CS 6120 - Information Retrieval Grade Report
Team: 8
Member 1: Devam Jariwala
Member 2: Tanay Pandya

DeepCite – Research Paper Recommendation System

Problem Statement:

Our project aims to recommend relevant research papers based on user queries. With the vast number of research papers available online, finding the most relevant papers for a specific query can be challenging. Not all papers are pertinent to the user's needs, and even when they are, there may be a lack of proven real-world use cases in the specific area of interest.

Our system addresses these factors by recommending papers that emphasize both credibility and recency. We assess credibility using semantic relevancy and citation counts, while recency is determined by the publication date. This approach ensures that the user receives research papers that are both highly relevant and up to date.

Progress Report:

We have completed the data acquisition phase and integrated both CrossRef and Semantic Scholar APIs. After experimenting with both sources, we evaluated the quality of their retrievals. We then implemented and tested BM25 and FAISS for retrieval, where BM25 showed weaknesses in handling semantic variations, leading us to explore FAISS for better relevance. Currently, we are working on building an in-memory vector database for FAISS. Next, we plan to focus on developing a ranking algorithm incorporating factors like recency and citation count.

Proposed Solution:

1. **Data Acquisition & API Integration:** Collect research papers using Semantic Scholar and CrossRef APIs, ensuring a diverse and high-quality dataset.
2. **Hybrid Retrieval Mechanism:** Implement a combination of BM25 (keyword-based retrieval) and FAISS (semantic search) to improve the relevance of retrieved research papers.
3. **Ranking & Weighted Factors:** Introduce weighted factors for citation count (α) and recency (β), applying them to retrieved papers to refine ranking. Optimize α and β through iterative tuning for optimal relevance.
4. **In-Memory Vector Database:** Build an in-memory FAISS-based vector database to enhance the efficiency and scalability of semantic search.
5. **Benchmarking & Evaluation:** Compare the algorithm’s output against retrieval results from Semantic Scholar, CrossRef, and Google Scholar to assess its effectiveness in recommending relevant research papers based on title and abstract.

Grade Contract:

Milestones to achieve B grade:

Successfully integrate the **Semantic Scholar and CrossRef APIs** for data acquisition, ensuring we can fetch and store all necessary metadata such as title, abstract, citation count, publication year, and DOI. Build a working corpus with at least 1000+ research papers for each user query and implement a basic keyword-based retrieval system, even if the retrieved papers are not highly relevant.

Milestones to achieve B+ grade:

Enhance keyword-based retrieval by implementing **BM25**, allowing us to rank papers based on term importance. At this stage, the system should be able to retrieve **somewhat related papers**, even if they are not always the most relevant. Additionally, we will introduce an initial FAISS-based vector search using sentence embeddings and optimize API queries to reduce redundancy and improve efficiency.

Milestones to achieve A- grade:

Develop a working hybrid retrieval system that combines **BM25** (sparse keyword matching) and **FAISS** (deep embeddings) to retrieve only the most relevant papers. This will involve building an **in-memory FAISS vector database** to support fast and scalable retrieval. We will fine-tune similarity thresholds to ensure that our system retrieves only highly relevant papers and effectively formats the results to include key metadata such as title, citation count, publication year, and DOI, making it user-friendly and informative.

Milestones to achieve A grade:

Successful implementation of a **Ranking Algorithm** that assigns weighted factors to relevance (BM25 + FAISS similarity) (α), citation count(β), and recency(γ) to optimize the final recommendations. The system will be designed to prioritize high-citation, recent, and highly relevant papers, making it a robust academic recommendation tool. Finally, we will evaluate our retrieval and ranking performance by comparing our results with Semantic Scholar and CrossRef, identifying strengths and potential improvements. If possible, we will refine our approach to **match or even surpass** the quality of their retrieval in terms of relevance and ranking effectiveness.

Example queries with relevancy judgements:

1) **"Supervised Machine Learning usage in healthcare"**
   We are exploring how machine learning can be applied in the healthcare industry. For example, using machine learning models to analyze blood test reports and detect the presence of certain bacteria in the human body.

   Relevant papers may cover topics like identifying bacteria through supervised machine learning models or determining mental health states based on medical reports and health factors.

2) **"Deep learning in agriculture"**
   We are looking at how deep learning models can benefit agriculture.

   Relevant papers might involve the use of models such as DCNN or ResNet-50 to detect pests in crops, measure moisture levels in plants, and improve agricultural practices.

3) **“Architecture in Apple Silicon”**
We are investigating the architecture of Apple’s silicon chips.

Relevant papers would describe their chip architecture, including details on transistors, GPU design, core structure, and how Apple has optimized their chips for performance and efficiency.

4) **“AI reshaping humankind”**
We are interested in how AI is impacting human life. Has AI improved our comfort levels, or is it making us lazy? Is it a potential threat to humanity?

Relevant papers could explore AI’s role in various industries, automation in manufacturing, and its use in large-scale machinery.

5) **“Reducing Latency in Distributed Databases Through Efficient Caching”**
We are researching different techniques for database caching which would reduce latency specifically in distributed DBs.

Relevant papers might discuss databases optimized for faster data retrieval and the caching strategies that improve performance.

6) **“Cloud storage services for sensitive data”**
We want to understand which cloud storage service—AWS, GCP, or Azure—is best suited specifically for private, sensitive data such as user client information.

Relevant papers should compare the performance, pricing, and features of these cloud platforms.

7) **“Nutritional benefits of eating fruits and nuts for breakfast”**
We are exploring the health benefits of including fruits and nuts in a breakfast diet.

Relevant papers would discuss how these foods can improve overall nutrition and well-being.

8) **“String theory vs Theory of Everything”**
We are investigating the fundamentals of string theory and comparing similarities with theory of everything.

Relevant papers would explain the basic concepts, recent developments, and its implications in theoretical physics.

9) **“Morse code and Sign Language in World Wars”**
We are looking into usage of Morse code and sign languages in world war communication.

Relevant papers may explore their histories, uses in communication, and potential overlaps in learning or application.

10) **“RAG pipelines in financial documents”**
We are researching how Retrieval-Augmented Generation (RAG) pipelines are applied to financial documents.

Relevant papers should discuss the use of RAG models for efficient retrieval and generation of financial information.