Programming Assignment № 1

Steven Lin R04922170 04/09/2016

VSM

In this assignment, we are asked to implement a Vector Space Model (VSM); that is, documents in the collection as well as query topics in the query file are represented by a vector each. In the implementation of my VSM, I let the number of dimensions of every vector be the number of terms.

Weights

The weight of a term, i.e., the value of the corresponding dimension in that vector, is determined by its Term Frequency (TF) and Document Frequency (DF). Additionally, the TF of every term with respect to a document is normalized, using the method: Okapi / BM25. For instance, the weight of term t_k in the vector of document with id doc_id should be:

$$(TF(t_k, doc\_id)/(1 - b + b \times DocLen(doc\_id)/AvgDocLen)) \times \log_2(DocNum/DF(t_k))$$

, where b is a variable parameter in Okapi normalization. And in my implementation, $b = 0.7$.

Rocchio Pseudo Feedback

If the $-r$ option is passed to $execute.sh$, Rocchio Pseudo Feedback is done by the following steps:

- run a normal search
- take out the IDs of the top 5 documents in the ranking list, say $\{d1, d2, d3, d4, d5\}$
- let the weight of term t double, for all t appears in any of $\{d1, d2, d3, d4, d5\}$
- run the search again with the updated weights

Unigram / Bigram

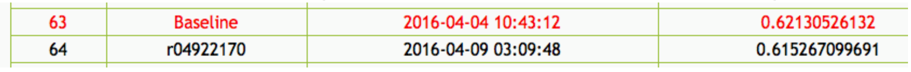
In my implementation, vectors are built in both perspectives of unigram and bigram. Therefore, the calculation of cosine similarity is also separated into 2 parts: uni_score and bi_score. The ultimate score $total\_score$ is a linear combination of the two, as follows.

$$total\_score \ = \ \lambda \ \times \ uni\_score \ + \ (1.0 - \lambda) \ \times \ bi\_score$$

, and λ is set to 0.2 in my program.

Results

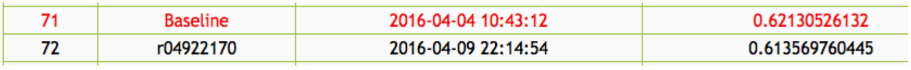
The best MAP score I got is: 0.615267099691, which is right below the baseline.



63	Baseline	2016-04-04 10:43:12	0.62130526132
64	r04922170	2016-04-09 03:09:48	0.615267099691

(This is **without Rocchio Pseudo Feedback**.)

If Rocchio Pseudo Feedback is enabled, the resulted MAP score is: 0.613569760445.



71	Baseline	2016-04-04 10:43:12	0.62130526132
72	r04922170	2016-04-09 22:14:54	0.613569760445

(slightly lower than the above one.)