CSE 535 INFORMATION RETREIVAL PROJECT 3

NAME: Sriram Venkataramanan

UB PERSON NUMBER : 50289666

The models that have been asked to implement in Solr are done as follows.

- **BM25**

  This is the default model of Solr 6.6.5. The following tags are added in schema-bm25.xml to tweak the paramters of BM25 model

  ```
  <similarity class="solr.BM25SimilarityFactory">
      <float name="k1">1.3</float>
      <float name="b">0.95</float>
  </similarity>
  ```

- **DFR**

  This is Divergence from Randomness model and this has been implemented in the Solr by giving the following tags in schema-dfr.xml

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

- **VSM**

  This is Vector Space model and this has been implemented in Solr by giving the following tags.

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

The Trec_eval results on different models are as follows:

The map all gives the average value for the 15 queries.

MODEL NAME	**MAP VALUE**
DFR	0.6645
BM25	0.6636
VSM	0.6600

Thus in terms of performance DFR is marginally better compared to other models

MAP value : DFR > BM25 > VSM

MEASURES TAKEN TO IMPROVE PERFORMANCE:

Filter factories that are used in schema are mentioned below:

1. **StopWordFilterFactory**
   PURPOSE: It is used for removing the commonly occurring terms.
   EFFECT ON PERFORMANCE AND END RESULT: It is useful for improving the relevance of result**s.**

2. **KeywordRepeatFilterFactory**
   PURPOSE: It emits two tokens for every input token. Any token which is not transformed later in the analysis chain will be in the document twice.
   EFFECT ON PERFORMANCE AND END RESULT: Improves recall.

3. **SynonymFilterFactory**
   PURPOSE: It is used to match string of tokens and replaces them with other strings of tokens
   EFFECT ON PERFORMANCE AND END RESULT: It is used to improve both the precision and recall.

4. **RemoveDuplicatesTokenFilterFactory**
   PURPOSE: It is used to remove duplicates which are at the same logical position in the tokenstream as previous token with the same text.
   EFFECT ON PERFORMANCE AND END RESULT: It is used to prevent idf-inflation at index time, or tf inflation at the time when we query i.e. query run time

   These have been added in Solr by adding the following tags

```
<fieldType name="text_en" class="solr.TextField" positionIncrementGap="100">
  <analyzer type="index">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.StopFilterFactory" words="lang/stopwords_en.txt"
ignoreCase="true"/>
    <filter class="solr.LowerCaseFilterFactory"/>
    <filter class="solr.EnglishPossessiveFilterFactory"/>
    <filter class="solr.KeywordMarkerFilterFactory" protected="protwords.txt"/>
    <filter class="solr.PorterStemFilterFactory"/>
  </analyzer>
  <analyzer type="query">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.SynonymGraphFilterFactory" expand="true" ignoreCase="true"
synonyms="synonyms.txt"/>
    <filter class="solr.StopFilterFactory" words="lang/stopwords_en.txt"
ignoreCase="true"/>
    <filter class="solr.KeywordRepeatFilterFactory"/>
    <filter class="solr.PorterStemFilterFactory"/>
    <filter class="solr.RemoveDuplicatesTokenFilterFactory"/>
    </analyzer>
  </fieldType>
```

- DISMAX PARSER: I have added Dismax parser in the Solr-config.xml
  PURPOSE: The DisMax query parser takes responsibility for building a good query from the user's input using Boolean clauses containing DisMax **queries** across fields and boosts specified by the user.
  EFFECT ON PERFORMANCE AND END RESULT: Thus it improves recall.

After improvement The Trec_eval results on different models are as follows:

MODEL NAME	**MAP VALUE**
DFR	0.7125
BM25	0.7100
VSM	0.6982

For BM25 we can tweak the parameters and change results. For k=1.2 b=0.5
We get MAP value = 0.7201

Thus again in terms of performance DFR is marginally better compared to other models.

MAP value : DFR > BM25 > VSM or BM25>DFR>VSM for certain tweaking paramters of BM25