Technische Universität Braunschweig
Institut für Informationssysteme
Prof. Dr. Wolf-Tilo Balke
M. Sc. Joachim Selke

Information Retrieval und Web Search Engines
Wintersemester 2008/2009

Übungsblatt 9

14. Januar 2009

Hinweis: Soweit nicht anders angegeben, gibt es für jede korrekt bearbeitete Teilaufgabe einen Punkt. Die Abgabe der Hausübungen ist bis spätestens zum Beginn der nächsten Vorlesung möglich – entweder persönlich direkt vor der Vorlesung oder per Einwurf in den Briefkasten des Instituts (Informatikzentrum, zweiter Stock, vor Raum 238).

Aufgabe 19 (Stemming)

Welche der folgenden Aussagen sind korrekt? Begründen Sie Ihre Antworten! (2 Punkte)

(a) In Booleschen IR-Systemen verringert die Verwendung von Stemming niemals die Precision.

(b) In Booleschen IR-Systemen verringert die Verwendung von Stemming niemals den Recall.

(c) Der Einsatz von Stemming vergrößert das indexierte Vokabular.

(d) Stemming sollte nur bei der Indexierung eingesetzt werden, nicht jedoch bei der Verarbeitung von Suchanfragen.

Aufgabe 20 (Positionsindexe)

Wir betrachten folgenden Ausschnitt aus einem invertierten Index, in dem auch Wortpositionen gespeichert sind:

Term			
angels	2: (26, 174, 252, 651)	4: (12, 22, 102, 432)	7: (17)
fools	2: (1, 17, 74, 222)	4: (8, 78, 108, 458)	7: (3, 13, 23, 193)
fear	2: (87, 704, 722, 901)	4: (13, 43, 113, 433)	7: (18, 328, 528)
in	2: (3, 37, 76, 444)	4: (10, 20, 110, 470, 500)	7: (5, 15, 25, 195)
rush	2: (2, 66, 194, 321, 702)	4: (9, 96, 149, 429, 569)	7: (4, 14, 404)
to	2: (47, 86, 234, 999)	4: (14, 24, 774, 944)	7: (199, 319, 599, 709)
tread	2: (57, 94, 333)	4: (15, 35, 155)	7: (20, 320)
where	2: (67, 124, 393, 1001)	4: (11, 41, 101, 421, 431)	7: (16, 36, 736)

Der Eintrag „2: (26, 174, 252, 651)“ für den Term „angels“ bedeutet dabei, daß dieser Term in Dokument 2 an den Wortpositionen 26, 174, 252 und 651 auftritt.

a) Welche Dokumente passen auf die phrase query „fools rush in“?

b) Welche Dokumente passen auf die phrase query „‚fools rush in‘ AND ‚angels fear to tread‘“ ?

Beschreiben Sie jeweils kurz, wie ein IR-System diese Anfragen auswerten würde.

Aufgabe 21 (Positionsindexe und Stopwords)

In einem IR-System soll sowohl ein Positionsindex zum Einsatz kommen (um pharse queries effizient zu unterstützen) als auch eine Elimination von Stopwords erfolgen (um Speicherplatz zu sparen). Welches Problem tritt hierbei auf und wie könnte man es lösen? (2 Punkte)

Aufgabe 22 (Kompression von Postinglisten)

Die Postingsliste eines Terms (nur die Liste der Dokumenten-IDs, also insbesondere keine Wortpositionen oder Termhäufigkeiten) wird wie folgt unter Verwendung der Gap-Technik mittels Gamma-Code kodiert:

$$1110001110101011111101101111011.$$

Dekodieren Sie diese Bitsequenz, indem Sie zunächst die Folge der Gaps bestimmen und daraus dann die Liste der Dokumenten-IDs berechnen. (2 Punkte)