# Research: Weighted Word Ranking in BPL Search System

*Jeremy Bui (Spring 2025)*

## Problem?

- The current system does not prioritize key metadata fields like titles or dates in retrieval.
- Users searching for specific historical events (e.g., “Boston 1919”) do not always get documents from that year at the top of their search results.
- All metadata fields contribute equally to retrieval, meaning a long abstract can outweigh a title match.
- BM25 reranking is applied post-retrieval, but it does not assign extra weight to title relevance.
- Queries structured in different ways affect retrieval quality

## Potential Solutions?

- Weighted Embeddings for Metadata
    - Instead of treating all metadata equally, embeddings should reflect field importance:
        - Title Weight: 1.5x
        - Date Weight: 1.2x
        - Abstract Weight (normal document stored): 1.0x
    - One way to implement this is by modifying the embedding input representation
- Multi-Vector Store Approach
    - Rather than storing all metadata in a single vector store, separate stores can be created for:
        - Title embeddings (prioritized for direct search)
        - Date embeddings (used for time-based relevance)
        - Abstract embeddings (fallback when no direct matches exist)
    - This allows a multi-stage retrieval process, where the order follows: title-based, date-based, and abstract-based matches.
- Adjusting BM25 Reranking
    - BM25 reranking is applied to the top 100 retrieved results, but it currently does not prioritize certain metadata fields. The weighting formula can be modified to emphasize more relevant fields (subject to change based on Sunday meeting):
        - Score = $(BM25_{title}\, x\, 1.5)\ +\ (BM25_{date} x\, 1.5)\ +\ (BM25_{abstract} x\, 1.0)$

- Optimized Query Processing
    - Rather than treating all queries equally, preprocessing should be applied to detect structured elements:
        - Identify titles, names, or historical periods within the query.
        - Adjust retrieval priority based on the query type.
        - If a query lacks a clear title or date, default to abstract-based retrieval.

## How can we implement it?

- Data Processing
    - Extract metadata from the Digital Commonwealth API.
    - Store metadata fields with assigned weighting values.
    - Normalize metadata for consistency.
- Algorithm Modifications
    - Modify embedding generation to incorporate field-based weighting.
    - Implement separate vector stores for titles, dates, and abstracts.
    - Adjust BM25 reranking to prioritize high-relevance fields.
- Performance Evaluation
    - Compare baseline vs. weighted retrieval using:
        - Mean Reciprocal Rank – How early does the first relevant document appear?
        - Retrieval latency – How much does the additional processing impact response time?

## Takeaways?

- Hopefully, by testing out different aspects and discussing with the clients, we can decide on how to best:
    - Improved search relevance for queries involving titles and dates.
    - Better ranking of historical documents, ensuring time-sensitive search results.
    - More structured retrieval, reducing the need for post-retrieval reranking.

## Resources:

- Embedding Meta-Textual Information for Improved Learning to Rank
- How do you handle multi-vector search with ranking?
- BM25-FIC: Information Content-based Field Weighting for BM25F
- Use Metadata in Your Search System
- *ChatGPT was used to expand on some research articles (BM25-FIC) to better understand concepts and methodologies.*