Fakultät für Informations- und Kommunikations-wissenschaften

DIS12/BdK2.4: Information Retrieval Probeklausur

Technology Arts Sciences TH Köln

Prof. Dr. Philipp Schaer

Name: _____

Matrikelnummer: _____

Beachten Sie bitte:

- 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 20 Punkte erreichen. Beachten Sie auch die Angabe zu den Punkten pro Aufgabe.
- Sie haben 30 Minuten Zeit!

Ich wünsche Ihnen für die Bearbeitung viel Erfolg!

Philipp Schaer

Aufgabe	**A1**	**A2**	**A3**	**Gesamt**
max. Punkte	8	4	8	20
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 in Ihren eigenen Worten den Grundgedanken von phonetischer Indexierung (z.B. Soundex). (4 Punkte)

Aufgabe 2

Bewerten Sie die folgenden Aussagen als wahr oder falsch. Falsche Antworten führen zu Punktabzug. 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 findet die Dokumentlänge in der Score-Berechnung keine Beachtung.
☐	☐	Ein Tokenizer zerlegt einen Text in einzelne Terme, die dann weiterverarbeitet werden.

Aufgabe 3

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

	Dok1	**Dok2**	**Dok3**
information	2	1	2
retrieval	1	0	2
support	1	0	0
through	1	0	0
better	1	0	0
search	0	1	0

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** basierende Ranking arbeitet hierbei mit einem **vereinfachten Scoring** mit **einfacher, unveränderter Termfrequenz**.

$$Score_{q,d} = \sum_{t \in q \cap d} tf_{t,d}$$

Zeigen Sie die einzelnen Schritte und die Berechnung bis zur finalen gerankten Ergebnisliste! (8 Punkte)