Project Milestone 2: Detailed description, approaches, datasets

Team: Kota Factory

Manoj & Dipanshu

Topic: Artificial Intelligence for Legal Assistance

Detailed Description: The problem is divided into 2 tasks. Let's analyze both tasks independently. We will be given a set of 50 queries, each of which describes a situation.

Task 1: Identify relevant prior cases: The first task is to identify relevant prior cases for a given situation. There will be around 3000 case documents that were judged in the Supreme Court of India. For each query, the task is to retrieve the most similar/relevant case document with respect to the situation in the given query.

We propose the following subtasks for task 1:

1. **Data pre-processing:** In this step, We will take care of the lower and upper case letters, stop word removal, stemming and remove punctuation and numbers from the case documents.
2. **Similarity score using TF-IDF:** We first calculate TF-IDF values and then compute the cosine of the angle between the two vectors.
3. **Word Embeddings:** We plan to use doc2vec/sent2vec embedding for the dataset as it is publicly available.

4. **BM25:** We plan to use the variant of BM25 given by Tortman which has proven to be effective in a number of scenarios. To retrieve the documents with the highest rank in both TF-IDF and BM25, we multiply the cosine similarity scores of a query-document pair with its BM25 score.
5. **SBERT:** We use Sentence-BERT to understand the context of a query with other documents.

Task 2: Identify the most relevant statutes for a given situation:

After applying task 1 on the corpus dataset, we retrieve the most relevant documents related to the query. Now in task 2, we need to re-rank them according to their relevance.

We plan to use Language Model for reranking: that is we will use the relevance models of Lavrenko and Croft's relevance models with a Unigram model with Dirichlet smoothing.

Our project implementation can be summarized in the following flowchart: