Project 3: Evaluation of IR models

Name : Jayaraj Channabasav Sajjanar
Purbarag Pathak Choudhury

UBIT:50208475, 50208670

Team:92

Implementation

Vector Space Model

ClassicSimilarityFactory provides the VSM similarity package in solr. This model requires no parameter. Figure 2a shows the implementation.

BM25

BM25SimilarityFactory provides the BM25 similarity package in solr. This model requires two parameter – k1 and b, which controls frequency normalization and document length normalization respectively. Figure 2b shows the implementation. The default value of the parameters are k1 = 1.2 and b = 0.75.

Divergence From Randomness

DFRSimilarityFactory provides the DFR similarity package in solr. This model requires three parameter – basicModel – basic model of information content, afterEffect – first normalization and normalization - second normalization. Figure shows all the parameters of DFRSimilarity class. Figure 2c shows the implementation. The default parameters are G, B, H2.

To construct a DFRSimilarity, you must specify the implementations for all three components of DFR:

1. `BasicModel`: Basic model of information content:
    - `BasicModelBE`: Limiting form of Bose-Einstein
    - `BasicModelG`: Geometric approximation of Bose-Einstein
    - `BasicModelP`: Poisson approximation of the Binomial
    - `BasicModelD`: Divergence approximation of the Binomial
    - `BasicModelIn`: Inverse document frequency
    - `BasicModelIne`: Inverse expected document frequency [mixture of Poisson and IDF]
    - `BasicModelIF`: Inverse term frequency [approximation of I(ne)]
2. `AfterEffect`: First normalization of information gain:
    - `AfterEffectL`: Laplace's law of succession
    - `AfterEffectB`: Ratio of two Bernoulli processes
    - `AfterEffect.NoAfterEffect`: no first normalization
3. `Normalization`: Second (length) normalization:
    - `NormalizationH1`: Uniform distribution of term frequency
    - `NormalizationH2`: term frequency density inversely related to length
    - `NormalizationH3`: term frequency normalization provided by Dirichlet prior
    - `NormalizationZ`: term frequency normalization provided by a Zipfian relation
    - `Normalization.NoNormalization`: no second normalization

Figure 1: DFR Similarity Parameters

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- Solr managed schema - automatically generated - DO NOT EDIT -->
<schema name="example-data-driven-schema" version="1.6">
  <uniqueKey>id</uniqueKey>
  <similarity class="solr.ClassicSimilarityFactory"/>
  <fieldType name="ancestor_path" class="solr.TextField">
    <analyzer type="index">
      <tokenizer class="solr.KeywordTokenizerFactory"/>
    </analyzer>
    <analyzer type="query">
      <tokenizer class="solr.PathHierarchyTokenizerFactory" delimiter="/"/>
    </analyzer>
  </fieldType>
  <fieldType name="binary" class="solr.BinaryField"/>
  <fieldType name="boolean" class="solr.BoolField" sortMissingLast="true"/>
  <fieldType name="booleans" class="solr.BoolField" sortMissingLast="true" multi
Valued="true"/>
  <fieldType name="currency" class="solr.CurrencyField" currencyConfig="currency
.xml" defaultCurrency="USD" precisionStep="8"/>
  <fieldType name="date" class="solr.TrieDateField" positionIncrementGap="0" doc
Values="true" precisionStep="0"/>
  <fieldType name="dates" class="solr.TrieDateField" positionIncrementGap="0" do
cValues="true" multiValued="true" precisionStep="0"/>
-- INSERT --                                                5,34          Top
```

(a) VSM Implementation

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- Solr managed schema - automatically generated - DO NOT EDIT -->
<schema name="example-data-driven-schema" version="1.6">
  <uniqueKey>id</uniqueKey>
  <similarity class="solr.BM25SimilarityFactory">
    <float name="k1">0.15</float>
    <float name="b">0.15</float>
  </similarity>
  <fieldType name="ancestor_path" class="solr.TextField">
    <analyzer type="index">
      <tokenizer class="solr.KeywordTokenizerFactory"/>
    </analyzer>
    <analyzer type="query">
      <tokenizer class="solr.PathHierarchyTokenizerFactory" delimiter="/"/>
    </analyzer>
  </fieldType>
  <fieldType name="binary" class="solr.BinaryField"/>
  <fieldType name="boolean" class="solr.BoolField" sortMissingLast="true"/>
  <fieldType name="booleans" class="solr.BoolField" sortMissingLast="true" multi
Valued="true"/>
  <fieldType name="currency" class="solr.CurrencyField" currencyConfig="currency
.xml" defaultCurrency="USD" precisionStep="8"/>
@
-- INSERT --                                                4,28          Top
```

(b) BM25 Implementation

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- Solr managed schema - automatically generated - DO NOT EDIT -->
<schema name="example-data-driven-schema" version="1.6">
  <uniqueKey>id</uniqueKey>
  <similarity class="solr.DFRSimilarityFactory">
    <str name="basicModel">G</str>
    <str name="afterEffect">B</str>
    <str name="normalization">H2</str>
    <float name="c">1.0</float>
  </similarity>
  <fieldType name="ancestor_path" class="solr.TextField">
    <analyzer type="index">
      <tokenizer class="solr.KeywordTokenizerFactory"/>
    </analyzer>
    <analyzer type="query">
      <tokenizer class="solr.PathHierarchyTokenizerFactory" delimiter="/"/>
    </analyzer>
  </fieldType>
  <fieldType name="binary" class="solr.BinaryField"/>
  <fieldType name="boolean" class="solr.BoolField" sortMissingLast="true"/>
  <fieldType name="booleans" class="solr.BoolField" sortMissingLast="true" multi
Valued="true"/>
@
-- INSERT --                                                10,16         Top
```

(c) DFR

Figure 2: schema.xml for the three models

Tables

VSM Model	
rows	MAP
20	0.6418
1000	0.6947

Table 1: MAP for VSM Similarity

DFR Model						
BM	AE	N	MAP	AE	N	MAP
Be	B	H1	0.6521	L	H1	0.6484
		H2	0.6550		H2	0.6561
		H3	0.6491		H3	0.6637
		Z	0.6572		Z	0.6568
D	B	H1	0.6452	L	H1	0.6371
		H2	0.6579		H2	0.6532
		H3	0.6125		H3	0.6452
		Z	0.6550		Z	0.6532
G	B	H1	0.6496	L	H1	0.6486
		H2	0.6554		H2	0.6562
		H3	0.6471		H3	0.6641
		Z	0.6567		Z	0.6565
P	B	H1	0.6439	L	H1	0.6403
		H2	0.6530		H2	0.6528
		H3	0.6130		H3	0.6478
		Z	0.6565		Z	0.6532

Table 2: MAP for DRF Similarity for rows=20

	VSM	BM25	DFR
For 20 rows	0.6418	0.6575	0.6554
For 1000 rows	0.6947	0.7107	0.7167

Table 3: MAP values with default settings

DFR Model						
BM	AE	N	MAP	AE	N	MAP
D	B	H1	0.6947	L	H1	0.6942
		H2	0.7066		H2	0.7043
		H3	0.6706		H3	0.6934
		Z	0.7056		Z	0.7081
G	B	H1	0.7002	L	H1	0.7041
		H2	0.7167		H2	0.7070
		H3	0.6944		H3	0.7155
		Z	0.7081		Z	0.7091
P	B	H1	0.6942	L	H1	0.6974
		H2	0.7019		H2	0.7042
		H3	0.6942		H3	0.6967
		Z	0.7059		Z	0.7081

Table 4: MAP for DRF Similarity for rows=1000

BM25 Model					
k1	b	MAP	k1	b	MAP
0.1	0.10	0.7168	1.7	0.50	0.7100
0.1	0.15	0.7175	1.7	0.75	0.7071
0.1	0.20	0.7175	1.0	0.25	0.7111
0.1	0.25	0.7175	1.0	0.50	0.7104
0.15	0.15	0.7167	1.0	0.75	0.7089
0.15	0.20	0.7167	2.0	0.25	0.7083
0.15	0.25	0.7166	2.0	0.50	0.7079
0.2	0.15	0.7160	2.0	0.75	0.7053
0.2	0.20	0.7082	2.5	0.25	0.7045
1.2	0.75	0.7107	2.5	0.50	0.7055
1.5	0.50	0.7107	2.5	0.75	0.6977
1.5	0.75	0.7089			

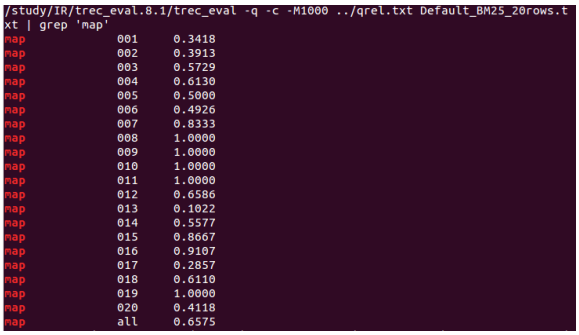
Table 5: MAP for BM Similarity for rows=1000

Discussion

The highlighted rows and columns shows the best value obtained for the model. We found that for DFR similarity the default values gave the best result when using 1000 rows. But for 20 rows the best result was given for Be, L, H3 parameters.

For BM25 model, even though range of value for 'b' and 'k1' is kept in range [0.0, 1.0] and [1, 5], better results were found far from those values at for k1 = 0.1 and b = [0.15, 0.25].

Table 3 shows our MAP value for default setting of the three models. Figure 3 shows two images of MAP calculation using the TREC program.



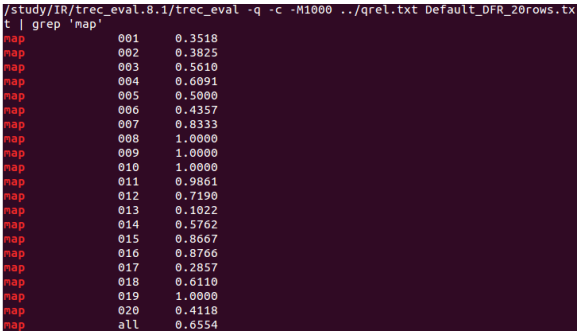
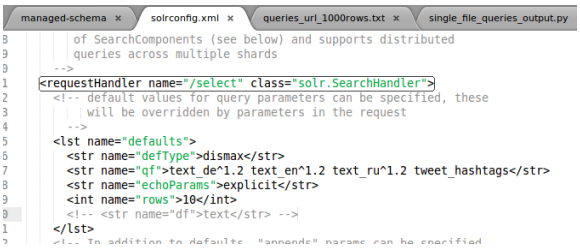
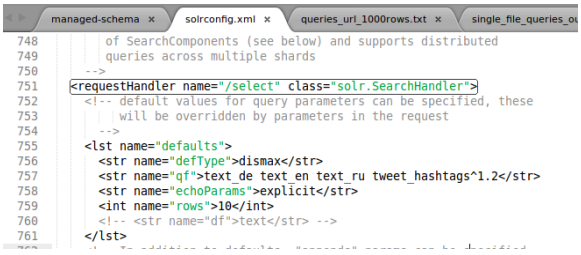


Figure 3: Screenshots of TREC run

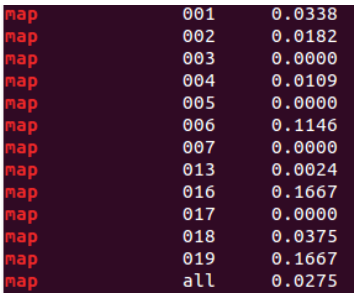
For improving the IR system we implemented DisMax, eDisMax and Query expansion. We moved the synonyms expansion from the analyzer's tokenizer chain to the query parser thus splitting the query into 'main query' and 'synonyms query' and combining with separate weights. The screenshots of the above is shown below.



(a) Boost 1



(b) Boost 2



(c) Boost Result

Figure 4: Configuration of solrconfig.xml for query boosting

However, we found the MAP value goes down which is shown in Figure 4 b. For this reason we discarded the implementation. We queried to solr after appending *defType=dismax* and *synonyms=true* to the url.

After much experimentation we found the better results were given for the given queries without using any synonyms expansion or field ranking.