TDT4117 Assignment 1

Information Retrieval

Erling Storaker Moen

Task 1: Basic Definitions

Explain the main differences between:

1. Information Retrieval vs Data Retrieval

Information retrieval deals with unstructured data or semi structured data while data retrieval deals with well structured data with well defined semantics. Querying a DBMS system produces exact results or no results at all while an information system produces multiple results with ranking from best to worst match, partial patch is also allowed in information retrieval.

2. Structured Data vs Unstructured Data

Structured data is comprised of clearly structured data with a pattern that makes them easily searchable. Unstructured data is basically everything else. This can be things like video and audio.

Task 2: IR Models

SubTask 1.1: Boolean Model and Vector Space Model

1. Which of the documents will be returned as the result for the above queries using the Boolean model? Explain your answers and draw a figure to illustrate.

q1= 1,2,4,7,8,10

q2= 1,3,4,7,10

q3= 2,5,8,9 I'm not sure if OR means "Or but not both" or "Either"

q4= 3

q5= 1,3,4,5,7,9,10

For every search query we look through all the documents looking for the keywords in the query, operators like AND, OR, NOT tell us what conditions the query puts in the search.

2. What is the dimension of the vector space representing this document collection when you use the vector model and how is it obtained?

With the vector model documents and queries are represented as vectors and we calculate the cosinus between the vectors to find a weight to which you can represent the document, to compare to other documents in the same query.

The dimensionality of the vector is the number of words in the vocabulary. In this case we have 3 distinct words appearing. Thus we have 3 dimensions.

3. Calculate the weights for the documents and the terms using tf and idf weighting. Put these values into a document-term-matrix. (Tip: use the equations in the book and state which one you used.)

	Winter	*Summer*	*Spring*	*Autumn*
Doc1	*1*	*0*	*1*	*1*
Doc2	*0*	*0*	*1*	*1*
Doc3	*1*	*0*	*0*	*1*
Doc4	*1*	*0*	*1*	*1*
Doc5	*1*	*0*	*0*	*0*
Doc6	*0*	*0*	*1*	*0*
Doc7	*1*	*0*	*1*	*1*
Doc8	*0*	*0*	*1*	*1*
Doc9	*1*	*0*	*1*	*0*
Doc10	*1*	*0*	*1*	*1*

Inverse document frequency: idf(t,D) = lg((|D|/t∈D))

idf(Spring,D) = lg(10/8) = 0.3219

idf(Summer,D) = lg(10/0) = 0

idf(Autumn,D) = lg(10/7) = 0.5145

idf(Winter,D) = lg(10/7) = 0.5147

4. Study the documents 1, 2, 4 and 10 and compare them to document 5. Calculate the similarity between document 5 and these four documents according to Euclidean distance. (Use tf-idf weights for your computations).

5-1: sqrt((1-1)^2 + (0-0)^2 + (1-0)^2 + (1-0)^2) = sqrt(2)

5-2: sqrt((0-1)^2 + (0-0)^2 + (1-0)^2 + (1-0)^2) = sqrt(3)

5-4: sqrt((1-1)^2 + (0-0)^2 + (1-0)^2 + (1-0)^2) = sqrt(3)

5-10: sqrt((1-1)^2 + (0-0)^2 + (1-0)^2 + (1-0)^2) = sqrt(3)

5. Rank the documents for query q5 using cosine similarity

We calculate the cosine angle between to vectors in a three dimensional space with the following formula. Cos(tetha) =((d1*d2)/(|d1| |d2)) Which gives us the ranking for query 5 of: 5>4>3>10>7>9>1

SubTask 1.2: Probabilistic Models

1. What are the main differences between BM25 model and the probabilistic model introduced by Robertson-Jones?

The BM25 model is a ranking function used by search engines to rank matching documents according to their relevance to a given search query. The probabilistic relevance model has the same function but uses a different method. The BM25 is the newer model with alterations and improvements. The main difference is that BM25 has the potential for giving negative scores for terms with very high document frequency. BM25 uses the trick of adding 1 to the value, before taking the log. Which makes it impossible to compute a negative value.

2. Rank the documents using the BM25 model. Set the parameters to k = 1.2 and b = 0.75. (Here we assume relevance information is not provided.) Hint: To avoid getting negative numbers, you need to use idf = log [N dft] in the BM25 model.

Task 3: Term Weighting

1. Term Frequency (tf)

Term frequency is a concept where you rank documents with a numerical value based on how many times a phrase is in document. Usually first eliminates all documents not containing the phrase and then counts how many times the phrase is in the remaining documents.

2. Document Frequency (df)

This is a method to determine how important a certain word or phrase is to a document.

3. Inverse Document Frequency (idf)

Inverse document frequency is like an addition to term frequency that solves the problem of some words coming up often and slowing down the search. Typically words like: the, be, to, of, and etc. IDF will be added to get rid of words that are used very often.

4. Why idf is important for term weighting

Without idf all the words that come up often in the English dictionary would skew the weight of these words or phrases containing these words way to much. This would result in queries working too hard and therefore taking too long and also the results would be weighted too hard on these common words and not the phrases we are actually looking for.