InsightEDGAR: Retrieval-Augmented Generative AI for Financial Document Insights

Problem Statement:

Financial analysts, investors, and researchers regularly need to extract actionable insights from SEC filings - documents that are lengthy, unstructured, and often difficult to search manually. Standard tools offer limited flexibility and rarely allow integration of real-time, user-provided context such as earnings call transcripts. To address this, we built **InsightEDGAR**, a Retrieval-Augmented Generation (RAG) application that combines public SEC filings with user-uploaded financial documents, enabling natural language Q&A over both data sources.

Project Name: InsightEDGAR

Solution Approach

InsightEDGAR combines the retrieval of relevant text segments from both SEC filings and user-uploaded documents with LLM models to provide instant answers to natural language queries. The system is built with LangChain, Hugging Face Embeddings, and Chroma, and supports flexible filtering using metadata.

Key Capabilities:

- Download and process SEC filings (10-K, 10-Q) for any user-specified US ticker; supports for uploads (e.g., earnings transcripts).
- Index all documents into a vector database with rich, consistent metadata to enable robust, filterable retrieval.
- Smart Retrieval - LangChain's SelfQueryRetriever (with GPT-3.5 Turbo) interprets queries, dynamically generates metadata filters, and retrieves only the most relevant chunks.
- Allow users to chat with all available documents using advanced retrieval and LLM-based answer generation.

Data Preparation & Metadata Consistency

Data Sources:

- **SEC EDGAR:** Automated downloading of 10-K and 10-Q filings for any U.S.-listed company via ticker symbol.
- **User Uploads:** Support for text-based transcripts, reports, or other custom financial documents.

Cleaning & Chunking:

- All text data, whether from EDGAR or user uploads, is cleaned with a unified function that removes tags and excessive whitespace.
- Documents are split into overlapping text chunks (~1000 characters each, with 200-character overlap) using *RecursiveCharacterTextSplitter*.

Consistent Metadata Integration:

- Each chunk is tagged with metadata from the start:
    - ✓ **ticker** (e.g., TSLA)
    - ✓ **doc_type** (e.g., 10-K, earnings_call_transcript)

- ✓ **year, quarter**
- ✓ **source** (edgar or user_upload)
- ✓ **section** (auto-detected where possible: e.g., "Item 1A. Risk Factors", "Q&A", or "Unknown")

- This consistent metadata allows robust filtering, enhances search accuracy, and underpins downstream retrieval logic.

Pipeline Architecture

1. **Data Collection & Cleaning:**
   SEC filings are downloaded using a custom pipeline and cleaned. User-uploaded files are similarly processed for noise removal and formatting.
2. **Chunking & Metadata Assignment:**
   Both data sources are split into manageable text chunks, each tagged with detailed metadata to maximize search and filtering precision.
3. **Embedding & Indexing:**
   Chunks are embedded using BAAI/bge-base-en-v1.5 (Hugging Face), and stored in a persistent Chroma vector database on Google Drive.
4. **Retrieval & Query:**
   - **Retriever:** LangChain's *SelfQueryRetriever* (using GPT-3.5 Turbo) interprets user queries, auto-generates filters based on metadata, and fetches the most relevant chunks.
   - **Q&A Chain:** Retrieved text is passed to an LLM for answer synthesis. Source documents and their metadata are shown for transparency.
5. **Sample Workflow:**
   - User specifies a ticker (e.g., TSLA), SEC filings are fetched and indexed.
   - User uploads an earnings call transcript, which is cleaned, chunked, and added to the index with detailed metadata.
   - User asks, "What are the main risk factors for Tesla in 2025?"—the retriever narrows the search to relevant filings and sections, and the LLM returns an instant, explainable answer.

Results & Evaluation

- **Performance:**
  The system delivers fast, relevant answers with supporting evidence. Manual checks confirm that answers align with the underlying filings or transcripts, and traceable metadata allows analysts to audit the sources.
- **User Experience:**
  Uploading and integrating transcripts is easy; all user data is searchable alongside official filings.
- **Metadata Use:**
  Metadata-driven retrieval (ticker, year, quarter, section, doc_type) provides fine-grained filtering and context-aware Q&A, significantly improving the relevance of responses.

Further Research & Recommendations

- **User Interface:** Build a Gradio-based web app for easy uploads, querying, and viewing results.
- **Cloud Hosting:** Deploy InsightEDGAR on a cloud platform (Hugging Face Spaces, AWS, or GCP) for public access.

- **Metadata Optimization:** Explore additional metadata fields (e.g., sentiment, entity tags) to improve search and filtering. Integrate semantic section mapping or use document layout parsers.
- **Enhanced Retrieval:** Explore combining retrievers (e.g., ParentDocumentRetriever) for deeper, multi-section answers.
- **Expanded Document Types:** Add support for more file types (investor presentations, press releases).
- **Continuous Improvement:** Implement user feedback and create benchmarks using real analyst queries to improve both retrieval and UI.

Conclusion

InsightEDGAR demonstrates a modern approach to financial document analysis—bridging unstructured public filings and user data with the power of retrieval-augmented language models. By focusing on consistent metadata from the outset and enabling seamless integration of user context, InsightEDGAR sets a new standard for explainable, interactive document Q&A. Next steps include building a user-facing web interface, enhancing metadata, and exploring open-source hosting options for a fully production-ready solution.