REPORT

1.1 INTRODUCTION:

This Information Retrieval system demonstrates advance methods of index construction.
Exact top K retrieval (Method 1)
Inexact top K retrieval (Method 2)

TF-IDF weighting of terms is used to construct the vectors for the documents and queries. Additionally the stop list is provided as an attachment to weed out words from dictionary. The system will then accept a free text query, generate the vectors for the documents and query. After vector generation it will compute the cosine similarity score for the documents.

The system will retrieve and display the names of the top K documents for each query in decreasing order of their score. Also displays time taken to retrieve the results for each query by each method.

1.2 Exact Top K Retrieval (Method 1):

The idea behind this method is pretty straight forward, calculate TF-IDF weight of terms then construct vectors. After that compute the cosine similarity scores for the documents. Then retrieve the names of the top K documents for each query.

1.3 Inexact Top K Retrieval (Method 2):

i. Champion List
ii. Index Elimination
iii. Cluster Pruning

Champion List: This method generates a list of r documents that is based on the weighted term frequency w t,d for each term. The list is created only once for a collection.

Index Elimination: This method uses only half the queries terms sorted in decreasing order of their IDF values.

Cluster Pruning: This method randomly pick √N leaders (where N is the number of documents in the collection) and then use them to implement the cluster pruning. The leader is selected only once. For each query it selects a leader closest to the query and then retrieve the top K results

1.4 Comparison of Exact & Inexact Methods:

Implementation of four methods on five queries whose evaluation data is represented in below table.

1.4.1 Performance Evaluation Table:

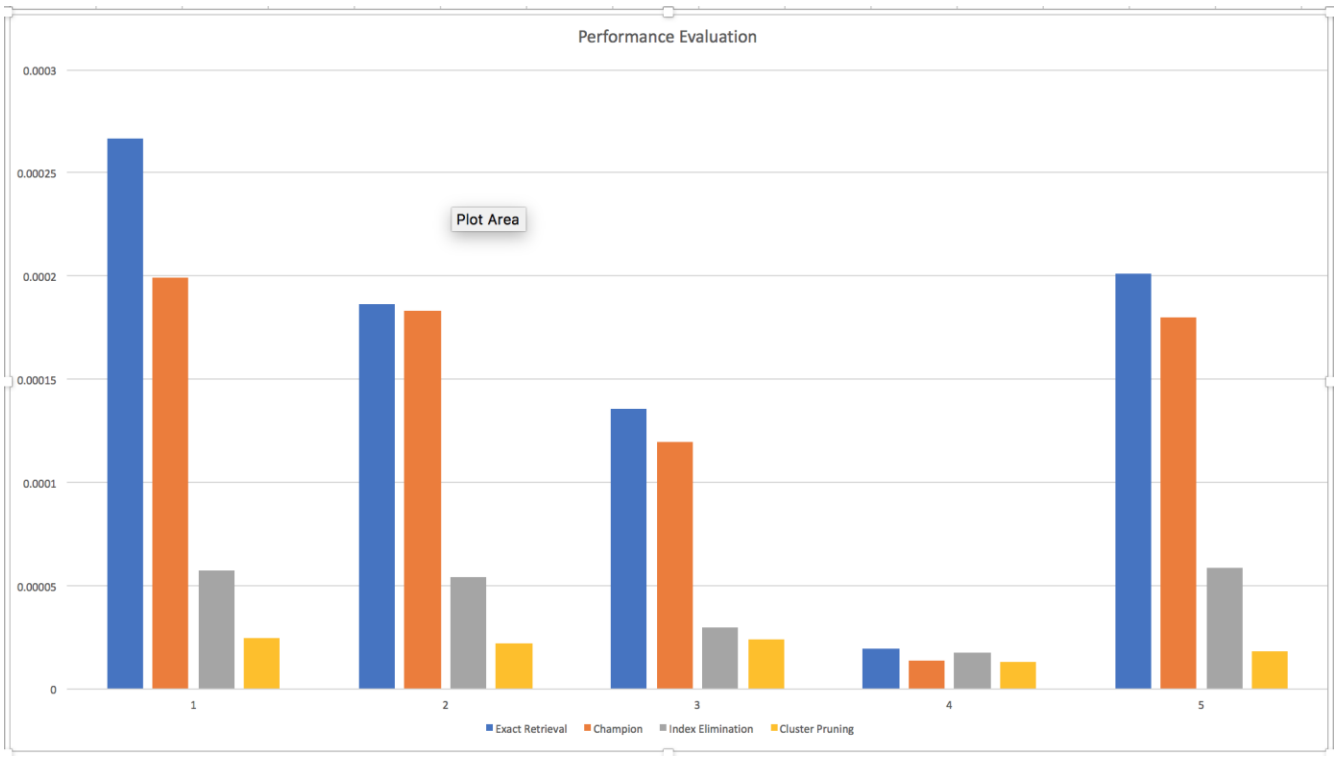
Query	**Exact Retrieval** (Time in Sec)	**Champion List** (Time in Sec)	**Index Elimination** (Time in Sec)	**Cluster Pruning** (Time in Sec)
with AND without AND yemen	0.0002391338	0.0001826286	0.0000541210	0.0000264645
with AND without AND yemen AND yemeni	0.0001749992	0.0002734661	0.0000803471	0.0000407696
berlin AND poland AND szczecin AND obacz AND plane	0.0002305508	0.0001878738	0.0000340939	0.0000374317
abc AND pqr AND xyz	0.0000343323	0.0000267029	0.0000350475	0.0000147820
million AND billion	0.0002093315	0.0001752377	0.0000569820	0.0000257492

Table 1.1

The performance evaluation statistics is shown in table 1.1 & Graph 1.1

On the basis of data in table 1.1 we can conclude that,
Exact Method is Most inefficient Retrieval method among rest of three.
Inexact method (Cluster Pruning) is most efficient Retrieval method among rest of three.

1.4.2 Performance Evaluation Graph:




```
Note:
Y-axis Time in Seconds.
X-axis Five queries implemented by four methods.
```

Graph 1.1

1.5 Conclusion:

1. Exact Method is Most inefficient Retrieval method.

2. Inexact method (Cluster Pruning) is most efficient Retrieval method.