CSE 435/535 Information Retrieval (Fall 2020)

REPORT

Project 3: Evaluation of IR models

By : Manasa Sai Challa (manasasa@buffalo.edu - 5035644)

Introduction

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. Twitter data in three languages - English, German and Russian, 15 sample queries and the corresponding relevance judgement was given. The given twitter data by Solr needs to be indexed, Vector Space Model and BM25 models need to be implemented on Solr, and the two sets of results obtained need to be evaluated using Trec_Eval program. Based on these evaluation results, improvement in the performance in terms of the measure Mean Average Precision (MAP) were asked to be discussed.

Methodology

The given twitter data is indexed by Solr, default configurations of Vector Space Model and BM25 are implemented on Solr, and obtained two sets of results are evaluated using Trec_Eval program.

Implementing the default configurations of IR Models :

A separate core for each model was created in Solr. Modifications to the respective schema.xml of each model were done to implement the default configurations for the models.

1.) VSM (Vector Space Model)

The following Similarity class was used in schema.xml to implement the default settings,

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

After the above similarity class is added to the schema.xml file of the core on solr, we reindex train.json for the above configured schema file and run the TREC_eval command to get MAP values.

```
num_ret                 001     13
num_rel                 001     20
num_rel_ret             001     3
map                     001     0.0875
Rprec                   001     0.1500
bpref                   001     0.1500
recip_rank              001     1.0000
iprec_at_recall_0.00    001     1.0000
iprec_at_recall_0.10    001     0.5000
iprec_at_recall_0.20    001     0.0000
iprec_at_recall_0.30    001     0.0000
iprec_at_recall_0.40    001     0.0000
iprec_at_recall_0.50    001     0.0000
iprec_at_recall_0.60    001     0.0000
iprec_at_recall_0.70    001     0.0000
iprec_at_recall_0.80    001     0.0000
iprec_at_recall_0.90    001     0.0000
iprec_at_recall_1.00    001     0.0000
P_5                     001     0.4000
```

2.) BM25 Model

The following Similarity class was used in schema.xml to implement the default settings,

```
<similarity class="solr.BM25SimilarityFactory">

<str name="b">0.75</str>
<str name="k1">1.2</str>

</similarity>
```

Similarly, After the above similarity class is added to the schema.xml file of the core on solr, we reindex train.json for the above configured schema file and run the TREC_eval command to get MAP values.

```
num_rel                 all     225
num_rel_ret             all     60
map                     all     0.2497
gm_map                  all     0.0242
Rprec                   all     0.3041
bpref                   all     0.2635
recip_rank              all     0.5263
iprec_at_recall_0.00    all     0.5844
iprec_at_recall_0.10    all     0.4489
iprec_at_recall_0.20    all     0.3906
iprec_at_recall_0.30    all     0.3822
```

MAP scores using the default settings for the IR models :

1.) VSM – 0.2430
2.) BM25 – 0.2497