Řešení příkladu 1

- $N = 3$
- tf_i - term frequency of term t in document d_i
- df_t - document frequency of term t
- $idf_t = log_{10}(\frac{N}{df_t})$ - inverted document frequency of term t
- tfq - term frequency of term t in query

Intermediate results

term	tf1	tf2	tf3	tfq	df	idf
Plzeň	1	0	1	0	2	0.17609
je	2	1	2	0	3	0
krásné	2	0	1	1	2	0.17609
město	1	0	1	1	2	0,17609
a	1	0	0	0	1	0.47712
to	1	0	0	0	1	0.47712
místo	1	1	0	0	2	0,17609
Ostrava	0	1	0	0	1	0.47712
ošklivé	0	1	0	0	1	0.47712
Praha	0	0	1	0	1	0.47712
také	0	0	1	0	1	0.47712
hezčí	0	0	1	0	1	0.47712

TF-IDF weights

- $w_i = (1 + log_{10}\ tf_i) * idf_t$ - *tf-idf* weight of term t in document i

term	w1	w2	w3	q
Plzeň	0.17609	0	0.17609	0
je	0	0	0	0
krásné	0.22910	0	0.17609	0.17609
město	0.17609	0	0.17609	0.17609
a	0.47712	0	0	0
to	0.47712	0	0	0
místo	0.17609	0.17609	0	0
Ostrava	0	0.47712	0	0
ošklivé	0	0.47712	0	0
Praha	0	0	0.47712	0
také	0	0	0.47712	0

term	w1	w2	w3	q
hezčí	0	0	0.47712	0

Normalized weights

- $||w_i||_2 = \sqrt{\sum w_{ij}^2}$

vector	norm
w1	0,77511
w2	0,69735
w3	0,88088
q	0,24903

Weights

term	w1	w2	w3	q
Plzeň	0.22718	0	0,19990	0
je	0	0	0	0
krásné	0.29557	0	0,19990	0,70710
město	0.22718	0	0,19990	0,70710
a	0.61555	0	0	0
to	0.61555	0	0	0
místo	0.22718	0,25251	0	0
Ostrava	0	0,68419	0	0
ošklivé	0	0,68419	0	0
Praha	0	0	0,54164	0
také	0	0	0,54164	0
hezčí	0	0	0,54164	0

Cosine similiarity

- $cos(\vec{q}, \vec{d}) = \frac{\vec{q} * \vec{d}}{|\vec{q}| * |\vec{d}|} = \frac{\sum q_i d_i}{\sqrt{\sum q_i^2} * \sqrt{\sum d_i^2}} = \sum q_i * d_i$ *(normalized)*
- q - krásné město
- $cos(\vec{q}, \vec{d_1}) = 0,70710 * 0,29557 + 0,70710 * 0,22718 = 0,36964$ ✅
- $cos(\vec{q}, \vec{d_2}) = 0,70710 * 0 + 0,70710 * 0 = 0$
- $cos(\vec{q}, \vec{d_3}) = 0,70710 * 0,19990 + 0,70710 * 0,19990 = 0,28270$

K dotazu q_1 je nejrelevantnější dokument d_1.