UNIVERSITÄT DES SAARLANDES
Fachrichtung 6.2 - Informatik
Lehrstuhl Prof. Weikum

Informationssysteme SS 2002

Übung 10 Abgabe: Dienstag,02.07.2002 (in der Vorlesung)

Aufgabe 1:

Zeigen Sie, dass für normalisierte Vektor (der Länge 1) das Cosinus-Ähnlichkeitsmaß und die Euklidische Distanz im Vektorraummodell dasselbe Anfrageresultats-Ranking produzieren.

Aufgabe 2:

Betrachten Sie den folgenden Korpus, der aus 4 Dokumenten besteht.

d1 : *Marcus tried to assissinate Caesar* .
d2 : *Marcus was a Roman* .
d3 : *Caesar was a ruler. All Romans were either loyal to Caesar or hated him*.
d4: *Everyone is loyal to someone. People only try to assissinate rulers they are not loyal to*.

Bei der Extraktion von Features dienen die folgenden Wörter als "Stoppwörter" (werden also nicht betrachtet):

- *a, all, and, are, either, everyone, her, him, is, not, only, or, someone, they, to, was, were, who*.

Ferner sollen alle restlichen Wörter nach der folgenden Abbildung auf ihre jeweilige Stammform reduziert werden, und Groß-/Kleinschreibung soll grundsätzlich unwesentlich sein:

- *assissinate assassin*
- *assissinated assassin*
- *assassination assassin*
- *loyality loyal*
- *rulers rule*
- *tried try*
- *hated hate*
- *Roman Rome*
- *Romans Rome*
- *ruler rule*

a) Bestimmen Sie die *idf*-Werte aller Terme (also der nach Stoppwortelimination und Stammformreduktion verbleibenden Wörter).

b) Bestimmen Sie für jedes der vier Dokumente den gewichteten Dokumentvektor aufgrund der *tf*idf*-Formel mit normalisierten *tf*-Werten und (mit Zweierlogarithmus) gedämpften *idf*-Werten:

$$\frac{tf_{ij}}{\max_k tf_{kj}} \log_2 \frac{N}{df_i} \qquad \text{für Term i in Dokument j}$$

c) Betrachten Sie die folgenden Anfragen:

q1: *Who assissinated Caesar?*
q2: *Loyality and assassination*.

Berechnen Sie die Resultatsranglisten für die beiden Anfragen gemäß Cosinus-Ähnlichkeit.

Aufgabe 3

Betrachten Sie die vier Dokumente *d1* bis *d4* der letzten Aufgabe. Nehmen Sie an, die Indexterme seien wie folgt geordnet:

1. *assassin*
2. *Caesar*
3. *hate*
4. *loyal*
5. *Marcus*
6. *people*
7. *Rome*
8. *rule*
9. *try*

Die 9x4-Term-Dokument-Ähnlichkeitsmatrix *A*, die sich aufgrund (einer Variante) der *tf*idf*-Formel ergibt, sieht folgendermaßen aus:

$$A = \begin{pmatrix} 0.5 & 0 & 0 & 0.318 \\ 0.5 & 0 & 0.539 & 0 \\ 0 & 0 & 0.637 & 0 \\ 0 & 0 & 0.318 & 0.539 \\ 0.5 & 0.707 & 0 & 0 \\ 0 & 0 & 0 & 0.637 \\ 0 & 0.707 & 0.318 & 0 \\ 0 & 0 & 0.318 & 0.318 \\ 0.5 & 0 & 0 & 0.318 \end{pmatrix}$$

$(0.707 = \frac{1}{\sqrt{2}})$

Die Singulärwertzerlegung von A in U, Δ und V liefert das folgende Resultat (mit Abrundungen):

$$\Delta = diag\ (1.318,\ 1.003,\ 0.861,\ 0.716)$$

$$U = \begin{pmatrix} -0.331 & -0.170 & 0.385 & 0.207 \\ -0.433 & -0.021 & -0.226 & 0.586 \\ -0.248 & -0.080 & -0.611 & 0.174 \\ -0.306 & -0.407 & -0.144 & -0.304 \\ -0.459 & 0.549 & 0.319 & -0.101 \\ -0.214 & -0.434 & 0.190 & -0.462 \\ -0.360 & 0.462 & -0.277 & -0.452 \\ -0.231 & -0.257 & -0.210 & -0.144 \\ -0.331 & -0.170 & 0.385 & 0.207 \end{pmatrix} \qquad V = \begin{pmatrix} -0.590 & 0.0934 & 0.500 & 0.627 \\ -0.439 & 0.713 & 0.034 & -0.546 \\ -0.513 & -0.126 & -0.826 & 0.196 \\ -0.444 & -0.684 & 0.257 & -0.520 \end{pmatrix}$$

Wenn Sie nur die zwei größten Singulärwerte betrachten, also $k=2$ setzen, erhalten Sie die folgende Matrix A_k:

$$A_k = \begin{pmatrix} 0.241 & 0.070 & 0.245 & 0.310 \\ 0.335 & 0.236 & 0.296 & 0.268 \\ 0.185 & 0.086 & 0.178 & 0.200 \\ 0.199 & -0.115 & 0.258 & 0.458 \\ 0.408 & 0.658 & 0.241 & -0.108 \\ 0.126 & -0.186 & 0.200 & 0.423 \\ 0.322 & 0.539 & 0.184 & -0.107 \\ 0.155 & -0.050 & 0.189 & 0.311 \\ 0.241 & 0.070 & 0.245 & 0.310 \end{pmatrix}$$

a) Vergleichen Sie Dokument d4 mit allen anderen Dokumenten mit der LSI-Methode. Welches ist das ähnlichste?

b) Vergleichen Sie den Term "*Marcus*" mit allen anderen Termen mit der LSI-Methode. Welches ist der ähnlichste Term, welches der zweitähnlichste? Begründen Sie diese Ähnlichkeitswerte intuitiv aufgrund der Termverteilung in den vier Dokumenten.

c) Betrachten Sie erneut die Anfrage aus *Aufgabe 2*:

- **q**: *Loyality and assassination*

Berechnen Sie das Anfrageresultatsranking mit der LSI-Methode. Vergleichen Sie das Ergebnis mit dem Ranking aus Aufgabe 2 (also mit *tf***idf*-Gewichten im vollständigen Termvektorraum und dem Cosinus-Ähnlichkeitsmaß).

d) Gegeben sei ein neues Dokument *d5*:

- *Almost all Romans were loyal to Caesar, but Marcus, a Roman, tried to assassinate Caesar.*

a) Behandeln Sie "*almost*" und "*but*" als zusätzliche Stoppwörter und benutzen Sie dieselben Wortstämme wie in *Aufgabe 2*. Stellen Sie einen neuen Vektor für *d5* auf der Basis der normalisierten *tf*idf*-Termgewichtung auf (wie in *Aufgabe 2*), wobei Sie die bisherigen idf-Werte verwenden können (also auf der Basis der bisherigen Dokumente *d1* bis *d4*). Berücksichtigen Sie das neue Dokument *d5* in der Themen-Dokument-Ähnlichkeitsmatrix *Vk*. sinnvoll ist.

Aufgabe 4:

Betrachten Sie das LSI-Beispiel mit den Backrezept-Dokumenten aus der Vorlesung. Berechnen Sie die approximative SVD auf der Basis der *k=2* größten Singulärwerte. Werten Sie die Anfragen "*baking*" und "*baking bread*" auf dieser Grundlage aus, also anhand der approximierten Term-Dokument-Matrix *A_2*.

Zur Vereinfachung können Sie das Skalarprodukt als Ähnlichkeitsmaß zwischen Vektoren verwenden (also nicht unbedingt das Cosinus-Maß). Vergleichen Sie das Ergebnis mit dem in der Vorlesung besprochenen Resultatsranking, das sich aufgrund der 3 größten Singulärwerte ergab.

Aufgabe 5:

Betrachten Sie den Graphen *G=(V,E)* mit der Knotenmenge *V={1,2,3,4}* und der Kantenmenge *E={(1,2),(2,3),(3,4),(4,2)}*. Bestimmen Sie die Transitionsmatrix *P* für den Random Walk eines Web-Surfers auf *G* mit dem Epsilon-Wert *0.1*. Berechnen Sie die Autoritätswerte von *V* nach der Methode von Page und Brin:

1. iterativ mit der Initialisierung *r(1)=r(2)=r(3)=r(4)=0.25* und 4 Iterationsschritten sowie
2. durch die entsprechende Eigenvektorberechnung bzw. das Lösen eines linearen Gleichungssystems.