Project 3: Evaluation of IR models

Anish Gadekar
UB Person No: 50291289
Department of Computer Science
University at Buffalo,
Buffalo, NY 14260
anishraj@buffalo.edu

1 Introduction

The project requires us to evaluate various IR models such as Vector Space Model(VSM), Okapi BM25 and Divergence-from-randomness(DFR) model using data extracted from twitter which is in English, German and Russian languages. The aim of the project is to improve the performance of the IR Model by tuning parameters and improving queries. The performance will be evaluated on the basis of MAP(Mean Average Precision) values obtained using the Trec Eval program.

2 Implementation

2.1 IR Models

Three cores were created on Solr – one for each model: BM25, VSM and DFR. The following similarity classes are added in the managedschema.xml file of each Solr core.
1) BM25: solr.BM25SimilarityFactory

BM25 is a probabilistic information retrieval model intended for short-length documents. It is the default for LuceneMatch Version 6.0 and above.

2) VSM: solr.ClassicSimilarityFactory

In VSM, queries are documents are represented as a multi-dimensional space where each term is represented and weights are evaluated on the basis of tf-idf values. VSM was the default similarity model for LuceneMatch version 6.0 and below.

3) DFR: solr.DFRSimilarityFactory

DFR is inversely related to probability of term frequency within the document which is obtained by a randomness model.

2.2 Obtaining map values

We used a python script to convert the query results of Solr in a format acceptable for the trec_eval function. The python script used the queries given

in queries.txt to query the solr index and the obtained result was converted into trec_eval acceptable input form. The trec_eval was used to evaluate each model and following values were obtained for default settings of the given models.

1) BM25: 0.6675 (b=0.75 and k1=1.2)
2) VSM: 0.6639
3) DFR: 0.6792

2.3 BM25 parameter tuning

We tuned parameters b and k1 in order to improve the MAP values for the BM25 model. Generally it is recommended that b should be between 0.5 and 0.8. b is used to evaluate whether the length of a document would hinder the relevance of a term. k1 is used to evaluate whether term saturation is occurring. We compared MAP values for fixed values of b and varied k1 and then again compared MAP values for fixed values of k and varied b.

We obtained the following values:

For b=0.75

k1 values	MAP values
2	0.6669
1.8	0.6701
1.5	0.6715
1.2	0.6675
1.0	0.6695
0.7	0.6720
0.4	0.6718

For k1 = 1.2

k values	MAP values
0.3	0.6662
0.5	0.6697
0.7	0.6686
0.9	0.6696
1.0	0.6674

We finally selected b=0.8 and k=0.3

2.3 VSM and DFR

For DFR we used BasicModel = Geometric Distribution(G), Bernoulli After Effect and H2 Normalization. VSM similarity was used as is.

2.4 Query Parser

We used Dismax query parser instead of Standard query parser. Dismax query parser is a more forgiving parser and is error tolerant. It is more apt for user based queries and does not rely on query accuracy.

We converted the field types of text_xx files to their respective _xx field types in order to provide the basic text_xx filters such as SynonymnGraphFilterFactory, StopFilterFactory, LowerCaseFilterFactory, EnglishPossessiveFilterFactory, KeywordMarkerFilterFactory and PorterStemFilterFactory for text_en.

For text_de we also used GermanNormalizationFilterFactory and GermanLightStemFilterFactory.

For text_ru we also used SnowballPorterFilterFactory for Russian language.

Along with Dismax query we used field boosting to improve the relevance of certain fields. The following fields were boosted:

1) Tweet_hashtags: Boost Factor = 3
2) Text_en: Boost Factor = 6
3) Text_de: Boost Factor = 6
4) Text_ru: Boost Factor = 6

The query boost improved the MAP value for all three models as follows:

1) BM25: 0.7073
2) VSM: 0.6862
3) DFR: 0.7002

2.5 Copy Fields and Synonyms

We generated copy fields for text_xx in order to improve exact case matching for terms where we needed to match the term with Russian with Russian since we were using lowercase filter in the text_xx fields. We also added a list of synonyms for words that were assumed to be occurring more frequently in order to improve relevance. The following MAP values were obtained for all three models:

1) BM25: 0.7266
2) VSM: 0.7193
3) DFR: 0.7031

3 Results

Along with parameter tuning, we also performed query field boosting,

implemented copy fields for exact case matching and added synonyms to improve relevance scores. We replaced the standard query parser with dismax query parser. Using the following changes we obtained the following results for all of our models.

BM25

b	0.8
k1	0.3
MAP	0.7266

VSM

MAP	0.7193

DFR

Basic Model	Geometric Distribution
After Effect	Bernoulli
Normalization	H2
MAP	0.7031

Since, we are working with a smaller dataset we find out that BM25 provides relatively the best value.