IR PROJECT 3

Evaluation of IR models

Group Number: 33

Sindhura Uppu (suppu) - 50206730

Anmisha Reddy Ramidi (anmishar) - 50208673

PROJECT OVERVIEW:

The goal of this project is to implement various project models, to evaluate the system and to improve the search result based on the students understanding and its implementation.

IMPLEMENTATION OF THE MODELS:

We implemented the default settings of the VSM, BM25 and DFR model.

The classes used for each model in the schema.xml file are as follows:

- Vector Space Model: org.apache.lucene.search.similarities.ClassicSimilarityFactory
- BM25 Model: org.apache.lucene.search.similarities.BM25SimilarityFactory
- DFR Model: org.apache.lucene.search.similarities.DFRSimilarityFactory

HISTORY OF THE EVALUATION:

- **Recall** is a measure of the ability of a system to present all relevant items.
  recall = number of relevant items retrieved / number of relevant items in collection
- **Precision** is a measure of the ability of a system to present only relevant items.
  precision = number of relevant items retrieved / total number of items retrieved
- The quality of the IR model was initially evaluated using these factors, but we use MAP (Mean Average Precision), GM_AP (Geometeric Mean Precision) and BPREF for this model.

Mean Average Precision: Mean average precision for a set of queries is the mean of the average precision scores for each query, which provides a single-figure measure of quality across recall levels. Among evaluation measures, MAP has been shown to have especially good discrimination and stability. For a single information need, Average Precision is the average of the precision value obtained for the set of top k documents existing after each relevant document is retrieved, and this value is then averaged over information needs. Here the value of k we considered is 20.

$$\mathrm{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \mathrm{Precision}(R_{jk})$$

Vector Space Model:

- Documents and queries are represented as vectors in a common vector space.
- ClassicSimilarity is Lucene's original scoring implementation, based upon the Vector Space Model.
- It can be implemented in Solr using the following declaration in managed-schema.xml in the core:

  <similarity class="solr.ClassicSimilartyFactory">
  </similarity>

  The MAP score for this default implementation is:

BM25 Model:

- Okapi BM25 (BM stands for Best Match) is based on the probabilistic retrieval framework developed by Stephen E. Robertson, Karen Sparck Jones and others.
- It ranks documents based on the query terms appearing in each document and is independent of their relative proximity.
- Given a query Q, containing keywords q1,..,qn, the BM25 score of a document D is:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)},$$

  Where $f(q_i, D)$ is q_i's term frequency in the term frequency in the document D, |D|, is the length of the document D in words, and avgdl is the average document length in the text collection.
  K_1 and b are free parameters, usually chosen, in absence of an advanced optimization.
- Inverse Document Frequency (IDF) weight of the query term q_i is usually computed as:

$$\text{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}$$

  Where N is the total number of documents in the collection, and $n(q_i)$ is the number documents containing q_i.
- It can be implemented in Solr using the following declaration in managed-schema.xml in the core:

  <similarity class="solr.BM25SimilartyFactory">
  <str name="k1">1.5</str>
  <str name="b">0.75</str>
  </similarity>
  K1 can be any value in the range [1.2,2.0] and b = 0.75 usually.

Divergence From Randomness Model:

- Term weights are computed by measuring the divergence between a term distribution produced by a random process and the actual term distribution.
- There are three components in DFR:
  - BasicModel
  - AfterEffect
  - Normalization
- The default values to be taken for these components are already mentioned in the scope of this project, as “BasicModelG”, “Bernoulli” first normalization and “H2” second normalization.
- It can be implemented in Solr using the following declaration in managed-schema.xml in the core:

<similarity class=”solr.DFRSimilarityFactory”>
<str name="c">6.75</str>
<str name="normalization">H2</str>
<str name="afterEffect">B</str>
<str name="basicModel">G</str>
</similarity>

The MAP score for this default implementation is:

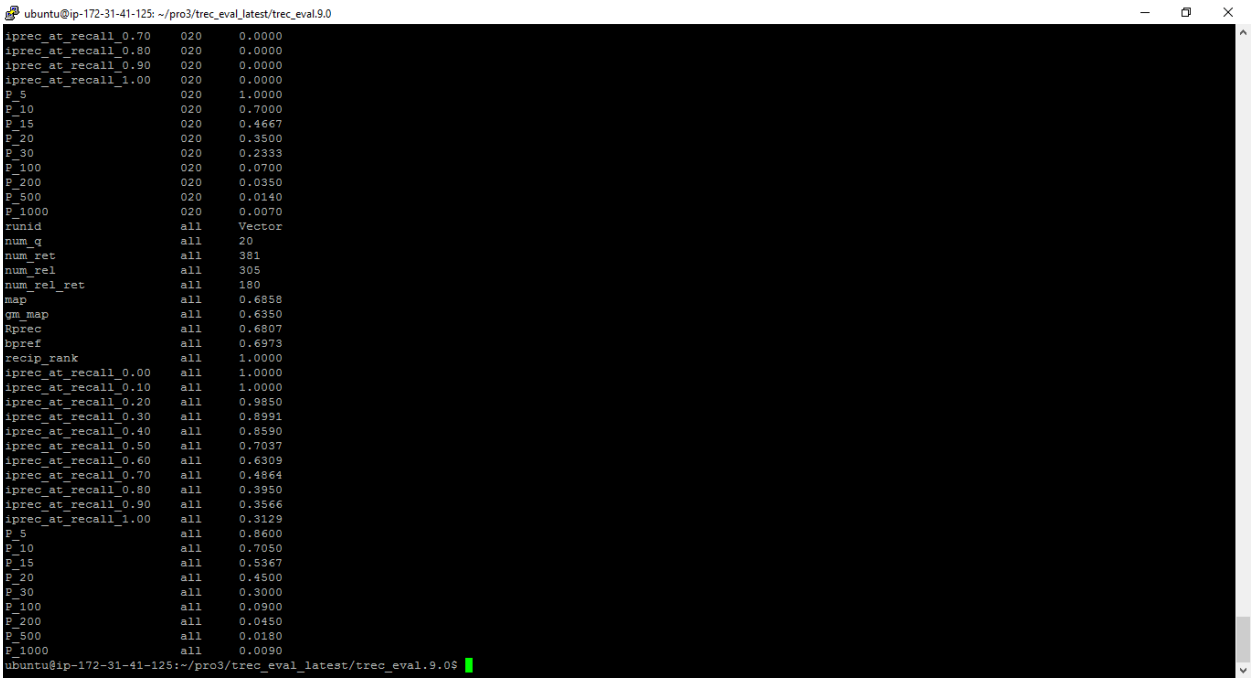
Optimized Scores:

Vector Space Model:

The MAP value for the given set of queries using the indexed tweets is found to be optimized when synonyms, stopwords and stemmers are removed from the data while being indexed.

```
<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"/>
  </analyzer>
  <analyzer type="query">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.KStemFilterFactory"/>
    <filter class="solr.StopFilterFactory" words="lang/stopwords_en.txt" ignoreCase="true"/>
    <filter class="solr.StopFilterFactory" words="lang/stopwords_de.txt" ignoreCase="true"/>
    <filter class="solr.StopFilterFactory" words="lang/stopwords_ru.txt" ignoreCase="true"/>
    <filter class="solr.SynonymFilterFactory" expand="true" ignoreCase="true" synonyms="synonyms.txt"/>
    <filter class="solr.LowerCaseFilterFactory"/>
  </analyzer>
</fieldType>
```

```
Last login: Fri Nov 11 00:41:26 2016 from 69.12.22.4
ubuntu@ip-172-31-41-125:~$ cd pro3
ubuntu@ip-172-31-41-125:~/pro3$ cd trec_eval_latest
ubuntu@ip-172-31-41-125:~/pro3/trec_eval_latest$ cd trec_eval.9.0
ubuntu@ip-172-31-41-125:~/pro3/trec_eval_latest/trec_eval.9.0$ ./trec_eval -q -c -M20 qrel.txt vsm2_1.txt
```



MAP for this implementation is 0.6858.

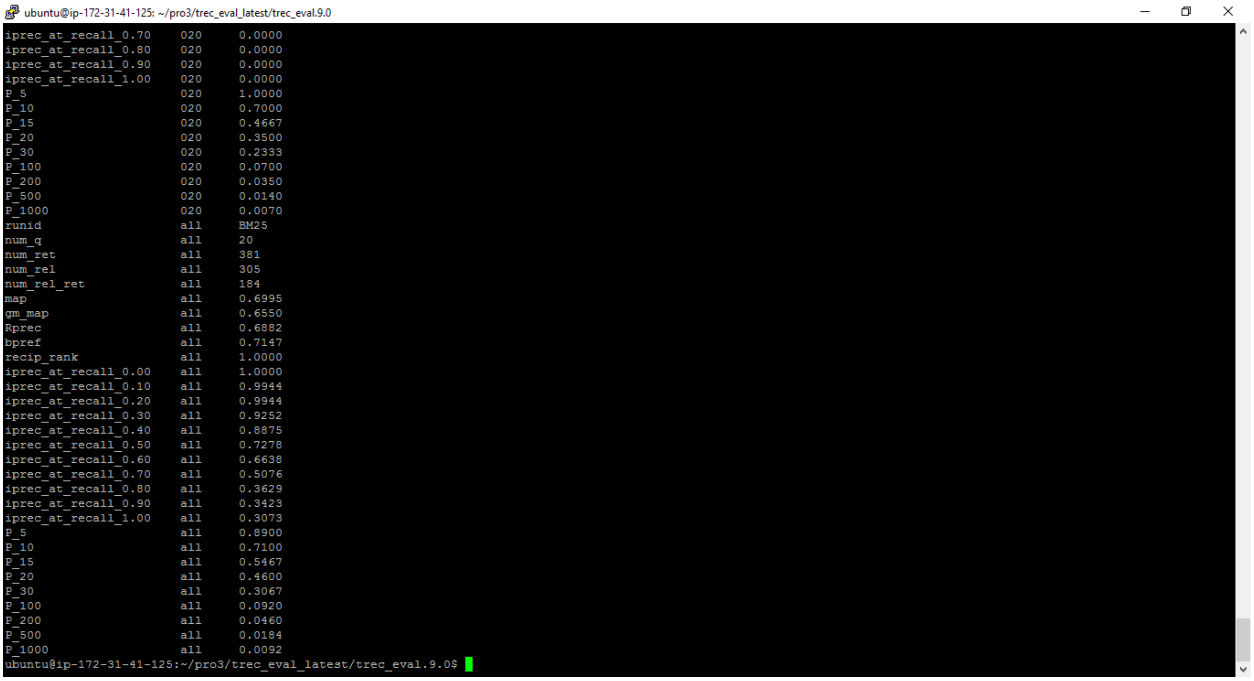
BM25 Model:

Synonyms, stopwords and stemming tokenizers are included in the optimized implementation of this model too so that query expansion will help improve the performance of the model.

The following parameters are considered to be optimized after a number of trials done to improve the performance of the model:

```
<schema name="example-data-driven-schema" version="1.6">
  <uniqueKey>id</uniqueKey>
  <similarity class="solr.BM25SimilarityFactory">
    <str name="k1">0.6</str>
    <str name="b">0.2</str>
  </similarity>
```

```
Solr schemaless example launched successfully. Direct your Web browser to http:/
/localhost:8984/solr to visit the Solr Admin UI
java -classpath /home/ubuntu/solr/solr-6.2.0/dist/solr-core-6.2.0.jar -Dauto=yes
 -Dport=8984 -Dc=core1 -Ddata=files org.apache.solr.util.SimplePostTool train.js
on
SimplePostTool version 5.0.0
Posting files to [base] url http://localhost:8984/solr/core1/update...
Entering auto mode. File endings considered are xml,json,jsonl,csv,pdf,doc,docx,
ppt,pptx,xls,xlsx,odt,odp,ods,ott,otp,ots,rtf,htm,html,txt,log
POSTing file train.json (application/json) to [base]/json/docs
1 files indexed.
COMMITting Solr index changes to http://localhost:8984/solr/core1/update...
Time spent: 0:00:02.885
ubuntu@ip-172-31-41-125:~/solr/solr-6.2.0$
ubuntu@ip-172-31-41-125:~/solr/solr-6.2.0$ cd
ubuntu@ip-172-31-41-125:~$ cd pro3
ubuntu@ip-172-31-41-125:~/pro3$ cd trec_eval_latest
ubuntu@ip-172-31-41-125:~/pro3/trec_eval_latest$ cd trec_eval.9.0
ubuntu@ip-172-31-41-125:~/pro3/trec_eval_latest/trec_eval.9.0$ ./trec_eval -q -c -M20 qrel.txt bm25_final.txt
num_ret                 001     20
num_rel                 001     20
```



The MAP score is 0.6995 for this set of parameters.

Divergence From Randomness (DFR) Model:

Synonyms, stopwords and stemming tokenizers are included in the optimized implementation of this model too so that query expansion will help improve the performance of the model.

The following parameters are considered to be optimized after a number of trials done to improve the performance of the model:

```
<schema name="example-data-driven-schema" version="1.6">
  <uniqueKey>id</uniqueKey>
  <similarity class="solr.DFRSimilarityFactory">
    <str name="c">8</str>
    <str name="normalization">H2</str>
    <str name="afterEffect">B</str>
    <str name="basicModel">G</str>
  </similarity>
```

```
Posting files to [base] url http://localhost:8984/solr/core1/update...
Entering auto mode. File endings considered are xml,json,jsonl,csv,pdf,doc,docx,ppt,pptx,xls,xlsx,odt,odp,ods,ott,otp,ots,rtf,htm,html,txt,log
POSTing file train.json (application/json) to [base]/json/docs
1 files indexed.
COMMITting Solr index changes to http://localhost:8984/solr/core1/update...
Time spent: 0:00:02.806
ubuntu@ip-172-31-41-125:~/solr/solr-6.2.0$ cd
ubuntu@ip-172-31-41-125:~$ cd pro3
ubuntu@ip-172-31-41-125:~/pro3$ cd trec_eval_latest
ubuntu@ip-172-31-41-125:~/pro3/trec_eval_latest$ cd trec_eval.9.0
ubuntu@ip-172-31-41-125:~/pro3/trec_eval_latest/trec_eval.9.0$ ./trec_eval -q -c -M20 qrel.txt dfr_final1.txt
num_ret                 001     20
num_rel                 001     20
```

```
ubuntu@ip-172-31-41-125: ~/pro3/trec_eval_latest/trec_eval.9.0
iprec_at_recall_0.70    020     0.0000
iprec_at_recall_0.80    020     0.0000
iprec_at_recall_0.90    020     0.0000
iprec_at_recall_1.00    020     0.0000
P_5                     020     1.0000
P_10                    020     0.7000
P_15                    020     0.4667
P_20                    020     0.3500
P_30                    020     0.2333
P_100                   020     0.0700
P_200                   020     0.0350
P_500                   020     0.0140
P_1000                  020     0.0070
runid                   all     DFR
num_q                   all     20
num_ret                 all     381
num_rel                 all     305
num_rel_ret             all     184
map                     all     0.7043
gm_map                  all     0.6589
Rprec                   all     0.6925
bpref                   all     0.7316
recip_rank              all     1.0000
iprec_at_recall_0.00    all     1.0000
iprec_at_recall_0.10    all     0.9944
iprec_at_recall_0.20    all     0.9944
iprec_at_recall_0.30    all     0.9278
iprec_at_recall_0.40    all     0.8758
iprec_at_recall_0.50    all     0.7292
iprec_at_recall_0.60    all     0.6608
iprec_at_recall_0.70    all     0.5093
iprec_at_recall_0.80    all     0.4300
iprec_at_recall_0.90    all     0.3473
iprec_at_recall_1.00    all     0.3123
P_5                     all     0.8800
P_10                    all     0.7100
P_15                    all     0.5500
P_20                    all     0.4600
P_30                    all     0.3067
P_100                   all     0.0920
P_200                   all     0.0460
P_500                   all     0.0184
P_1000                  all     0.0092
ubuntu@ip-172-31-41-125:~/pro3/trec_eval_latest/trec_eval.9.0$
```

The MAP score for this set of parameters is 0.7043.

Improving the performance of the models:

The following methods are used to improve the performance of the model:

- KStemFilterFactory: Using this factory reduced any of the forms of a verb/word to its elemental root increasing number of true matches.
- SynonymFilterFactory: Various query terms were expanded using relevant and synonyms to improve the result set. Below are a few terms added:

- GB – gigabyte, gib, gigabytes
- Television – TV, Televisions
- Million – Mio
- War – Krieg, война
- civil war - Bürgerkrieg, гражданская война
- унисских, Tunisian, tunesische
- Tech – Airbnb, Instacart, Kickstarter

- Mapping variations in this way increased the term weighting calculated by BM25 and DFR models. In all the models, equating Tech to the company names Airbnb, Instacart, Kickstarter boosted documents which did not mention these names explicitly but used the generic term Tech.
- Dismax Query Parser: It searches for individual terms across several fields using different weights based on the significance level.
- Apart from these parsers, we have changed the values of the different parameters involved and obtained specific results that are documented below for each model:

DFR Similarity Model:

Basic Model	After Effect	Normalization	c	MAP
P	L	H2	7	0.6581
Be	L	H2	1	0.6561
G	B	H2	0.7	0.6500
G	B	H2	1.2	0.6596
G	B	H2	0.7	0.6499
G	B	H2	8	0.6618
G	B	H2	9	0.6615
G	B	H2	13	0.6595
G	B	H2	8.35	0.6618
I(n)	B	H2	6.75	0.6653
I(ne)	B	H2	6.75	0.6639
I(n)	L	H2	6.75	0.6584
I(n)	B	H1	6.75	0.6654
G	B	H2	6.75	0.6930 (A)
G	B	H2	8	0.7043 (A)

- Although it is recommended to use “BasicModelG”, “Bernoulli” first normalization and “H2” second normalization, we have tried out various other parameters such as “BasicModelP”, Poisson approximation of the Binomial,

"BasicModelBE", Limiting form of Bose – Einstein, "BasicModelIn", Inverse document frequency, "BasicModelIne", Inverse expected document frequency and "Laplace" effect.

- In the last 2 trials, we have included an analyzer in the text_general that indexes the data by removing stopwords and after performing stemming and synonym filtering.
- The score for the model is found to be optimized using the last set of parameters, and the score is found to be 0.7043.

BM25 Similarity Model:

K1	b	MAP
0.4	0.2	0.6585
0.6	0.7	0.6581
1	0.2	0.6571
1.9	0.2	0.6564
0.6	0.3	0.6575
0.65	0.35	0.6604
0.45	0.15	0.6580
0.68	0.15	0.6569
0.1	0.0	0.6637
0.4	0.0	0.6676
0.93	0.1	0.6579
0.63	0.05	0.6558
0.33	0.05	0.6569
0.0	0.4	0.6622
0.01	0.4	0.6940 (A)
0.6	0.2	0.6995 (A)

- After including the stemmers, synonym and stopword filters, the MAP score is found to be the highest at $k_1 = 0.6$ and $b = 0.2$, i.e., MAP = 0.6995.

Vector Space Model:

- First, we have included synonym filter and KStemFilterFactory, Stopwords for language en. The MAP score is obtained to be 0.6776.
- After including stopwords for languages 'de' and 'ru', the MAP score is obtained to be 0.6858.
- We have tried implementing with various parameters in dismax parser and copy fields, but the scores seem to be in the same range. Hence the optimized value is 0.6858.