

# IR Competition Project Proposal

---

John James      jtjames2

This project aims to leverage the information retrieval competition to explore state-of-the-art learn to rank system configuration document retrieval models.

Learn to rank document retrieval system configuration is an area of active research that seeks to improve document retrieval performance by predicting system configuration parameters for a query that maximizes the likelihood of relevant documents.

## Approach

The proposed approach is adapted from Deveaud et. al [1] and is comprised of the following four phases:

- **Pre-processing:** Given a collection, the first step is to determine all feasible system configurations  $C$  and conduct document indexing. Next, the documents are ranked and the effectiveness of each query/system configuration pair is stored to file. The query and system configuration features serve as the feature set and the effectiveness scores represent the labels of the training set.
- **Training:** This step makes use of the training examples constructed from the query features, system configuration features, and relevance labels measured by the respective evaluation metric. The learning to rank algorithm trains a model to maximize the effectiveness metric (e.g. nDCG). Once the training is completed, a learned model is generated.
- **Document Ranking:** Taking as input, an unseen test set of queries, the trained model produces a ranked list of system configurations for each query.
- **Evaluation:** Finally, the top ranked system configuration for each query is applied and the overall system performance is computed.

## Features

The learn to rank features include:

- Retrieval Model Features such as:
  - Absolute Discount Smoothing
  - Dirichlet Prior Smoothing
  - Jelinek-Mercer Smoothing
  - Okapi BM25
  - Pivoted Length Normalization
- Query Features including:
  - Query Statistics e.g., Mean and standard deviation variants of IDF
  - Linguistic Features: Synonyms Hyponyms, Meronyms, etc...

I look forward to exploring these techniques in the IR competition!