Übungen zu Multimedia-Datenbanken
Aufgabenblatt 5

Prof. Dr. Steffen Staab,
Olaf Görlitz, Christoph Ringelstein
Fachbereich Informatik, Universität Koblenz–Landau

Ausgabe: 01.06.2006
Abgabe: 15.06.2006

1. TF-IDF Verfahren (18 Punkte)

Gegeben seien vier Dokumente deren Inhalt aus Newsartikeln extrahiert wurde, indem das Datum und der erste Satz des jeweiligen Artikels herausgefiltert wurde.

`http://www.zdnet.de/news/software/0,39023144,39143535,00.htm`:
ZDNet 15. Mai 2006. Mit der am Wochenende veröffentlichten Firefox-Alpha-2-Testversion haben die Browser-Entwickler von Mozilla einen weiteren Schritt Richtung Finalversion 2.0 getätigt.

`http://www.zdnet.de/news/software/0,39023144,39143283,00.htm`:
ZDNet 3. Mai 2006. Nur rund zwei Wochen nach Firefox 1.5.0.2 steht jetzt der Nachfolger 1.5.0.3 zum kostenlosen Download breit.

`http://www.zdnet.de/news/software/0,39023144,39143131,00.htm`:
ZDNet 26. April 2006. Einem Blogeintrag des Firefox-Entwicklers Ben Goodger zufolge wird der neue Firefox 2.0 ohne das mit Spannung erwartete neue Lesezeichen- und Verlaufsystem "Places" auskommen müssen.

`http://www.zdnet.de/news/software/0,39023144,39142882,00.htm`:
ZDNet 18. April 2006. Pünktlich zu Ostern hat Mozilla die neue Version 1.5.0.2 von Firefox zum kostenlosen Download bereit gestellt.

Termgewichte von Dokumenten können auch mit Hilfe des $tf * idf$-Verfahrens bestimmt werden. Die entsprechende Formel lautet $w_{t,d} = tf_{t,d} \cdot idf_t$, wobei die Termfrequenz $tf_{t,d} = \frac{n_{t,d}}{max_l\, n_{l,d}}$ und die inverse Dokumentenfrequenz $idf_t = \log_2 \frac{N}{df_t}$ ist[1].

[1] N ist die Anzahl aller Dokumente, df_t die Anzahl der Dokumente in denen Term t enthalten ist, $n_{t,d}$ die Erwähnungen von Term t in Dokument d und $max_l\, n_{l,n}$ die häufigste Erwähnung eines Terms.

1. Berechnet für die obigen Dokumente d_1-d_4 und die Terme {*Entwickler, Wochen, Download, Browser, Version, neue, 2006*} zunächst die inverse Dokumentenfrequenz (*idf*) und dann die ensprechenden Termgewichte.

2. Welche Gewichtung erhalten Terme, die in allen Dokumenten auftauchen?

2. Relevance Feedback (10 Punkte)

1. Was ist unter Relevance Feedback zu verstehen und was bedeutet dessen Einsatz für den Anwender?

2. Gegeben seien die Dokumente d_1-d_5 sowie der Anfragevektor q_0. Erstellt ein Diagramm der entsprechenden Feature-Vektoren, mit folgenden Gewichtungen:

	t_1	t_2
d_1	0,8	0,5
d_2	0,5	0,4
d_3	0,6	0,8
d_4	0,3	0,4
d_5	0,1	0,2
q_0	0,7	0,3

3. Vom Benutzer wurden die Dokumente d_2 und d_4 als relevant eingestuft. Berechnet mit dem Rocchio-Verfahren[2] die resultierende Anfragemodifikation und stellt sie ebenfalls graphisch dar. Als Parameter seien $\alpha = 1$ und $\beta = 0,5$ gegeben.

4. Wie lautet der modifizierte Anfragevektor, wenn für die Parameter $\alpha = \beta = 1$ gilt. Ergibt sich daraus eine bessere oder schlechtere Anfrage? (Begründung!)

[2] $q_{neu} = q_{alt} + \frac{\alpha}{|D_r|} \sum_{d_r \in D_r} (d_r - q_{alt}) - \frac{\beta}{|D_i|} \sum_{d_i \in D_i} (d_i - q_{alt})$, wobei D_r, D_i für die Menge der (ir)relevanten Dokumente steht und d_r, d_i für entsprechende Elemente daraus.

3. Bewertung von Retrieval Systemen (12 Punkte)

1. Erklärt die Begriffe *Precision*, *Recall* und *Fallout*.

2. Eine Datenbank mit 80 Fußballerportraits beinhaltet diese zu einem bestimmten Interesse passende Bilder[3]:

Zwei verschiedene Anfragen, die an das System gestellt werden, liefern folgende unterschiedliche Ergebnisse:

1. Anfrage:

2. Anfrage:

Bewertet das System hinsichtlich Precision, Recall und Fallout (Durchschnitt).

3. *Bonusaufgabe* (+6 Punkte)
Ermittelt wie sich Precision und Recall bei der Ausgabe von jeweils nur k Werten (mit k=1..6) verändern.

[3]Quelle: `http://www.promikatur.de/fussballer/fussballer_karikaturen.html`