Assignment 06

Digital Libraries and Foundations of Information Retrieval

Winter semester 2022

1542011 Franka Brunen, 1365848 Andreas Schneider

Task 1: Vector Space Model 2+3+3+2+1+4 Points

(a)
$$\begin{aligned}
idf_{t_1} &= \log\tfrac{4}{2} &= 1\\
idf_{t_2} &= \log\tfrac{4}{4} &= 0\\
idf_{t_3} &= \log\tfrac{4}{2} &= 1\\
idf_{t_4} &= \log\tfrac{4}{2} &= 1\\
idf_{t_5} &= \log\tfrac{4}{1} &= 2\\
idf_{t_6} &= \log\tfrac{4}{2} &= 1\\
idf_{t_7} &= \log\tfrac{4}{1} &= 2
\end{aligned}$$

(b)

	d_1	d_2	d_3	d_4
biscuits (t_1)	0	3	0	4
stollen (t_2)	0	0	0	0
lebkuchen (t_3)	1	7	0	0
macaroons (t_4)	0	1	0	2
brownies (t_5)	0	2	0	0
cookies (t_6)	0	1	0	4
pastries (t_7)	0	0	8	0

(c)

	d_1	d_2	d_3	d_4
biscuits (t_1)	0	$\frac{3}{8}$	0	$\frac{4}{6}$
stollen (t_2)	0	0	0	0
lebkuchen (t_3)	1	$\frac{7}{8}$	0	0
macaroons (t_4)	0	$\frac{1}{8}$	0	$\frac{2}{6}$
brownies (t_5)	0	$\frac{2}{8}$	0	0
cookies (t_6)	0	$\frac{1}{8}$	0	$\frac{4}{6}$
pastries (t_7)	0	0	1	0

(d) $q = (0, 0, \frac{1}{\sqrt{3}}, 0, 0, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}})$

(e) Summing up the normalized tf-idf scores on query terms, documents 1 through 3 will score either identical on a highest sum of 1, or document 2 will score better/worse, since its sum is distributed.

(f)
$$\begin{aligned}
sim(d_1, q) &= 0 + 0 + 1 * \tfrac{1}{\sqrt{3}} + 0 + 0 + 0 + 0 &= 1 * \tfrac{1}{\sqrt{3}}\\
sim(d_2, q) &= 0 + 0 + \tfrac{7}{8} * \tfrac{1}{\sqrt{3}} + 0 + 0 + \tfrac{1}{8} * \tfrac{1}{\sqrt{3}} + 0 &= 1 * \tfrac{1}{\sqrt{3}}\\
sim(d_3, q) &= 0 + 0 + 0 + 0 + 0 + 0 + 1 * \tfrac{1}{\sqrt{3}} &= 1 * \tfrac{1}{\sqrt{3}}\\
sim(d_4, q) &= 0 + 0 + 0 + 0 + 0 + \tfrac{4}{6} * \tfrac{1}{\sqrt{3}} + 0 &= \tfrac{4}{6} * \tfrac{1}{\sqrt{3}}
\end{aligned}$$

Having equal weights, it did not make a difference that the sum is distributed.

Task 2: Vector Space Model 1+2+3+4+5 Points

(a) It is irrelevant except if the query consists of a term with $idf = 0$. Then the query will return all documents equally, instead of a ranking by term frequency.

(b)

$\lvert d_1\rvert$	$\lvert\lvert d_1\rvert\rvert$	$\lvert d_2\rvert$	$\lvert\lvert d_2\rvert\rvert$	$\lvert d_3\rvert$	$\lvert\lvert d_3\rvert\rvert$	$\lvert d_4\rvert$	$\lvert\lvert d_4\rvert\rvert$
=27	$=\sqrt{427}$	=27	$=\sqrt{269}$	=105	$=\sqrt{10025}$	=1015	$=\sqrt{1000125}$

(c) $sim(d_1, q) = \frac{1}{\sqrt{427}}$
$sim(d_2, q) = \frac{10}{\sqrt{269}}$
$sim(d_3, q) = 0$
$sim(d_4, q) = 0$

(d) $sim(d_1, q) = \frac{5}{\sqrt{427}}$
$sim(d_2, q) = \frac{5}{\sqrt{269}}$
$sim(d_3, q) = \frac{5}{\sqrt{10025}}$
$sim(d_4, q) = \frac{5}{\sqrt{1000125}}$

The documents do not have the same score since each document vector has a different length. It could be reasonable to believe, that a document is more relevant if the relative frequency of the query term is higher, even though the absolute frequency is lower, e.g. one paragraph about the term alone versus in a long text (including other topics).

(e) $sim(d_1, q) = \frac{1}{\sqrt{427}}$
$sim(d_2, q) = 0$
$sim(d_3, q) = 0$
$sim(d_4, q) = \frac{10}{\sqrt{1000125}}$

In task (c), document 1 was less relevant than the other document, document 2. In this task, document 1 is more relevant than the other document, document 4.

This is because document 4 is shorter than document 2, making the identical absolute number of occurrences less impactful.