Stefan Bunk 756006
Franz Liedke 757081
Cornelius Bock 754432

# Week 3: Retrieval Models I

## Assignment 1: Boolean Retrieval

*a) Evaluate the query: q1 = (t1 or t5) but not t2.*
D1, D4, D5

*b) Evaluate the query: q2 = (t1 and t5) or (t3 and t2).*
D1, D2

*c) One of the drawbacks of the Boolean retrieval model lies in the size of the returned result set. Why is the size typically difficult to control?*
The use of the "NOT" operator typically leads to large result sets, as documents typically contain less words than they do contain. Thus, use of this operator usually only makes sense in combination with other operators or terms, to exclude unwanted results from an already small result set.

*d) Does Google support Boolean search? Which operators?*
Google supports Boolean search, especially the Boolean "OR" and "BUT NOT" search operators - with the latter being expressed through a hyphen in front of the word or phrase to exclude. When specifying multiple search words, Google understands this as an implicit "AND" search, looking for results that contain all of the specified words. The advanced search (https://www.google.com/advanced_search) offers a step-by-step interface to specify search conditions in detail without entering commands such as "OR".

## Assignment 2: Vector Space Model

*a) Can the tf-idf weight of a term in a document exceed 1?*
No.
$Idf_k$ is always bigger as one (because N is always bigger than $n_k$), so is $\log(f_{ik})$. Thus, both nominator and denominator are positive. Because of the sum in the denominator it is always bigger than the nominator which makes everything smaller as one.

*b) What is the purpose of normalizing a documents vector representation for document length?*
By normalizing the vectors for document length, we achieve a more accurate relevance ranking. Because two occurrences of a word in a short document point to higher relevance of that word than two occurrences of that same word in a much longer document, the length of each document should be taken into account when normalizing the vector lengths.

Stefan Bunk 756006
Franz Liedke 757081
Cornelius Bock 754432

*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?*
Because all terms are orthogonal in the vector space model, there is no understanding of any relations between terms. This is not a problem when compared to simpler models, but making meaning of words that only make sense together, phrases etc. is desirable.

## Assignment 3: (Programming) Boolean Queries

*Print the list of Wikipedia titles in your development set matching the queries:*
*a) "Artikel AND Smithee"*
```
Alan Smithee
```

*b) "Artikel OR Reaktion"*
```
Alan Smithee
Actinium
Ang Lee
```

*c) "Art* BUT NOT Artikel"*
Keine Ergebnisse.

*d) "'Filmfestspiele in Venedig'"*
```
Ang Lee
```
Tatsaechlich kommt im Dokument nur "Filmfestspiele von Venedig" vor. Da sowohl "von" als auch "in" aber von der Stopword-List gefiltert werden, wird es dennoch gefunden.