

T		beijing	dish	duck	rabbit	recipe	roast	Σ
bin.TF	D_1	0	0	1	0	0	0	
	D_2	1	1	1	0	0	0	
	D_3	0	0	1	1	1	0	
	D_4	0	0	0	1	1	0	
	D_5	1	1	1	0	1	0	
	Q	1	0	1	0	1	0	
RTF	D_1	0	0	3	0	0	0	$3 = DL(D_1)$ $4 = DL(D_2)$ $4 = DL(D_3)$ $2 = DL(D_4)$ $4 = DL(D_5)$ $3 = DL(Q)$
	D_2	1	1	2	0	0	0	
	D_3	0	0	2	1	1	0	
	D_4	0	0	0	1	1	0	
	D_5	1	1	1	0	1	0	
	Q	1	0	1	0	1	0	
TF	D_1	0	0	1	0	0	0	1
	D_2	0,25	0,25	0,5	0	0	0	
	D_3	0	0	0,5	0,25	0,25	0	
	D_4	0	0	0	0,5	0,5	0	
	D_5	0,25	0,25	0,25	0	0,25	0	
	Q	1/3	0	1/3	0	1/3	0	
IDF		$\log \frac{5}{2} = 0,398$	$\log \frac{5}{2} = 0,398$	$\log \frac{5}{4} = 0,097$	$\log \frac{5}{2} = 0,398$	$\log \frac{5}{3} = 0,222$	$:= 0$	
TF_IDF	D_1	0	0	0,097	0	0	0	
	D_2	0,100	0,100	0,049	0	0	0	
	D_3	0	0	0,049	0,100	0,056	0	
	D_4	0	0	0	0,199	0,111	0	
	D_5	0,100	0,100	0,024	0	0,055	0	
	Q	0,133	0	0,032	0	0,074	0	

T		beijing	dish	duck	rabbit	recipe	roast
RTF	Q	1	0	1	0	1	0
TF	Q	1/3	0	1/3	0	1/3	0
IDF		0,398	0,398	0,097	0,398	0,222	0
TF_IDF	Q	0,133	0	0,032	0	0,074	0

	$\text{sim}_{\text{RTF}}(Q, D_i)$	$\text{sim}_{\text{TF}}(Q, D_i)$	$\text{sim}_{\text{TF_IDF}}(Q, D_i)$
D_1	$\frac{3}{\sqrt{3 \cdot 1^2 \cdot \sqrt{3^2}}} \approx 0,58$	$\frac{1/3}{\sqrt{3 \cdot (1/3)^2 \cdot \sqrt{1}}} \approx 0,58$	$\frac{0,003}{0,156 \cdot 0,097} \approx 0,21$
D_2	$\frac{3}{\sqrt{3 \cdot 1^2 \cdot \sqrt{2 \cdot 1^2 + 2^2}}} \approx 0,71$	$\frac{0,25}{\sqrt{3 \cdot (1/3)^2 \cdot \sqrt{0,375}}} \approx 0,55$	$\frac{0,015}{0,156 \cdot 0,150} \approx 0,64$
D_3	$\frac{3}{\sqrt{3 \cdot 1^2 \cdot \sqrt{2^2 + 2 \cdot 1^2}}} \approx 0,71$	$\frac{0,25}{\sqrt{3 \cdot (1/3)^2 \cdot \sqrt{0,375}}} \approx 0,55$	$\frac{0,006}{0,156 \cdot 0,125} \approx 0,29$
D_4	$\frac{1}{\sqrt{3 \cdot 1^2 \cdot \sqrt{2 \cdot 1^2}}} \approx 0,41$	$\frac{1/6}{\sqrt{3 \cdot (1/3)^2 \cdot \sqrt{0,5}}} \approx 0,41$	$\frac{0,008}{0,156 \cdot 0,228} \approx 0,23$
D_5	$\frac{3}{\sqrt{3 \cdot 1^2 \cdot \sqrt{4 \cdot 1^2}}} \approx 0,87$	$\frac{0,25}{\sqrt{3 \cdot (1/3)^2 \cdot \sqrt{0,25}}} \approx 0,87$	$\frac{0,018}{0,156 \cdot 0,154} \approx 0,75$

Aufgabe 3-2 Vector Space Model, TF_IDF

Gegeben eine Dokumentensammlung mit 20.000 Dokumenten, so dass:

- „Maximilian“ kommt in 300 Dokumenten vor.
- „Ludwig“ kommt in 40 Dokumenten vor.
- „University“ kommt in 10.500 Dokumenten vor.
- „in“ kommt in 19.500 Dokumenten vor.
- „Munich“ kommt in 6.000 Dokumenten vor.
- „Germany“ kommt in 10.000 Dokumenten vor.

Nehmen Sie an, daß „in“ als Stopwort behandelt wird.

- a) Berechnen Sie den TF_IDF Vektor für das folgende Dokument. Stopwörter werden entfernt. Die Terme seien alphabetisch geordnet.

D_0 : „Ludwig-Maximilian University Munich, in Munich, Germany“

Lösungsvorschlag:

N = 20.000 (Anzahl der Dokumente in der Dokumentensammlung).

$DF(T)$ = Anzahl der Dokumente, in denen der Term T vorkommt.

$DL(D)$ = Länge des Dokuments D gemessen in Termvorkommen.

$RTF(D, T)$ = Anzahl der Vorkommen von Term T in Dokument D .

$TF(D, T) = \frac{\log(RTF(D, T) + 1)}{\log(DL(D) + 1)}$ logarithmisch-gedämpfte normalisierte Termfrequenz.

$IDF(T) = \log \frac{N}{DF(T)}$

T		germany	ludwig	maximilian	munich	university	Σ
RTF	D_0	1	1	1	2	1	$6 = DL(D_0)$
TF	D_0	$\frac{\log 2}{\log 7} \approx 0,356$	$\frac{\log 2}{\log 7} \approx 0,356$	$\frac{\log 2}{\log 7} \approx 0,356$	$\frac{\log 3}{\log 7} \approx 0,5656$	$\frac{\log 2}{\log 7} \approx 0,356$	
IDF		$\log \frac{20.000}{10.000} \approx 0,301$	$\log \frac{20.000}{40} \approx 2,699$	$\log \frac{20.000}{300} \approx 1,824$	$\log \frac{20.000}{6.000} \approx 0,523$	$\log \frac{20.000}{10.500} \approx 0,28$	
TF_IDF	D_0	0,11	0,96	0,65	0,3	0,1	

Wir verwenden \log_{10} .

Ob die Termfrequenz TF logarithmisch gedämpft wird oder nicht, hat keine signifikante Auswirkung auf das Ranking.

(Man berechne zum Vergleich die ungedämpften TF-Werte: die Reihenfolge $0,356 < 0,5656$ bzw. $\frac{1}{6} < \frac{2}{6}$ bleibt erhalten.)

- b) Ist die *inverse document frequency* (IDF) eines Terms immer endlich? Falls ja, begründen Sie. Falls nein, geben Sie ein Gegenbeispiel!

Lösungsvorschlag:

Ja:

Wenn T nicht in der Dokumentensammlung vorkommt, ist $IDF(T) := 0$ oder undefiniert.

Wenn T in der Dokumentensammlung vorkommt, ist $1 \leq DF(T) \leq N$

Wegen $IDF(T) = \log \frac{N}{DF(T)}$ gilt $IDF(T) \in \left\{ \log \frac{N}{1}, \log \frac{N}{2}, \log \frac{N}{3}, \dots, \log \frac{N}{N} \right\}$

- c) Welchen Wert nimmt der IDF eines Terms an, wenn dieser Term in *allen* Dokumenten vorkommt?

Lösungsvorschlag:

$IDF(T) = \log \frac{N}{N} = \log 1 = 0$.

- d) Wie wirkt sich die Basis des Logarithmus auf die Berechnung des IDF aus?

Lösungsvorschlag:

Für zwei Basen b, b' mit $0 < b < b'$ ist $\log_b b' = c > 1$ eine Konstante.

Wegen $\log_b x = \log_b b' \cdot \log_{b'} x$ ist $IDF_b(T) = c \cdot IDF_{b'}(T) > IDF_{b'}(T)$.

Größere Basis ergibt kleinere IDF-Werte.

Aber für zwei Terme T_1, T_2 ist $\frac{IDF_b(T_1)}{IDF_b(T_2)} = \frac{c \cdot IDF_{b'}(T_1)}{c \cdot IDF_{b'}(T_2)} = \frac{IDF_{b'}(T_1)}{IDF_{b'}(T_2)}$

Das Verhältnis der IDF-Werte von zwei Termen bleibt unverändert.

- e) Vergleichen Sie die Verwendung von TF_IDF mit der Verwendung von Stopwörtern. Lohnt es sich, die beiden Techniken zu kombinieren?

Lösungsvorschlag:

Wenn ein Term T in sehr vielen Dokumenten vorkommt,

ist $DF(T)$ wenig kleiner N und $\frac{N}{DF(T)}$ wenig größer 1 und $IDF(T)$ wenig größer 0.

Stopwörter kommen normalerweise in sehr vielen Dokumenten vor.

Da TF_IDF mit $IDF(T)$ multipliziert, gibt es Stopwörtern automatisch ein Gewicht nahe 0.

Mit dem Maß TF_IDF macht es also mathematisch wenig Unterschied, ob man Stopwörter vorher streicht oder nicht.

Trotzdem hätte es Nachteile, die Stopwörter drinzulassen:

- Effizienz (Stopwörter müssten gespeichert und Maße dafür berechnet werden)
- Intuitiv „gleichwertige“ Terme wie „der“, „die“, „das“ hätten unterschiedliche Gewichte (abhängig von der Dokumentensammlung)
- Dokumente, die nur Stopwörter enthalten (und deshalb irrelevant sind), wären in den Antwortmengen enthalten (wenn auch mit schlechtem Ranking).

Preis für die Stopwortlisten: sie sind ein Mal festgelegt und daher unflexibel.

Stopwörter, die in speziellen Fällen doch relevant wären, bleiben auch in diesen Fällen unberücksichtigt.

Aufgabe 3-4 Lineare Algebra: Wiederholung

Im folgenden betrachten wir der Einfachheit halber nur 2×2 -Matrizen.

Lösungsvorschlag:

Visualisierung von Linearen Transformationen. <http://www.math.duke.edu/education/webfeatsII/Lite-Applets/Eigenvalue/transformations.html>

Matrixoperationen.

$$\lambda \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \lambda a & \lambda b \\ \lambda c & \lambda d \end{pmatrix}$$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} + \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} = \begin{pmatrix} a+a' & b+b' \\ c+c' & d+d' \end{pmatrix}$$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} - \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} = \begin{pmatrix} a-a' & b-b' \\ c-c' & d-d' \end{pmatrix}$$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} = \begin{pmatrix} aa'+bc' & ab'+bd' \\ ca'+dc' & cb'+dd' \end{pmatrix}$$

P_{ij} : Summe der Komponentenprodukte der Zeilen aus der i -ten Zeile der ersten Matrix mit der j -ten Spalte der zweiten Matrix.

Determinante, Inverse Matrix.

Die Identitätsmatrix ist $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Für jede 2×2 -Matrix A gilt $|A| = A$.

Zu einer Matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ ist

- die Determinante $\det A = ad - bc$
- die inverse Matrix $A^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

Die inverse Matrix existiert (offensichtlich) nur, falls $\det A \neq 0$ ist. Dann gilt $AA^{-1} = A^{-1}A = I$.

- a) Sei $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, $B = \begin{pmatrix} 4 & 3 \\ 2 & 1 \end{pmatrix}$. Berechnen Sie AB und BA .

Lösungsvorschlag:

$$AB = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 4 & 3 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 8 & 5 \\ 20 & 13 \end{pmatrix} \quad BA = \begin{pmatrix} 4 & 3 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 13 & 20 \\ 5 & 8 \end{pmatrix}$$

- b) Sei $A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$, $B = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$. Berechnen Sie AB und BA . Was fällt auf?

Lösungsvorschlag:

$AB = BA = I$, da $B = A^{-1}$.

- c) Bestimmen Sie die inversen Matrizen zu $A = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}$ und $B = \begin{pmatrix} 1 & b \\ 1 & b \end{pmatrix}$.

Lösungsvorschlag:

$$A^{-1} = \begin{pmatrix} -1 & 1 \\ -1 & 0 \end{pmatrix}$$

B hat keine Inverse da Determinante = 0.

- d) Sei $A^2 = AA$, $A^3 = AAA$ etc. Berechnen Sie A^2 , A^3 , A^4 , A^{100} für $A = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}$.

Lösungsvorschlag:

$$A^2 = \begin{pmatrix} -1 & 1 \\ -1 & 0 \end{pmatrix} = A^{-1}, \quad A^3 = I, \quad A^4 = A, \quad A^{100} = AA^{99} = A(A^3)^{33} = A^{33} = A.$$

- e) Zeigen Sie: $A = \begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}$ kommutiert mit jeder Diagonalmatrix.

Für $a \neq d$ kommutiert A nur mit Diagonalmatrizen.

Lösungsvorschlag:

Sei $B = \begin{pmatrix} x & y \\ z & w \end{pmatrix}$ eine beliebige Matrix. Dann gilt

$$AB = \begin{pmatrix} ax & ay \\ dz & dw \end{pmatrix} \quad BA = \begin{pmatrix} ax & dy \\ az & dw \end{pmatrix}$$

Wenn B Diagonalmatrix ist, ist $y = z = 0$, also $AB = BA = \begin{pmatrix} ax & 0 \\ 0 & dw \end{pmatrix}$

Wenn $a \neq d$ und $AB = BA$ ist, ist $dy = ay$ und $dz = az$.

Wegen $a \neq d$ ist $y = z = 0$.

Also ist B eine Diagonalmatrix: $B = \begin{pmatrix} x & 0 \\ 0 & w \end{pmatrix}$.

Bemerkung: Mit der Einschränkung $a \neq d$ ist A keine Skalarmatrix, welche per Definition mit allen Matrizen kommutieren.

- f) Zeigen Sie: $(AB)^T = (B^T)(A^T)$.

Lösungsvorschlag:

Sei $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $B = \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix}$. Dann ist

$$AB = \begin{pmatrix} aa' + bc' & ab' + bd' \\ ca' + dc' & cb' + dd' \end{pmatrix} \quad (AB)^T = \begin{pmatrix} aa' + bc' & ca' + dc' \\ ab' + bd' & cb' + dd' \end{pmatrix}$$

$$(B^T)(A^T) = \begin{pmatrix} a' & c' \\ b' & d' \end{pmatrix} \begin{pmatrix} a & c \\ b & d \end{pmatrix} = \begin{pmatrix} a'a + c'b & a'c + c'd \\ b'a + d'b & b'c + d'd \end{pmatrix}$$

$$\begin{pmatrix} 0 & r & u \\ 0 & 0 & r \\ 1 & 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 0 & u & r \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 0 & u & r \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 0 & r & u \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}$$

$$\begin{pmatrix} u & 0 & r \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$