EMBEDDING METADATA RESEARCH NOTES

- Purpose of Metadata Embedding
    - Increases relevance by aligning search with fields like author, date, or format.
    - Enables more refined or filtered searches (e.g., date range, format type).
    - Improves reranking, providing additional signals for which documents best answer a query.
    - Reduces hallucination by grounding LLM responses in structured data.
- Metadata Fields to Consider
    - Author/Contributor
    - Title/Alternate Titles
    - Date or Time Period
    - Subject Headings/Keywords
    - Format/Medium (text, image, audio, etc.)
    - Location/Geographic
    - Rights/License
    - Collection/Series Name
    - Publisher/Printing Details
    - Prioritize fields heavily used in queries or that differentiate documents meaningfully.
- Approaches
    - Concatenate Text + Metadata
        - Combine metadata with main text into a single “augmented text” string.
        - Simple to implement but can blur important metadata signals.
    - Separate Metadata Vectors
        - Create standalone vectors for each metadata field.
        - Offers granular control over retrieval but increases complexity.
    - Hybrid Retrieval (BM25 + Embeddings)
        - For short metadata fields (author, date), a keyword-based approach like BM25
        - For free-form text (abstract, descriptions), use semantic embeddings.
        - Combines statistical and semantic approaches.
- Metadata in the RAG Workflow
    - Chunking: Split main text and metadata consistently; short metadata may not need chunking.
    - Index Creation:
        - If using Pinecone or another vector DB, store fields in `metadata={...}` for direct retrieval.
        - Check indexing/storage limits to ensure metadata is properly captured.
    - Retrieval + Reranking:
        - Retrieve top documents by vector similarity.
        - Apply metadata-based filtering or reranking (e.g., by date, format).
    - Prompt Assembly:

        - Insert retrieved metadata into the prompt to ground LLM output.
        - Keep a structured format (e.g., “Title: [title], Author: [author]”) for clarity.
- Implementation
    - Metadata Normalization: Standardize date formats, author names, etc.
    - Multi-Index Strategy:
        - Possible separate indices for text-based content vs. metadata fields.
        - Or combine into one index but store key fields in the metadata object.
    - Caching:
        - Cache frequently used metadata to reduce repeated API calls (major latency saver).
    - Structured Prompts:
        - Use consistent, parseable formats (JSON/XML) for the LLM.
        - Improves LLM’s ability to reference metadata accurately.
    - Metadata Pre-filtering:
        - Check queries for date/author/etc. references before embedding-based retrieval.
        - A small parser or entity-recognition model can direct queries to the right metadata filters.
- Advanced Considerations
    - Query Parsing:
        - Extract date ranges, authors, or formats to inform retrieval.
        - Tools like spaCy or a lightweight LLM can handle basic entity recognition.
    - Knowledge Graphs:
        - If data is highly structured, a knowledge graph can enrich semantic understanding.
    - Hybrid Matching:
        - Exact matching for critical fields (e.g., specific date)
        - Semantic matching for broad concepts (abstract, main text).
    - Performance & Scalability:
        - Pinecone can handle large datasets, but be mindful of costs.
        - Explore local or cloud hosting options for trade-offs in speed and storage.
- Immediate Next Steps
    - Decide on a strategy (concatenation vs. separate vectors vs. hybrid).
    - Implement a test subset with metadata embedding to measure improvements.
    - Optimize latency (caching, local database for metadata) to handle large queries.
    - Experiment with structured prompt design for more accurate LLM outputs.

References

- LangChain Documentation
- Pinecone Documentation: https://docs.pinecone.io/
- Digital Commonwealth API Info: https://www.digitalcommonwealth.org/developers
- https://arxiv.org/abs/2005.11401