# 2. Übung zur Vorlesung "Internet-Suchmaschinen" im Sommersemester 2009

Prof. Dr. Gerd Stumme, M.Sc. Wi-Inf. Beate Krause

06. Mai 2009

## 1 Boolesches Retrieval (2)

Eine Erweiterung des booleschen Retrieval Models kann ein Ranking durch die Einbeziehung von Termhäufigkeiten in Dokumenten sein. Je höher die Vorkommen eines Terms im Dokument, desto weiter oben in der Rangliste kann das Dokument eingeordnet werden.

- x AND y: $tf(x, D) * tf(y, D)$
- x OR y: $tf(x, D) + tf(y, D)$
- NOT x: 0 if $tf(x, D) > 0$, 1 if $tf(x, D) = 0$

Betrachten Sie die beiden folgenden Dokumente:

$D_1$ max sagt fischers fritz fischt frische fische
$D_2$ moritz sagt fischers fritz fischt frische fische frische fische fischt fischers fritz

1. Bestimmen Sie ein Ranking für die folgenden Anfragen: (fritz OR max) AND fische, max OR (moritz AND fische), fritz AND NOT fische.
2. Können Sie sich weitere Abwandlungen des booleschen Modells überlegen, die dessen Ausdrucksmächtigkeit erhöhen?

## 2 TF-IDF Gewichtung

1. Betrachten Sie wieder die folgenden Dokumente:

   $D_1$ max sagt fischers fritz fischt frische fische
   $D_2$ moritz sagt fischers fritz fischt frische fische frische fische fischt fischers fritz

   Stellen Sie drei Term-Dokument-Matrizen auf. Die erste Matrix enthält als Gewicht die Termfrequenz ohne Normierung, die zweite Matrix als Gewicht die Termfrequenz mit Normierung und die dritte Matrix enthält eine TF/IDF-Gewichtung.

2. Wie sieht das IDF aus, wenn ein Term in allen Dokumenten erscheint? Was bedeuted dies für das Ranking?
3. Wie sieht das IDF aus, wenn ein Term in genau einem Dokument erscheint? Was bedeutet das für das Ranking?
4. Warum ist der IDF Wert eines Terms immer endlich?

## 3 Vektorraummodel

1. In der Vorlesung wurde ein Maß $cosSim(a, b)$ für die Ähnlichkeit zweier Dokumente $a$ und $b$ eingeführt. Dementsprechend sei $d_c(a, b) := 1 - cosSim(a, b)$ als Abstandsmaß definiert.

   Weiterhin kann man die euklidische Distanz $d_e(a, b) := \|a - b\|_2 := \sqrt{\sum_i (a_i - b_i)^2}$ definieren.

   Berechnen Sie den euklidischen und den Kosinusabstand von $D_1$ und $D_2$ für die Matrix mit den einfachen Häufigkeiten. (Sie brauchen nicht die numerischen Werte auszurechnen, Ausdrücke der Art $3 + \sqrt{19}$ reichen.) Was beobachten Sie?
2. Rechnen Sie nach, in welchem Zusammenhang die beiden Maße stehen, wenn man mit normierten Dokumenten ($\|a\| = \|b\| = 1$) arbeitet!

## 4 Praxisübung

*Abgabe: 19.05.2009*

Implementieren Sie einen invertierten Index mit TF-IDF-Gewichtung entsprechend dem Interface `InvertedIndex`! Die Termgewichte sollen wie in der Vorlesung skizziert normiert sein. Es gibt wieder eine Testklasse `IndexText`, mit der Sie Ihren Index ausprobieren können. Folgende Dokumente sind die am höchsten gerankten für die jeweils genannten Terme:

| Term | Datei → Gewicht, ... |
|---|---|
| november | 8683 → 0.5246, 3639 → 0.1749, ... |
| shipbuilding | 1902 → 0.2623, 6541 → 0.2623, 5818 → 0.1749, ... |
| sugarcane | 10306 → 0.2736, 11173 → 0.2736, 4630 → 0.1824, 259 → 0.1824, ... |

Tip: Implementieren Sie die Postinglisten als absteigend sortierte `Collection` von `TokenOccurence`-Objekten. Es ist etwas trickreich, dazu das Interface `Comparable<TokenOccurrence>` in `TokenOccurrence` korrekt (!) zu implementieren. Lesen Sie zuerst die Java-Dokumentation zu `Comparable`!