Evaluation of IR Models

Deboshree Mazumdar - 50291550-
UBITname - deboshre

November 15, 2018

1 Introduction

The project deals with the evaluation of different IR models using a set of queries on a corpus indexed on Solr. The three different models used are:

1. Vector Space Model
2. BM25 Model
3. Divergence from Randomness Model

2 Implementing the default configurations of VSM

With the default configuration of similarity class in schema.xml, we can implement the vector space model.

```
<similarity class="solr.ClassicSimilarityFactory"/>
```

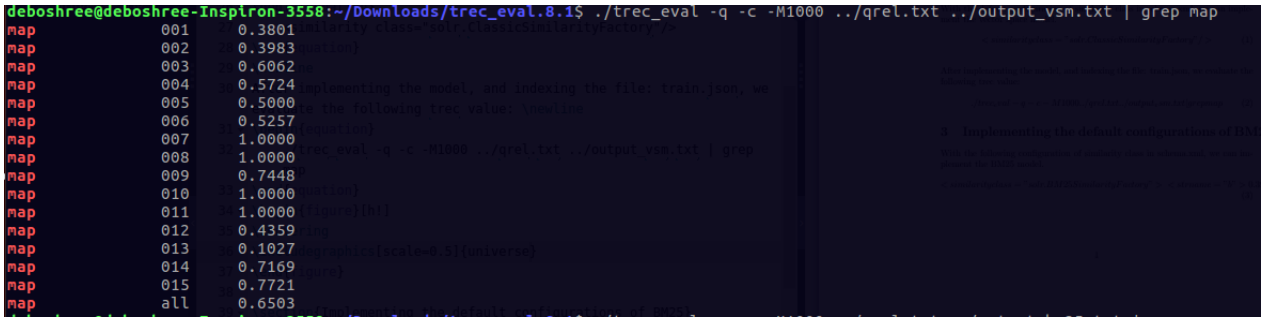
After implementing the model, and indexing the file: train.json, we evaluate the following trec value:

```
./trec_eval -q -c -M1000 ../qrel.txt ../output_vsm.txt | grep map
```

3 Implementing the default configurations of BM25

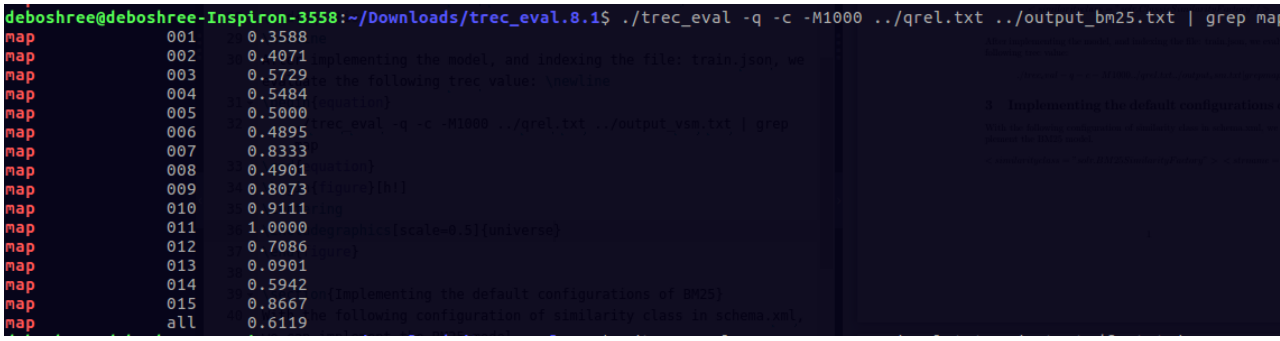
With the following configuration of similarity class in schema.xml, we can implement the BM25 model.

```
<similarity class="solr.BM25SimilarityFactory">
<str name="b">0.80</str>
<str name="k1">1.3</str>
</similarity>\newline
```



After implementing the model, and indexing the file: train.json, we evaluate the following trec value:

```
./trec_eval -q -c -M1000 ../qrel.txt ../output_bm25.txt | grep map
```



4 Implementing the default configurations of DFR

With the following configuration of similarity class in schema.xml, we can implement the DFR model.

```
<similarity class="solr.DFRSimilarityFactory">
  <str name="c">3.0</str>\newline
  <str name="normalization">H2</str>
  <str name="afterEffect">B</str>
  <str name="basicModel">G</str>
</similarity>\newline
```

After implementing the model, and indexing the file: train.json, we evaluate the following trec value:

```
./trec_eval -q -c -M1000 ../qrel.txt ../output_dfr.txt | grep map
```

```
deboshree@deboshree-Inspiron-3558:~/Downloads/trec_eval.8.1$ ./trec_eval -q -c -M1000 ../qrel.txt ../output_dfr.txt | grep map
map                   001     0.3735
map                   002     0.4160
map                   003     0.5471
map                   004     0.5804
map                   005     0.5000
map                   006     0.4914
map                   007     0.8333
map                   008     0.4901
map                   009     1.0000
map                   010     1.0000
map                   011     1.0000
map                   012     0.7303
map                   013     0.0911
map                   014     0.5942
map                   015     0.8667
map                   all     0.6343
```

5 Improvement of models

5.1 VSM

Experiment1: Applying dismax query parser:

```
<requestHandler name="/select" class="solr.SearchHandler">
    <lst name="defaults">
      <str name="defType">edismax</str>
      <str name="qf">text_en^1.5 text_de^1.2 text_ru^0.2</str>
      <str name="echoParams">explicit</str>
      <int name="rows">10</int>
      <!-- <str name="df">text</str> -->
    </lst>
```

The dismax query parser is used for query boosting and to provide different weights to different fields thorugh qf. By assigning different weights, we could see the following changes:

wt(text-en)	wt(text-de)	wt(text-ru)	MAP-initial	MAP-modified
0.8	1.2	0.2	0.6503	0.6641
1.5	1.2	0.2	0.6503	0.6951
1.5	1.2	1.2	0.6503	0.6541
1.0	1.7	0.2	0.6503	0.6832

Experiment2: Applying different filter class:

```
<analyzer type="index">
<charFilter class="solr.PatternReplaceCharFilterFactory"
pattern="([@#])" replacement=""/>
</analyzer type>
```

```
<analyzer type="query">
<charFilter class="solr.PatternReplaceCharFilterFactory"
pattern="([@#])" replacement=""/>
</analyzer type>
```

MAP value almost remained unchanged after applying the filter though it did show a little deflection.

Model	Initial	Optimized
MAP	0.6503	0.6560

5.2 BM25

BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document (e.g., their relative proximity). It is not a single function, but actually a whole family of scoring functions, with slightly different components and parameters.

Experiment1: Changing the values of parameter k1 and b:

k1 - Controls non-linear term frequency normalization (saturation).

b - Controls to what degree document length normalizes the tf values.

```
    <similarity class="solr.BM25SimilarityFactory">\\
    <str name="b">0.69</str>\\
    <str name="k1">1.6</str>\\
  </similarity>
```

Experiment 2: Using URLTokenizer instead of standard tokenizer for text-en for analyzer type query

```
    <analyzer type='query'>
    <tokenizer class="solr.UAX29URLEmailTokenizerFactory"/>
    </analyzer>
```

Model	Initial	Optimized
MAP	0.6119	0.6261

Experiment 3: Using different filters and tokenizers

```
<analyzer type='query'>
<filter class="solr.PatternReplaceFilterFactory"
pattern="([^A Z][^a z])" replacement="" replace="all"/>
</analyzer>
```

The usage of this tokenizer decreased the value of MAP, therefore we refrain from using this.After testing with various tokenizers, we obtained better results with URL Email Tokenizer.

Model	Initial	Optimized
MAP	0.6261	0.6125

5.3 DFR

The DFR has three parameters BasicModel which is the basic model of the information content, AfterEffect specifies the first normalization of information gain and Normalization refers to the second normalization. A parameter 'c' that controls the term frequency normalization with respect to the document length which is specified for normalization H1 and H2.

Experiment1: Tuning the parameters

Normalization	AfterEffect	Basic Model	C	MAP
H2	B	G	3	0.6463
H2	B	G	2	0.6263
H2	B	G	4	0.6342
H2	B	G	5	0.6333

From the above we notice that by decreasing value of c parameter for the H2 normalization parameter, the MAP value increases but when it touches 2.0, the value starts decreasing. Hence we adapt a smaller value of c but not too small.

Experiment2: Using URL tokenizer

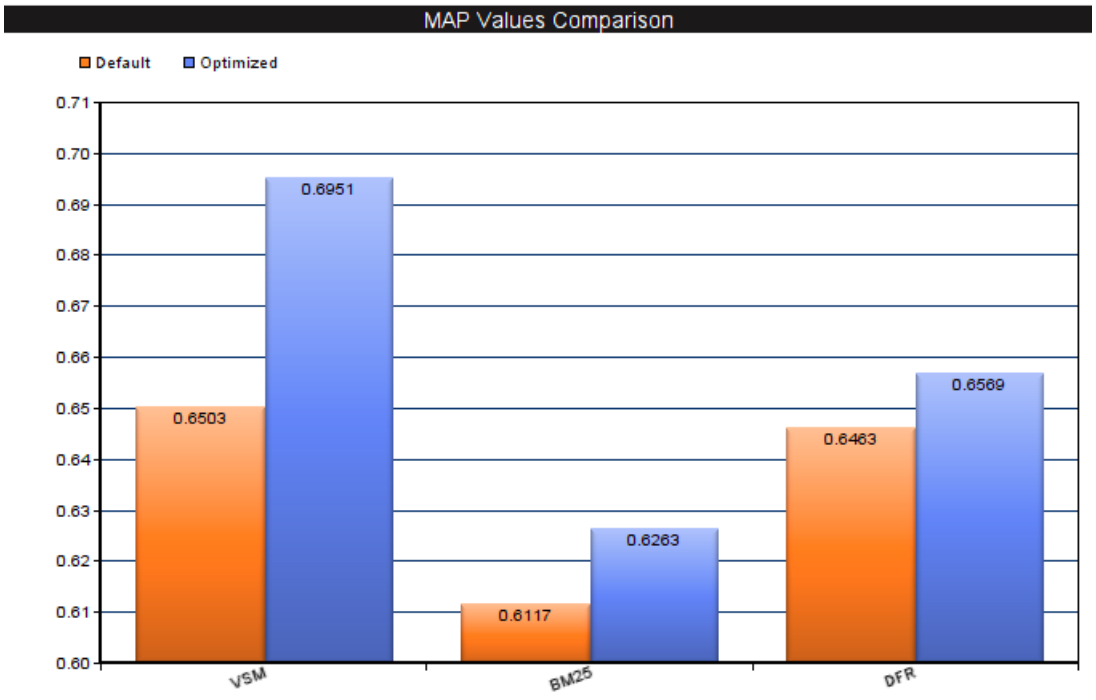
```
<analyzer type=    query    >
<tokenizer class="solr.UAX29URLEmailTokenizerFactory"/>
</analyzer>
```

Model	Initial	Optimized
MAP	0.6463	0.6569

6 Conclusion

After optimization of the default model settings, we obtain the following MAP values:

Model	Initial Value	Optimized Value
VSM	0.6503	0.6951
BM25	0.6117	0.6263
DFR	0.6463	0.6569


MAP Values Comparison
Default
Optimized
0.71
0.70
0.69
0.68
0.67
0.66
0.65
0.64
0.63
0.62
0.61
0.60
0.6503
0.6951
0.6117
0.6263
0.6463
0.6569
VSM
BM25
DFR