Week 5 – Information Retrieval WS 15/16

Johannes Jasper(755292)
Norman Rzepka(754644)

November 18, 2015

Assignment 1

a)

Can the tf-idf weight of a term in a document exceed 1?

That depends on the specific formula used to compute tf-idf. If for example $tf_{t,d}$ is simply the number of occurrences of term t in document d then this can easily exceed 1. Thus, $tf_{t,d} \cdot idf_{t,D}$ could exceed 1 as well since the range of $idf_{t,D}$ is in $[0, 1]$. If, however, some form of normalization is used this does not apply. If, for instance, we divide by the frequency of the most frequent term in the document, i.e. $tf_{t,d} = \frac{f(t,d)}{\max\{f(w,d):w \in d\}}$, the range of $tf_{t,d}$ is in $[0, 1]$.

b)

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

Similarity measures that use the distance of two vectors (e.g. Eucliadean distance) would be influenced by the length of a document (and the resulting counts). To lower the effect term counts are being normalized.

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. Why do we normalize the vector representation of documents in the vector space model? Is it always a good idea?

The model assumes that the occurrence of terms is independent. It is more likely, though, that the occurrence of a term depends on the occurrence of others, i.e. topically related terms co-occur. Of course this assumption does not hold, in practice, however, working with this model turns out to produce good results despite the false assumption. In the standard VSM, the feature vectors consists

of term occurrences. If such a vector is normalized term occurrence is lowered according to the length of the document. However, repetition can be a sign of emphasis. Therefore, tf-idf values might work better for feature vectors.

Assignment 2

a)

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

The model only considers the occurrence or non-occurrence of a term in the document. No counting or weighting takes place.

b)

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

BIM uses the naïve Bayes assumption which states that the values of all features are independent from each other. It ignores any corelation between those features.

c)

What are the differences between standard vector space tf-idf weighting and the BIM probabilistic retrieval model (in the case where no document relevance information is available?

The vector space model uses vectors of numeric features to represent a document. The standard VSM model uses a bag of words as the document vector. Rather than counts or occurrence tf-idf values for each term can be computed to employ their discriminating power. Ranking can be realized through some similarity measure between two documents, i.e. cosine similarity. BIM only uses binary values for occurrence or non-occurrence of a term in the document.

d)

What is parameter b good for in the BM25 model? As explained before term frequency should be normalized such that long documents do not get too large a boost. In the given formula b sets K which is used to normalize by document length. $b \in [0, 1]$ where 0 would mean no normalization and 1 would mean full normalization.

Assignment 3

Given the following document collection:

docID	document text
1	click click test click
2	click click
3	foo bar
4	click here foo bar test

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 $k_1 = 1.2$, $k_2 = 100$ and $b = 0.75$. Compute the ranking of the four documents for the queries

a)

docID	score
2	-0.1139
1	-0.3677
4	-1.3771
3	-2.3826

(a) QL Model

docID	score
3	0.0
4	-0.6943
1	-1.2687
2	-1.3063

(b) BM25

docID	score
1	2.0
2	1.5849
4	1.0
3	0.0

(c) TF-IDF

Figure 1: Scores for query 'click'

b)

docID	score
1	-1.4663
4	-1.6566
2	-3.4812
3	-3.4812

(a) QL Model

docID	score
1	0.0
2	0.0
3	0.0
4	0.0

(b) BM25

docID	score
1	1.0
4	1.0
2	0.0
3	0.0

(c) TF-IDF

Figure 2: Scores for query 'test'

For the query 'test' on BM25 the first factor is always 0 since

$$\frac{0.5/0.5}{(2+0.5)/(4-2+0.5)} = 0$$

. For TF-IDF we used the more complex formula including normalization.