PROJECT 3 REPORT

Information Retrieval CSE535

BY

Kapindran Kulandaivelu

UB Person#: 50316983

UBIT Name: kapindra

Table of Contents

BM25 MODEL

Tweaks in Similarity Parameter

BM25 Model Similarity Parameters w.r.t Map score		
'b'	**'k1'**	**map (all)**
0.2	0.5	0.5465
0.5	1.0	0.5502
1.0	1.0	0.5472
1.2	1.0	error
0.75	1.2	0.5511
0.8	1.2	0.5508
0.75	1.25	0.5513
0.85	1.4	0.5482
0.2	1.5	0.5485
0.5	1.5	0.5498
1.0	1.5	0.5473
1.5	1.5	error

```
<similarity class="solr.BM25SimilarityFactory">
    <str name="b">0.75 </str>
    <str name="k1">1.25</str>
  </similarity>
```

```
runid                   all     BM25
num_q                   all     15
num_ret                 all     255
num_rel                 all     225
num_rel_ret             all     94
map                     all     0.5513
gm_map                  all     0.2324
Rprec                   all     0.5410
bpref                   all     0.5500
recip_rank              all     0.9333
iprec_at_recall_0.00    all     0.9333
iprec_at_recall_0.10    all     0.9017
iprec_at_recall_0.20    all     0.7889
iprec_at_recall_0.30    all     0.6598
iprec_at_recall_0.40    all     0.5376
iprec_at_recall_0.50    all     0.5376
iprec_at_recall_0.60    all     0.4900
iprec_at_recall_0.70    all     0.4395
iprec_at_recall_0.80    all     0.3333
iprec_at_recall_0.90    all     0.2926
iprec_at_recall_1.00    all     0.2926
```

The best result is achieved when 'k1'= 1.25 and 'b'=0.75. The resulting MAP(all) score is very close to the MAP(all) for the default parameters [k1 = 1.2; b= 0.75]. (Given in Project_03_Lab_02.pdf).

Default Filters with Changes to stopwords.txt

The schema file wasn't tweaked, rather the language specific stop words were used in StopFilterFactory to enhance the MAP(all) score to 0.5513.

The stopwords in "stopwords_en.txt", "stopwords_ru.txt" and "stopwords_de.txt" from location {E:\UB Fall'19\IR\Project 3\solr-7.4.0\server\solr\BM25\conf\lang\} were copied to the stopword.txt file in \conf.

Results and Screenshots:

runid	all	BM25
num_q	all	15
num_ret	all	255
num_rel	all	225
num_rel_ret	all	94
map	all	0.5513
gm_map	all	0.2324
Rprec	all	0.5410
bpref	all	0.5500
recip_rank	all	0.9333

```
<fieldType name="text_general" class="solr.TextField" positionIncrementGap="100" multiValued="true">
  <analyzer type="index">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.StopFilterFactory" words="stopwords.txt" ignoreCase="true"/>
    <filter class="solr.LowerCaseFilterFactory"/>
    <!-- <filter class="solr.EnglishMinimalStemFilterFactory"/>
    <filter class="solr.EnglishPossessiveFilterFactory"/>
    <filter class="solr.PorterStemFilterFactory"/> -->
  </analyzer>
  <analyzer type="query">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.StopFilterFactory" words="stopwords.txt" ignoreCase="true"/>
    <filter class="solr.SynonymGraphFilterFactory" expand="true" ignoreCase="true" synonyms="synonyms.txt"/>
    <filter class="solr.LowerCaseFilterFactory"/>
    <!-- <filter class="solr.EnglishMinimalStemFilterFactory"/>
    <filter class="solr.EnglishPossessiveFilterFactory"/>
    <filter class="solr.PorterStemFilterFactory"/> -->
  </analyzer>
</fieldType>
```

*** used with Similarity Parameters: k1 = 1.25 and b= 0.75

Schema with EnglishMinimalStemFilter, EnglishPossessiveFilter, PorterStemFilter

We can find an increase of 0.0025 in MAP(all) score.

Results and Screenshots:

runid	all	BM25
num_q	all	15
num_ret	all	255
num_rel	all	225
num_rel_ret	all	94
map	all	0.5538
gm_map	all	0.2341
Rprec	all	0.5384
bpref	all	0.5536
recip_rank	all	0.9333

```
<analyzer type="index">
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.StopFilterFactory" words="stopwords.txt" ignoreCase="true"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.EnglishMinimalStemFilterFactory"/>
  <filter class="solr.EnglishPossessiveFilterFactory"/>
  <filter class="solr.PorterStemFilterFactory"/>
</analyzer>
```

```
<analyzer type="query">
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.StopFilterFactory" words="stopwords.txt" ignoreCase="true"/>
  <filter class="solr.SynonymGraphFilterFactory" expand="true" ignoreCase="true" synonyms="synonyms.txt"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.EnglishMinimalStemFilterFactory"/>
  <filter class="solr.EnglishPossessiveFilterFactory"/>
  <filter class="solr.PorterStemFilterFactory"/>
</analyzer>
```

Schema with EnglishMinimalStemFilter

This gives a better score than the previous BM25 model with 3 filters. Hence, we can conclude that the combination of EnglishMinimalStemFilter with PorterStemFilter is not the best. Further, it is learnt that the positioning of filters with respect to other filters has a significant impact on the MAP(all) score.

Results and Screenshots: (Refer next page)

```
runid                   all     BM25
num_q                   all     15
num_ret                 all     255
num_rel                 all     225
num_rel_ret             all     96
map                     all     0.5539
gm_map                  all     0.2340
Rprec                   all     0.5345
bpref                   all     0.5498
recip_rank              all     0.9333
```

```
<fieldType name="text_general" class="solr.TextField" positionIncrementGap="100" multiValued="true">
  <analyzer type="index">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.StopFilterFactory" words="stopwords.txt" ignoreCase="true"/>
    <!-- <filter class="solr.PorterStemFilterFactory"/> -->
    <filter class="solr.LowerCaseFilterFactory"/>
    <filter class="solr.EnglishMinimalStemFilterFactory"/>
    <!-- <filter class="solr.DaitchMokotoffSoundexFilterFactory" inject="true"/> -->
  </analyzer>
  <analyzer type="query">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.StopFilterFactory" words="stopwords.txt" ignoreCase="true"/>
    <!-- <filter class="solr.PorterStemFilterFactory"/> -->
    <filter class="solr.LowerCaseFilterFactory"/>
    <filter class="solr.EnglishMinimalStemFilterFactory"/>
    <filter class="solr.SynonymGraphFilterFactory" expand="true" ignoreCase="true" synonyms="synonyms.txt"/>
    <!-- <filter class="solr.DaitchMokotoffSoundexFilterFactory" inject="true"/> -->
  </analyzer>
</fieldType>
```

Schema with PhoneticFilterFactory

The Phonetic Filter Factory is used with “soundex” encoder. There are other available encoders such as Metaphone, DoubleMetaphone, CaverPhone and etc. Other filter factories had more or less the similar effect on the MAP(all) score.

```
runid                   all     BM25
num_q                   all     15
num_ret                 all     270
num_rel                 all     225
num_rel_ret             all     91
map                     all     0.5533
gm_map                  all     0.3351
Rprec                   all     0.5514
bpref                   all     0.5804
recip_rank              all     0.9368
```

```
<analyzer type="index">
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.StopFilterFactory" words="stopwords.txt" ignoreCase="true"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.PhoneticFilterFactory" encoder="soundex"/>
</analyzer>
```

Schema with PatternReplaceFilter and EnglishMinimalStemFilter

We use Pattern Replace Filter Factory to replace “@”, “#” and “-“ in the tweets while indexing and also during querying.

Results and Screenshots:

```
<tokenizer class="solr.StandardTokenizerFactory"/>
<filter class="solr.StopFilterFactory" words="stopwords.txt" ignoreCase="true"/>
<filter class="solr.LowerCaseFilterFactory"/>
<filter class="solr.PatternReplaceFilterFactory" pattern="@" replacement="at_"/>
<filter class="solr.PatternReplaceFilterFactory" pattern="#" replacement="hash_"/>
<filter class="solr.PatternReplaceFilterFactory" pattern="-" replacement="_"/>
<filter class="solr.EnglishPossessiveFilterFactory"/>
<filter class="solr.EnglishMinimalStemFilterFactory"/>
```

runid	all	BM25
num_q	all	15
num_ret	all	259
num_rel	all	225
num_rel_ret	all	97
map	all	0.5550
gm_map	all	0.2355
Rprec	all	0.5412
bpref	all	0.5523
recip_rank	all	0.9333

Even though we could get an increase of 0.002 in the MAP(all) score, this increase is not significant enough.

Schema with PhoneticFilter and EnglishMinimalStemFilter

The MAP(all) score has been increased to 0.5649. We can conclude that, since different language are being used, phonetic filter is able to map queries to the indexed documents(tweets) in a improved way compared to previous attempts. ***[BEST SO FAR]***

Results and Screenshots:

runid	all	BM25
num_q	all	15
num_ret	all	270
num_rel	all	225
num_rel_ret	all	91
map	all	0.5649
gm_map	all	0.2340
Rprec	all	0.5656
bpref	all	0.5758
recip_rank	all	0.9333

```
<analyzer type="index">
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.StopFilterFactory" words="stopwords.txt" ignoreCase="true"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.EnglishMinimalStemFilterFactory"/>
  <filter class="solr.PhoneticFilterFactory" encoder="soundex"/>
</analyzer>
<analyzer type="query">
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.StopFilterFactory" words="stopwords.txt" ignoreCase="true"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.SynonymGraphFilterFactory" expand="true" ignoreCase="true" synonyms="synonyms.txt"/>
  <filter class="solr.EnglishMinimalStemFilterFactory"/>
  <filter class="solr.PhoneticFilterFactory" encoder="soundex"/>
</analyzer>
```

Query Optimization

Schema (Schema.xml)

1. Create a custom field of type “text_general”. We are going to refer this field for querying on the documents.

```
<field name="text_unified_lang" type="text_general" multiValued="true" inde
xed="true" stored="true"/>
```

2. Set “text_unified_lang” as the pool of documents(tweets) of all the languages. (“text_en”, “text_de” and “text_ru”)

```
<copyField source="text_de" dest="text_unified_lang"/>
 <copyField source="text_en" dest="text_unified_lang"/>
 <copyField source="text_ru" dest="text_unified_lang"/>
```

Python File (Json_to_Trec.py)

The user query should be modified in such a way that the query irrespective of the language will refer the document collection indexed in “text_unified_lang” field.

```
inurl = 'http://localhost:8983/solr/'+IRModel+'/select?q=text_unified_lang%3A'
 + query + '&qs=100&fl=id%2Cscore%2Ctext_unified_lang&wt=json&indent=true&rows
=20&defType=edismax&qf=text_unified_lang^5'
```

{Before optimization:

```
inurl = 'http://localhost:8983/solr/'+ IRModel +'/select?q=text_en:'+ query +'
&fl=id%2Cscore&wt=json&indent=true&rows=20'
```

}

Query Optimization with PhoneticFilter and EnglishMinimalStemFilter [BEST MAP(all) SCORE]

```
runid                   all     BM25
num_q                   all     15
num_ret                 all     288
num_rel                 all     225
num_rel_ret             all     120
map                     all     0.6810
gm_map                  all     0.6037
Rprec                   all     0.6869
bpref                   all     0.7040
recip_rank              all     1.0000
iprec_at_recall_0.00    all     1.0000
iprec_at_recall_0.10    all     0.9778
```

DFR MODEL

The schema file is unchanged from the BM25 Model. (with only changes to the similarity)

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

The model name should be changed in the python file.

```
#Output File
IRModel ='DFR'                                    #Change according to model
output_filename = IRModel + '_Model.txt'
outf = open(output_filename, 'a+')
```

Result:

```
runid                   all     DFR
num_q                   all     15
num_ret                 all     288
num_rel                 all     225
num_rel_ret             all     124
map                     all     0.6903
gm_map                  all     0.6126
Rprec                   all     0.6885
bpref                   all     0.7198
recip_rank              all     1.0000
iprec_at_recall_0.00    all     1.0000
iprec_at_recall_0.10    all     0.9800
iprec_at_recall_0.20    all     0.9244
iprec_at_recall_0.30    all     0.8978
iprec_at_recall_0.40    all     0.8366
iprec_at_recall_0.50    all     0.7597
iprec_at_recall_0.60    all     0.6055
iprec_at_recall_0.70    all     0.4700
iprec_at_recall_0.80    all     0.4593
iprec_at_recall_0.90    all     0.4238
iprec_at_recall_1.00    all     0.3333
P_5                     all     0.8400
P_10                    all     0.6533
P_15                    all     0.4889
P_20                    all     0.4133
P_30                    all     0.2756
P_100                   all     0.0827
P_200                   all     0.0413
P_500                   all     0.0165
P_1000                  all     0.0083
```

MAP(all)= 0.6903

LM MODEL

Tweaks in Similarity Parameter

LM Model Similarity Parameter w.r.t Map score	
"mu"	**Map(all)**
10	0.5437
50	0.5428
250	0.5379
500	0.5430
1000	0.5375
1500	0.5375
2000	0.5375
2500	0.5375

```
<similarity class="solr.LMDirichletSimilarityFactory">
  <str name="mu">10.0</str>
</similarity>
```

```
runid                   all     LM
num_q                   all     15
num_ret                 all     255
num_rel                 all     225
num_rel_ret             all     91
map                     all     0.5437
gm_map                  all     0.2297
Rprec                   all     0.5458
bpref                   all     0.5429
recip_rank              all     0.9333
iprec_at_recall_0.00    all     0.9333
iprec_at_recall_0.10    all     0.9022
iprec_at_recall_0.20    all     0.7889
iprec_at_recall_0.30    all     0.6598
iprec_at_recall_0.40    all     0.5365
iprec_at_recall_0.50    all     0.5365
iprec_at_recall_0.60    all     0.4796
iprec_at_recall_0.70    all     0.3928
iprec_at_recall_0.80    all     0.3000
iprec_at_recall_0.90    all     0.3000
iprec_at_recall_1.00    all     0.3000
```

Default Schema(with no filter factory)

Optimized Query with PhoneticFilter and EnglishMinimalStemFilter

The schema file is unchanged expect for the similarity parameters.

```
<similarity class="solr.LMDirichletSimilarityFactory">
    <str name="mu">10.0</str>
  </similarity>
```

The model name should be changed in the python file. (json_to_trec.py)

```
#Output File
IRModel ='LM'                                    #Change according to model
output_filename = IRModel + '_Model.txt'
outf = open(output_filename, 'a+')
```

Result:

runid	all	LM
num_q	all	15
num_ret	all	288
num_rel	all	225
num_rel_ret	all	118
map	all	0.6733
gm_map	all	0.5942
Rprec	all	0.6673
bpref	all	0.6936
recip_rank	all	1.0000
iprec_at_recall_0.00	all	1.0000
iprec_at_recall_0.10	all	0.9778
iprec_at_recall_0.20	all	0.9244
iprec_at_recall_0.30	all	0.8828
iprec_at_recall_0.40	all	0.7800
iprec_at_recall_0.50	all	0.7350
iprec_at_recall_0.60	all	0.5972
iprec_at_recall_0.70	all	0.4578
iprec_at_recall_0.80	all	0.4207
iprec_at_recall_0.90	all	0.3689
iprec_at_recall_1.00	all	0.3222
P_5	all	0.8133
P_10	all	0.6467
P_15	all	0.4978
P_20	all	0.3933
P_30	all	0.2622
P_100	all	0.0787
P_200	all	0.0393
P_500	all	0.0157
P_1000	all	0.0079

MAP(all)= 0.6733

RESULT AND OBSERVATION

From the extensive list of filters and analysers, few were implemented and the corresponding MAP(all) score was recorded. From the above results, it can be inferred that the MAP score can be increased by tweaking the schema file and also through query optimization.

Without any query optimization, it is found that the combination of Phonetic Filter and EnglishMinimalStem Filter has the best result (compared with other observation).
By coupling query optimization and the inclusion of FilterFactory in the schema file gives us the best score for MAP(all). The best overall MAP score for each model is given in the table below. It is possible that the MAP score could be further increased by using different combinations of filters or through better query optimization. (due to time constraint, it wasn't possible to experiment with all the solr analysers and filters)

MODEL	MAP(all) Score
BM25	0.6810
DFR	0.6903
LM	0.6733
Dataset: training_tweet.json | Queries: queries.txt	