ÜBUNGSKLAUSUR ZUR VORLESUNG
INFORMATION RETRIEVAL
SS 2011
CHRISTOPH RINGLSTETTER

Verwendet werden darf in der Klausur ein Taschenrechner. Aufgabe 8 ist eine Zusatzaufgabe. Zu erreichen sind also die Punkte der Aufgaben 1-7. Etwa fehlende Punkte können mit Punkten aus Aufgabe 8 ausgeglichen werden.

Aufgabe 1 (5 Punkte)

In einem IR-System werden folgende Dokumente in einer Rankingliste zurückgeliefert.
Die mit einem Pfeil markierten Dokumente sind relevante Dokumente. Berechnen Sie schrittweise die jeweiligen Werte für Recall und Precision. Gehen Sie davon aus, dass insgesamt 8 Dokumente relevant sind.

Ranking	Recall	Precision
1. d_{76}		
2. d_{23} ←		
3. d_{298} ←		
4. d_{412}		
5. d_{99}		
6. d_{87} ←		
7. d_{723} ←		
8. d_{615}		
9. d_{187}		
10. d_{399} ←		
11. d_{12}		
12. d_{54}		

Aufgabe 2 (Indexfile, 15 Punkte)

Gegeben sei die Dokumentenkollektion D. Erstellen Sie einen blockbasierten Index für alle Terme. Dabei sollen die Posting-Listen zu den einzelnen Termen erstellt werden.
Dokumentenkollektion D:
Dokument 1: Für das Autofahren ist eine Lenkberechtigung nötig.
Dokument 2: In Frankreich ist es nicht nötig, eine Kreditkarte zu besitzen.
Dokument 3: Die Hauptstadt von Frankreich ist Paris.
Dokument 4: Der Autoverkehr in Paris ist der Schlimmste in Frankreich.
Dokument 5: Paris ist eine große Stadt.
Dokument 6: Deutschland ist Frankreichs Nachbarland.

1) Konstruieren Sie einen Index durch folgende Zwischenschritte: Tokenisierung, Normalisierung(Lemmatisierung), Sortieren und Gruppieren.
2) Wieweit hängen die Ergebnisse von der Normalisierung der Token (linguistische, Bearbeitung wie Lemmatisierung etc.) ab?

3) Wieweit hängen die Ergebnisse von der Tokenisierung ab? Inwiefern würde sich das Ergebnis ändern, wenn ein Kompositasplitter eingebaut ist.

Aufgabe 3 (Retrievalmodell, 10 Punkte)

Vergleichen Sie das Boolesche Retrievalmodell mit dem Vektorraummodell. Erläutern Sie die Prinzipien beider Modelle und stellen Sie anschließend Gemeinsamkeiten und Unterschiede heraus.

Aufgabe 4 (Vektorraummodell, 20 Punkte)

(a) Beantworten Sie die folgenden Fragen zum Vektorraummodell mit jeweils einem Satz:
- Wie werden die Dokumente repräsentiert?
- Wie werden Anfragen repräsentiert?
- Wodurch ergibt sich die Rangfolge der Dokumente zu einer Anfrage?
- Wodurch ist die Dimension des Vektorraums bestimmt?

(b) Geben Sie die Definition des Cosinus-Maßes an und nennen Sie minimale und maximale Werte für das Cosinus-Maß. In welchen Fällen treten diese Extremwerte auf?

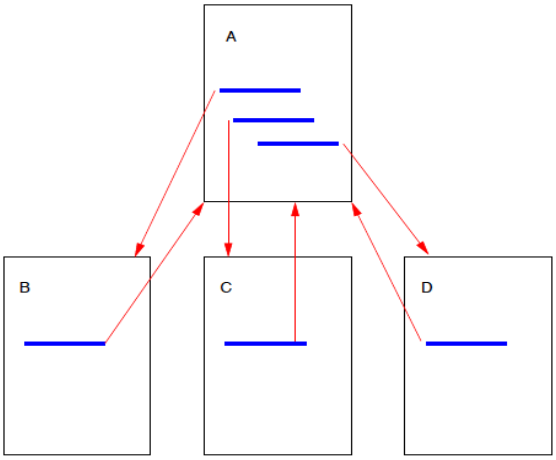
(c) Gegeben seien die Terme T = {a, b, c, d} und die Dokumente
D_1 : (a; b; b; a)
D_2 : (a; c; d; c)
D_3 : (a; d; a; c)
D_4 : (b; b; d; d)
D_5 : (c; d; d; c)
Berechnen Sie die Dokumentvektoren w_i und verwenden Sie dazu $TF_{i;k}/DF_k$ als Termgewicht für Term k in Dokument i.
Stellen Sie eine Rangfolge der Dokumente auf Basis des Cosinusmaßes für folgende Anfrage auf:
Q : (c; d; d) (nehmen sie für die Queryterme jeweils Gewicht wq = 1 an)

Aufgabe 5 (Relevanzfeedback 10 Punkte)

Angenommen es werden Dokumente für die Anfrage *cheap CDs cheap DVDs extremely cheap CDs* gesucht. Zwei Dokumente werden gefunden, D1: *CDs cheap software cheap CDs* und D2: *cheap thrills DVDs*. Der User beurteilt D1 als relevant und D2 als nicht relevant. Für die Berechnung der Termgewichte wird die Termfrequenz verwendet (keine Normalisierung und Verzicht auf Dokumentfrequenz). Es ist nicht notwendig, die Länge der Vektoren zu normalisieren. Berechnen sie zwei Relevance Feedback Durchgänge mit der Methode von Rocchio mit Parametern $\alpha=1$, $\beta=0.75$ und $\gamma=0.25$

Aufgabe 6 (Suchmaschinen, 5 + 5 = 15 Punkte)

(a) Erläutern Sie kurz das "Page Rank" Verfahren (ohne mathematische Details).
(b) Gegeben sei die nachfolgende Verweisstruktur. Geben Sie Übergangsmatrix der Wahrscheinlichkeiten zur Berechnung des "Page Rank" der Dokumente an.



Aufgabe 7 (5 Punkte)

Beschreiben Sie kurz die Idee zum Shingling-Verfahren zur Erkennung von Fast-Duplikaten.

Aufgabe 8 (Zusatzaufgabe Probabilistisches Retrieval – BIR-Modell, 15 Punkte)

Gegeben seien die folgenden Dokumente mit den Termen a,b,c,d,e,f,g,h,i,j,k,l:

d1 = a d i
d2 = a b c i k
d3 = a d f i
d4 = a b c e f g i l
d5 = a b i j k l
d6 = a d l
d7 = a c d e f h j
d8 = a d e h
d9 = a b c d e f g j k l
d10 = a c h k

Zu den Anfragen q1 = (e; f; g), q2 = (d; e; f), q3 = (b; e; f; g), und q4 = (a; c; i) gibt der Benutzer folgende Relevanzbeurteilungen ab:

d_i	1	2	3	4	5	6	7	8	9	10
$r(q_1, d_i)$	$\bar{R}$	$\bar{R}$	$\bar{R}$	R	$\bar{R}$	$\bar{R}$	R	R	R	$\bar{R}$
$r(q_2, d_i)$	$\bar{R}$	$\bar{R}$	R	R	$\bar{R}$	$\bar{R}$	R	R	$\bar{R}$	$\bar{R}$
$r(q_3, d_i)$	$\bar{R}$	$\bar{R}$	$\bar{R}$	R	$\bar{R}$	$\bar{R}$	R	R	R	R
$r(q_4, d_i)$	$\bar{R}$	R	$\bar{R}$	R	R	R	R	$\bar{R}$	$\bar{R}$	$\bar{R}$

(a) Berechnen Sie die Termgewichte c_{ik} (Formel siehe unten) für q1.

$$c_t = \log \frac{|V_t| + \frac{1}{2}}{|V| - |V_t| + 1} + \log \frac{N}{df_t}$$

(b) In welcher Reihenfolge werden die Dokumente für q1 auf Grundlage dieser Werte gerankt?
(c) Welche Reihenfolge ergibt sich wenn man über alle Queries aggregiert und jetzt die c_t auf der Grundlage aller 4 Queries berechnet.