Fakultät für Informations- und Kommunikations-wissenschaften

Information Retrieval Probeklausur

Technology Arts Sciences TH Köln

Prof. Dr. Philipp Schaer

Name: _____

Matrikelnummer: _____

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

- Es sind **keine Hilfsmittel** zur Prüfung zugelassen, außer einem Taschenrechner (auch, wenn Sie diesen eigentlich nicht benötigen).
- Geben Sie **auf jeder Seite** 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, die Sie auf Anfrage bekommen, auf denen Sie **Ihren Namen, Ihre Matrikel-Nr. sowie die Aufgaben-Nr**. vermerken.
- **Lösen Sie auf keinen Fall die Klammerung** der Klausurbögen!
- Schreiben Sie bitte **leserlich**; Lösungen, die ich nicht lesen kann, kann ich nicht bewerten.

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.
- Insgesamt können in dieser Prüfung 22 Punkte erreichen. Beachten Sie auch die Angabe zu den Punkten pro Aufgabe.

Ich wünsche Ihnen für die Bearbeitung viel Erfolg!

Philipp Schaer

Aufgabe	**A1**	**A2**	**A3**	**Gesamt**
max. Punkte	8	4	10	22
erreichte Punkte				

Aufgabe 1

a) Erklären Sie in Ihren eigenen Worten den Zusammenhang zwischen Zipfs Gesetz und der inversen Dokumentfrequenz. (4 Punkte)

b) Erklären Sie den Grundgedanken von phonetischer Indexierung (z.B. Soundex). (4 Punkte)

Aufgabe 2

Bewerten Sie die folgenden Aussagen als wahr oder falsch. Richtige und eine falsche Antwort heben sich gegenseitig auf. Nicht beantwortete Fragen werden nicht gezählt. (4 Punkte)

Wahr	Falsch	
☐	☐	Ranked Retrieval hilft beim Problem des „Feast“.
☐	☐	Die Entfernung von Stoppwörtern verkleinert den Index.
☐	☐	Im Vektorraummodell spielt die Dokumentlänge in der Score-Berechnung keine Rolle.
☐	☐	Ein Tokenizer zerlegt einen Text in einzelne Terme, die dann weiterverarbeitet werden.

Aufgabe 3

Sie haben einen Dokumentenkorpus, der aus vier Dokumenten besteht. Die entsprechende Term-Dokument-Matrix sieht wie folgt aus:

	Dok1	**Dok2**	**Dok3**	**Dok4**
information	2	1	2	1
retrieval	1	0	2	0
support	1	0	0	0
through	1	0	0	0
better	1	0	0	0
search	0	1	0	0
from	0	1	0	1
the	0	1	0	1
web	0	1	1	1
library	0	0	1	0
and	0	0	1	0
retrieve	0	0	0	1

Wie würde das Ranking bei einem **erweiterten Booleschen Retrieval** (also nicht dem Vektorraummodell!) aussehen, wenn die Anfrage „web OR information“ lauten würde? Das auf **tf-idf** basierende Ranking arbeitet hierbei mit einem **vereinfachten Scoring** mit

- einfacher, unveränderter Termfrequenz und
- einer inversen Dokumentfrequenz von 10/df.

Zeigen Sie die einzelnen Schritte und die Berechnung bis zur finalen gerankten Ergebnisliste! (10 Punkte)