CSE 435/535 Information Retrieval (Fall 2020)

REPORT

Project 3: Evaluation of IR models

By : Nishant Kapoor (nkapoor2@buffalo.edu - 50354536)

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.

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.) BM25 Model

Scoring of BM25 model depends on various parameters. One of the parameter is k1 and b measure. Model in schema is decided by its similarity class. The class used in BM25 model is :-

<similarity class="solr.BM25SimilarityFactory">

K	B1	Output
1.2	0.75	0.2497
1.5	0.75	0.2466
2	0.75	0.2470
0.1	1	0.2541
1.25	0.75	0.2589
1.5	0.75	0.2464
1.6	0.75	0.2471

The best result is achieved at **K=1.25** and **b1=0.75**

num_rel	all	225
num_rel_ret	all	62
map	all	0.2589
gm_map	all	0.0238
Rprec	all	0.3109
bpref	all	0.2733
recip_rank	all	0.4841
iprec_at_recall_0.00	all	0.5422
iprec_at_recall_0.10	all	0.4556
iprec_at_recall_0.20	all	0.4222
iprec_at_recall_0.30	all	0.3956
iprec_at_recall_0.40	all	0.3870
iprec_at_recall_0.50	all	0.2933
iprec_at_recall_0.60	all	0.1881
iprec_at_recall_0.70	all	0.1550
iprec_at_recall_0.80	all	0.1133
iprec_at_recall_0.90	all	0.0667
iprec_at_recall_1.00	all	0.0667
P_5	all	0.4400
P_10	all	0.3533
P_15	all	0.2711
P_20	all	0.2067
P_30	all	0.1378
P_100	all	0.0413
P_200	all	0.0207
P_500	all	0.0083
P_1000	all	0.0041

Fig1: At k=1.25 and b1=0.75

```
<analyzer type="index">
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.StopFilterFactory" words="stopwords.txt" ignoreCase="true"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.EnglishMinimalStemFilterFactory"/>
  <filter class="solr.PhoneticFilterFactory" encoder="soundex"/>
</analyzer>
<analyzer type="query">
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.StopFilterFactory" words="stopwords.txt" ignoreCase="true"/>
  <filter class="solr.SynonymGraphFilterFactory" expand="true" ignoreCase="true" synonyms="synonyms.txt"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.EnglishMinimalStemFilterFactory"/>
  <filter class="solr.PhoneticFilterFactory" encoder="soundex"/>
</analyzer>
```

Schema is improved by using Phonetic Filter. It increased the MAP score by 0.0021. Also, a custom field in schema is created that helped in querying data in 3 different languages English, German and Russian.

```
<copyField source="text_de" dest="text_unified_lang"/>
<copyField source="text_en" dest="text_unified_lang"/>
<copyField source="text_ru" dest="text_unified_lang"/>
```

Trec_eval program is used the calculate results of training set.

2.) VSM Model

For implementing the default settings, the similarity class used for VSM model is as follows:-,

<similarity class="solr.ClassicSimilarityFactory"/>

```
num_rel                 all     225
num_rel_ret             all     60
map                     all     0.2430
gm_map                  all     0.0237
Rprec                   all     0.2893
bpref                   all     0.2599
recip_rank              all     0.4930
iprec_at_recall_0.00    all     0.5622
iprec_at_recall_0.10    all     0.4489
iprec_at_recall_0.20    all     0.3906
iprec_at_recall_0.30    all     0.3822
iprec_at_recall_0.40    all     0.3108
iprec_at_recall_0.50    all     0.2552
iprec_at_recall_0.60    all     0.1881
iprec_at_recall_0.70    all     0.1550
iprec_at_recall_0.80    all     0.1133
iprec_at_recall_0.90    all     0.0667
iprec_at_recall_1.00    all     0.0667
P_5                     all     0.3867
P_10                    all     0.3267
P_15                    all     0.2578
P_20                    all     0.2000
P_30                    all     0.1333
P_100                   all     0.0400
P_200                   all     0.0200
P_500                   all     0.0080
P_1000                  all     0.0040
```

For improving the model, Pattern Replacement filter is implemented.

```
<charFilter class="solr.PatternReplaceCharFilterFactory" pattern="([@#])" replacement=""/>
```

It helped us gain MAP score by 0.0011.

Result and Observation

From various types of filters, few were implemented and respective MAP results were obtained. The trial of choosing k and b values helped us gain better results. The Phonetic filter gave a better efficiency and result in case of the bm25 model.

The MAP value obtained for the default setting is

1. BM25 – 0.2497
2. VSM - 0.2430

The results obtained after modification for both the models are listed below:-

MODEL	**MAP SCORE**
BM25	0.2589
VSM	0.2438