Information Retrieval 2011, Aufgabenblatt 1, Abgabe Montag 11. Mai 8.00 Uhr

Aufgabe 1

Gegeben sei eine Datenbank, die die Dokumente d1 bis d6 enthält und diese wie folgt durch die
Indexterme t1 bis t8 repräsentiert:
d1 : {t1, t4, t6, t7}
d2 : {t2, t4, t8}
d3 : {t1, t3, t4}
d4 : {t2, t6, t7}
d5 : {t1, t4}
d6 : {t1, t3, t6}.
a) Bestimmen Sie den zu dieser Datenbank gehörigen invertierten Index („Inverted File Index“).
b) Welche Treffer liefert das Boolesche Modell für die folgenden Anfragen?
„t1 AND t4“
„(t3 OR t4) AND NOT t6“
„(t1 AND t6) OR NOT t6“
c) Formen Sie die Anfragen aus dem vorherigen Aufgabenteil jeweils zu äquivalenten Anfragen in konjunktiver und disjunktiver Normalform um.
d) Geben Sie zu jeder der folgenden Treffermengen eine (möglichst kurze) Anfrage an, die die jeweilige Treffermenge liefert.
{d2} , {d6}, {d3, d5}

Aufgabe 2

Es sollen w1,w2, . . . unterschiedliche Wörter sein. Die Dokumente
d1, . . . , d5 sollen folgende Wortfolgen darstellen:

d1 : w5,w1,w9,w3,w8,w2
d2 : w9,w8,w3
d3 : w2,w3,w8,w7
d4 : w9,w1,w8,w2,w3

(a) Welche Antwortmengen ergeben sich beim Booleschen Retrieval für folgende Anfragen:
q1 : w2 AND (w8 -> w1)
q2 : w7 OR (w1 -> ¬w3)
q3 : (w8 -> w2) -> (w1 -> ¬w3)

(b) Glauben Sie dass die Implikation in realen Systemen üblicherweise implementiert wird?

Aufgabe 3

Schreiben Sie einen Algorithmus für das Zusammenführen zweier Postinglisten zur Auswertung einer OR-Query - analog dem AND-Algorithmus aus der Sitzung.

Aufgabe 4

Implementieren Sie folgenden ersten Teil eines sehr einfachen IR Systems.

(a) Ein Teilprogramm soll den Shakespearetextkorpus von der Webseite in einzelne Stücke zerlegen, jedem Dokument eine ID zuordnen, den Text ausgeben und den Dokumententitel ausgeben können

(b) Eine Methode für eine lineare Suchmöglichkeit über die Texte mit Rückgabe der Dokumenten-id

(c) Implementieren Sie eine Klasse, die eine binäre Term-Dokumentenmatrix für einen Textkorpus wie er in a eingelesen wurde aufbaut.