Project 3: Evaluation of IR models

Author : Sai Ram Prasanth Tatiraju

Abstract:

We were given twitter data in three languages -English, German and Russian. We indexed the given twitter data by Solr and implemented Vector Space Model, BM25 and Divergence from Randomness Model based on Solr, and evaluated the three sets of results using Trec_eval program. Then I improved the performance in terms of the measure Mean Average Precision (MAP) by making some changes to those models. Finally a comparison between the performance of all the three models before and after tweaking were discussed.

Implementation of BM25 Model in Solr:

Implementation with Default settings:

Solr 6.6.5 implements BM25 model by default. We'll add a similarity class to the schema.xml

```
<similarity class="solr.BM25SimilarityFactory">
   <str name="b">.75</str>
   <str name="k1">1.2</str>
 </similarity>
```

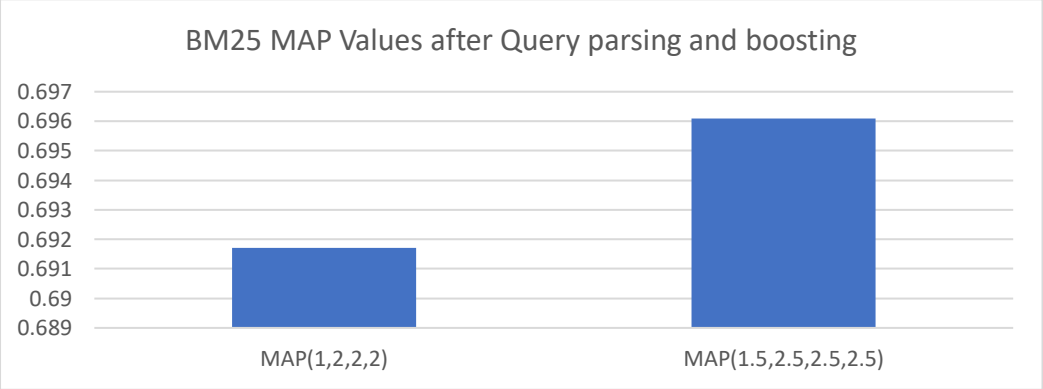
The MAP value for the default implementation of BM25 is : 0.6675

Implementation with Tweaking:

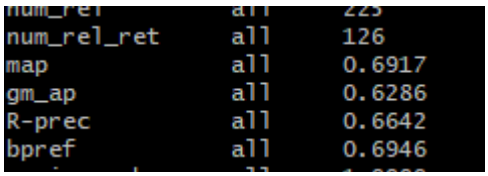
- We did query boosting using various parameters using the query parser Type= dismax.

Below are the various results for different parameters we tweaked.

BM25	Hashtag	text_en	text_ru	text_de	MAP
BM25	1	2	2	2	0.6917
BM25	1.5	2.5	2.5	2.5	0.6961



- We also tweaked the synonyms.txt using different synonyms in all the three languages. But the performance got decreased.
- Screenshot for Max MAP value in BM25



Implementation of DFR Model:

Implementation with Default settings:

Solr 6.6.5 implements DFR model by adding a similarity class to the schema.xml

```
<similarity class="solr.DFRSimilarityFactory">
  <str name="basicModel">P</str>
  <str name="afterEffect">L</str>
  <str name="normalization">H2</str>
  <float name="c">7</float>
</similarity>
```

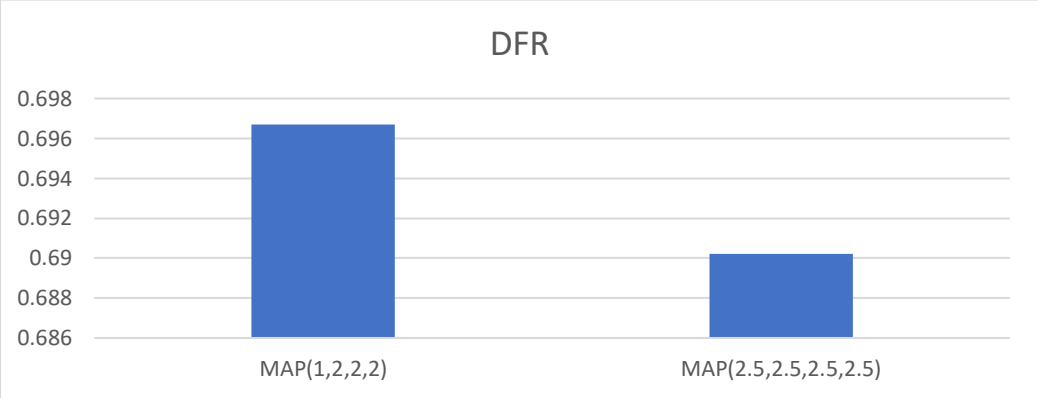
The MAP value for the default implementation of DFR is : 0.6792

Implementation with Tweaking:

- We did query boosting using various parameters using the query parser Type= dismax.

Below are the various results for different parameters we tweaked.

DFR	Hashtag	text_en	text_ru	text_de	MAP
DFR	1	2	2	2	0.6967
DFR	2.5	2.5	2.5	2.5	0.6902



- We also tweaked the synonyms.txt using different synonyms in all the three languages. But the performance got decreased.

Implementation of Vector Space Model

Implementation with Default settings:

Solr 6.6.5 implements VSM model by adding a similarity class to the schema.xml

<similarity class="solr.ClassicSimilarityFactory"/>

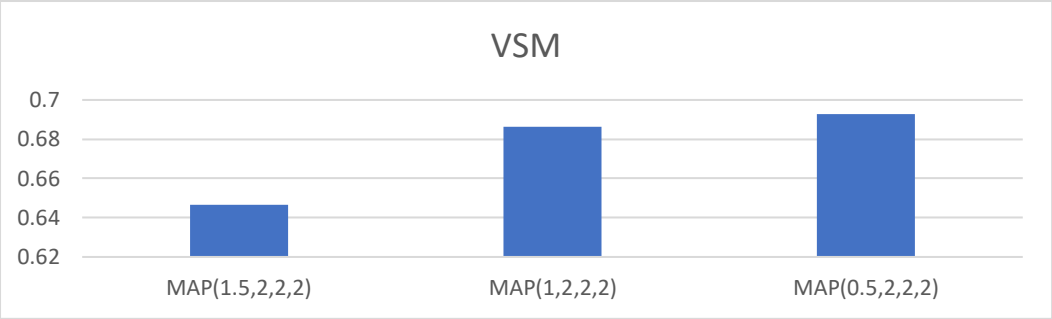
The MAP value for the default implementation of VSM is : 0.6639

Implementation with Tweaking:

- We did query boosting using various parameters using the query parser Type= dismax.

Below are the various results for different parameters we tweaked.

VSM	Hashtag	text_en	text_ru	text_de	MAP
VSM	1.5	2	2	2	0.6467
VSM	1	2	2	2	0.6862
VSM	0.5	2	2	2	0.6928



- We also tweaked the synonyms.txt using different synonyms in all the three languages. But the performance got decreased.

Conclusion:

IR Model	Original MAP	Final MAP
DFR Model	0.6792	0.6967
BM25	0.6675	0.6961
VSM	0.6639	0.6928