Fachhochschule Köln
Cologne University of Applied Sciences

institut für
informationswissenschaft

Prof. Winfried Gödert
Prof. Dr. Klaus Lepsky

Klausuraufgaben

Modulprüfung 6400 - Information Retrieval (IR4)
Studiengang Angewandte Informationswissenschaft

13.07.2015

Name .....

Matrikel-Nr. .....

Lesen Sie bitte den nachstehenden Text ***vor*** der Bearbeitung der Klausur aufmerksam durch!

Zugelassene Hilfsmittel: **Keine, außer Taschenrechner**

Geben Sie **deutlich** an:

- Ihren Namen
- Ihre Matrikel-Nummer

Beantworten Sie die Fragen direkt auf jedem Blatt unterhalb des Aufgabentextes.
Falls der Platz nicht ausreicht, benutzen Sie die Rückseite.
Falls der Platz immer noch nicht ausreicht, verwenden Sie separate Blätter, auf denen Sie dann **unbedingt** Ihren Namen, Ihre Matrikel-Nr. sowie die Aufgaben-Nr. vermerken.

Schreiben Sie bitte **leserlich**; unserer Fähigkeit zur Entzifferung von Handschriften sind Grenzen gesetzt.

Beachten Sie bitte auch:
Das Bestehen der Klausur erfordert nicht die Bearbeitung aller Aufgaben. Sorgfältige Bearbeitung einiger Aufgaben kann sinnvoller sein, als das flüchtige Bearbeiten aller Fragen

Wir wünschen Ihnen für die Bearbeitung viel Erfolg

W. Gödert / K. Lepsky

Aufgabe 1

a) Ist es bei einer *indexbasierten* Suche *nach Einzelwörtern* möglich, alle Dokumente zu ermitteln, in denen ein Wort, das auf **„en“** endet, von einem Wort gefolgt wird, das mit **„Be“** beginnt?
Begründen Sie die Antwort!

2 Punkte

b) Über welche *Suchfunktionalität* müsste ein Retrievalsystem verfügen, um eine Suche durchführen zu können, wie sie in Teil a) der Aufgabe beschrieben wird?

2 Punkte

Aufgabe 2

Welche Rückschlüsse auf die Behandlung der nachstehenden Sucheingaben durch die *Retrieval-Funktionalitäten* der jeweiligen Suchumgebung*) können aus den Ergebnismengen für die nachstehend angegebenen Suchanfragen gezogen werden?

3 Punkte

a) *Suchumgebung 1*

(1)	facet classification	3.590.000
(2)	facet and classification	3.680.000
(3)	“facet classification“	4.760
(4)	facet and classification or classifications	438.000
(5)	facet and (classification or classifications)	438.000

b) *Suchumgebung 2*

3 Punkte

(1)	facet classification	7
(2)	facet and classification	402
(3)	“facet classification“	7
(4)	facet and classification or classifications	605
(5)	facet and (classification or classifications)	6

*) Es handelt sich um zwei verschiedene Kollektionen mit einer jeweils unterschiedlichen Gesamtzahl von Dokumenten.

Aufgabe 3

Welchen Einfluss auf die Suchgeschwindigkeit hat es, ob ein Suchvorgang nach dem Konzept der *Sequenziellen Suche* oder nach dem *Konzept einer Suche in Binärbäumen* für Invertierte Listen durchgeführt wird?
Erläutern Sie die jeweiligen Konzepte durch Angabe eines Beispiels. **5 Punkte**

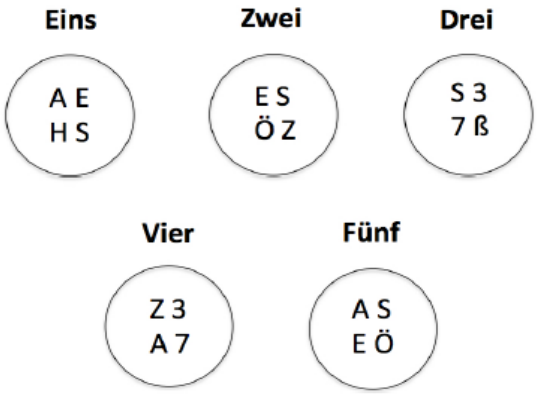
Aufgabe 4

a) Wie lassen sich *inhaltliche Gewichtungen* von Wörtern (z. B. der Unterschied zwischen dem Vorkommen eines Wortes als Titel-Stichwort, als Wort des Abstracts oder als indexierter Deskriptor) in einem Ranking-Algorithmus berücksichtigen? **4 Punkte**

b) Wie lassen sich die Gesichtspunkte einer *Gewichtung von Wörtern nach inhaltlichen Gesichtspunkten* mit einer *statistischen Gewichtung nach Häufigkeiten* im Rahmen eines *TF*IDF*-Ansatzes miteinander verbinden? **4 Punkte**

Aufgabe 5

Gegeben seien die folgenden Dokumentmengen:



Welche Ergebnisse erzielen jeweils die folgenden Suchanfragen:

a) (*Drei* NOT *Vier*) OR (*Zwei* AND *Fünf*) **2 Punkte**

b) ((*Eins* OR *Fünf*) NOT (*Zwei*)) OR (*Drei* NOT *Eins*) **3 Punkte**

Aufgabe 6

a) Erhalten *Hochfrequenzterme* im Rahmen der Anwendung eines *TF*IDF*- Ansatzes ein besonders hohes oder ein besonders niedriges Gewicht? Geben Sie eine Begründung! **3 Punkte**

b) Sollte man bei der Anwendung einer *Automatischen Indexierung mit Grundform-Bestimmung* und einer statistischen *Termgewichtung* zunächst das Termgewichtung und danach die Grundform-Bestimmung durchführen, oder eher umgekehrt? Begründen Sie die jeweiligen Vor- und Nachteile! **4 Punkte**

Aufgabe 7

Gegeben seien die folgenden Dokumente mit Titeln und gewichteten Deskriptoren:

(1) **TI**: Entwicklung eines akustischen Feedbacksystems als ruderspezifisches Trainingsgerät
DE: Akustische Lernhilfe (5); Bewegungsablauf; Bewegungslehre (3); Entwicklung; Feedback; Kinematik; Rudern (3); Sportinformatik; Sportliches Training; Sporttechnologie; Sportwissenschaft; Trainingsgerät (4)

(2) **TI**: Entwicklung eines Mess- und Analysesystems zur Optimierung der Bootsbewegung im Wassertraining und Ruderrennen
DE: Belastungsintensität; Belastungsumfang; Bewegungsanalyse; Leistungsdiagnostik (3); Messgerät; Rudern; Rudersport (4); Sportinformatik (2); Sportliche Technik; Sportliches Training; Sporttechnologie; Sportwissenschaft (2); Trainingssteuerung; Trainingswissenschaft (4)

(3) **TI**: Aktuelle Probleme der Renngestaltung mit trainingsmethodischen Schlussfolgerungen für die olympische Regatta im Rudern
DE: Olympische Spiele (3); Rudern (5); Rudersport; Sportliches Training (2); Trainingsmethode (2); Trainingswissenschaft (4); Wassersport; Wettkampfgestaltung (2); Wettkampfvorbereitung

(4) **TI**: Übung macht den Meister : gilt das auch für das Rudern?
DE: Leistungssteigerung (3); Rudern (4); Sportliches Training; Superkompensation; Trainingslehre (2); Übertraining

(5) **TI**: Techniktraining nach Gehör : Steigerung der Bootsgeschwindigkeit durch akustische Wahrnehmung
DE: Akustik (4); Auditive Wahrnehmung (4); Beschleunigung; Bewegung; Bewegungslehre; Boot; Feedback (2); Mentales Training (3); Ruderboot (3); Rudern; Rudersport (2); Techniktraining; Trainingsmittel (4); Visuelle Wahrnehmung; Wassertraining

Erstellen Sie für die *Suchanfrage*:

Belastungsintensität; Belastungsumfang; Bewegungsanalyse; Rudern; Sportliches Training; Techniktraining; Trainingssteuerung; Wettkampfvorbereitung

den *Anfragevektor* und berechnen Sie das *ähnlichste Dokument* auf der Basis des vereinfachten Skalarprodukts:

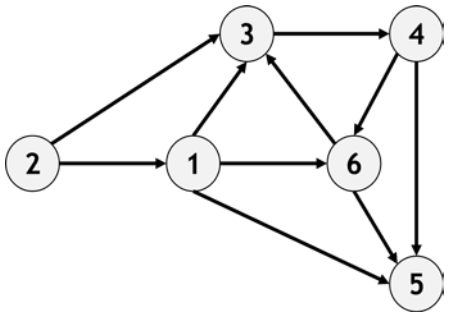
$$\ddot{A}(A\mathbf{d}_i) = \frac{1}{M+n}\sum_{k=1}^{n} a_k d_{jk}, M = \max_{1\le i\le m}(\sum_{j=1}^{n} d_{ij}, A), i = 1,...,m$$

8 Punkte

Aufgabe 8

a) Erstellen Sie für die nachstehend abgebildete Linkstruktur die *Matrix der Übergangswahrscheinlichkeiten* für die iterative Berechnung des *PageRank*.

5 Punkte



b) Welche Rolle spielt es für das Ergebnis der Iteration beim PageRank-Verfahren, ob für die Elemente des Startvektors die Zahl „1“ oder die Zahl „1/n“ gewählt wird? Geben Sie eine Begründung!

2 Punkte