Project Proposal

Title:
Automated Extraction and Reloading of Stocks News into a Vector Database

Domain:
Data Engineering, Artificial Intelligence, Information Retrieval, Financial Technology

Problem Statement:
The financial industries generate a huge volume of unstructured data continuously from multiple data sources such as APIs, stock news portals, and market reports. Traditional storage systems make it difficult to efficiently retrieve and analyze this data for insights due to data redundancy, lack of semantic search capabilities, and poor scalability. Without an optimized approach to vectorize and organize this content, financial analysts and AI-driven applications face a lot of challenges in quickly accessing relevant and context-aware information.

Objectives:

- To design a pipeline that extracts a financial content from structured and unstructured data sources such as APIs and websites.
- To preprocess the extracted data for noise reduction, cleaning and summarization to improve its quality.
- To transform textual content into semantic vector using embedding strategy to efficiently store and manage the vectors in a scalable vector database (e.g., ChromaDB, Qdrant) along with metadata.
- To enable fast, context-aware retrieval for downstream applications such as stock prediction models, recommendation systems, or real-time financial analysis.

Solution Approach:
The proposed solution includes several stages. Initially, web scraping tools and APIs (such as requests, BeautifulSoup, and Playwright/Selenium) will be used to fetch and filter content from both static and dynamic data sources. Extracted content will undergo preprocessing steps such as cleaning and summarizing, to enhance consistency and usability. The processed content will then be converted into dense vector embeddings using modern NLP models (OpenAI embeddings, Sentence Transformers). These embeddings, along with relevant metadata, will be reloaded into a vector database (ChromaDB or Qdrant) for semantic search and analysis. This architecture ensures that future queries can efficiently retrieve related content across large, heterogeneous datasets for financial decision-making and downstream AI tasks.

Objectives:
The expected outcome of this proposal is an automated pipeline that extracts, preprocess and summarizes financial as well as stock news, transforms the text into vector embeddings and stores it in a vector database.

1. **Efficient Data Pipeline** – A functional pipeline that can automatically extract financial news and stock-related content from APIs and websites.

2. **Clean & Usable Data** – Preprocess, noise-reduction, and summarizes financial news data that is ready for analysis.
3. **Vectorized Content** – Text transformed into embeddings (vectors) stored in a vector database like **ChromaDB** or **Qdrant**, enabling semantic search and retrieval.
4. **Scalable Knowledge Base** – A centralized system that can handle large volumes of financial news and scale as new data sources are added.
5. **Improved Retrieval Capabilities** – Fast and context-aware retrieval of financial information for downstream tasks, like stock prediction models, recommendation engines, or real-time analytics.
6. **Support for AI/ML Applications** – A solid foundation for advanced AI-driven financial applications, enabling analysts and systems to query news data semantically rather than through keyword-based methods.