Information Retrieval SS2012, Aufgabenblatt 5, Abgabe 12.Juli 2012.

Besprechung 12. Juli 12.00 Uhr

Aufgabe 1

In Aufgabe 3 Übungsblatt 4 sollten aus der Tabelle zu einer Dokumentensammlung die Termgewichte berechnet werden. Alle Wörter ausser w1,w2,w3 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
w1	3	0	0	5	12	0	0	2	8	1
w2	8	6	0	12	0	0	9	1	3	10
w3	0	1	7	0	1	5	12	0	2	0

In einer Anfrage an die Dokumentenkollektion tritt w1 einmal und w3 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 2

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 characteristisch 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 Hintergrdund den Unterschied zwischen der Verwendung des Kosinus Abstandsmaßes und des Skalarprodukts im Vektorraummodell.

Aufgabe 3

In den invertierten Files einer Dokumentenkollektion mit 10 Dokumenten d_0,....,d_9 sollen die Wörter w1,...w4 die folgenden nach Termhäufigkeit geordneten Postinglisten haben.

w1: ((d4:14),(d2:9),(d8:4),(d1:1))

w2: ((d2:14),(d5:6),(d7:5),(d1:2) ,(d9:1), (d0:1)

w3: ((d9:3),(d2:1),(d6:1))

w4: ((d3:11),(d6:11),(d7:6),(d0:3) ,(d1:3), (d2:1), (d5:1))

Verwenden Sie als Anfrage die Liste mit den Termen w1,...w4. Führen Sie die Berechnung der Ähnlichkeiten mittels Akkumulatorlisten durch wie es im Verfahren in Kapitel 6 gezeigt wurde.

Aufgabe 4

Precision und Recall

Es seien in der Rückgabemenge eines IR Systems bereits 8 relevante Dokumente gefunden worden und es ist das Precision-Recall-Paar (0.44, 0.4) gegeben.

a) wie viele Dokumente sind bereits untersucht worden?
b) wie viele relevante Dokumente befinden sich in der Kollektion?

Aufgabe 5

Es soll in Web-Suchmaschinen nach Informationen zu Jaguar (eine indische Automobilmarke) gesucht werden.

1. Zunächst bei Google(http://www.google.de) nach dem Sichtwort „Jaguar “ suchen. Man nehme die ersten 10 Suchergebnisse (Ergebnisse die nur Bilder, Maps und Video etc. enthalten sollen nicht berücksichtigt werden) und entscheide zu jedem gefundenen Dokument, ob es relevant ist. Anhand dessen soll die Präzision berechnet werden.
2. Wiederholen der gleichen Anfrage mit Bing(http://www.bing.com/?cc=de) und ebenfalls die Präzision anhand der ersten 10 Suchergebnisse berechnen.
3. Annahme, dass die Vereinigung der Mengen der relevanten Dokumente aus den ersten beiden Teilaufgaben der Gesamtmenge aller relevanten Dokumente entspricht. Berechnen des Recall für beide Fälle.

Links: http://ls6-www.informatik.uni-dortmund.de/uploads/tx_ls6ext/is11_08h.pdf