CSE 435/535 Information Retrieval (Fall 2016)

Project 3: Evaluation of IR models

Objective:

To implement three IR models namely Vector Space model (VSM), Best Matching (BM25) and Divergence from Randomness model (DFR) and evaluate the IR system and improve the search results based on our understanding of the models, the implementation and the evaluation.

Steps implemented:

We were provide with Twitter data and relevance score for all the documents. Below are the steps which we followed to imple0ment various models

- Modified the schema and indexed the corpus according to the model
- Using relevance score for each document and the output from the queries, we recorded original TREC evaluation data as a reference
- Implemented various technique to improve MAP like query booster, synonym and so on. Recorded values for each improvement.
- Implemented the techniques and obtained the relevant documents for the test queries.

Techniques used:

Optimizing the default parameters:

By defau

Addition of Synonyms:

Adding a set of pre-determined synonyms for all languages would help improve the performance of the IR system. For instance when checking for the query *002 US air dropped 50 tons of Ammo on Syria*, our IR system should also look for tweets having words like USA, America which are synonymous to US. Also for the query *013 Airbnb, Instacart, Kickstarter* launch campaigns to fund refugee relief, all the three terms Airbnb, Instacart, Kickstarter refer to tech companies and hence our IR system should expand the query with the synonym search word: *Tech Companies* as well.

Addition of Cross Lingual Synonyms:

Usage of cross lingual synonyms while querying as well as indexing so that the relevant documents in both the languages would be retrieved and thereby increasing the recall. It could also even improve the precision as there could be more number of relevant documents now retrieved from all languages. For instance, in the query *РФ в Сирии вынудили 250 тунисских боевиков бежать* translates to the phrase - *Russia in Syria forced the 250 Tunisian militants* flee in English. Therefore our IR model should expand the query terms to all the specified languages in the corpus which in our case is to English, German and Russian and then retrieve all the documents relevant in all the three languages. Since implementing query translation is a complicated procedure, we could translate some of the keywords like Russia, Syria, refugee etc. to all the three languages and use them as synonyms. This way we would improve the recall and in turn improve the precision.

Disjunction Max search (dismax):

For a given query our IR model should search in all the available languages in the corpus to avoid missing out on documents from other languages. In order to achieve this we use Disjunction max

Boosting Hashtags in a query:

Improving weights on certain terms in the documents would improve the MAP. For instance for the query 010 #Syria #SALMA #LATAKIA, the tweets having Syria and Salma and Latakia as hashtags should be retrieved with a higher relevancy score than those which are don't have these words as hashtags.

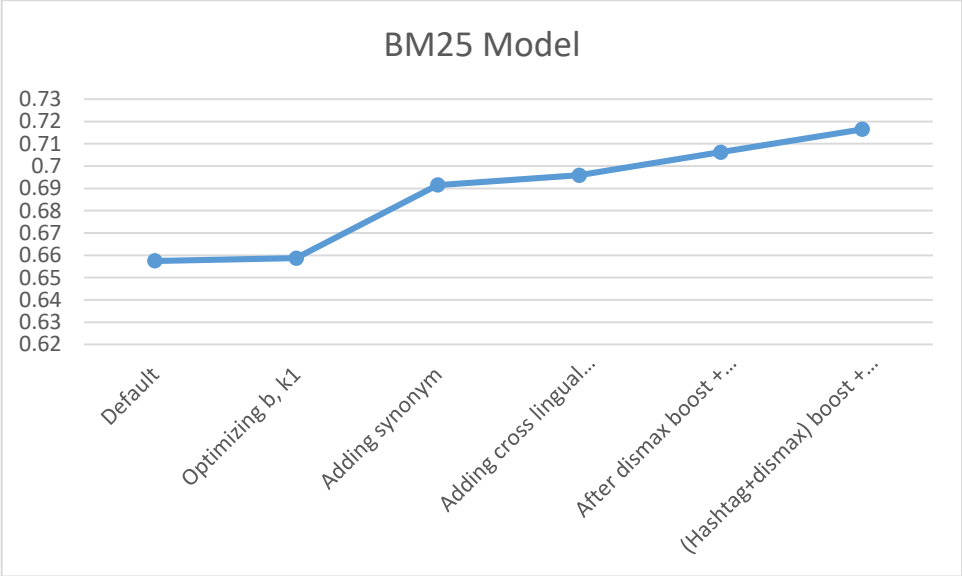
1.0 Best Matching 25 Model:

BM25 is a ranking function used by search engines for the ranking of documents retrieved for a certain query based on their relevance. It is a bag-of-words retrieval function which ranks a document set based on the query terms present in each document regardless of the inter-relationship between the query terms within a document. It a cumulative scoring function of various different components and parameters.

The Mean Average Precision of this model depends on the two below parameters**:**

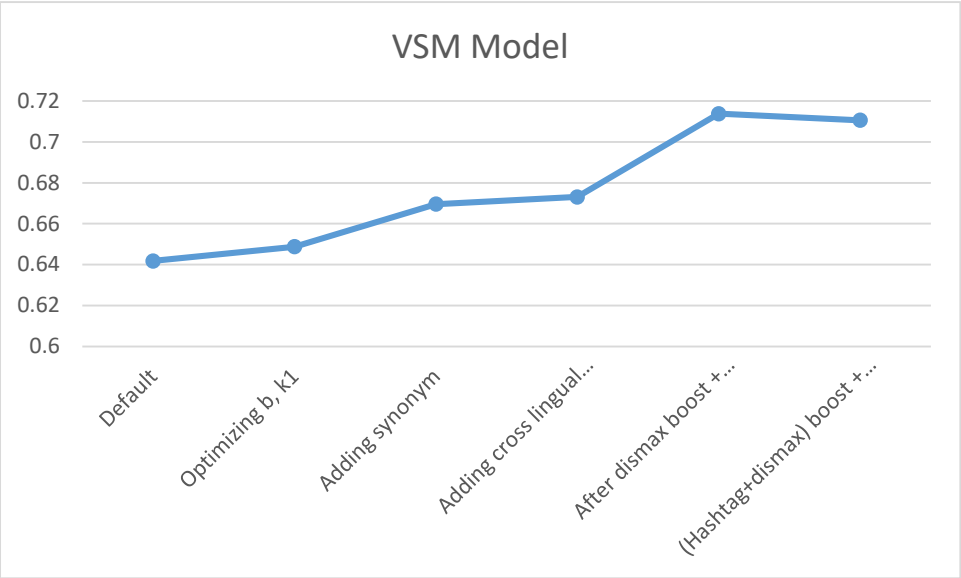
- k1 - Controls non-linear term frequency normalization (saturation)
- b - Controls to what degree document length normalizes tf values

By default the values of these two fields are set to - k1 = 1.2, b = 0.75. Depending on the corpus we need to optimize these values so that we achieve the maximum MAP value.



2.0 Vector Space Model:

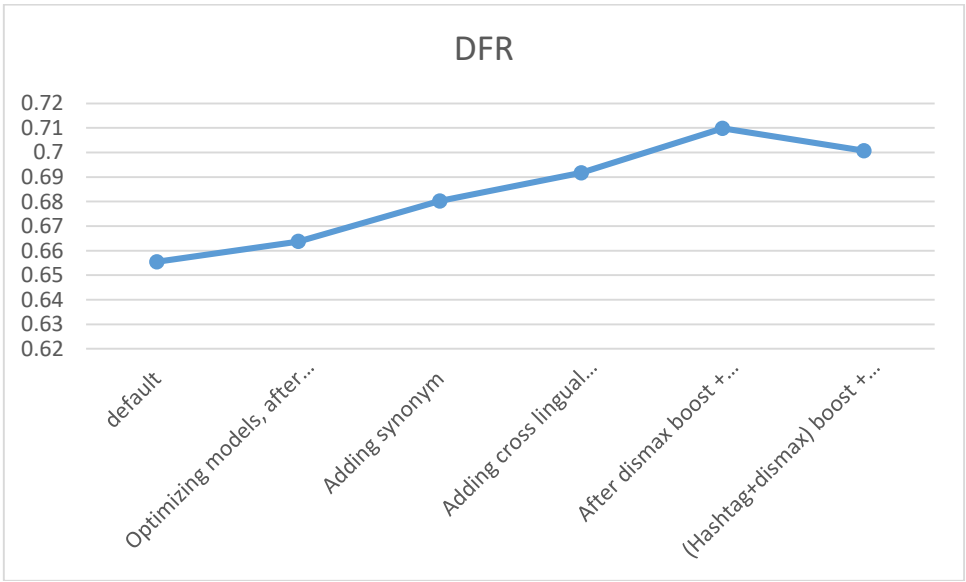
Vector space model is an algebraic model in which text documents are represented as vectors of identifiers with each term as an axis. In this model, we assign a similarity score to each query term based on the cosine similarity of the query term with the documents i.e. the dot product of the query terms with the document. The documents are then retrieved depending on the cosine similarity scores.



From the above graph we see that the MAP value increases with the addition of synonyms or cross lingual synonyms. It also improves by adding dismax query boosters.

3.0 Divergence from Randomness Model:

Divergence from Randomness is a type of probabilistic model in which term weights are computed by measuring the divergence between a term distribution produced by a random process and the actual term distribution.



Adding more weight to hashtags for BM25 model
Adding lesser weight to hashtags for VSM and DFR models

References:

https://en.wikipedia.org/wiki/Vector_space_model
https://en.wikipedia.org/wiki/Divergence-from-randomness_model