

Information Retrieval SS2011, Aufgabenblatt 6, Abgabe Mittwoch 22. Juni 8.00 Uhr.

Besprechung 22. Juni 12.00 Uhr

### Aufgabe 23

In Aufgabe 22 sollten aus der Tabelle zu einer Dokumentensammlung die Termgewichte berechnet werden. Alle Wörter ausser  $w_1, w_2, w_3$  sollten Stopwörter sein.

Termgewichte  $W(t,d) = tf(t,d) * idf(t) = tf(t,d) * \log N/df(t)$  ( Hier: log Basis 2)

| $tf(w,D)$ | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | d10 |
|-----------|----|----|----|----|----|----|----|----|----|-----|
| $w_1$     | 3  | 0  | 0  | 5  | 12 | 0  | 0  | 2  | 8  | 1   |
| $w_2$     | 8  | 6  | 0  | 12 | 0  | 0  | 9  | 1  | 3  | 10  |
| $w_3$     | 0  | 1  | 7  | 0  | 1  | 5  | 12 | 0  | 2  | 0   |

In einer Anfrage an die Dokumentenkollektion tritt  $w_1$  einmal und  $w_3$  zweimal auf. Welche Ähnlichkeit im Sinne des Kosinusmaßes ergibt sich zwischen Anfragevektor und den Dokumentenvektoren, wenn die tf-idf Gewichte wie in 22 berechnet verwendet werden.

### Aufgabe 24

Sei  $q$  eine Anfrage an ein IR System, die für ein bestimmtes Thema  $X$  beispielsweise

Radfahren steht. Was macht ein Dokument  $d$  intuitiv „sehr ähnlich“ zur Anfrage  $q$ ? Man wird auf jeden Fall erwarten, dass das Thema  $X$  charakteristisch für  $d$  ist. Man kann sich aber streiten, ob es die Ähnlichkeit verringern würde, wenn in  $d$  neben  $X$ (Radfahren) auch noch weitere Themen etwa  $Y$ (Schwimmen) auftreten. Diskutieren Sie vor diesem Hintergrund den Unterschied zwischen der Verwendung des Kosinus Abstandsmaßes und des Skalarprodukts im Vektorraummodell.

## Aufgabe 25

In den invertierten Files einer Dokumentenkollektion mit 10 Dokumenten  $d_0, \dots, d_9$  sollen die Wörter  $w_1, \dots, w_4$  die folgenden nach Termhäufigkeit geordneten Postinglisten haben.

$w_1$ : ((d4:14),(d2:9),(d8:4),(d1:1))

$w_2$ : ((d2:14),(d5:6),(d7:5),(d1:2) ,(d9:1), (d0:1))

$w_3$ : ((d9:3),(d2:1),(d6:1))

$w_4$ : ((d3:11),(d6:11),(d7:6),(d0:3) ,(d1:3), (d2:1), (d5:1))

Verwenden Sie als Anfrage die Liste mit den Termen  $w_1, \dots, w_4$ . Führen Sie die Berechnung der Ähnlichkeiten mittels Akkumulatorlisten durch wie es im Verfahren in Kapitel 6 gezeigt wurde.