

Aufgabe 23

Termgewichte in Aufgabe 22

tf-idf(t,d)	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
w1	2,21	0	0	3,68	8,84	0	0	1,48	5,9	0,74
w2	4,12	3,09	0	6,17	0	0	4,63	0,51	1,54	5,15
w3	0	0,74	5,16	0	0,74	3,68	8,84	0	1,48	0

Beim Cosinus-Maß kann der Beitrag der Anfrage-Länge zum Nenner ignoriert werden:

$$\cos(q, d) = \frac{\sum(W(q,t) \cdot W(d,t))}{\sqrt{\sum(W(d,t) \cdot W(d,t))}}$$

{q für Query, t für Term, d für Dokument}

d.h. Skalarprodukt durch euklidische Länge

Der Tf-Idf Wert von Terme in Anfrage:

$$W_{q,1} = 1 \cdot \log_{10} 6 = 0.74$$

$$W_{q,2} = 2 \cdot \log_{10} 6 = 1.48$$

hier auch andere Varianten denkbar etwa: Query Terme nur mit 1 bzw 2 Gewichten ohne idf.

$$\cos(q, d1):$$

$$= 0.74 \cdot 2.21 / \sqrt{(2.21 \cdot 2.21) + (4.12 \cdot 4.12)}$$

$$= 0.35$$

$$\cos(q, d2):$$

$$= 1.48 \cdot 0.74 / \sqrt{(3.09 \cdot 3.09) + (0.74 \cdot 0.74)}$$

$$= 0.34$$

$$\cos(q, d3):$$

$$= 1.48$$

$$\cos(q, d4):$$

$$= 0.74 \cdot 3.68 / \sqrt{(3.68 \cdot 3.68) + (6.17 \cdot 6.17)}$$

$$= 0.38$$

$$\cos(q, d5):$$

$$= (0.74 \cdot 8.84 + 1.48 \cdot 0.74) / \sqrt{8.84 \cdot 8.84 + 0.74 \cdot 0.74}$$

$$= 0.86$$

$$\cos(q, d6):$$

$$= 1.48$$

$$\cos(q, d7):$$

$$= 1.48 \cdot 8.84 / \sqrt{(4.63 \cdot 4.63 + 8.84 \cdot 8.84)}$$

$$= 1.31$$

$$\cos(q, d8):$$

$$= 0.74 \cdot 1.48 / \sqrt{1.48 \cdot 1.48 + 0.51 \cdot 0.51}$$

$$= 0.70$$

$$\cos(q, d9):$$

$$= (0.74 \cdot 5.9 + 1.48 \cdot 1.48) / \sqrt{5.9 \cdot 5.9 + 1.54 \cdot 1.54 + 1.48 \cdot 1.48}$$

$$= 1.04$$

$$\cos(q, d10)$$

$$= 0.74 \cdot 0.74 / \sqrt{0.74 \cdot 0.74 + 5.15 \cdot 5.15}$$

$$= 0.11$$

Ergebnisse:

cos(q,d)	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
q	0,35	0,34	1,48	0,38	0,86	1,48	1,31	0,7	1,04	0,11

Aufgabe24

Angenommen $q = \{\text{Radfahren}\}$ und in D1 kommen 10 mal $\{\text{Radfahren}\}$ und 1 mal $\{\text{Schwimmen}\}$ vor und in D2 10 mal $\{\text{Radfahren}\}$ und 10 mal $\{\text{Schwimmen}\}$.

tf(t,d)	D1	D2
Radfahren	10	10
Schwimmen	1	10

tf(q,t)	Query
Radfahren	1
Schwimmen	0

Skalarprodukt von Anfrage und Dokumente:

$$q \cdot D1 = 10 \cdot 1 + 0 \cdot 10 = 10$$

$$q \cdot D2 = 10 \cdot 1 + 0 \cdot 10 = 10$$

Kosinus-Abstandsmaes von Anfrage und Dokumente:

$$\cos(q, D1) = 10 / \sqrt{10 \cdot 10 + 1 \cdot 1} = 0.995$$

$$\cos(q, D2) = 10 / \sqrt{10 \cdot 10 + 10 \cdot 10} = 0.707$$

Wenn man das Skalarprodukt als Ma auswhlt, dann bleibt die hnlichkeit der D1, D2 zur Query gleich, die zustzliche irrelevante Information in D2 schadet also nichts. Wenn man hingegen den Kosinus-Abstand als Ma nimmt, dann ist die hnlichkeit zwischen D1 und Query grer als die hnlichkeit zwischen D2 und der Query. Warum: der Winkel ist durch die zustzliche Information vergrert.

Aufgabe 25

Das Verfahren zeigt folgende Schritte:

1. Bilde eine leere Akkumulatorliste: $\text{Scores}[N] = 0$
2. Sortiere die Anfrage-Terme nach Gewicht $W_{t,q}$.
3. Fr jeden Term t der Anfrage: Fr jeden Eintrag $\langle d, \text{tf}(d,t) \rangle$ setzt $\text{Scores}[d] := \text{Scores}[d] + W_{t,q} \cdot \text{tf}(d,t)$
4. Teile jeden Akkumulatorwert durch die Dokumentlnge $\text{Length}[d]$.
5. Identifiziere die k hchsten Werte und gebe die korrespondierenden Dokumente aus.

$$w1: ((d4:14), (d2:9), (d8:4), (d1:1))$$

$$w2: ((d2:14), (d5:6), (d7:5), (d1:2), (d9:1), (d0:1))$$

$$w3: ((d9:3), (d2:1), (d6:1))$$

$$w4: ((d3:11), (d6:11), (d7:6), (d0:3), (d1:3), (d2:1), (d5:1))$$

Step 2

Berechnen zuerst die Termgewichte von Anfrage: (log Basis 2)

$$W_{1,q} = 1 * \log 10/4 = 1.32$$

$$W_{2,q} = 1 * \log 10/6 = 0.74$$

$$W_{3,q} = 1 * \log 10/3 = 1.74$$

$$W_{4,q} = 1 * \log 10/7 = 0.51$$

Sortieren die Termgewichte von Anfragen (wie Postinglisten von Dokumenten: nach absteigendem Gewicht) und bekommen die Reihenfolge: $W_{3,q}$ $W_{1,q}$ $W_{2,q}$ $W_{4,q}$

Step 3

Term W_3

$w_3: ((d_9:3), (d_2:1), (d_6:1))$

$$\text{Scores}(d_9) = 0 + W_{3,d_9} * W_{3,q} = 3 * \log 10/3 * \log 10/3 = 9.05$$

$$\text{Scores}(d_2) = 0 + W_{3,d_2} * W_{3,q} = 1 * \log 10/3 * \log 10/3 = 3.02$$

$$\text{Scores}(d_6) = 0 + W_{3,d_6} * W_{3,q} = 1 * \log 10/3 * \log 10/3 = 3.02$$

Term W_1

$w_1: ((d_4:14), (d_2:9), (d_8:4), (d_1:1))$

$$\text{Scores}(d_4) = 0 + W_{1,d_4} * W_{1,q} = 14 * \log 10/4 * \log 10/4 = 24.46$$

$$\text{Scores}(d_2) = 3.02 + W_{1,d_2} * W_{1,q} = 3.02 + 15.73 = 18.75$$

$$\text{Scores}(d_8) = 0 + W_{1,d_8} * W_{1,q} = 4 * \log 10/4 * \log 10/4 = 6.90$$

$$\text{Scores}(d_1) = 0 + W_{1,d_1} * W_{1,q} = 1 * \log 10/4 * \log 10/4 = 1.75$$

Term W_2

$w_2: ((d_2:14), (d_5:6), (d_7:5), (d_1:2), (d_9:1), (d_0:1))$

$$\text{Scores}(d_2) = 18.75 + W_{2,d_2} * W_{2,q} = 18.75 + 14 * \log 10/6 * \log 10/6 = 26.35$$

$$\text{Scores}(d_5) = 0 + W_{2,d_5} * W_{2,q} = 6 * \log 10/6 * \log 10/6 = 3.26$$

$$\text{Scores}(d_7) = 0 + W_{2,d_7} * W_{2,q} = 2.72$$

$$\text{Scores}(d_1) = 1.75 + W_{2,d_1} * W_{2,q} = 2.84$$

$$\text{Scores}(d_9) = 9.05 + W_{2,d_9} * W_{2,q} = 9.59$$

$$\text{Scores}(d_0) = 0 + W_{2,d_0} * W_{2,q} = 0.54$$

Term W_4 :

$w_4: ((d_3:11), (d_6:11), (d_7:6), (d_0:3), (d_1:3), (d_2:1), (d_5:1))$

Für jeden Eintrag $\langle d, \text{tf}_d, t \rangle$ setzt $\text{Scores}[d] := \text{Scores}[d] + W_{t,d} * W_{t,q}$

$$\text{Scores}(d_3) = 0 + W_{4,d_3} * W_{4,q} = 11 * \log 10/7 * \log 10/7 = 2.91$$

$$\text{Scores}(d_6) = 3.02 + W_{4,d_6} * W_{4,q} = 3.02 + 11 * \log 10/7 * \log 10/7 = 5.93$$

$$\text{Scores}(d_7) = 2.72 + W_{4,d_7} * W_{4,q} = 2.72 + 6 * \log 10/7 * \log 10/7 = 4.31$$

$$\text{Scores}(d_0) = 0.54 + W_{4,d_0} * W_{4,q} = 0.54 + 3 * \log 10/7 * \log 10/7 = 1.33$$

$$\text{Scores}(d_1) = 2.84 + W_{4,d_1} * W_{4,q} = 3.63$$

$$\text{Scores}(d_2) = 26.35 + W_{4,d_2} * W_{4,q} = 26.61$$

$$\text{Scores}(d_5) = 3.26 + W_{4,d_5} * W_{4,q} = 3.52$$

	d0	d1	d2	d3	d4	d5	d6	d7	d8	d9
Scores(d)	1,33	3,63	26,61	2,92	24,46	3,52	5,93	4,31	6,9	9,59

Step 4

	d0	d1	d2	d3	d4	d5	d6	d7	d8	d9
w1	0	1	9	0	14	0	0	0	4	0
w2	1	2	14	0	0	6	0	5	0	1

w3	0	0	1	0	0	0	1	0	0	3
w4	3	3	1	11	0	1	11	6	0	0

1. Variante: length(d) ist Anzahl der Terme in Dokument. Scores(d) = Scores(d)/length(d)

$\text{Scores}(d_0) = \text{Scores}(d_0)/\text{length}(d_0) = 1.33/4 = 0.3325$
 $\text{Scores}(d_1) = \text{Scores}(d_1)/\text{length}(d_1) = 3.63/6 = 0.605$
 $\text{Scores}(d_2) = \text{Scores}(d_2)/\text{length}(d_2) = 26.61/25 = 1.0644$
 $\text{Scores}(d_3) = 2.92/11 = 0.2655$
 $\text{Scores}(d_4) = 24.46/14 = 1.7471$
 $\text{Scores}(d_5) = 3.52/7 = 0.5029$
 $\text{Scores}(d_6) = 5.93/12 = 0.4942$
 $\text{Scores}(d_7) = 4.31/11 = 0.3918$
 $\text{Scores}(d_8) = 6.9/4 = 1.725$
 $\text{Scores}(d_9) = 9.59/4 = 2.3975$

2. Variante: length(d) ist euklidische Länge. Scores(d) = Scores(d)/ sqrt(sum(Wt,d*Wt,d))

$\text{Scores}(d_0) = 1.33/\sqrt{W_{2,d_0}^2 + W_{4,d_0}^2} = 1.53$
 $\text{Scores}(d_1) = 3.63/\sqrt{W_{1,d_1}^2 + W_{2,d_1}^2 + W_{4,d_1}^2} = 1.45$
 $\text{Scores}(d_2) = 26.61/\sqrt{W_{1,d_2}^2 + W_{2,d_2}^2 + W_{3,d_2}^2 + W_{4,d_2}^2} = 1.68$
 $\text{Scores}(d_3) = 2.92/\sqrt{W_{4,d_3}^2} = 0.52$
 $\text{Scores}(d_4) = 24.46/\sqrt{W_{1,d_4}^2} = 1.32$
 $\text{Scores}(d_5) = 3.52/\sqrt{W_{2,d_5}^2 + W_{4,d_5}^2} = 0.79$
 $\text{Scores}(d_6) = 5.93/\sqrt{W_{3,d_6}^2 + W_{4,d_6}^2} = 1.00$
 $\text{Scores}(d_7) = 4.31/\sqrt{W_{2,d_7}^2 + W_{4,d_7}^2} = 0.90$
 $\text{Scores}(d_8) = 6.9/\sqrt{W_{1,d_8}^2} = 1.30$
 $\text{Scores}(d_9) = 9.59/\sqrt{W_{2,d_9}^2 + W_{3,d_9}^2} = 1.82$

Step 5

Angenommen K= 10:

1. Variante: D9,D4,D8,D2,D1,D5,D6,D7,D0,D3

2. Variante: D9,D2,D0,D1,D4,D8,D6,D7,D5,D3