Midterm Probeklausur von 12.01.2012

Hilfsmittel: Taschenrechner
Bearbeitungszeit: **45 Minuten**
Viel Erfolg!

I. Anfrage-Optimierung und Tolerant Retrieval (15 Punkte)

1. Eine Dokumentsammlung mit insgesamt 50 000 Dokumenten beinhaltet Wetterberichte. Gegeben sei die Anfrage: [**5 P.**]

(Frühling) AND (Sonne OR Wind) AND NOT (Regen OR Gewitter)

Geben Sie die effizienteste Abfrage-Reihenfolge an, die sich aus folgender Tabelle bestimmen lässt:

Term	**DF**
Frühling	25 000
Sonne	15 000
Wind	30 000
Regen	2 000
Gewitter	11 000

Beschreiben Sie eine mögliche Verteilung der Terme für welche die gefundene Reihenfolge nicht optimal ist.

2. Beschreiben Sie den Trigram-Index. Geben Sie die Anfragen an einen Permuterm-Index sowie an einen Trigram-Index für die Suchanfrage S*warzeneg*er (Schwarzenegger). [**5 P.**]

3. Gegeben sind zwei Terme „duck“ und „quack“. Berechnen Sie [**5 P.**]
die Levenshtein-Distanz sowie die Bigram-basierte Ähnlichkeit unter Anwendung der Jaccard Koeffizient zwischen den beiden Termen.

II. Web Information Retrieval (15 Punkte)

1. Gegeben sei eine Dokumentsammlung mit insgesamt zwei Dokumenten:
[5(3a)+5(3b)=**10 P.**]

Dokument 1:
Eclipse – Biss zum Abendrot ist ein Film, der auf dem Roman „Biss zum Abendbrot“ von Stephenie Meyer basiert.
Dokument 2:
Eclipse (von engl.: eclipse = Sonnenfinsternis, Finsternis, Verdunkelung) ist eine integrierte Entwicklungsumgebung.

1a. Erzeugen Sie aus den Dokumenten einen invertierten Index.
Tokenisierung-Regeln: Wort-weise (Satzzeichen ignorieren), alles klein geschrieben.
Stopwörter: {der, die, das, dem, den, ein, eine}. Geben Sie dabei jeweils **an passender Stelle** für jeden Term die entsprechenden Werte für TF und DF an.

1b. Die Funktion für eine Suchanfrage-Dokument-Relevanz sei

$$w_{Q,d} = \sum_{q \in Q,d} \left(1 + \log\left(TF_{q,d}\right)\right) * IDF_q$$

Berechnen Sie die jeweilige Relevanz für die beiden Dokumente im Bezug auf die beiden folgenden Suchanfragen:

Q_1= *Die Eclipse IDE*
Q_2= *Abendbrot Rezepte*
Erläutern Sie die Ergebnisse!

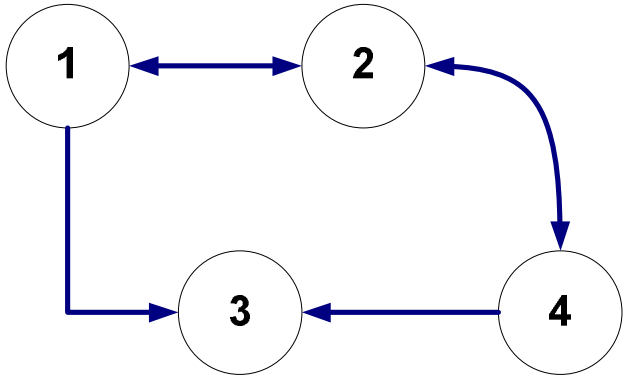
2. Von einer Dokumentensammlung mit insgesamt 100 Dokumenten seien 5% für die Suchanfrage relevant. Für diese Anfrage liefert die Suchmaschine folgende Reihe von relevanten (T) sowie nicht relevanten (F) Suchergebnissen:

FFFFF FFTTT TFFFF FFFFT

Berechnen Sie die Top-10 Precision und Recall sowie das nicht interpolierte Precision bei 20% Recall. **[5 P.]**

III. Ranking (15 Punkte)

1. Gegeben sei folgender Graph.



1a. Geben Sie für diesen Graphen die Link-Matrix *A'* ***mit Teleportation*** an. Die Teleportationswahrscheinlichkeit sei 10%. **[5 P.]**

1b. Gegeben sei die PageRank-Formel: **[10 P.]**

$$\vec{x}^{k+1} = (1 - c)\vec{x}^{k}A + \frac{c}{N}\vec{e}$$

$\vec{e}$ sei $\vec{1}$. Im x_0 befinden sich alle Zufallssurfer auf dem Knoten 3. Berechnen Sie für den gegebenen Graphen den Vektor $\vec{x}$ für die ersten 4 Iterationen der PageRank-Formel (k = 0..3). Geben sie die Werte nicht-normalisiert und auf fünf Nachkommastellen genau an!