Information Retrieval and Web Search Term Project Summary

Shrikrishn Bansal 19MCS005

Deepa Tilwani 19MCS002

April 13, 2020

Step-wise Summary

The summary and steps for the term project are as follows:

1. First we have collected 200 documents under five categories as Sports, Technology, Politics, Science and Entertainment. 40 documents per category.

2. Then we have created a set of 50 benchmark queries, 10 queries per category.

3. Relevance judgment for each query-document pair is created.

   In its original form, relevance feedback refers to an interaction cycle in which the user selects a small set of documents that appear to be relevant to the query, and the system then uses features derived from these selected relevant documents to revise the original query. This revised query is then executed and a new set of documents is returned. Documents from the original set can appear in the new results list, although they are likely to appear in a different rank order.

   We have stored these query-document pairs in a JSON file where each key represents query and value represents an array of relevant documents for that query.

4. Then we have written a program to retrieve documents based on the query input. This program is written in C# language with .net framework console application.

In this program first we have created Term-document-incidence matrix and then based on this matrix we have retrieved documents for the query. These retrieved documents are stored in list and then based on the query input we have extracted the relevant documents of the query from the JSON file and stored these relevant documents in another list.

5. Another list is created in the program which stores intersection of these two list.

6. Finally Precision, Recall and F1 measures are then calculated as a system performance evaluation metrics.
   Precision tells us how relevant are the positive detection.
   Recall tell us what percentage of actual positives are detected.
   and the F_1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision and the recall of the test to compute the score