International Cartographic Association
Association Cartographique Internationale

Semantic search engine for geospatial datasets using large language models and AWS OpenSearch

Xinli Cai [a,*], Bo Lu [a,*], Sajad Ramezani [b], Anthony Fok [a], Yannick Lanthier [a], Joost Van Ulden [a]

[a] *Natural Resources Canada – xinli.cai@nrcan-rncan.gc.ca, bo.lu@nrcan-rncan.gc.ca, anthony.fok@nrcan-rncan.gc.ca, yannick.lanthier@nrcan-rncan.gc.ca, joost.vanulden@nrcan-rncan.gc.ca*
[b] *University of Alberta, srameza1@ualberta.ca*

* Corresponding author

Keywords: semantic search, natural language processing, user experience, infrastructure as code, geo.ca

Abstract:

The GeoDiscovery team, part of the Canada Centre for Mapping and Earth Observations (CCMEO), has developed an advanced semantic search engine to enhance the relevance and accuracy of geospatial dataset searches on GEO.ca, CCMEO's public dissemination platform. By integrating natural language processing (NLP) and machine learning (ML) techniques, this search engine addresses the limitations of traditional keyword-based search, which relies on exact word matches and fails to interpret context and synonyms. In contrast, semantic search enables the processing of natural language queries, understands complex context and account for synonyms, delivering more relevant and accurate search results.

Our approach uses Sentence-Transformer (SBERT) on Hugging Face for vector embeddings. We fine-tuned three widely used SBERT models: all-MiniLM-L6-v2, all-mpnet-base-v2 and paraphrase-multilingual-MiniLM-L12-v2 using geospatial metadata on GEO.ca. To evaluate the performance of these models, we created a benchmark dataset of curated queries and responses informed by domain expertise. Evaluation results showed that all-mpnet-base-v2 outperformed the others with a Mean Reciprocal Rank at 5 (MRR@5) score of 0.33 and Accuracy at 5 score (ACC@5) of 0.43. These scores indicate the model performs reasonably well by recommending relevant results within the top 5, with 43% of the queries returning at least one relevant result in this range. Additionally, we created a semantic search demonstration website to compare the semantic search engine and existing keyword search results. An internal survey indicated that the semantic search engine greatly outperformed the keyword search, improving search accuracy and enhancing the overall user experiences.

The deployment architecture for this solution is shown in Figure 1. We used Shared Services Canada's High Performance Computing (HPC) resources to fine-tune the NLP model and selected the best-performing model for deployment on Amazon Web Services (AWS). Key components include the use of AWS SageMaker for real-time model inference, AWS OpenSearch for indexing spatial and non-spatial data alongside its web application interface capabilities (such as OpenSearch Dashboards), and AWS Lambda to facilitate communication between SageMaker and OpenSearch. To streamline deployment and maintain consistency, the entire infrastructure was implemented using AWS CloudFormation, enabling infrastructure-as-code practices.

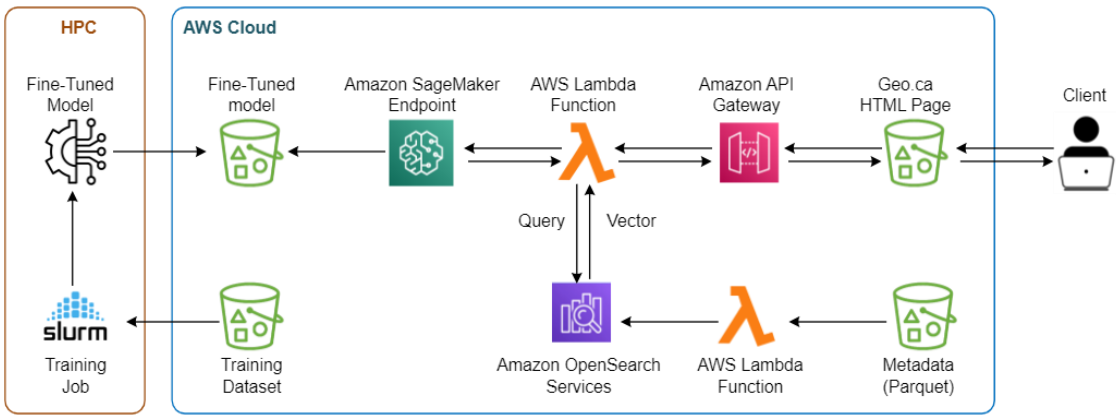

Figure 1. Deployment architecture of the geospatial semantic search engine