CSE 535: INTRODUCTION TO INFORMATION RETRIEVAL

EVALUATION OF INFORMATION RETRIEVAL MODELS (BM25, DFR & LANGUAGE MODEL)

Sriparna Chakraborty
Department of Computer Science
University at Buffalo
Buffalo, NY 14214
UBIT name: sriparna
sriparna@buffalo.edu

INTRODUCTION:

In this project, we deal with the implementation of different IR models such as the BM25 model, DFR models and the Language Models based on Solr and using the twitter data. The results are evaluated using the TREC_eval tool. We are given 15 training queries and 10 test queries in languages – English, German and Russian. Our main goal in this project is to improve the performance of the IR systems by considering primarily the MAP (Mean Average Precision) score as the evaluation measure.

EXPERIMENTATION:

DEFAULT SETUP:

We have created 3 cores – one for each model – BM25, DFR (Divergence from Randomness) and the LM (Language Models) by modifying the schema.xml file for each model. The following similarity classes have been used for each model:

1. BM25 MODEL:

   Okapi BM25 model is a probabilistic information retrieval model which was originally designed for short-length documents. In Solr, the similarity class for this is solr.BM25SimilarityFactory.

2. DFR (DIVERGENCE FROM RANDOMNESS) MODEL:

   Divergence from Randomness is a framework including multiple models and normalization techniques. They all share the same principle: the term may occur in a document randomly, following a certain distribution. The more a document diverges from our configured random distribution, the higher would

be the score. The term-weight is inversely related to the probability of the term-frequency within the document obtained by a model of randomness. In Solr, the similarity class for this is given by solr.DFRSimilarityFactory.

3. LM (LANGUAGE MODELS):

A Language model computes the probability that a query is generated by a document. Language models basically revolve over the idea of smoothing scores based on unseen words (i.e. document length). The similarity class used to implement this model in Solr is given by solr.LMDirichletSimilarityFactory .

Using the default settings which has the standard query parser, we obtained the following MAP values:

i) BM25 - 0.6985 - default k1 =1.2 and b=0.75
ii) DFR - 0.7055 – given defaults = H2 normalization, Basic model G and Bernoulli
iii) LM - 0.6299 - default mu = 2000

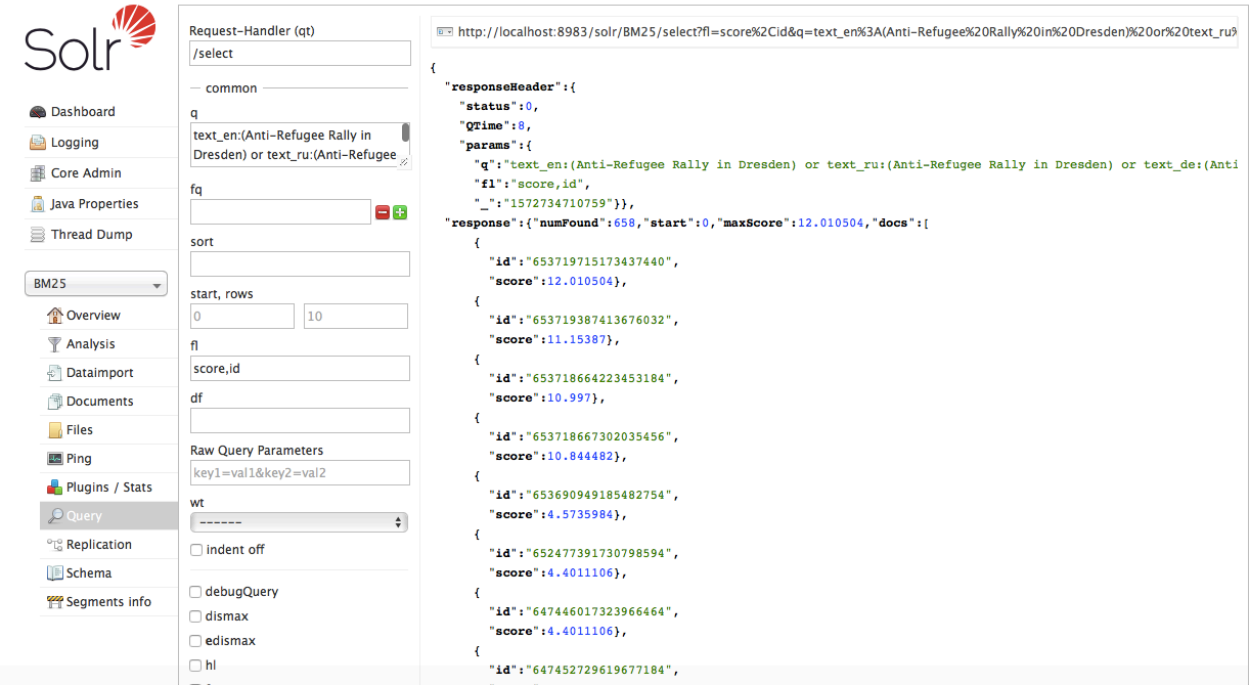
We developed a python script which parses the queries in the query text file sequentially and returns the query results into a new text file for each of the models by running the query url. Three output text files are produced corresponding to each of the cores (BM25, DFR, LanguageModel) and query list.

Trec_eval program is invoked from the terminal (macOS) using the following command:

./trec_eval –q –c –M 1000 qrel.txt <output_query_file.txt>

This command gave the measure result for each query followed by overall performance.

Initially we queried against the given sets of training and testing queries. We observed the tweets which were returned as results and we noticed the changes in results returned and impact on scores by making minor changes to the queries. So query parsing plays a crucial role in the improvement of system performance. In order to enhance the MAP, the settings and parameters of the models are modified. The standard query parser, which is the default query parser, is intolerant of syntax errors as it expects the query to be well-formed. On the other hand, DisMax query parser is known to be a more forgiving parser as it is useful for directly passing in a user-supplied query string. However, we have used the standard query parser in our query parsing tool.



TUNING THE PARAMETER VALUES FOR EACH MODEL:

The tuning process involves the tweaking of parameters to improve the performance of the models.

- TUNING B AND K1 VALUES IN BM25:
  We tried out tuning the parameters of BM25 model which has the default setting: b = 0.75 and k1 = 1.2. For the default setting we got a MAP = 0.6985. Initially we varied the values of b keeping k1 = 1.2 and we observed that the MAP value remained unchanged at MAP = 0.6819. It is generally recommended that the values of b range from 0.5 to 0.8. We chose b = 0.8 and varied k1 from 0.4 to 1.8 as shown below in the table. At k1 = 0.4, A peak in the graph can be observed and the MAP value corresponding to this peak value is 0.7014. We chose the values of b and k1 as 0.8 and 0.4 for good MAP value.

BM25 (value of b is taken as 0.8)	
K1	MAP value
0.4	0.7014
0.6	0.6988
0.8	0.6941
1.0	0.6992
1.2	0.6987
1.4	0.7000
1.6	0.6985
1.8	0.6986

- TUNING NORMALIZATION, AFTEREFFECT AND BASIC MODEL VALUES IN DFR:

  For DFR, we tried out various combinations of the three parameter (Normalization, Aftereffect and Basic Model) values. For example, for Normalization we used H2, H3, for Aftereffect we used Bernoulli(B) and G and I(F) as the Basic models. The default values are H2, B and G for which we got the best MAP score of 0.7055. For all other combinations, MAP score was lesser than the default score.

DFR (AFTEREFFECT B IS CONSTANT)		
NORMALIZATION	BASIC MODEL	MAP VALUE
H2	G	0.7055
H3	I(F)	0.6897
H2	I(F)	0.6982
H3	G	0.6952
Z	I(F)	0.6999
Z	G	0.7019

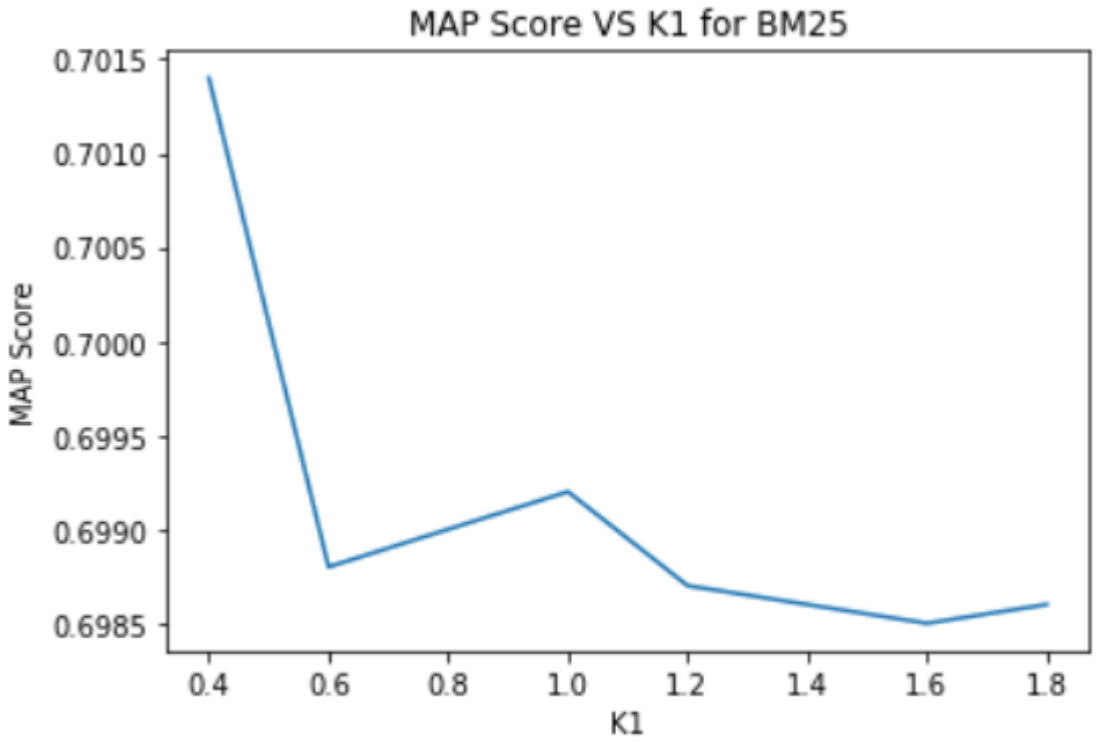
- TUNING MU VALUES IN LEARNING MODEL:

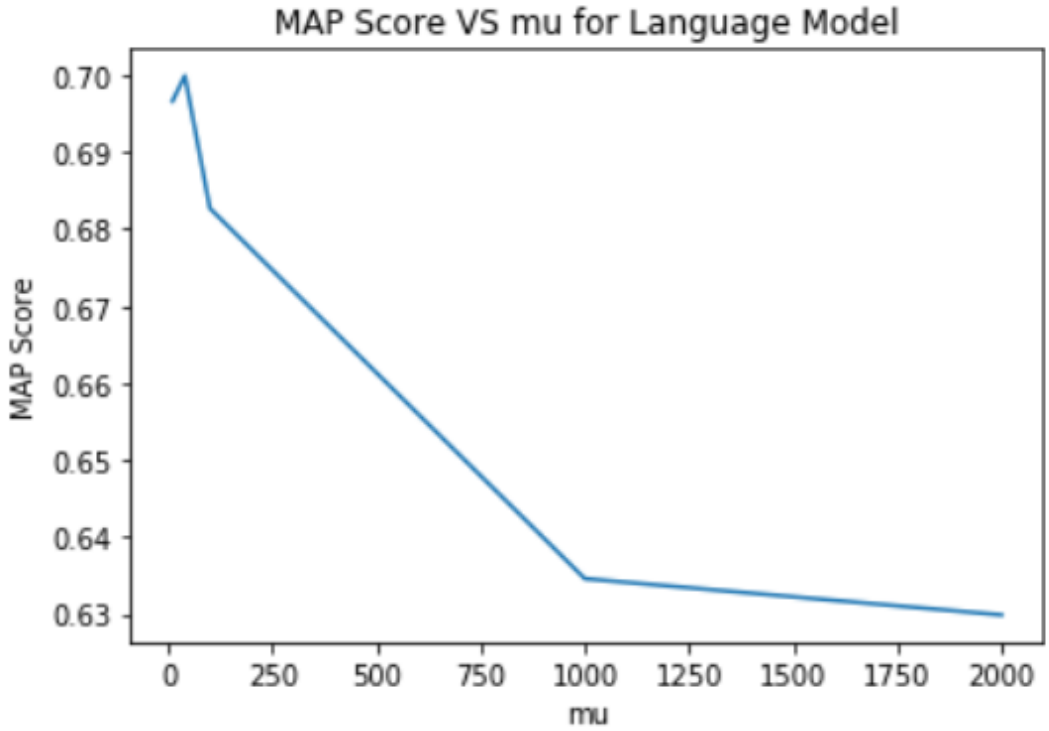
  For Language Model, mu(μ) is the tuning parameter and 2000 is its default value. When mu = 2000, the MAP score has been found to be around 0.6299 which attained a better score as we started decreasing the value of mu. At mu=40, we got the best MAP score of 0.6977.

Language Model	
mu(μ)	MAP values
2000	0.6299
1000	0.6346
100	0.6827
40	0.6977
10	0.6966

Visual Representation:

The graph plots showing the overall performances of BM25 and LM models with parameter tuning are given below:





CONCLUSION:

After enhancing the search engine performance of the models, we have been obtained the following results:

Models	Tuning Parameters		MAP final scores
BM25:	K1	0.4	0.7014
	b	0.8	
DFR:	Normalization	h2	0.7055
	Aftereffect	b	
	Basic Model	g	
LM:	Mu	40	0.6977

The screenshots of the final MAP scores for each model (for training queries) have been provided below:

BM25:

```
P_500                   015     0.0260
P_1000                  015     0.0130
runid                   all     BM25
num_q                   all     15
num_ret                 all     280
num_rel                 all     225
num_rel_ret             all     131
map                     all     0.7014
gm_map                  all     0.6341
Rprec                   all     0.7012
bpref                   all     0.7100
recip_rank              all     1.0000
iprec_at_recall_0.00    all     1.0000
iprec_at_recall_0.10    all     0.9667
iprec_at_recall_0.20    all     0.9286
iprec_at_recall_0.30    all     0.8859
iprec_at_recall_0.40    all     0.8537
iprec_at_recall_0.50    all     0.8242
```

DFR:

```
P_500                   015     0.0260
P_1000                  015     0.0130
runid                   all     DFR
num_q                   all     15
num_ret                 all     280
num_rel                 all     225
num_rel_ret             all     130
map                     all     0.7055
gm_map                  all     0.6382
Rprec                   all     0.6890
bpref                   all     0.7124
recip_rank              all     1.0000
iprec_at_recall_0.00    all     1.0000
iprec_at_recall_0.10    all     0.9667
iprec_at_recall_0.20    all     0.9286
iprec_at_recall_0.30    all     0.9009
```

Language Model/LM:

```
P_100                   015     0.1300
P_200                   015     0.0650
P_500                   015     0.0260
P_1000                  015     0.0130
runid                   all     LanguageModel
num_q                   all     15
num_ret                 all     280
num_rel                 all     225
num_rel_ret             all     130
map                     all     0.6977
gm_map                  all     0.6256
Rprec                   all     0.7011
bpref                   all     0.7036
recip_rank              all     1.0000
iprec_at_recall_0.00    all     1.0000
iprec_at_recall_0.10    all     0.9649
iprec_at_recall_0.20    all     0.9286
iprec_at_recall_0.30    all     0.8802
iprec_at_recall_0.40    all     0.8387
```