

# CSE 535: INFORMATION RETRIEVAL

## Project 3: Evaluation of IR Models

---

Sumedha Salil Nashte – sumedhas

Prashanth Anuradha Balu - panuradh

### Objective

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

Different IR models to be used in this project were Vector Space Model (VSM), Best Matching (BM25) - Okapi, Divergence From Randomness (DFR). The way to go about this is to first index the given data in either schema or schema less modes. In our implementation we have used the schema less mode, for its advantages of rapidly creating all the fields by simply indexing the data, without any manual editing. The next step is to add similarity classes in the managed-schema, re-index the data and run the provided python code to generate scores for a particular query in the TREC format. Next, the score file is the passed to the trec\_eval program to evaluate the TREC scores for the retrieved data.

### Implementation of the given 3 IR models

#### 1) VSM Model:

VSM Model or Vector Space Model is implemented as Classic Similarity in Solr 6.2.0. We declared the Classic similarity globally in the **managed-schema** file.

XML Format:

```
<similarity class="org.apache.lucene.search.similarities.ClassicSimilarity" />
```

Screenshot:



```
<similarity class="org.apache.lucene.search.similarities.ClassicSimilarity" />
```

#### 2) BM25 Model:

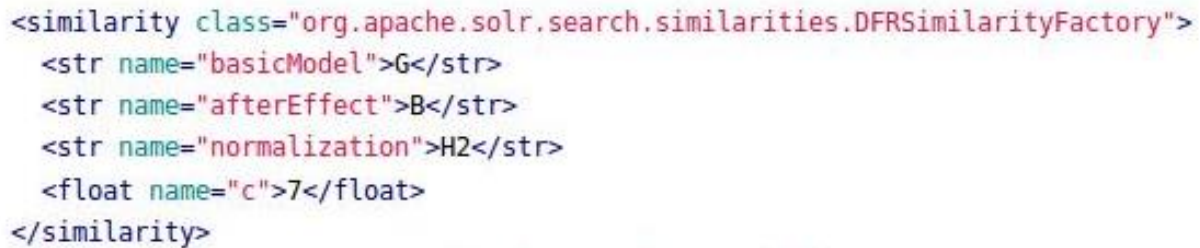
BM25 model is implemented by default by the Solr 6.2.0. Even if we do not implement BM25 explicitly in the managed-schema XML file, Solr uses this similarity factory by default with the values of  $k_1=1.2$  and  $b=0.76$ . Ideally we do need to change the values of  $b_1$  and  $k$  for boosting the relevancy across documents,  $k_1$  value should be decreased and  $b$  value should be increased. A higher  $b$  value adds more document length to the scoring process. Setting it to 0 results in document length not being considered at all which

might prove inefficient. A higher k value causes the term frequency to reach saturation longer. Hence we experimented with lower k value and higher b values.

XML Format:

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

Screenshot:

A screenshot of an XML configuration snippet for a similarity factory. The root element is <similarity class="org.apache.solr.search.similarities.DFRSimilarityFactory">. It contains four child elements: <str name="basicModel">G</str>, <str name="afterEffect">B</str>, <str name="normalization">H2</str>, and <float name="c">7</float>. The XML is displayed in a monospaced font with syntax highlighting: tags are in red, attribute names in green, and values in blue.

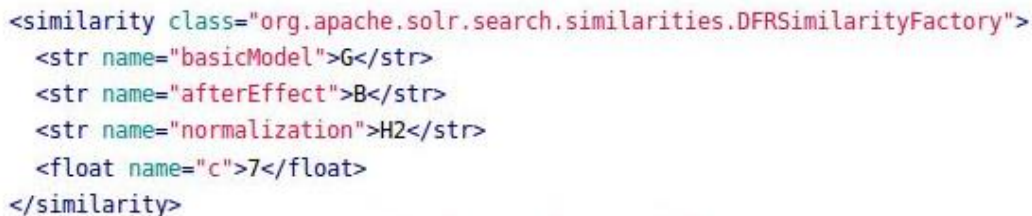
```
<similarity class="org.apache.solr.search.similarities.DFRSimilarityFactory">
  <str name="basicModel">G</str>
  <str name="afterEffect">B</str>
  <str name="normalization">H2</str>
  <float name="c">7</float>
</similarity>
```

### 3) DFR Model:

DFR Model is by far the best performing model and returned the best Mean Average Precision values for the training data provided. It has the following parameter values including basicModel, afterEffect and Normalization. We had the following default model in DFR to implement using Geometric approximation of Bose- Einstein model (G), ratio of two Bernoulli processes as afterEffect (B) and H2 normalization. Following is the XML format of DFR similarity.

```
<similarity class="solr.DFRSimilarityFactory">
  <str name="basicModel">G</str>
  <str name="afterEffect">B</str>
  <str name="normalization">H2</str>
  <float name="c">7</float>
</similarity>
```

Screenshot:

A screenshot of an XML configuration snippet for a similarity factory, identical to the one above. It shows the XML structure for DFRSimilarityFactory with parameters basicModel (G), afterEffect (B), normalization (H2), and c (7).

```
<similarity class="org.apache.solr.search.similarities.DFRSimilarityFactory">
  <str name="basicModel">G</str>
  <str name="afterEffect">B</str>
  <str name="normalization">H2</str>
  <float name="c">7</float>
</similarity>
```

## Approaches Leading to Optimization

- Tokenizers

We had tried various tokenizers like Classic Tokenizer, Ngram tokenizer but they did not result in any noticeable change so we proceeded with Standard Tokenizer.

- Filter factories

Filter factories including spell checking for commonly misspelt words reduced the overall MAP value so we did not continue with it to our final optimized model. We tried using Soundex filters namely Daitch Mokotoff soundex filter but it didn't contribute to increased scores as well.

## Optimization Techniques Implemented

- Dismax Query Parser

Dismax means Maximum Disjunction. It finds out the maximum disjunction i.e. OR of the input query terms so that it would return the maximum results relevant to the query. We have also specified the query fields in which we are doing the querying using the qf parameter. This aided in increase of MAP value significantly by a factor of 0.02.

- Addition of Synonyms for Improving Relevance

Synonyms proved to be crucial to improving the relevancy scores for documents. For instance, if users queried usa, US, u.s.a, USA, U.S. etc. they are all meaning the same word U.S.A. Furthermore Airbnb, Instacart, Kickstarter were added as US Tech Companies in the synonyms.txt file in the Solr core. Since we do not want to miss out on the relevant data in the corpus, because of punctuations synonyms aided in increasing the MAP value by a factor of 0.03-0.04.

- Removed Stopwords

Initially, we tried using stop words, since stop word removal would eliminate unwanted common words and increase our overall MAP. But we noticed that our Map got reduced due to their usage. Hence, we removed stopwords.

## Implementation of the given 3 IR models

### 1) VSM Model:

Sr. No.	Improvement	MAP
1	Implementing VSM	0.6418
2	Added dismax	0.6635
3	Added synonyms	0.7045

### 2) BM25

Sr. No.	Improvement	MAP
---------	-------------	-----

1	Implementing BM25 (with default values)	0.6575
2	Added dismax	0.6846
3	Added parameters $K1 = 0.2$ and $b = 0.4$	0.6893
4	Added parameters $k1 = 0.7$ and $b = 0.76$	0.6883
5	Added parameters $k1 = 0.4$ and $b = 0.0$	0.6830
6	Added Synonyms and $k1 = 0.2$ and $b = 0.4$	0.6984

### 3) DFR

Sr. No	Improvement	MAP
1	Implementing DFR (with default values $G, B, H2, 7$ )	0.6618
2	Added dismax	0.6951
3	Changed values of DFR ( $I(ne), L, H3$ )	0.6777
4	Added synonyms( $I(ne), L, H3$ )	0.6960
5	Changed back to $G, B, H2, 7$	0.6927
6	Changed to $G, B, H1, 2.6$	0.6840
7	Changed values of $DF(I(n), L, H2)$	0.6968
8	Changed values of $DF(I(ne), L, H1)$	0.6944
9	Changed values of $DF(I(n), L, H2)$	0.6977

