, continuaremos explorando ferramentas e tecnologias que podem aprimorar suas habilidades em Ciência de Dados.

Saudações,

Prof. Dr. Dilermando Piva Jr
Coordenador de Ciência de Dados para Negócios / Fatec Votorantim
E-mail: f301.cdn@fatec.sp.gov.br