Information Retrieval

Blatt 4

Aufgabe 11

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	R
d_1	1	1	1	1	0	0	0	0	0	0
d_2	1	0	0	0	1	0	0	0	0	1
d_3	0	0	0	1	1	1	1	1	0	1
d_4	0	1	1	0	0	1	1	0	1	0
n_i	2	2	2	2	2	2	2	1	1	
r_i	1	0	0	1	2	1	1	1	0	
p_i	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{5}{6}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{6}$	
q_i	$\frac{1}{2}$	$\frac{5}{6}$	$\frac{5}{6}$	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{2}$	

- $q = \begin{pmatrix}0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1\end{pmatrix}$
- $d_5 = \begin{pmatrix}1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1\end{pmatrix}$.
- $q \cap d_5 = \begin{pmatrix}0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1\end{pmatrix}$

$$\begin{aligned} sim(q, d_5) &= \sum_i \log \frac{p_i}{1-p_i} + \sum_i \log \frac{1-q_i}{q_i} \\ &= \log \frac{\frac{1}{6}}{\frac{5}{6}} + \log \frac{\frac{1}{6}}{\frac{5}{6}} + \log \frac{\frac{1}{6}}{\frac{5}{6}} + \log \frac{\frac{1}{2}}{\frac{1}{2}} \\ &= -3 \log 5 \end{aligned}$$

Aufgabe 12

$$U_2 = \begin{pmatrix} 0.1261 & -0.1237 \\ 0.2856 & 0.2940 \\ 0.2856 & 0.2940 \\ 0.4818 & -0.5959 \\ 0.2208 & -0.3834 \\ 0.3803 & 0.0343 \\ 0.5684 & 0.3538 \\ 0.1922 & -0.2852 \\ 0.1881 & 0.3195 \end{pmatrix}, \; S_2 = \begin{pmatrix} 3.4785 & 0 \\ 0 & 2.2721 \end{pmatrix}, \; V_2 = \begin{pmatrix} 0.3390 & -0.0579 \\ 0.0997 & -0.2232 \\ 0.6685 & -0.6480 \\ 0.6544 & 0.7259 \end{pmatrix}$$

- $q_1 = \begin{pmatrix}0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0\end{pmatrix}$
- $q_2 = \begin{pmatrix}0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0\end{pmatrix}$

Über $U_2^t \cdot q_i$ werden die Anfragen in den Dokument-Topic-Raum überführt. Dort kann die Ähnlichkeit zu den Dokumenten $d_1', \ldots, d_4'$, die Zeilen in V_2, über die Cosinusähnlichkeit bestimmt werden.

Als Ergebnisse erhalten wir für q_1

- $sim(q_1, d_1') = 0.3366$

- $sim(q_1, d'_2) = 0.1745$
- $sim(q_1, d'_3) = 0.8593$
- $sim(q_1, d'_4) = 0.3433$

und für q_2

- $sim(q_2, d'_1) = 0.2350$
- $sim(q_2, d'_2) = -0.0554$
- $sim(q_2, d'_3) = 0.1410$
- $sim(q_2, d'_4) = 0.9601$

Somit erhalten wir als Trefferlisten für q_1 und q_2:

q_1**:** d3, d4, d3, d2

q_2**:** d4, d1, d3, d2