Write your proposed project title here
Write your group name underneath

First Author
email1@domain

Second Author
email2@domain

Abstract

In the abstract, describe your project in a few sentences. State what the expected application is going to be. Some general guidelines for this proposal document: one page, due April 14 on odtuclass. Use this LaTeX template to write your proposal.

1 Introduction

Tell us about the information retrieval system that you want to implement. Describe what features will be incorporated in this system. Which data set will be used? State the size of the collection, size of the set of evaluation topics (e.g. queries), the number of relevance judgements.

2 Overview of the Proposed System

State the main modules of your project in this section. A typical IR project would include the following modules:

- Data parsing and indexing: Describe preprocessing of the corpus (e.g. tokenization, stop word removal, stemming etc.). Describe how you will implement the inverted index. You can use an existing IR system (Lucene or PyTerrier) and adapt it to work on your document collection and queries (this system can use the vector space model or a more advanced model). Also indicate the number and type of indices (i.e. positional index, field/zone index, permuterm index etc.)
- Query processing: Describe the allowed query formats, parsing query. Are you planning to support boolean queries, wildcards, phrase queries, proximity queries etc.?
- Ranking and retrieval: You must implement a basic vector space model as a baseline system (e.g. vector space model with BM25). Then, implement an improved version of this baseline system. You can add one or more optimizations and/or components that could improve the performance of your baseline system. Some ideas for improvements: You can use pseudo relevance feedback or neural information retrieval models to re-rank the results you obtained from the baseline system. You can produce vectors for the query and each of the selected documents using various versions of word2vec, sent2vec, doc2vec, BERT, or an open version of GPT models or other LLMs that you can run. You can also use pre-trained word embeddings and assemble them to produce query/document embeddings.
- User Interface: An aesthetic user interface for smooth interaction is one of the key things for the success of an IR system. Describe how you are planning to implement the user interface.

Note: In the final project report you will include a section for the evaluation of your system. You will test the performance of your baseline system with evaluation metrics like MAP, Precision@10 etc. Then evaluate the performance of the improved version(s) of your system. You will present a precision-recall plot (e.g. Averaged 11-point precision-recall graph) that shows the comparison of different approaches that you implemented in your project.