IA GENERATIVA NA CURADORIA DOS METADADOS DA IDE DA EMBRAPA

DAVI DE OLIVEIRA CUSTÓDIO[1]
JAUDETE DALTIO[1]

[1] EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA - EMBRAPA/CNPM
CAMPINAS - SP
{DAVI.CUSTODIO, JAUDETE.DALTIO}@EMBRAPA.BR

A Empresa Brasileira de Pesquisa Agropecuária (Embrapa) atua com soluções de pesquisa, desenvolvimento e inovação para a sustentabilidade da agricultura e, neste contexto, representa uma importante instituição governamental produtora de dados geoespaciais que subsidiam estudos e políticas públicas do setor agropecuário. A gestão de dados espaciais na Embrapa começou a ser estruturada em 2012, com a implantação de um processo de gestão dos dados apoiado por um repositório e infraestrutura interoperável de publicação e compartilhamento. Este arcabouço, denominado de GeoInfo, foi estruturado a partir do uso de softwares livres (Geonode e Geonetwork) e lançado em 2018. Atualmente, o GeoInfo é uma solução corporativa consolidada na Embrapa, amplamente adotada e essencial no processo de entrega de dados espaciais e ativos cartográficos, resultados dos projetos de pesquisa conduzidos pela instituição. Ele é utilizado tanto para o armazenamento de dados restritos ao público interno como para a publicação de dados ao público externo por meio da INDE. O GeoInfo também é utilizado como parâmetro para monitoramento de adoção destes resultados, visando mensurar seu impacto e alcance. O volume de dados que atualmente compõem a plataforma é considerável. Entre dados públicos e privados, que atualmente estão em processo de migração para uma nova arquitetura tecnológica [1], há cerca de 7.000 planos de informação, cadastrados pelos mais de 20 centros de pesquisa produtores de dados. Por se tratar de uma ferramenta de auto depósito, apesar dos esforços na padronização das orientações de preenchimento dos metadados, observa-se uma expressiva heterogeneidade na execução do processo de catalogação, que impacta diretamente a recuperação e reuso destes dados. O objetivo deste trabalho é apresentar uma das frentes de atuação da Embrapa para endereçar esta questão. Considerando o grande volume de dados e metadados envolvidos, está em desenvolvimento um algoritmo de automação voltado para a curadoria de metadados do GeoInfo. Utilizando ferramentas de IA generativa e modelos de linguagem de grande escala, conhecidos como LLMs (Large Language Models), o algoritmo foi desenvolvido em Python e faz uso de frameworks específicos para interagir com as APIs das LLMs, que são servidas pelo servidor Ollama. O algoritmo visa elencar os planos de informação do acervo que demandam uma revisão para aprimorar a qualidade de seus metadados. A meta é assegurar que esses metadados se mantenham relevantes e consistentes, facilitando a busca, o acesso e a utilização dos dados espaciais disponíveis na plataforma.

Nesta etapa do processo, optou-se por modelos LLM *open source* em detrimento de opções pagas, como GPT-4 ou Google Gemini. Entre os modelos escolhidos para implementação estão Gemma, Llama3, Mistral, OpenChat, Phi3, Mixtral e Wizardlm2. A escolha por LLMs de código aberto deve-se à possibilidade de customizações futuras, utilizando a técnica de Fine Tuning, e ao menor custo de utilização. Embora esses modelos apresentem menos parâmetros ajustados (7 a 14 bilhões) em comparação com modelos maiores, sua execução requer hardwares potentes, incluindo GPUs com considerável quantidade de

memória (VRAM), bem como máquinas com um número mínimo de núcleos de processamento (CPUs) e memória RAM. A necessidade desse tipo de hardware pode ser um desafio,

especialmente em ambientes com recursos limitados. Para os testes utilizando as LLMs abaixo de 8 bilhões de parâmetros, um ambiente com 128GB de memória RAM e 2 placas GPUs de 8GB de memória VRAM atendeu de forma satisfatória aos testes. No entanto, para as demais LLMs (acima de 8 bilhões de parâmetros), o processo de resposta aos prompts enviados se mostrou inviável devido ao tempo e delay no processamento. Esse problema de desempenho evidencia a necessidade de um planejamento cuidadoso e, possivelmente, de investimentos adicionais em infraestrutura para garantir que os modelos mais avançados possam ser utilizados de forma eficiente, resultando em melhores resultados em todo o processo. Em linhas gerais, o processo automatizado seguiu os seguintes passos: utilizando a API do Geonode, o algoritmo extrai todos os metadados da plataforma e gera uma lista de títulos, resumos e palavras-chave, conforme exigido pelo Perfil de Metadados Geoespaciais do Brasil (MGB). Em seguida, cria-se um prompt com esses dados, solicitando ao modelo a verificação gramatical, ortográfica e de coerência semântica entre os atributos, para apontamento de problemas e sugestões de correção e melhoria. Um documento PDF é gerado, agrupando estas sugestões de acordo com o proprietário do dado e, em seguida, envia-se estas sugestões por e-mail. Atualmente, o algoritmo foca apenas na verificação dos campos de título, resumo e palavras-chave e foram empregados apenas para analisar aspectos de gramática, ortografia e coerência semântica entre estes três atributos. A proposta é apresentar um diagnóstico comparativo da eficiência desses modelos na análise dos metadados, em termos de precisão, consistência e capacidade de identificar e sugerir correções. O algoritmo poderá ser estendido para uma análise mais ampla, incluindo mais campos de metadados do perfil ISO19115 e MGB. Esta frente de trabalho é um passo importante para garantir a qualidade dos metadados no Geoinfo a longo prazo, além de ser aplicável a outras Infraestruturas de Dados Espaciais ou outros repositórios de dados de propósito geral. A utilização de modelos de IA generativa *open source* não só promove a inovação e a flexibilidade no desenvolvimento do processo, mas também se alinha com os princípios de transparência e acessibilidade aos dados da uma IDE.

REFERÊNCIAS

[1] MARIA, A. C. M.; DALTIO, J.; CRISCUOLO, C. Cinco anos de gestão de dados espaciais na Embrapa: diagnóstico e perspectivas. In: CONGRESSO INTERINSTITUCIONAL DE INICIAÇÃO CIENTÍFICA, 17., 2023, Campinas. Anais [...]. Campinas: Embrapa Territorial, 2023. 11 p. 2965-2812 CIIC 2023. Nº 23509. Disponível em: https://ainfo.cnptia.embrapa.br/digital/bitstream/doc/1156897/1/6161.pdf