Evaluation of IR Models

Sudarshan Pol UBIT: **spol**

14th November 2020

Highest Achieved Map Scores
for Individual Models

- BM25: ***0.7094***
- VSM: ***0.6514***

Optimizations (Max value reached)

Initial + parameter tweaking	BM25 - 0.**2497** VSM - 0.**2952**
Added Language Appropriate Stemmers (Porter/Snowball/Light)	BM25 - 0.**2939** VSM - 0.**3112**
Adding Delimiter Filter and a Separate field for split and Case Sensitive words.	BM25 - 0.**4424** VSM - 0.**3467**
Added Synonyms and Implemented Query Expansion	BM25 - 0.**5817** VSM - 0.**5556**
Used Dismax Parser with Varying weights and Matching Parameters	BM25 - 0.**6577** VSM - 0.**6307**
Tweaked Queries and Changed Model parameters	BM25 - 0.**7094** VSM - 0.**6514**

Approach

BM25: Created two separate fields one with Word Delimiter Filter Factory and one without; for each language and ran each query over all of them.

VSM: Created separate fields with Case Preserved words and Lowercase Words for each language and ran each query over all of them and the Tweet Hashtag field.

Parameters Used

BM25

Model Parameters: b=**0.2** k1=**0.3**
Query weights

Russian: **tweet_hashtags^2.3 text_en^1.0 text_de^1.0 text_ru^1.5**

English: **tweet_hashtags^2.3 text_en^1.5 text_de^1.0 text_ru^1.0**

German: **tweet_hashtags^2.3 text_en^1.0 text_de^1.5 text_ru^1.0**

VSM

Dismax Minimum Matching used: **~15%**
Query weights

Russian: **text_en^1.0 text_de^1.0 text_ru^1.5**

English: **text_en^1.5 text_de^1.0 text_ru^1.0**

German: **text_en^1.0 text_de^1.5 text_ru^1.0**

Filters, Tokenizers and Parsers

solr.StandardTokenizerFactory

solr.WhitespaceTokenizer

solr.SynonymGraphFilterFactory

solr.WordDelimiterGraphFilterFactory

solr.StopFilterFactory

solr.LowerCaseFilterFactory

solr.GermanLightStemFilterFactory

solr.GermanNormalizationFilterFactory

solr.EnglishPossessiveFilterFactory

solr.FlattenGraphFilterFactory

solr.RemoveDuplicatesTokenFilterFactory

solr.SnowballPorterFilterFactory

Dismax Parser