INLS 609: Experimental Information Retrieval

Project Proposal Description

The proposal should be about 2-3 pages and should answer the following **four** questions.

The format of the proposal is up to you. In general, it should be comprehensive and clear.

1. What research questions or hypothesis do you plan to explore?

Your project should consider a number of different research questions or hypotheses.

> **Example hypothesis:** Automatically removing non-medical terms from the Electronic Heath Record (EHR) "note" will improve retrieval performance.

You should describe your research questions or hypotheses as clearly as possible so that I know exactly what you plan to explore.

2. How are your research questions <u>interesting</u> in light of prior research?

Your research questions should build upon prior work. In other words, you should be trying to address a challenge that has already been recognized by prior researchers, and your proposed solution should be reasonable in light of approaches others have tried in the past. In other words, each of your proposed approaches should not be a "shot in the dark". Based on your readings of the notebook papers associated with your chosen TREC Track (in most cases, the Clinical Decision Support Track), you should be able to argue that each of your proposed approaches has a "greater than zero" chance for success.

> **Example:** EHR "notes" are very verbose and contain non-technical terms that are not central to the topic and are likely to hurt retrieval performance when included in the query.
>
> In fact, several groups who participated in the TREC 2016 Clinical Decision Support Track tried to automatically reduce the EHR "note" by automatically dropping non-medical or non-technical terms. For example, University of ABC tried to automatically identify non-medical/non-technical terms using approach XYZ (citation). Similarly, University of DEF tried to dropped terms from the EHR "note" using IDF values from the collection---terms with an IDF value less than threshold *T* were dropped from the EHR "note" (citation).

3. What experiments will you run in order to answer each of your research questions?

Your proposal should describe the experiments you plan to run. I will be looking for clear explanations of how you will test each of your hypotheses. Please describe the baseline experiment that you will use to compare against. Also, please describe the evaluation metric you plan to focus on.

Example: We will develop an algorithm to automatically drop non-medical terms by comparing each term's frequency in the CDS target corpus (1.3 million PubMed articles) with its frequency in a non-medical corpus of equal size (e.g., a sample of 1.3 Wikipedia articles). Our approach will drop every term whose frequency in the target medical corpus is greater than its frequency in the non-medical corpus.

We will evaluate against the baseline approach of not reducing the EHR "note" in any way.

We will evaluate based on P@10 (i.e., precision at rank 10) averaged across all 30 CDS queries.

4. What some of the risks associated with your proposed approaches?

Please spend some time contemplating the different risks associated with your proposed plan. Please describe these risks clearly. I encourage you to take risks. However, your project should contain results. It is acceptable to try something that does not work. So-called "negative results" are interesting and valuable. It is not acceptable to get to the end of the semester and not be able to complete an experiment.

Example: In order to implement our approach, my partner and I will need to figure out how to extract term frequency information from an Indri index and how to parse and index a set of 1.3 million Wikipedia articles. However, we have found some code on Github that can do part of this and we believe this risk is manageable.