Project 3: Evaluation of IR Models

Tannu Priya
UBIT name: tannupri
Department of Computer Science & Engineering
University at Buffalo
Buffalo, PA 14214
tannupri@buffalo.edu

Abstract

The goal of this project is to implement various IR models, evaluate the IR system and improve the search result based on your understanding of the models, the implementation and the evaluation. The twitter data that is used in this project is in three languages- English, German and Russian. The twitter data is indexed using Solr and the three models are evaluated using Trec_eval program.

1. Introduction

The project aims to compare the performance of the same dataset when different models are employed. First, Solr is used to index the tweets that are in three different languages. Then, the performance of the IR system on the following three models is evaluated:

1. Vector Space Model(VSM)
2. BM25
3. Divergence from Randomness Model(DFR)

1.1 Implementation of VSM in Solr

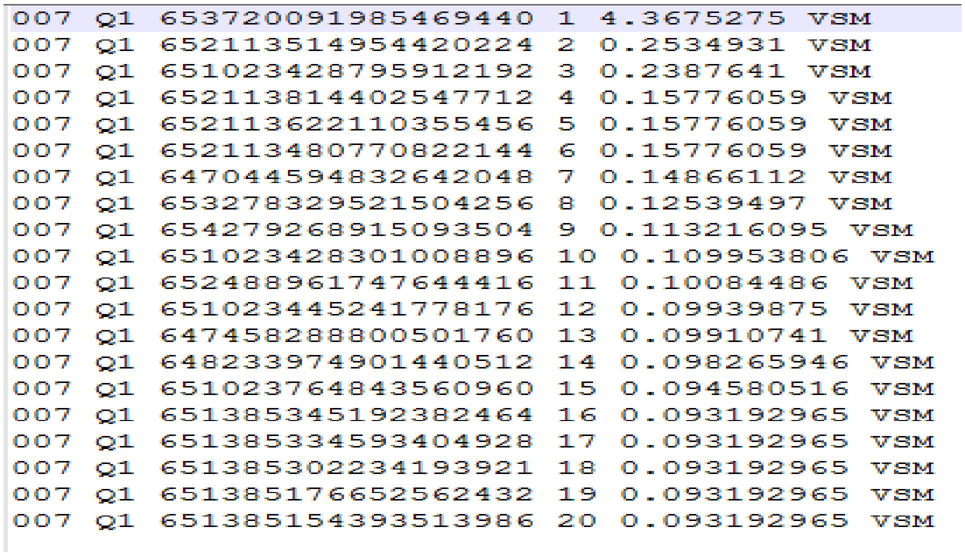
In VSM, documents and queries are represented as weighted vectors in a multi-dimensional space, where each distinct index term is a dimension, and weights are Tf-idf values.

Key implementation steps:

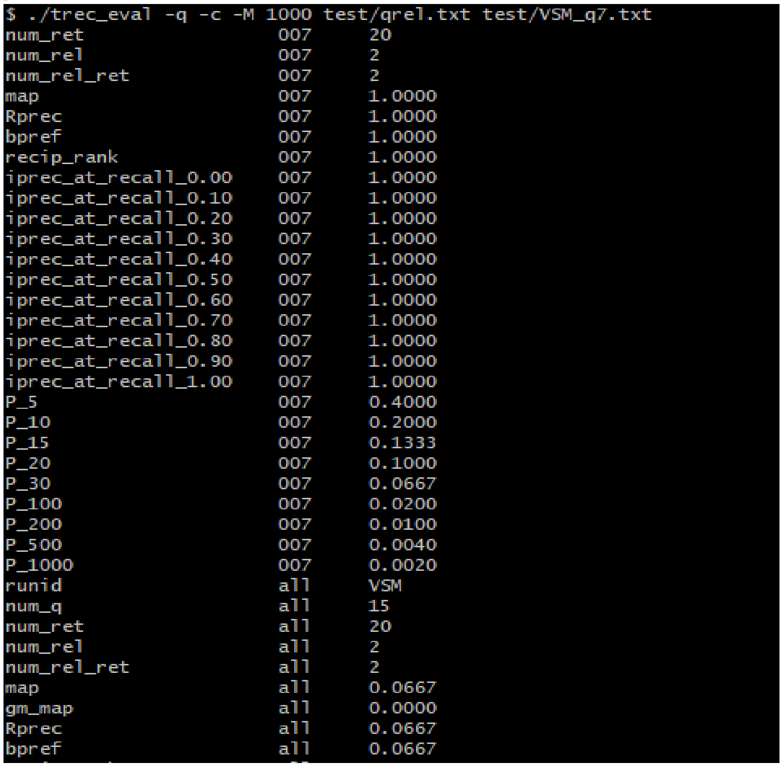
- We created core name VSM in SOLR and indexed train.json data provided.
- In Schema.xml file we are updating the model so Solr uses VSM instead of BM25, which is a default model.
- We converted our queries to solr url format which was to fetch the resultsfrom python code.
- We run a python code (part of project documents in src folder) to fetch the results and write them into an output file.
- We use this file as the input for our Trec_eval program along with qrel.txt file to generate MAP value for that corresponding query result pair.
- To update the model, we are adding the below code in our schema file.

<similarity class="solr.ClassicSimilarityFactory"/>
By updating Similarity class to classic model that Solr used before BM25.

Results:
By updating the model we get different scores and MAP values computed



```
007 Q1 6537200919854694401 1 4.3675275 VSM
007 Q1 652113514954420224 2 0.2534931 VSM
007 Q1 651023428795912192 3 0.2387641 VSM
007 Q1 652113814402547712 4 0.15776059 VSM
007 Q1 652113622110355456 5 0.15776059 VSM
007 Q1 652113480770822144 6 0.15776059 VSM
007 Q1 647044594832642048 7 0.14866112 VSM
007 Q1 653278329521504256 8 0.12539497 VSM
007 Q1 654279268915093504 9 0.113216095 VSM
007 Q1 651023428301008896 10 0.109953806 VSM
007 Q1 652488961747644416 11 0.10084486 VSM
007 Q1 651023445241778176 12 0.09939875 VSM
007 Q1 647458288800501760 13 0.09910741 VSM
007 Q1 648233974901440512 14 0.098265946 VSM
007 Q1 651023764843560960 15 0.094580516 VSM
007 Q1 651385345192382464 16 0.093192965 VSM
007 Q1 651385334593404928 17 0.093192965 VSM
007 Q1 651385302234193921 18 0.093192965 VSM
007 Q1 651385176652562432 19 0.093192965 VSM
007 Q1 651385154393513986 20 0.093192965 VSM
```



```
$ ./trec_eval -q -c -M 1000 test/qrel.txt test/VSM_q7.txt
num_ret                 007     20
num_rel                 007     2
num_rel_ret             007     2
map                     007     1.0000
Rprec                   007     1.0000
bpref                   007     1.0000
recip_rank              007     1.0000
iprec_at_recall_0.00    007     1.0000
iprec_at_recall_0.10    007     1.0000
iprec_at_recall_0.20    007     1.0000
iprec_at_recall_0.30    007     1.0000
iprec_at_recall_0.40    007     1.0000
iprec_at_recall_0.50    007     1.0000
iprec_at_recall_0.60    007     1.0000
iprec_at_recall_0.70    007     1.0000
iprec_at_recall_0.80    007     1.0000
iprec_at_recall_0.90    007     1.0000
iprec_at_recall_1.00    007     1.0000
P_5                     007     0.4000
P_10                    007     0.2000
P_15                    007     0.1333
P_20                    007     0.1000
P_30                    007     0.0667
P_100                   007     0.0200
P_200                   007     0.0100
P_500                   007     0.0040
P_1000                  007     0.0020
runid                   all     VSM
num_q                   all     15
num_ret                 all     20
num_rel                 all     2
num_rel_ret             all     2
map                     all     0.0667
gm_map                  all     0.0000
Rprec                   all     0.0667
bpref                   all     0.0667
```

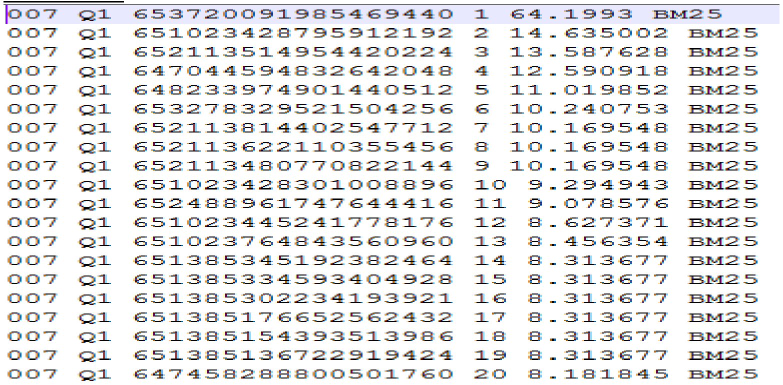
1.2 Implementation of BM25 Model in Solr

BM25 is one of the most established probabilistic term weighting models. Here BM stands for Best Matching, itis a ranking function used by search engines to rank matching documents according to their relevance to a given search query. It is based on the probabilistic retrieval framework.

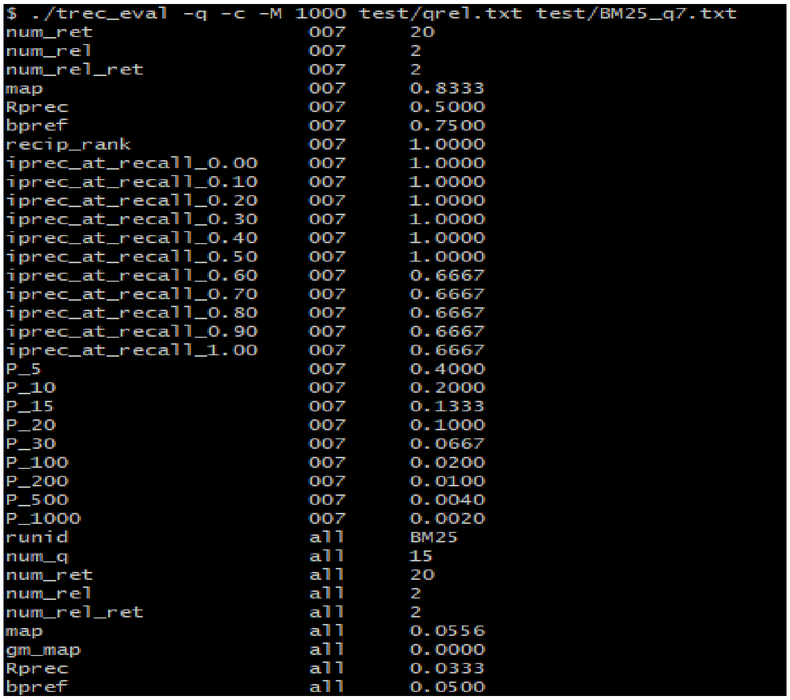
Implementation steps:

- We created core name BM25 in SOLR and indexed train.json data provided.
- We converted our queries to solr url format which was to fetch the resultsfrom python code.
- We run a python code (part of project documents in src folder) to fetch the results and write them into an output file.
- We use this file as the input for our Trec_eval program along with qrel.txt file to generate MAP value for that corresponding query result pair.
- As we are using luceneMatchVersion 6.0 and later, BM25 is used as the default model and hence no model updates were done. Similarity remains the same for this Model.

Results:



```
007  Q1  653720091985469440  1  64.1993  BM25
007  Q1  651023428795912192  2  14.635002  BM25
007  Q1  652113514954420224  3  13.587628  BM25
007  Q1  647044594832642048  4  12.590918  BM25
007  Q1  648233974901440512  5  11.019852  BM25
007  Q1  653278329521504256  6  10.240753  BM25
007  Q1  652113814402547712  7  10.169548  BM25
007  Q1  652113622110355456  8  10.169548  BM25
007  Q1  652113480770822144  9  10.169548  BM25
007  Q1  651023428301008896  10  9.294943  BM25
007  Q1  652488961747644416  11  9.078576  BM25
007  Q1  651023445241778176  12  8.627371  BM25
007  Q1  651023764843560960  13  8.456354  BM25
007  Q1  651385345192382464  14  8.313677  BM25
007  Q1  651385334593404928  15  8.313677  BM25
007  Q1  651385302234193921  16  8.313677  BM25
007  Q1  651385176652562432  17  8.313677  BM25
007  Q1  651385154393513986  18  8.313677  BM25
007  Q1  651385136722919424  19  8.313677  BM25
007  Q1  647458288800501760  20  8.181845  BM25
```



```
$ ./trec_eval -q -c -M 1000 test/qrel.txt test/BM25_q7.txt
num_ret                 007     20
num_rel                 007     2
num_rel_ret             007     2
map                     007     0.8333
Rprec                   007     0.5000
bpref                   007     0.7500
recip_rank              007     1.0000
iprec_at_recall_0.00    007     1.0000
iprec_at_recall_0.10    007     1.0000
iprec_at_recall_0.20    007     1.0000
iprec_at_recall_0.30    007     1.0000
iprec_at_recall_0.40    007     1.0000
iprec_at_recall_0.50    007     1.0000
iprec_at_recall_0.60    007     0.6667
iprec_at_recall_0.70    007     0.6667
iprec_at_recall_0.80    007     0.6667
iprec_at_recall_0.90    007     0.6667
iprec_at_recall_1.00    007     0.6667
P_5                     007     0.4000
P_10                    007     0.2000
P_15                    007     0.1333
P_20                    007     0.1000
P_30                    007     0.0667
P_100                   007     0.0200
P_200                   007     0.0100
P_500                   007     0.0040
P_1000                  007     0.0020
runid                   all     BM25
num_q                   all     15
num_ret                 all     20
num_rel                 all     2
num_rel_ret             all     2
map                     all     0.0556
gm_map                  all     0.0000
Rprec                   all     0.0333
bpref                   all     0.0500
```

1.3 Implementation of DFR in Solr

In the field of information retrieval, divergence from randomness, one of the very first models, is one type of probabilistic model. It is basically used to test the amount of information carried in the documents.

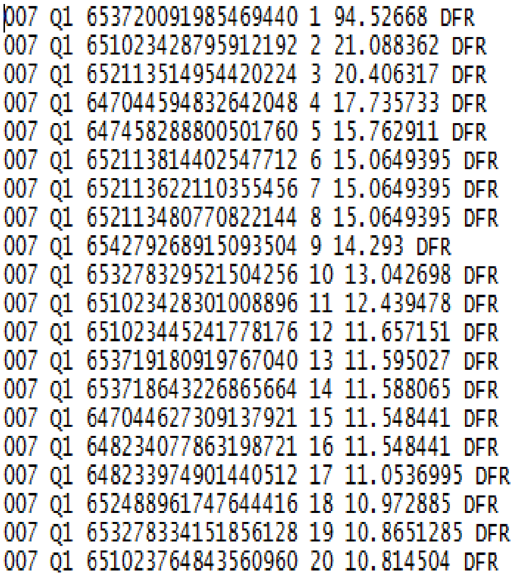
Implementation steps:

- We created core name DFR in SOLR and indexed train.json data provided.
- In Schema.xml file we are updating the model so Solr uses DFR instead of BM25, which is a default model.
- We converted our queries to solr url format which was to fetch the resultsfrom python code.
- We run a python code (part of project documents in src folder) to fetch the results and write them into an output file.
- We use this file as the input for our Trec_eval program along with qrel.txt file to generate MAP value for that corresponding query result pair.
- To update the model, we are adding the below code in our schema file.

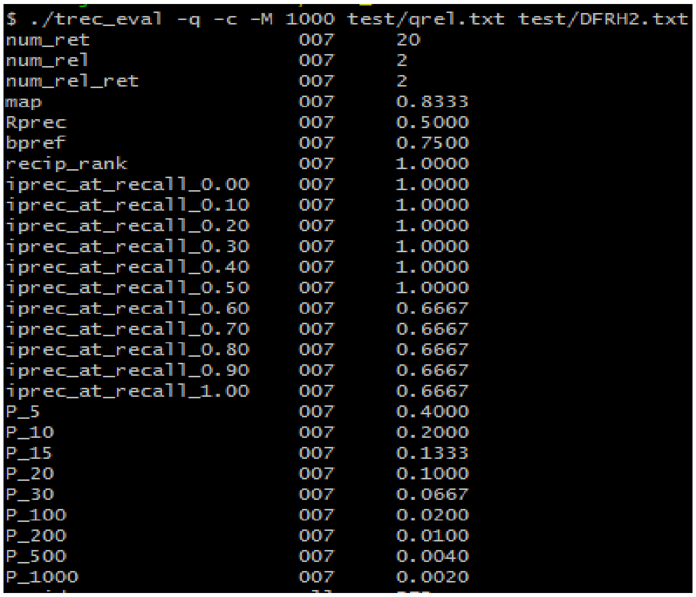
```
<similarity class="solr.DFRSimilarityFactory">
  <str name="basicModel">I(ne)</str>
      <str name="afterEffect">B</str>
      <str name="normalization">H1</str>
  <float name="c">7</float>
</similarity>
```

By updating Similarity class to DFRSimilarityFactory solr uses DFR instead of BM25 which is a default model. Also, we are keeping the basic model as Inverse Document Frequency Model, which we could update to improve our results.

Results:



```
007 Q1 653720091985469440 1 94.52668 DFR
007 Q1 651023428795912192 2 21.088362 DFR
007 Q1 652113514954420224 3 20.406317 DFR
007 Q1 647044594832642048 4 17.735733 DFR
007 Q1 647458288800501760 5 15.762911 DFR
007 Q1 652113814402547712 6 15.0649395 DFR
007 Q1 652113622110355456 7 15.0649395 DFR
007 Q1 652113480770822144 8 15.0649395 DFR
007 Q1 654279268915093504 9 14.293 DFR
007 Q1 653278329521504256 10 13.042698 DFR
007 Q1 651023428301008896 11 12.439478 DFR
007 Q1 651023445241778176 12 11.657151 DFR
007 Q1 653719180919767040 13 11.595027 DFR
007 Q1 653718643226865664 14 11.588065 DFR
007 Q1 647044627309137921 15 11.548441 DFR
007 Q1 648234077863198721 16 11.548441 DFR
007 Q1 648233974901440512 17 11.0536995 DFR
007 Q1 652488961747644416 18 10.972885 DFR
007 Q1 653278334151856128 19 10.8651285 DFR
007 Q1 651023764843560960 20 10.814504 DFR
```



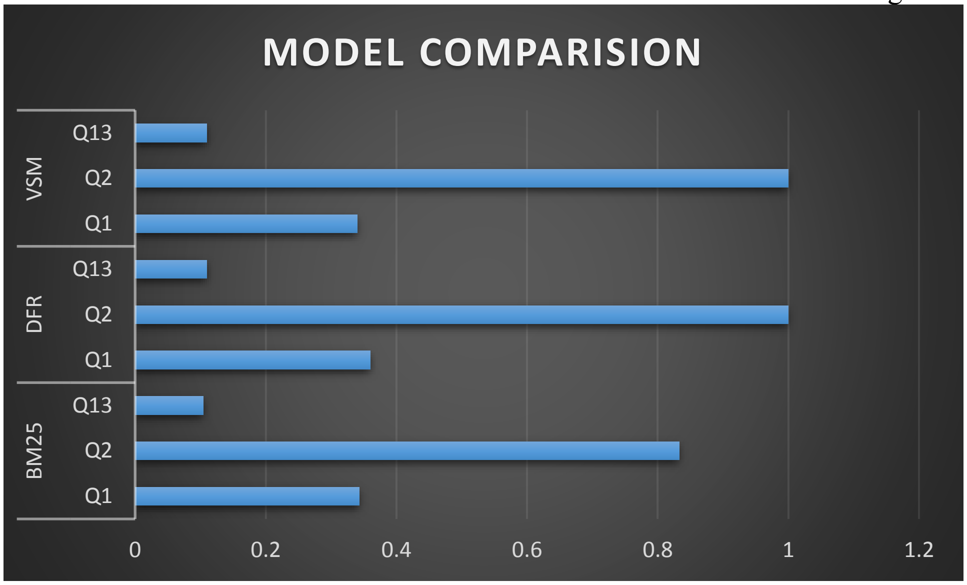
```
$ ./trec_eval -q -c -M 1000 test/qrel.txt test/DFRH2.txt
num_ret                 007     20
num_rel                 007     2
num_rel_ret             007     2
map                     007     0.8333
Rprec                   007     0.5000
bpref                   007     0.7500
recip_rank              007     1.0000
iprec_at_recall_0.00    007     1.0000
iprec_at_recall_0.10    007     1.0000
iprec_at_recall_0.20    007     1.0000
iprec_at_recall_0.30    007     1.0000
iprec_at_recall_0.40    007     1.0000
iprec_at_recall_0.50    007     1.0000
iprec_at_recall_0.60    007     0.6667
iprec_at_recall_0.70    007     0.6667
iprec_at_recall_0.80    007     0.6667
iprec_at_recall_0.90    007     0.6667
iprec_at_recall_1.00    007     0.6667
P_5                     007     0.4000
P_10                    007     0.2000
P_15                    007     0.1333
P_20                    007     0.1000
P_30                    007     0.0667
P_100                   007     0.0200
P_200                   007     0.0100
P_500                   007     0.0040
P_1000                  007     0.0020
```

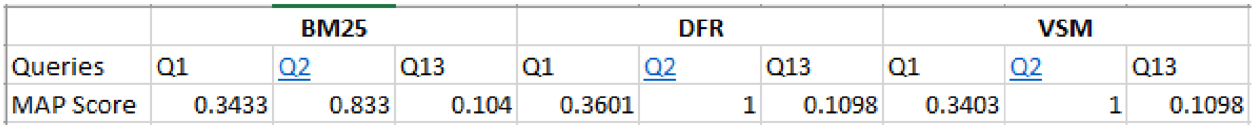
Model Comparisons with MAP Values:

We can compare the efficiency of the three models based on its Mean Average Precision.

STEPS:

- We are executing different queries on the 3 models.
- Result generated(Score and Ranks computed) are then fed to the Trec for evaluation.
- Based on the MAP score evaluated form TREC we get the below results:





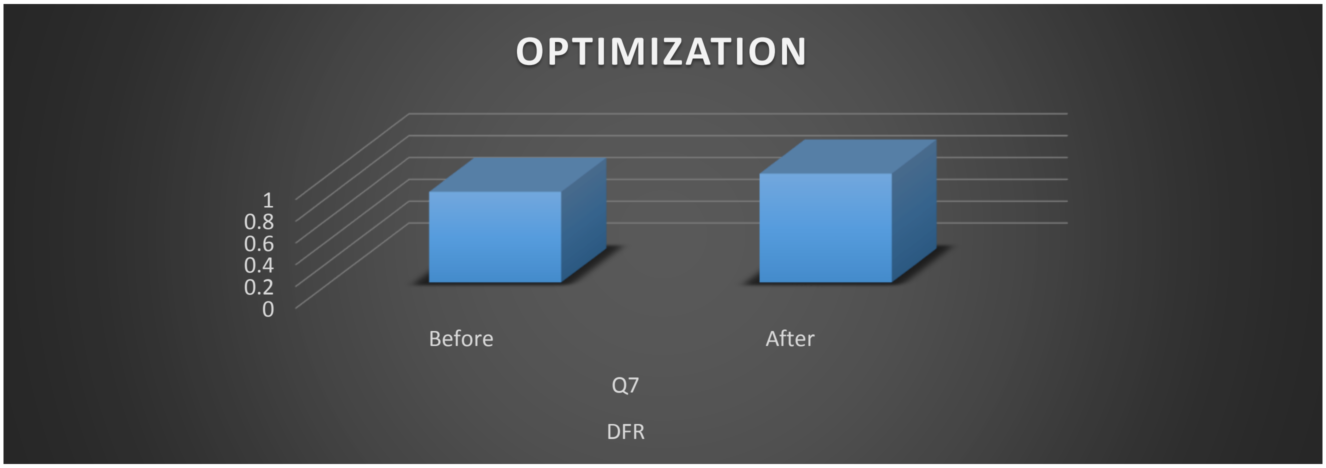
	BM25			DFR			VSM		
Queries	Q1	Q2	Q13	Q1	Q2	Q13	Q1	Q2	Q13
MAP Score	0.3433	0.833	0.104	0.3601	1	0.1098	0.3403	1	0.1098

Measures for Improvement of the result

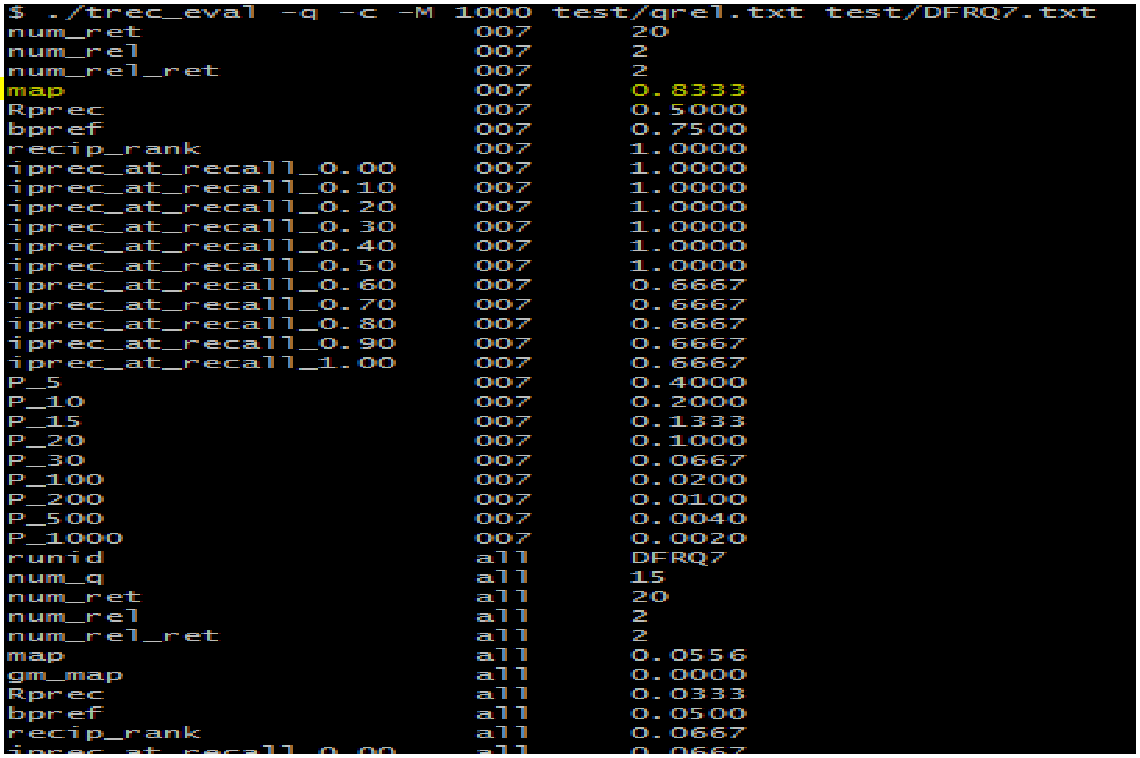
Measures for Improving DFR:

Measure: We are updating the basic model from “Inverse Document Frequency Model” to “Approximation of the binomial” and H2 normalization from H1.We are also updating Laplace to Bernoulli.

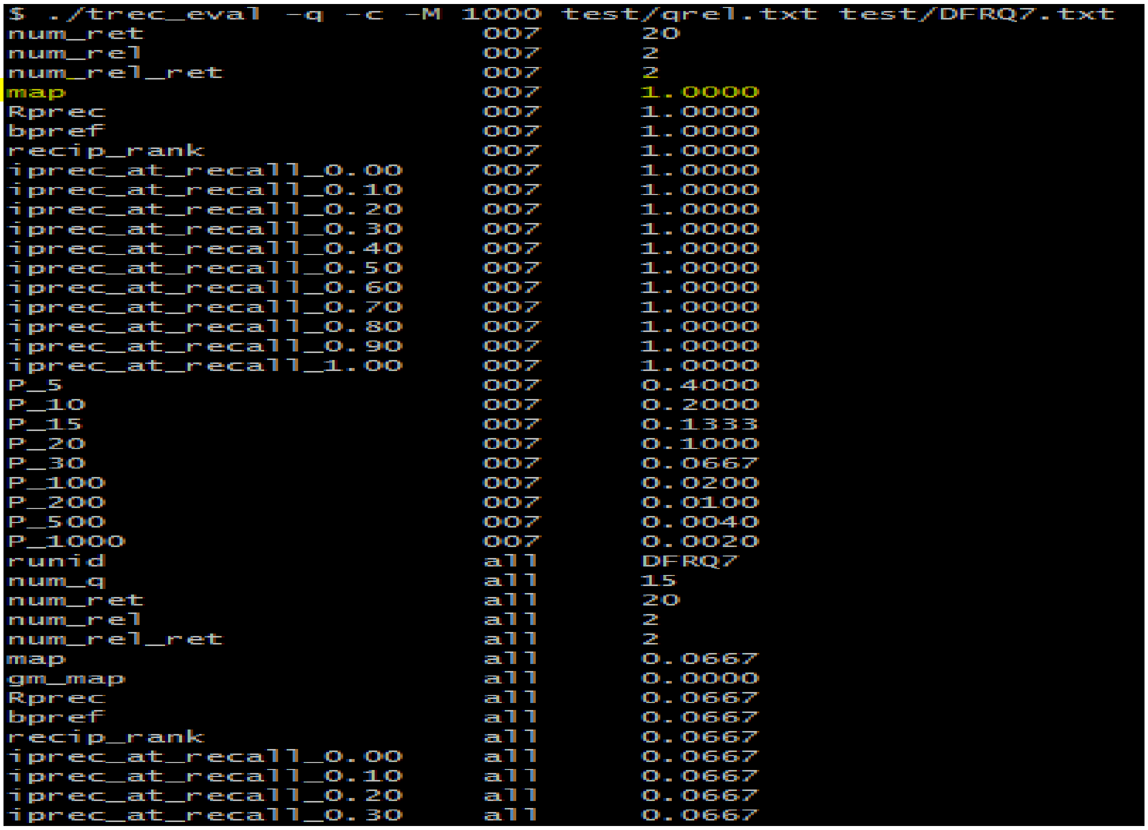
Observation: The MAP precision improved from 0.8333 to 1.0000.



MAP VALUE UPDATES:



```
$ ./trec_eval -q -c -M 1000 test/qrel.txt test/DFRQ7.txt
num_ret                 007     20
num_rel                 007     2
num_rel_ret             007     2
map                     007     0.8333
Rprec                   007     0.5000
bpref                   007     0.7500
recip_rank              007     1.0000
iprec_at_recall_0.00    007     1.0000
iprec_at_recall_0.10    007     1.0000
iprec_at_recall_0.20    007     1.0000
iprec_at_recall_0.30    007     1.0000
iprec_at_recall_0.40    007     1.0000
iprec_at_recall_0.50    007     1.0000
iprec_at_recall_0.60    007     0.6667
iprec_at_recall_0.70    007     0.6667
iprec_at_recall_0.80    007     0.6667
iprec_at_recall_0.90    007     0.6667
iprec_at_recall_1.00    007     0.6667
P_5                     007     0.4000
P_10                    007     0.2000
P_15                    007     0.1333
P_20                    007     0.1000
P_30                    007     0.0667
P_100                   007     0.0200
P_200                   007     0.0100
P_500                   007     0.0040
P_1000                  007     0.0020
runid                   all     DFRQ7
num_q                   all     15
num_ret                 all     20
num_rel                 all     2
num_rel_ret             all     2
map                     all     0.0556
gm_map                  all     0.0000
Rprec                   all     0.0333
bpref                   all     0.0500
recip_rank              all     0.0667
iprec_at_recall_0.00    all     0.0667
```



```
$ ./trec_eval -q -c -M 1000 test/qrel.txt test/DFRQ7.txt
num_ret                 007     20
num_rel                 007     2
num_rel_ret             007     2
map                     007     1.0000
Rprec                   007     1.0000
bpref                   007     1.0000
recip_rank              007     1.0000
iprec_at_recall_0.00    007     1.0000
iprec_at_recall_0.10    007     1.0000
iprec_at_recall_0.20    007     1.0000
iprec_at_recall_0.30    007     1.0000
iprec_at_recall_0.40    007     1.0000
iprec_at_recall_0.50    007     1.0000
iprec_at_recall_0.60    007     1.0000
iprec_at_recall_0.70    007     1.0000
iprec_at_recall_0.80    007     1.0000
iprec_at_recall_0.90    007     1.0000
iprec_at_recall_1.00    007     1.0000
P_5                     007     0.4000
P_10                    007     0.2000
P_15                    007     0.1333
P_20                    007     0.1000
P_30                    007     0.0667
P_100                   007     0.0200
P_200                   007     0.0100
P_500                   007     0.0040
P_1000                  007     0.0020
runid                   all     DFRQ7
num_q                   all     15
num_ret                 all     20
num_rel                 all     2
num_rel_ret             all     2
map                     all     0.0667
gm_map                  all     0.0000
Rprec                   all     0.0667
bpref                   all     0.0667
recip_rank              all     0.0667
iprec_at_recall_0.00    all     0.0667
iprec_at_recall_0.10    all     0.0667
iprec_at_recall_0.20    all     0.0667
iprec_at_recall_0.30    all     0.0667
```

Results:

On updating the fields, we get a better precision. Our Map values improves from 0.8 to 1.0.