Assignment 4

Information Retrieval and Web Search

Ahmed Rekik - 790063 & Tim Henning - 789242

December 3, 2017

1. Vector Space Modell

1.a. Can the tf-idf weight of a term in a document exceed 1? Why? Why not?

The tf-idf metric denotes to the term frequency-inverse document frequency and could be calculated by multiplying the "Term Frequency" with the "Inverse document frequency". While the "Term frequency" metric depends on the occurrence of a term in a query or a document (either =0 or ≤ 1), the "Inverse document frequency" could be potentially a big value since it depends on the document quantity in which occur certain specified term. This quantity of documents is mostly of times much less than the number of all documents in the corpus (since all most occured words in a corpus are classified as stop-word). Thus the "Inverse document frequency" could potentially exceed 1, although it is a logarithmic function. Since both "Term frequency" and "Inverse document frequency" could not be negative, so the metric "td-idf" could exceed 1.

1.b. What is the purpose of normalizing a documents vector representation for document length?

The normalization is usually described as an efficient algebric formula to assign weight to the normalized value in order to distinguish the dominant values (object) among the others. In our case, normalizing a documents vector representation for document length could be a tool to decriminalize the important term(s) from the others with taking into consideration the document length.

In this way, we could avoid the fact that "long" documents have a better priority than the short one. As an example: let the term "X" occurs 1000 times in a certain document which is very long and "X" has no big relevance in the document context. On the other side, the term "X" occurs only 20 times in a short document in which "X" is the main context of the document. By processing without normalizing the document vector representation, the long document could be more suitable than the short one. But in fact, with the normalization process, we frost the weightiest terms in a document, thus the similarity with ahort document will be as high as in the long one.

1.c. If each term represents a dimension in a t-dimensional space, the vector space model is making an assumption that the terms are orthogonal. Explain this assumption and discuss whether you think it is reasonable.

The orthogonality in the vector space model has been presumed as a problem in the retrieval model for the terms independencies factor, such that each term do not have any relationship with the other terms in the vector. Such assumption could be seen as meaningful when the model is implemented rawly and besides that term positions cannot be taken into consideration such that the n-grams similarity will not be possible which frost the orthogonality assumption.

But in other point of view and with some studiousness in the implementation of the retrieval model, the vector space model could include dependencies between occurred terms in the document such that each lexical field of terms will be clustered at its own in order to classify a document in terms of its occurred topic (topic-based vector space model). At that rate, the terms orthogonality will not more be considered as an issue in the vector space model and an association will be modeled between certain terms.

2. Probabilistic Model

2.a. What is 'binary' in the binary independence model (BIM)?

The term "binary" in "Binary Independance Model" denotes the absolute property of defining the occurrence or non-occurrence of a term in a document by the retrieval model.

2.b. What is 'independent' in the binary independence model (BIM) and is this a reasonable assumption? Explain.

"Independence" denotes to the independent representation of each term in the document vector and each of them has no modeled relationship with the other ones (no position or terms similarity are taken into consideration). In a sense, this assumption is equivalent to the orthogonality assumption in the vector space model. In this context, the independence assumption is exaggerated and absolute which could mislead the result, as in real life (in our information retrieval framework) it is almost impossible to find a document, in which all its value have no relationship with the other. (either by their position or in the document context) A term classification (topic, synonym, similarity...) could be a more reasonable way to proceed with this formula and thus elaborating class-independence could lead to a safer result.

2.c. What are the differences between standard vector space tf-idf weighting and the BIM probabilistic retrieval model?

While the tf-idf vector space is a standard which denotes to how much weight has the term in a document in terms of its frequency by the model, the BIM by the probabilistic retrieval model gives a probabilistic representation of the document relevance, without taking in consideration the frequency of a term in the document.

The similarity score metric in the vector space model is always represented between 0 and 1. If the similarity is not 0, so there is at least some similarity which will be taken into consideration. The BIM otherwise is more severe in his "relevance" classification such that a document is designed as relevant by the model only if the relevance of the document probability divided by the non-relevance probability of the document exceeds the likelihood ratio. In that context, the ranking method will obviously also differ.

2.d. What are the differences between standard vector space tf-idf weighting and the BIM probabilistic retrieval model?

The b parameter in BM25 in a scaled normalization tool of the K parameter and thus the document length. The higher is b, the more important role the document length will play in the BM25 score similarity between the query and the document.

Thus, by considering b is high (between 0.75 and 1), the score gap will be regulated and reduced between the short and the long document. (if the document is short, K will be less thus the score increase, the inverse for long documents)

3. Comparing Models

Build a query likelihood model using maximum likelihood estimates, a BM25 model and a tf-idf model. Use Jelinek-Mercer smoothing with $\lambda = 0.2$ for the query likelihood model. For BM25 assume that there is no relevance information and that k1=1.2, k2=100 and b=0.75. Compute the ranking of the four documents for the queries

Table 1: "Click"-Model

	Query: Click			
		BM25	Likelihood	tf-idf
Doc1		$1,497 * Log(N/n) = 0.187$	0,692	0.374
Doc2		$1,541 * Log(N/n) = 0.192$	0,892	0.249
Doc3		0	0	0
Doc4		$0,819 * Log(N/n) = 0.102$	0.252	0.124

Table 2: "Test"-Model

	Query: test			
		BM25	Likelihood	tf-idf
Doc1		$0,913 * Log(N/n) = 0.274$	0.230	0.301
Doc2		0	0	0
Doc3		0	0	0
Doc4		$0,819 * Log(N/n) = 0.246$	0.190	0.301

4. (Programming) Different Models for Ranked Retrieval

Will be added till Tuesday.