Proposal for System Project: Scientific Paper Ranking System for Natural Language Queries

Introduction

The exponential growth of scientific literature has made it increasingly difficult for researchers to identify the most relevant and impactful papers for their research questions. Traditional keyword-based search engines often fail to interpret the context of complex natural language queries, leading to irrelevant or suboptimal results. This project aims to build a Scientific Paper Ranking System that leverages a two-stage retrieval process combining BM25, Cross-Encoder Re-ranking, and Locality-Sensitive Hashing (LSH) techniques. The system will utilize the arXiv dataset, a vast repository of preprint scientific articles (~3M), to retrieve, rank, and highlight the most contextually relevant scientific papers.

The arXiv dataset is ideal for this project due to its openness, extensive scientific coverage, and structured format, enabling efficient preprocessing and ranking. Key steps include cleaning duplicates, tokenizing text, and generating semantic embeddings with models like SciBERT or Sentence-BERT. These embeddings facilitate precise similarity computations, allowing the system to deliver highly relevant results for natural language queries in scientific research.

By combining traditional information retrieval methods with advanced NLP techniques, the system overcomes the limitations of keyword-based search, ensuring high precision and relevance.

Methods

The proposed system uses a two-stage retrieval framework:

1. **First Stage**: BM25 for Initial Retrieval
   - BM25 (Best Matching 25) will be used as the first-pass retrieval model to fetch a broad set of potentially relevant papers from the arXiv dataset.
   - It will rank documents based on term frequency, inverse document frequency, and query-document matching, ensuring fast retrieval.
2. **Second Stage**: Cross-Encoder Re-ranking
   - The top-ranked papers retrieved by BM25 will be passed to a fine-tuned Cross-Encoder model.
   - The Cross-Encoder will process the query and candidate papers jointly, using a transformer-based model (e.g., SciBERT or Sentence-BERT fine-tuned on scientific text) to compute semantic relevance.
   - This stage captures nuanced relationships between the query and the document, significantly improving ranking precision.
3. **Optimization**: with Locality-Sensitive Hashing (LSH)
   - For scalability, Locality-Sensitive Hashing (LSH) will be applied to pre-cluster papers based on semantic embeddings.
   - This reduces computational overhead in the initial BM25 retrieval, making the system efficient for large-scale datasets like arXiv.
4. **Result Presentation**

- The ranked list of papers will be presented along with abstracts, citation counts, and highlighted query-relevant sections for better user comprehension.

Team Members

Linus Stuhlmann

- Responsibilities: Designing and implementing the retrieval pipeline (BM25 and Cross-Encoder re-ranking), fine-tuning the NLP models, and evaluating system performance, data preprocessing (cleaning and embedding the arXiv dataset).

Robert Angerer

- Responsibilities: implementing Locality-Sensitive Hashing (LSH), and optimizing system scalability. Building the user interface to display ranked papers with detailed metadata and integrating feedback mechanisms for iterative improvement.

Preferred Time Slot

Preferred time slot: 18. December 2024