Srirakshith Betageri

A20414667

Homework 2

The axes of the vector space are the terms and the value is the tf.idf of the given term and document pair.

The query is converted to a vector in the same vector space. This

**Exact Top K**

In this method, we compare the cosine similarity between the given query and all the given documents.

**Inexact Champion List**

Instead of comparing the query with all the documents, we instead do the following.

For each term, we keep track of the documents with the highest IDF. Using only these top IDF documents, we reconstruct them into vector space and perform exact topk search again. The only difference is in changing the search vectors.

**Inexact Query Index Elimination**

We initially sort the given query terms based on their IDF values. The terms are sorted in descending order. Having sorted them thus, we eliminate half of the lower valued terms. We then perform cosine similarity to obtain the top k ranked documents.

**Cluster Pruning**

In cluster pruning, we initially assign the role of a leader to sqrt(N) documents, where N is the total number of documents.

The remaining documents are assigned to a cluster and each cluster has only one leader. The assignment is based on the similarity between the remaining documents and the leaders.

Having performed this pruning on the clusters, we then find the leader which is most similar to the given query. Having found this leader, we then search in the cluster for the documents. We stop the search when we find k documents.

Performance:

It is expected that each of the subsequent queries provide for a smaller search time.