Final Project, Deliverable 1

Submission. For your submission, each student will upload their group's proposal as a PDF document in Canvas.

The culminating activity in this course is the final project. In groups of three to four students, you will design a project that establishes a research question and executes some type of experimental procedure around it. This work involves two key components: a software component and a presentation component.

All groups will work on **one of two** TREC (Text REtrieval Conference) problems: the CRISISFacts track or the NeuCLIR (Neural Cross-Language Information Retrieval) track. TREC serves as a forum for researchers and practitioners to share and assess the effectiveness of information retrieval systems.

CRISISFacts focuses on developing summarization technologies working over a variety of media types to be used in emerging crisis events. The CRISISFacts dataset encompasses a wide range of data points, including social media posts, news articles, and official statements related to various crises. The dataset is structured to help researchers analyze how information is shared and perceived in real-time during critical events, allowing for insights into public response and the effectiveness of communication strategies. CRISISFacts is described in this 2023 paper.

NeuCLIR focuses on evaluating and advancing retrieval systems that bridge language barriers. The NeuCLIR track aims to explore the capabilities of neural models in retrieving information across different languages. By working with multilingual datasets, participants can evaluate how well systems can interpret queries in one language and retrieve relevant documents in another. TREC has a long history of fostering research in information retrieval, and both tracks have seen submissions where retrieval systems and approximate nearest neighbor search are at the heart of successful approaches. The track organizers have published papers on the 2022 and 2023 NeuCLIR tracks, which you should review for an overview of the effort.

Given a choice of one of the above problems, your team will imagine that you are building a search application similar to this demo. Putting aside the details of quantization, the aim is to focus on building an application that a user can provide input into. Your frontend will be supported not by Gradio, but something more sophisticated. Your goal will be to provide a query as input, where the data you search over will be taken from CRISIFacts, NeuCLIR, or something else (see below for more). As you read, in CRISISFacts, the queries are disaster-specific, while in NeuCLIR, the queries are in English but the collections you search over are in other languages. For a target collection size, aim for a collection of about 1 million documents. Notice that the NeuCLIR collections already fit this requirement. The demo above from Hugging Face uses a collection of about 40 million documents.

In short, the final deliverable is an interface that allows the user to search over a very large collection both effectively and efficiently. We are using these datasets because you rank and get scores to see how well your systems are actually ranking relevant information. If you choose to use another dataset, you are welcome to skip the evaluation, but then the dataset size must be much larger (e.g., 40 million documents as we saw above). This allows you to choose a project that doesn't have you quantitatively evaluate but instead focus on engineering.

Internally, the scoring methods you choose to use can come from any model family of interest. We suggest that you read through the survey papers to get a good idea of the open-source tools you might want to use. In class, we'll discuss some neural retrievers, like ColBERT and ColBERTv2, which are frequently used in modern retrieval problems.

For the proposal, you will write a one- to two-page document summarizing what your group expects to accomplish in the final project. This means you will explicitly outline the steps you expect to take to meet your goals. Your proposal must address the following questions:

- What problem have you chosen?
- What subsets of data from the track your team chose will you use for training and evaluation?
- What recent work has been done in the area? You do not need to complete a thorough literature review at this point, but you should identify two or three key papers and developments relevant to your experimental approach. In particular, it is likely enough to read through the survey papers as a first pass.
- What methods will you use? These may include some of what we've covered in class, but you may optionally include approaches outside of class, too. Your team should be thinking generally about retrieval systems, many of which include approximate nearest neighbor routines as subcomponents.
- How will you evaluate these methods on the track data?

The final project is worth 40 points, broken into two deliverables:

- Deliverable 1 (Proposal) is worth 8 points.
- Deliverable 2 (Presentation and Software Submission) is worth 32 points, 12 points for a presentation component and 20 points for a software component.