International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Search Engine for the US Federal Registry

Faizan Naviwala
Department of BECHLOR OF VOCATIONAL IN ARTIFICIAL INTELLIGENCE AND DATA SCIENCE
Anjuman-I-Islam's AbduL Razzaq Kalsekar Polytechnic, New Panvel, Maharashtra, India

---

Abstract - The continuous growth of public regulatory data, particularly within the US Federal Registry, presents significant challenges for efficient information retrieval. Traditional keyword-based search mechanisms often prove inadequate, yielding broad and irrelevant results due to a lack of semantic understanding. This paper introduces an AI-powered semantic search engine designed to overcome these limitations. Our system integrates a robust data pipeline for automated ingestion and transformation of raw JSON documents into structured, metadata-rich formats, coupled with a vector embedding approach for advanced semantic search. We detail the system's architecture, comprising an asynchronous data downloader, a resilient data processor, a ChromaDB-backed vector store, and a FastAPI-driven web interface. Through rigorous experimentation and comparative analysis against the official FederalRegister.gov portal, our system demonstrates marked improvements, achieving a Precision 5 of 78% and a Recall@10 of 89%. These results underscore the effectiveness of leveraging modern Natural Language Processing (NLP) and vector database technologies in enhancing the accessibility and utility of complex governmental datasets.

Key Words: Semantic Search, Information Retrieval, Vector Embeddings, ChromaDB, FastAPI, Natural Language Processing.

1.INTRODUCTION

The US Federal Register serves as the authoritative source for federal agency regulations and public notices, acting as a cornerstone for legal compliance, policy development, and public oversight. Published daily, its vast and ever-expanding corpus of documents, often presented in semi-structured or unstructured formats (e.g., JSON, PDF), poses considerable challenges for efficient and accurate information retrieval. Users, ranging from legal professionals and compliance officers to academics and journalists, frequently struggle to extract specific, relevant information from this voluminous dataset using conventional keyword-based search tools. These tools often suffer from synonymy and polysemy issues, leading to low precision and recall, as they fail to grasp the contextual meaning or semantic relationships between terms.

This research addresses these critical limitations by proposing and implementing a novel AI-powered semantic search engine. Our primary objective is to transform the existing paradigm of regulatory document access from a cumbersome, keyword-driven process to an intuitive, semantically aware inquiry system. The core contributions of this project include:

Development of an automated, scalable data pipeline capable of ingesting, parsing, and transforming raw Federal Register JSON data into a structured format suitable for advanced indexing.

Implementation of a vector-based semantic search engine utilizing state-of-the-art sentence transformer models and a persistent vector database (ChromaDB) to enable conceptual understanding of queries and documents.

Integration of a hybrid search architecture combining semantic search with robust metadata filtering to facilitate highly granular and precise queries.

Empirical validation of performance gains, demonstrating a significant improvement in information retrieval metrics compared to the current official search mechanisms.

The remainder of this paper is structured as follows: Sec. 2 reviews existing solutions and related work. Sec. 3 details the architectural design and implementation methodology. Sec. 4 describes the experimental setup and evaluation metrics. Sec. 5 presents and discusses the results. Finally, Sec. 6 outlines limitations and future work, and Sec. 7 concludes the paper.

2. Body of Paper

2.1 Related Work

Access to government documents typically relies on official portals and commercial services, each with inherent limitations.

Official Government Portals: FederalRegister.gov is the primary web interface, offering basic keyword search and limited filtering options. Its functionality is fundamentally lexical and often misses conceptually relevant documents that use different terminology. The GovInfo.gov Application Programming Interface (API) provides programmatic access to raw data but lacks built-in search relevance ranking or semantic capabilities.

Commercial Platforms: Platforms like Bloomberg Law offer advanced analytics but are typically subscription-based and cost-prohibitive for many users. Moreover, their search methodologies are often proprietary.

Academic and Open-Source Alternatives: Academic initiatives have focused on specific sub-domains but do not offer a comprehensive search solution across the entire Federal Register. Generic Elasticsearch implementations can provide full-text search but require significant customization to incorporate AI-powered relevance ranking.

Our solution addresses the critical gaps identified in existing systems by incorporating semantic understanding, automating the creation of an analysis-ready dataset, allowing for dynamic and granular filtering, and ensuring scalability through an asynchronous data pipeline.

2.2 Methodology

The proposed system operates on a modular, end-to-end architecture categorized into three core components: the Data Ingestion Pipeline, the Semantic Search System, and the Web Service Interface.

Fig -1: System Architecture Overview Code snippet

```
graph TD
   A[START:FEDERALREGISTERAPI]>B(DOWNLOADER MODULE: AIOHTTP);
   B --> C(RAW JSON DATA);
   C --> D(PROCESSOR MODULE: JSON TO CSV);
   D --> E(PROCESSED CSV DATA);
   E --> F(CHROMADB PIPELINE MODULE);
   F --> G(CHROMADB VECTOR STORE);
   G --> H(USER QUERY);
   H --> I(FASTAPI WEB SERVICE);
   I --> J(SEARCH LOGIC: QUERY CHROMADB);
   J --> K(RANKED RESULTS);
   K --> L(DISPLAY RESULTS TO USER);
```

2.2.1 Data Ingestion Pipeline

This component collects and prepares raw data from the Federal Register. The Downloader Module uses aiohttp and asyncio for concurrent Hypertext Transfer Protocol (HTTP) GET requests, accelerating data retrieval while handling pagination and rate limiting. Raw JSON data is saved into organized subdirectories. The Processor Module transforms this raw data into a clean, flat Comma-Separated Values (CSV) format. This involves dynamic ID generation for records missing a document_number, concatenation of textual fields for optimal embedding, and normalization of metadata fields.

2.2.2 Semantic Search System

At the core of the intelligent retrieval, this system transforms textual data into searchable vectors. ChromaDB, an open-source vector database, is used for persistent storage of document embeddings. Documents are converted into dense numerical vectors (embeddings) using the SentenceTransformerEmbeddingFunction with the sentence-transformers/all-MiniLM-L6-v2 model, chosen for its efficiency and performance. The search logic converts user queries into embeddings and uses cosine similarity to find the closest matching documents in ChromaDB, while also supporting advanced metadata filtering.

2.2.3 Web Service Interface

A lightweight web application built with FastAPI serves the user interface. It includes RESTful API endpoints for rendering the home page, handling search requests, and securely proxying PDF documents to the client to avoid cross-origin issues. The frontend uses Jinja2 for templating and provides an interactive search experience with an inline PDF viewer.

2.3 Experimental Setup

A comparative analysis was conducted against the official FederalRegister.gov search interface. A diverse set of 100 sample queries was manually curated, covering keyword-only, conceptual, and metadata-filtered searches. Performance was quantified using standard information retrieval metrics:

Precision: The proportion of relevant documents among the top 5 retrieved documents.

Recall: The proportion of relevant documents retrieved within the top 10 results out of all known relevant documents.

Filtered Query Support: A qualitative assessment of the system's ability to combine semantic queries with multiple metadata filters.

Relevance judgments were determined manually by human evaluators.

2.4 Results and Discussion

The experimental evaluation demonstrates a compelling advantage for our AI-powered system.

Table -1: Performance Comparison of Search Systems

Metric	FederalRegister.gov (Baseline)	Our System (Proposed)
Precision	42%	78%
Recall	55%	89%
Filtered Query Support	3 basic filters	6+ combined

A striking improvement was seen in Precision@5, which increased from 42% to 78%, indicating a near doubling in the accuracy of top-ranked results. This gain is attributed to the use of vector embeddings, which capture semantic similarity beyond exact keywords. Similarly, Recall@10 improved from 55% to 89%, showing a superior ability to identify and retrieve a larger proportion of all relevant documents.

Beyond these metrics, our system offers a qualitative advantage in filtered query support, allowing users to formulate highly granular queries by combining semantic search with multiple metadata filters, such as agency, presidential administration, and date ranges. The observed gains validate the project's core technical decisions, including the robust data pipeline, the choice of embedding model, and the use of ChromaDB with metadata filtering.

2.5 Limitations and Future Work

While the system offers significant advancements, several limitations have been identified.

Near-Term Enhancements (0-6 Months): We plan to integrate Elasticsearch to create a hybrid search architecture combining vector and keyword search. An interactive dashboard using Plotly Dash for visualizing regulatory trends is also planned.

Mid-Term Roadmap (6-12 Months): We aim to integrate advanced Optical Character Recognition (OCR) technologies for comprehensive text extraction from PDFs and implement user-centric features like saved searches with notifications and collaborative annotation.

Long-Term Vision (1-3 Years): A crucial long-term goal is to automatically detect and track regulatory changes over time by fine-tuning large language models. We also plan to expand the system to include state-level registries by developing a modular "State Adapter Interface".

Research Opportunities: Future research will explore training domain-specific embedding models on legal texts and implementing Explainable AI (XAI) techniques to provide transparency for search results.

3. CONCLUSIONS

The "US Federal Registry Search Engine" project successfully addresses a critical need for efficient and intelligent access to complex public regulatory documents. By implementing a robust data pipeline and an AI-powered semantic search system, we have demonstrated a significant leap in information retrieval capabilities. The empirical results, showcasing a 78% Precision@5 and 89% Recall@10, validate the effectiveness of integrating modern Natural Language Processing (NLP) techniques with a flexible web service. The system transforms the user's ability to interact with this information, enabling nuanced, semantic queries combined with powerful metadata filtering. This project serves as a compelling proof-of-concept and, with a clear roadmap for future enhancements, is poised to become a vital tool for the legal, policy, and research communities

3. CONCLUSIONS

The online version of the volume will be available in LNCS Online. Members of institutes subscribing to the Lecture Notes in Computer Science series have access to all the pdfs of all the online publications. Non-subscribers can only read as far as the abstracts. If they try to go beyond this point, they are automatically asked, whether they would like to order the pdf, and are given instructions as to how to do so.

REFERENCES

1. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (2019) 3969–3979
2. Chroma.: ChromaDB: The AI-native open-source vector database. Retrieved from https://www.trychroma.com/
3. FastAPI.: FastAPI: A modern, fast (high-performance) web framework for building APIs with Python 3.7+. Retrieved from https://fastapi.tiangolo.com/
4. aiohttp.: aiohttp: Asynchronous HTTP client/server for asyncio and Python. Retrieved from https://docs.aiohttp.org/
5. US Government Publishing Office.: Federal Register API. Retrieved from https://www.federalregister.gov/developers/api/v1