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 3

19. November 2008

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 236).

Aufgabe 6 (Probabilistic Indexing)

In dieser Aufgabe betrachten wir eine Dokumentensammlung, die die Dokumente d_1 bis d_4 enthält. Diese Dokumente werden wie folgt durch die Terme t_1 bis t_4 repräsentiert:

	t_1	t_2	t_3	t_4
d_1	0,7	0,1	0,9	1
d_2	0,2	1	0,8	0,2
d_3	0	0,7	0,3	0,5
d_4	1	1	0	0

Zudem ist durch Untersuchungen des Nutzerverhaltens bekannt, daß bei den letzten 1000 gestellten Anfragen ...

- ... 200mal das Dokument d_1 als relevant bezüglich der jeweiligen Anfrage beurteilt wurde,
- ... 700mal das Dokument d_2 als relevant bezüglich der jeweiligen Anfrage beurteilt wurde,
- ... 400mal das Dokument d_3 als relevant bezüglich der jeweiligen Anfrage beurteilt wurde,
- ... 100mal das Dokument d_4 als relevant bezüglich der jeweiligen Anfrage beurteilt wurde.

Beantworten Sie die folgenden Anfragen mit Hilfe des Probabilistic Indexing:

$$q_1 \colon t_1, t_2, \qquad q_2 \colon t_3, \qquad q_3 \colon t_2, t_4.$$

(3 Punkte)

Aufgabe 7 (Binary Independence Retrieval)

a) In diesem Aufgabenteil betrachten wir eine Dokumentensammlung, die die Dokumente d_1 bis d_4 enthält. Diese Dokumente werden wie folgt durch die Terme t_1 bis t_4 repräsentiert:

$$d_1 \colon t_2, t_3, t_4,$$
$$d_2 \colon t_2, t_3,$$
$$d_3 \colon t_2, t_4,$$
$$d_4 \colon t_1, t_2.$$

Beantworten Sie die folgenden Anfragen mit Hilfe des Binary Independence Retrievals:

$$q_1 : t_1, t_2, \qquad q_2 : t_3, \qquad q_3 : t_2, t_4.$$

Schätzen sie dazu den Wert von $P(D_i = 1 \mid D \in R_q)$ mit 0,9 ab. (3 Punkte)

b) Worin ähneln sich das Vektorraummodell mit TF-IDF-Gewichtung und das Binary-Independence-Retrieval-Modell, wenn man nur die schließlich zum Ranking der Dokumente verwendeten Formeln betrachtet?