**Lab Report**

**Idea:** Query Expansion using an LLM for Search Engine Optimization as well as a basic Stemming and Stopword Feature

**Project Outline:** Our objective in this project is to enhance the information retrieval efficiency of a search engine built on Solr by using query expansion techniques through Large Language Models

**Problem Analysis:** A search engine is most effective when it matches user queries with the most relevant documents. However, simple keyword matching often fails to capture the full context of a user’s information need or might confuse some terms that are similar. Here, traditional methods might miss relevant documents due to vocabulary mismatch.

**Planned Approach:**

* Understanding Query Expansion and how they can be improved by using LLMs
* Deciding on which LLM to use
* Setting up the environment in which we will implement it
* Deleting Stopwords
* Implementing Stemming to reduce a basic shape
* Use LLM to generate relevant keywords for every document
* Keywords are combined with the document in a separate column inside Solr
* Search terms are also Stemmed and Stopwords are deleted
* Use LLM to generate additional and new terms terms to the search query
* Since the LLM output may be verbose, we repeat the original query terms 5 times to upweight their relative importance using concat
* Potentially adding PRF documents to the prompt to help top-heavy ranking metrics

**Strategies:** There are many strategies for LLMs, such as zero-shot, few-shot, and Chain-of-Thought (CoT) prompts. CoT prompts, are mentioned to be better and instruct the model to decompose queries step by step, giving us a large number of relevant terms for expansion. Pseudo-Relevance Feedback is another feature that can be added to further enhance the results

**Empirical Validation:** Experiments on the MS-MARCO and BEIR datasets demonstrate that query expansions generated by LLMs can outperform traditional PRF methods, this can be seen in both recall and precision metrics.

**Model Evaluation:** The study evaluates model performance across different sizes of the Flan-T5 and Flan-UL2 models, finding that larger models generally perform better

**Meeting on the 18.12:**

* Discussion on how to structure the Lab Book and when to have a new meeting during the holidays
* We decided that everyone will add their research information in by Friday