Informationssysteme SS 2017
Prof. Dr.-Ing. Stefan Deßloch
Dipl.-Inf Weiping Qu
TU Kaiserslautern, FB Informatik – Lehrgebiet Informationssysteme
http://wwwlgis.informatik.uni-kl.de
Übungsblatt 13: Ausgabe 11.07.2017, Abgabe 17.07.2017, 16:00 Uhr

Aufgabe 1: Distanzmaße auf Strings und Dokumenten (1 P.)

a) Geben Sie für die beiden Worte "Rederei" und "Redezeit" die Hamming-Editier-Distanz, Längste-Gemeinsame-Teilsequenz-Distanz und die Levenshtein-Editier-Distanz an. Benutzen Sie für letztere den in der Vorlesung vorgestellten DP-Algorithmus.

b) Wieso sind die folgenden Varianten der Berechnung der Längste-Gemeinsame-Teilsequenz-Distanz nicht sinnvoll?

- $d(x,y) = \max_{s \in S(x,y)} |s|$
- $d(x,y) = min(|x|,|y|) - \max_{s \in S(x,y)} |s|$

c) Gegeben folgende Dokumente:

$d_1 =$ `Frau Neu ist morgen nicht im Büro.`
$d_2 =$ `Frau Neu, ist Prof. Michel morgen im Büro?`

- Betrachten Sie im folgenden die Dokumente ohne Satzzeichen und geben Sie jeweils für $k = 3$ und $k = 4$ die Menge der Shingles der Dokumente an.
- Berechnen Sie jeweils die Ähnlichkeit der Dokumente basierend auf den Mengen der Shingles unter Verwendung des Jaccard-Koeffizienten.
- Berechnen Sie ebenfalls die Ähnlichkeit der Dokumente basierend auf den einzelnen Worten unter Verwendung des Jaccard-Koeffizienten und vergleichen Sie die Resultate.

Aufgabe 2: Precision und Recall (1 P.)

Stellen Sie sich eine Suchmaschine vor, die eine Anfrage über einem Dokumentenkorpus mit 20 000 Einträgen ausführt und eine geordnete Liste mit 80 Dokumenten zurückgibt. Ein Experte beurteilt die Liste der Ergebnisse und stellt fest, dass die Dokumente auf den folgenden Positionen relevant sind:

$$2, 3, 5, 6, 7, 8, 10, 12, 14, 19,\ \ 21, 23, 29, 30, 37, 42, 46, 48, 50, 55,\ \ 59, 60, 61, 62, 64, 65, 71, 72, 75, 76$$

Weiterhin gibt dieser Experte an, dass der Gesamtkorpus über 400 relevante Dokumente verfügt.

a) Geben Sie für diese Ergebnisliste Precision, Recall, F_1-measure, Precision@10 und Precision@20 an.

b) Wie verhalten sich Precision und Recall generell bzgl. Anzahl der Ergebnisse

c) Die Entwickler der Suchmaschine können Ihnen folgende Angebote machen:

- Recall wird erhöht, wenn mehr Zeit für das initiale Sammeln und Verarbeiten der Dokumente aufgewendet wird.
- Precision wird erhöht, wenn für jede einzelne Suchanfrage eine längere Antwortzeit in Kauf genommen wird.

Diskutieren Sie, für welche Einsatzgebiete von Suchmaschinen diese Angebote akzeptabel sind.

Aufgabe 3: TF*IDF, MMR und LSI (1 P.)

Bei der Analyse klassischer Märchen stellen Sie folgende Verteilung von Termen auf Dokumente $d_1 \dots d_7$ fest:

Term	d_1	d_2	d_3	d_4	d_5	d_6	d_7
Vater	0	5	0	0	0	1	0
Mutter	2	2	3	2	0	0	3
Königin	0	0	0	0	8	1	0
Zwerge	0	0	0	0	4	0	0
Königstochter	0	0	0	0	1	1	0
Wolf	0	0	0	6	0	0	6
Gold	2	0	1	0	1	0	0
Haus	2	5	1	3	4	1	1

a) Sortieren Sie die Dokumente für eine Suche nach {Mutter, Haus} nach dem TF*IDF-Modell.

b) Folgende Tabelle enthält die paarweisen Ähnlichkeiten der Dokumente untereinander. Passen Sie die Sortierung aus Teil a) entsprechend diesen Ähnlichkeiten mittels der in der Vorlesung vorgestellten MMR-Methode an, für $\lambda = 0.5$. Berechnen Sie die top-3 Treffer.

	d_1	d_2	d_3	d_4	d_5	d_6	d_7
d_1	1.0	0.16	0.14	0.18	0.16	0.14	0.16
d_2	0.16	1.0	0.16	0.40	0.18	0.18	0.16
d_3	0.14	0.16	1.0	0.16	0.15	0.16	0.12
d_4	0.18	0.40	0.16	1.0	0.17	0.16	0.17
d_5	0.16	0.18	0.15	0.17	1.0	0.18	0.15
d_6	0.14	0.18	0.16	0.16	0.18	1.0	0.13
d_7	0.16	0.16	0.12	0.17	0.15	0.13	1.0

c) Wenden Sie den in der Vorlesung vorgestellten LSI-Ansatz auf die oben angegebene Term-Dokument-Matrix an, für $k = 3$. Auf der Vorlesungsseite finden Sie weitere Informationen zur entsprechenden Singulärwertzerlegung.

- Berechnen Sie die besten Treffer für die Anfrage {Mutter, Haus} und diskutieren Sie die Unterschiede zum Ergebnis aus Teil a).
- Betrachten Sie die Term-Topic-Matrix U_3 und diskutieren Sie, in welche Topics LSI die Märchenwelt aufgeteilt ist.

Aufgabe 4: PageRank (1 P.)

Ein Webcrawler sammelt folgende Informationen über Webseiten:

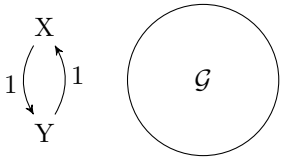
Seite	Enthält Links zu
A	E
B	C
C	D
D	B, F
E	A
F	C, D

a) Geben Sie die passende Zustandsübergangsmatrix P für die Berechnung von PageRank an und berechnen Sie den PageRank-Vektor mit der "Power Iteration"-Methode für $\varepsilon = 0.1$. Wählen Sie ein geeignetes Terminierungskriterium. Wie wirkt sich die Änderung von ε auf die Verteilung des PageRanks aus?

b) Neben PageRank gibt es auch den HITS Algorithmus von J. Kleinberg zur Linkanalyse. HITS unterscheidet zwischen Hubs und Authorities. Gute Hubs sind Seiten, die auf viele gute Authorities verweisen, und gute Authorities sind Seiten, auf die von guten Hubs verwiesen wird. Für eine Adjazenzmatrix A ist der Vektor mit den Hub-Scores gegeben durch den (dominanten) Rechtseigenvektor von AA^T. Der Vektor mit den Authority-Scores ist der (dominante) Rechtseigenvektor von A^TA.

Berechnen Sie dementsprechend für die Adjazenzmatrix A (basierend auf den Daten der obigen Tabelle) den Hub- und Authority-Vektor direkt durch Aufruf der Funktion `eigen(A %*% t(A))$vectors[,1]` für den Fall des Hub-Vektors (analog für den Authority-Vektor) in R, bzw. in einem Programm Ihrer Wahl. Interpretieren Sie die absoluten Werte anhand des Webgraphen und vergleichen Sie insbesondere den Authority-Vektor mit dem PageRank-Vektor aus a). Berechnen Sie ferner die Singulärwertzerlegung für A und vergleichen Sie die erste Spalte von U und V mit dem Hub- bzw. Authority-Vektor; was fällt dabei auf?

c) Gegeben der folgende Linkgraph:



Hierbei steht $\mathcal{G}$ für einen vollständigen gerichteten Graphen mit $|n| = 48$ Knoten, X und Y sind nur gegenseitig verbunden. Welchen Pagerank haben X und Y für $\varepsilon = 0.2$?