Vorlesung Information Retrieval
Übungsblatt 5 (Aufgaben 13-15)

Sommersemester 2009

Ausgabe: **25.05.2009**
Abgabe: **08.05.2009**
Besprechung: **15.06.2009**

Aufgabe 13:

Betrachten Sie die Anfrage mit m = 3 Suchbegriffen, bei der der Benutzer in k = 2 top Treffern interessiert ist. Die aggregierten Scores der Dokumente werden als (ungewichtete) Summe der Scores für Einzel-Terme berechnet. Die Indexlisten im System sehen wie folgt aus:

L1	**L2**	**L3**
d1 0.9	d3 0.7	d1 0.8
d7 0.6	d4 0.7	d6 0.7
d3 0.3	d7 0.4	d7 0.6
d2 0.3	d1 0.3	d4 0.4
d4 0.3	d6 0.2	d2 0.3
d5 0.2	d5 0.2	d3 0.2
d6 0.1	d2 0.2	d5 0.1

Wenden Sie die top-k NRA Methode (d.h. die Variante ohne Random Access) auf diese Problemstellung an. Dokumentieren Sie alle Zwischenschritte bei der Berechnung der top-k Ergebnisse. Wie viele Indexzugriffe sind notwendig?

Aufgabe 14:

In welchen Fällen bringt top-k **keine** Leistungsverbesserung gegenüber konventionellen Indexjoins einer Datenbank? Welche Nachteile kann top-k gegenüber einem Datenbankjoin (z.B. Hash-Join) haben?

Aufgabe 15:

Welche Maße für die (Dis-)Similarity zwischen den Dokumenten, die durch probabilistische Vektoren (d.h. Wahrscheinlichkeiten, die in der Summe 1 ergeben, wie bei LDA) dargestellt werden, sind möglich? Verwenden Sie Ihre Grundkenntnisse aus der Informationstheorie (Kapitel Technical Basics).