IR Project 3 Report

By:-
Shubham Sharma
Anurag Dixit

Introduction

The scope of this document is to demonstrate the life cycle of implementation of this project. This project requires the implementation and evaluation of 3 different IR models (Vector Space Model, Best Matching 25 and Divergence From Randomness) to enhance the performance of our IR system. The projects involves twitter tweets collected and indexed over topic “Russian intervention in refugee crisis” in 3 different languages. The objective of this project is to improve the performance of IR system via different models and measure the performance via TREC Eval system.

In order to complete the project various approaches were taken. They have been enumerated and illustrated as follows:

Before we proceed further, we would like to state the initial state of our system upon which we will make improvements.

1. The key configuration of our schema are as listed below. (Please find the entire file in src folder)

```
<fieldType name="text_ru" class="solr.TextField" positionIncrementGap="100">
<analyzer>
<tokenizer class="solr.StandardTokenizerFactory"/>
</analyzer>
</fieldType>
<fieldType name="text_de" class="solr.TextField" positionIncrementGap="100">
<analyzer>
<tokenizer class="solr.StandardTokenizerFactory"/>
</analyzer>
</fieldType>
<fieldType name="text_en" class="solr.TextField" positionIncrementGap="100">
<analyzer>
<tokenizer class="solr.StandardTokenizerFactory"/>
</analyzer>
</fieldType>
```

```
<field name="created_at" type="date" indexed="true" stored="true"/>
<field name="id" type="string" multiValued="false" indexed="true" required="true" stored="true"/>
<field name="lang" type="string" indexed="true" stored="true"/>
<field name="text_de" type="text_de" indexed="true" stored="true"/>
<field name="text_en" type="text_en" indexed="true" stored="true"/>
<field name="text_ru" type="text_ru" indexed="true" stored="true"/>
<field name="tweet_hashtags" type="strings"/>
<field name="tweet_urls" type="strings"/>
```

1. Initially the query being fired is of format:-

```
defType=dismax&fl=score,id&indent=on&q=' + query +
'&rows=100&wt=json&qf=text_en%20text_ru%20text_de%20tweet_urls%20tweet_hashtags'
```

So

1. The overall MAP values under the initial states of the model (Refer Section 1) are:-

IR Model	MAP Value (1000 rows)
BM25	0.6031
VSM	0.5569
DFR	0.6024

Section 1

Implementations of IR Models in Apache Solr:-

1. VSM Model:-

Apache Solr sets VSM as the model of similarity comparisons if the following global directive is put in schema.xml file:-

<similarity class="ClassicSimilarityFactory" /> -- extends ClassicSimilarity class

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

1. BM25 Model:-

Apache Solr sets BM25 as the model of similarity comparisons if the following global directive is put in schema.xml file:-

<similarity class="solr.BM25SimilarityFactory" > -- extends
<float name="k1">1.2</float> -- default values set for initial setup
<float name="b">0.75</float> --default values set for initial setup
</similarity>

```
<similarity class="solr.BM25SimilarityFactory" >
<float name="k1">1.2</float>
<float name="b">0.75</float>
</similarity>
```

1. DFR Model:-

Apache Solr sets DFR as the model of similarity comparisons if the following global directive is put in schema.xml file:-

<similarity class="solr.DFRSimilarityFactory">
<str name="basicModel">G</str>
<str name="afterEffect">B</str>
<str name="normalization">H2</str>
<float name="c">1</float>
</similarity>

```
<similarity class="solr.DFRSimilarityFactory">
<str name="basicModel">G</str>
<str name="afterEffect">B</str>
<str name="normalization">H2</str>
<float name="c">1</float>
</similarity>
```

Section 2

Improving IR System:

In order to improve the IR system in terms of MAP value following approaches have been attempted during the course of this phase. The mentions are as follows:

1. **Query matching after Lemmatization, Stemming, lowercasing**

   Both the queries and fields (text_xx and hashtags) were subjected to filters like stopwords removal, lowercasing etc to reduce the contribution of less relevant words in scoring (correctly predicting the about-ness of a tweet).

IR Model	Previous MAP Value	New MAP Value
BM25	0.6031	0.7338

VSM	0.5569	0.7257
DFR	0.6024	0.7393

2. Boosting tweet fields - hashtags

We thought about positively boosting field tweet_hashtags with the intuitive idea that if a tweet contains a hashtag which matches with query term(s), then that tweet must be more relevant. However, any boost value greater than 1 was actually reducing the MAP. On the other hand, a boost value of 0.5, i.e. a negative boost, increased the MAP of the system pertaining to all the models. We attribute this behavior to the difference in lengths of tweet_xx fields and the tweet_hashtags field.

IR Model	Previous MAP Value	New MAP Value
BM25	0.7338	0.7401
VSM	0.7257	0.7385
DFR	0.7393	0.7450

3. Tuning of k1, b hyperparameter of BM25 model

For the values k1 = 1.56, b= 0.45 the best MAP value for BM25 Model is obtained.

IR Model	Previous MAP Value	New MAP Value
BM25	0.7401	0.7445
VSM	0.7385	0.7385
DFR	0.7450	0.7450

4. Tuning of H1 parameter of normalization for DFR Model

For c=0.95 where c represents uniform distribution of term frequency, we obtain the best MAP value for DFR Model.

IR Model	Previous MAP Value	New MAP Value
BM25	0.7445	0.7445
VSM	0.7385	0.7385
DFR	0.7450	0.7456

5. Thesaurus/Synonyms

In our experimentation, we found that using synonyms is substantially advantageous for VSM model. However, this methodology turns out to be counter-productive for the DFR and BM25 models. So we apply synonyms only for VSM models.

IR Model	Previous MAP Value	New MAP Value
BM25	0.7445	0.7445
VSM	0.7385	0.7502
DFR	0.7456	0.7456

6. Some approaches to increase MAP which did not work:-

Similar to tweet_hashtags, we tried boosting tweet_urls but we didn't find any major change in MAP values for any of the 3 models. We tried to get more weight to a tweet which matched the query term in exact case (by copying the text_xx and tweet_hashtags to respective duplicate fields whilst preserving the case and boosting weights for these fields). However, the overall MAP value actually decreased.

The Following table illustrates our best effort MAP values for the 3 IR Models:-

(Please note that for training we are returning 1000 rows in our query, but for test_queries we have returned 20 rows as instructed)

IR Model	MAP values
BM25	0.7445
VSM	0.7502
DFR	0.7456

For 20 rows in training set, following are the MAP values:-

IR Model	MAP values
BM25	0.6969
VSM	0.6927
DFR	0.6913