Lehrstuhl Praktische Informatik III
Universität Mannheim
Prof. Dr. Guido Moerkotte
Sven Helmer

August 2003

# Hauptdiplomklausur
# Einführung in Information Retrieval
# Sommersemester 2003

Name: .....
Vorname: .....
Matrikelnummer: .....
Studienfach: .....

Wichtige Hinweise:

1. Prüfen Sie Ihr Klausurexemplar auf Vollständigkeit (6 Seiten).
2. Es sind keine Hilfsmittel zugelassen.
3. Die Klausur dauert 33 Minuten.
4. Jede Aufgabe ist auf dem zugehörigen Aufgabenblatt (und ggf. auf separaten Lösungsblättern) zu bearbeiten.
5. Vermerken Sie Ihren Namen und Ihre Matrikelnummer auf jedem Aufgaben- (bzw. Lösungsblatt). Blätter ohne Namens- und Matrikelnummerangabe werden nicht bewertet.
6. Das Deckblatt sowie alle Aufgabenblätter (evtl. Lösungsblätter) sind abzugeben.

| | maximale Anzahl Punkte | erreichte Anzahl Punkte |
|---|---|---|
| Aufgabe 1 | 6 | |
| Aufgabe 2 | 6 | |
| Aufgabe 3 | 5 | |
| Aufgabe 4 | 4 | |
| Aufgabe 5 | 4 | |
| Aufgabe 6 | 5 | |
| Aufgabe 7 | 3 | |
| | 33 | |

1. (6 Punkte)
   In einem IR-System werden folgende Dokumente in einer Rankingliste zurückgeliefert. Die mit ← markierten Dokumente sind relevante Dokumente. Berechnen Sie schrittweise die jeweiligen Werte für Recall und Precision. Gehen Sie davon aus, daß insgesamt 8 Dokumente relevant sind. (Sie können die Angaben in Brüchen machen, also z.B. $\frac{1}{8}$ statt 12,5%.)

| Ranking | Recall | Precision |
|---|---|---|
| 1. $d_{76}$ | | |
| 2. $d_{23}$ ← | | |
| 3. $d_{298}$ ← | | |
| 4. $d_{412}$ | | |
| 5. $d_{99}$ | | |
| 6. $d_{87}$ ← | | |
| 7. $d_{723}$ ← | | |
| 8. $d_{615}$ | | |
| 9. $d_{187}$ | | |
| 10. $d_{399}$ ← | | |
| 11. $d_{12}$ | | |
| 12. $d_{54}$ | | |

2. Welche der folgenden Formeln sind gültige TFxIDF Heuristiken? (Dabei gibt $f_{j,i}$ die Anzahl des Terms $t_i$ in Dokument $d_j$, $f_i$ die Anzahl der Dokumente in denen der Term $t_i$ auftritt und $n$ die Gesamtzahl der Dokumente an.) Begründen Sie Ihre Antwort kurz.

   (a) (2 Punkte)
   TF: $r_{j,i} = \sqrt{f_{j,i}}$, IDF: $w_i = \cos(\frac{n}{f_i})$

(b) (2 Punkte)
TF: $r_{j,i} = |f_{j,i} - 5|$ , IDF: $w_i = e^{-f_i}$

(c) (2 Punkte)
TF: $r_{j,i} = \sqrt{f_{j,i} + f_{j,i}}$, IDF: $w_i = \frac{\max_i(f_i)}{f_i}$

3. Eine Dokumentsammlung wird mit Hilfe einer Signaturdatei indexiert. Dabei gibt es folgende Suchterme mit ihren jeweiligen Signaturen:

| | |
|---|---|
| $t_a$ | 100001 |
| $t_b$ | 011000 |
| $t_c$ | 001010 |
| $t_d$ | 100100 |
| $t_e$ | 010001 |
| $t_f$ | 000110 |

(a) (3 Punkte)
Das Dokument $d_1$ enthalte die Terme $t_b, t_c, t_d$, das Dokument $d_2$ die Terme $t_a, t_b$ und das Dokument $d_3$ die Terme $t_d, t_e, t_f$. Berechnen Sie die Signaturen der drei Dokumente.

Signatur von $d_1$:

Signatur von $d_2$:

Signatur von $d_3$:

(b) (2 Punkte)
Jetzt wird eine Anfrage gestellt in der alle Dokumente gesucht werden in denen der Suchterm $t_f$ auftritt. Welche Dokumente sind in der Antwortmenge, wenn nur die Signaturen verglichen wurden? Sind darunter auch „false drops“ zu finden?

4. (4 Punkte)
Berechnen Sie eine Huffmankodierung für folgendes Alphabet (die Zahlen unter den Zeichen geben die Häufigkeit des Auftretens des Zeichens an).

| A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|
| 2 | 8 | 4 | 1 | 5 | 9 | 7 | 11 |

5. (a) (2 Punkte)
Was sind die Hauptprobleme der klassischen Retrievalmodelle?

(b) (2 Punkte)
Warum werden heutzutage trotzdem hauptsächlich die klassischen Retrievalmodelle eingesetzt?

6. (a) (3 Punkte)
Was sind die Vorteile des Shift-Or String-Matching Algorithmus?

(b) (2 Punkte)
Wie verändert sich die Laufzeit des Shift-Or Algorithmus bei größer werdendem Suchmuster?

7. (3 Punkte)
Komprimieren Sie folgende Lückenliste (gap list) einer invertierten Liste mit Hilfe des unären Codes (unary coding): 1,2,1,1,2,3,1