Tony Barboza
Bobby Chisholm

Installation Notes

Part 1 run instructions

Download the TREC Complex Answer Retrieval “test200“ dataset from http://trec-car.cs.unh.edu/ and unpack and place it into src/main/java/data

To run Maven build

```
mvn clean install
```

*Compile

```
mvn clean compile assembly:single
```

NOTE: You can specify an index as a command line argument but if you don’t it will be written to the default directory. Default index: src/main/java/index

*How to run the indexer:

```
java -Xmx50g -cp target/IR_program2-0.1-jar-with-dependencies.jar lucene.Indexer
```

NOTE: First command line argument is path to index, second is path to input file, if no arguments it will use default index: src/main/java/index and the file: train.pages.cbor-outlines.cbor. The output files can be found in src/main/java/output

*How to run the searcher:

```
java -Xmx50g -cp target/IR_program2-0.1-jar-with-dependencies.jar lucene.SearchFiles
```

Part 2 run instructions

1. Put the Trec_eval 9.0 version tool in the base directory of the project.
2. Cd trec_eval.90
3. Compile the program using make
4. Run the command:

```
./trec_eval -m Rprec -m map -m ndcg_cut
../src/main/java/data/test200/test200-train/train.pages.cbor-article.qrels
../src/main/java/output/DefaultRankingOutput.txt
```

5. And:

```
./trec_eval -m Rprec -m map -m ndcg_cut
../src/main/java/data/test200/test200-train/train.pages.cbor-article.qrels
../src/main/java/output/CustomRankingOutput.txt
```

*Note ndcg_20 would not work so this prints out all ndcg values!

Part 3

*To run the Prescision at R file run:

```
java -Xmx50g -cp target/IR_program2-0.1-jar-with-dependencies.jar lucene.PrecisionatR
```

Output:

```
Custom Rank Output Precision at R: 0.5260762
Default Rank Output Precision at R: 0.5963763
```

*Our result is .0003 off from the trec_eval score. After hours of evaluating our precision at R we came to the conclusion it must be a rounding difference.

Part 6:

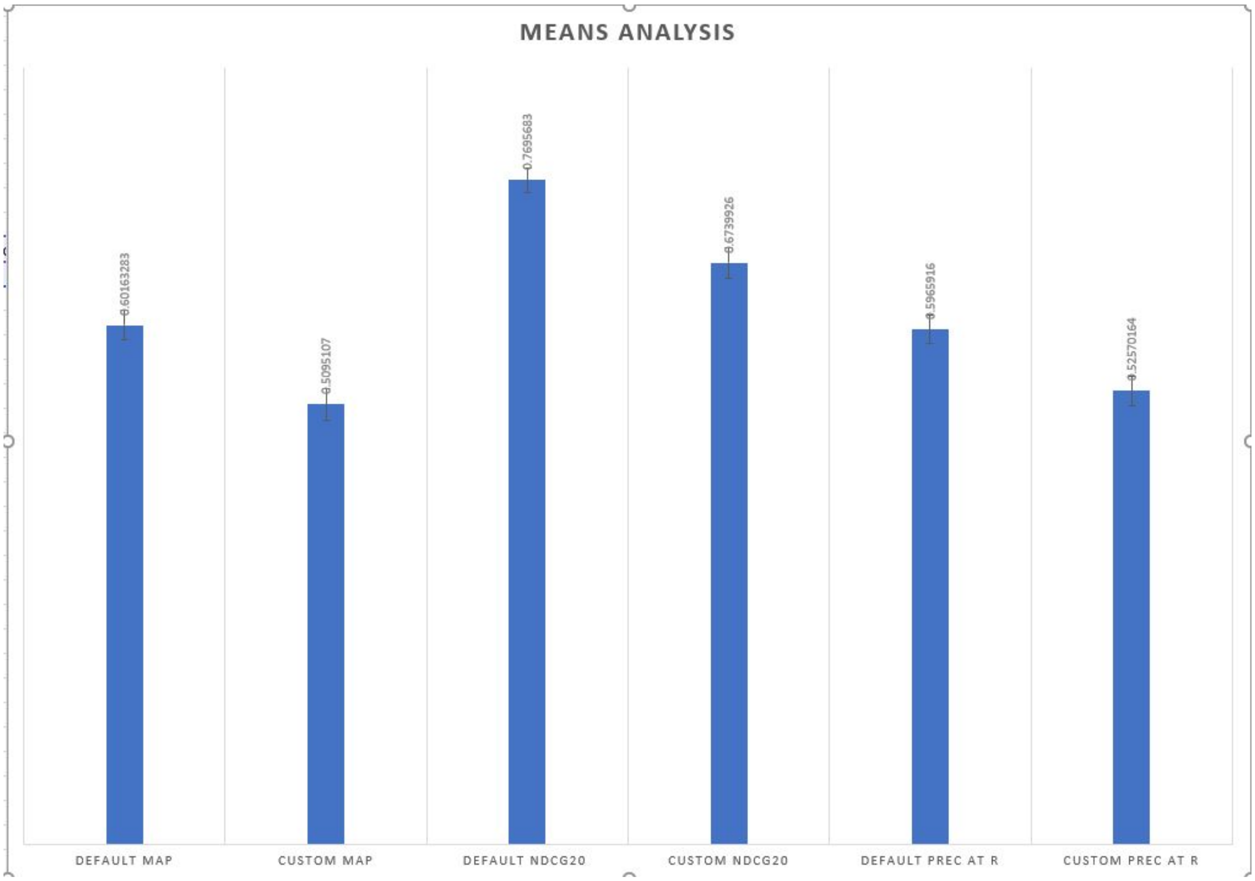
Default-

```
        map             all     0.6016
        Rprec           all     0.5966
        ndcg_cut_20     all     0.7696
```

Custom-

```
        map             all     0.5095
        Rprec           all     0.5257
        ndcg_cut_20     all     0.6740
```

The graph comparing the two ranking functions is bellow. We created a program that crunched the mean and standard deviation for each function and query. We put that data into an excel file and then created the bar graph from their. As seen in the table on the next page the default ranking algorithm is consistently higher than the custom. This difference is significant because the standard deviations do not overlap, the default ranking is entirely above ours. The default ranking definitely performs better.


MEANS ANALYSIS
0.60163283
0.5095107
0.7695683
0.6739926
0.5965916
0.52570164
DEFAULT MAP
CUSTOM MAP
DEFAULT NDCG20
CUSTOM NDCG20
DEFAULT PREC AT R
CUSTOM PREC AT R