Exercise 1

a) Compute idf

The value of the idf (inverse document frequency) for a particular term T is given by

$$\mathrm{idf}_T = \log\left(\frac{n}{\mathrm{df}_T}\right),$$

where n is the *total number of documents in the corpus*, in our example $n = 10'000$, and df_T denotes the *document frequency* of term T, i.e. the number of documents that contain T. From our example we have the following document frequencies:

$$\begin{aligned}
\mathrm{df}_{T_1} &= 100,\\
\mathrm{df}_{T_2} &= 200,\\
\mathrm{df}_{T_3} &= 200,\\
\mathrm{df}_{T_4} &= 100.
\end{aligned}$$

From this we can directly compute the idf-values for all our 4 terms:

$$\begin{aligned}
\mathrm{idf}_{T_1} = \mathrm{idf}_{T_4} &= \log_{10}\left(\frac{10'000}{100}\right) = 2,\\
\mathrm{idf}_{T_2} = \mathrm{idf}_{T_3} &= \log_{10}\left(\frac{10'000}{200}\right) \approx 1.69897
\end{aligned}$$

b) Compute tf-idf weighting

The tf-idf for a term T_k and a document d_j is given by

$$\text{tf-idf}_{kj} = \mathrm{tf}_{kj} \cdot \mathrm{idf}_{T_k}$$

where t_{kj} denotes the number of occurrences of term T_k in document d_j, i.e. the numbers that are given in the table. With the above formula we get the following tf-idf weights:

Document \Term	T_1	T_2	T_3	T_4
D_1	$4 \cdot 2 = 8$	$4 \cdot 1.69897 = 6.79588$	0	$1 \cdot 2 = 2$
D_2	8	$2 \cdot 1.69897 = 3.39794$	$10 \cdot 1.69897 = 16.9897$	$5 \cdot 2 = 10$
D_3	8	3.39794	$2 \cdot 1.69897 = 3.39794$	$30 \cdot 2 = 60$

Table 1: tf-idf values

c) Order of documents for query (T_3, T_4)

Assume we have a document D_4 consisting only of the two terms T_3 and T_4. We can compute the tf-idf value for this document as follows:

Term	T_3	T_4
tf-idf	$1 \cdot 1.69897 = 1.69897$	$1 \cdot 2 = 2$

The terms T_1 and T_2 will of course have a tf-idf of 0 because neither of them occur in the given document. Representing each document as a vector of the above computed tf-idf weights gives the following vectors:

$$\begin{aligned}
D_1 &= (8, 6.79588, 0, 2), \\
D_2 &= (8, 3.39794, 16.9897, 10), \\
D_3 &= (8, 3.3979, 3.39794, 60), \\
D_4 &= (0, 0, 1.69897, 2)
\end{aligned}$$

We can compare two documents by comparing their relativ angle – or rather by looking at the cosine between two vector, which can be found by building the dot product between two normalized vectors:

$$\begin{aligned}
(v_1, v_2) &= \cos(\theta) \cdot |v_1| \cdot |v_2| \\
\iff \cos(\theta) &= \frac{(v_1, v_2)}{|v_1| \cdot |v_2|},
\end{aligned}$$

where (v_1, v_2) denotes the dot product between two vectors v_1 and v_2, i.e.

$$\sum_i v_1^{(i)} \cdot v_2^{(i)},$$

and $|v_i]$ is given by $\sqrt{(v_i, v_i)}$.

Using this comparison scheme, we can compute scores for the comparison of D_4 with all the other documents. We start by computing the lengths of the single vectors:

$$\begin{aligned}
|D_1| &= \sqrt{8^2 + 6.79588^2 + 0^2 + 2^2} \approx 10.6857 \\
|D_2| &\approx 21.5452 \\
|D_3| &\approx 60.7214 \\
|D_4| &\approx 2.62421
\end{aligned}$$

The cosine similarities are now given by

$$\begin{aligned}
\text{cosSim}(D_1, D_4) &= \frac{(D_1, D_4)}{|D_1| \cdot |D_4|} = \frac{8 \cdot 0 + 6.79588 \cdot 0 + 0 \cdot 1.69897 + 2^2}{10.6857 \cdot 2.62421} = \frac{2^2}{10.6857 \cdot 2.62421} \approx 0.1426 \\
\text{cosSim}(D_2, D_4) &= \frac{16.9897 \cdot 1.69897 + 10 \cdot 2}{21.5452 \cdot 2.62421} \approx 0.8643 \\
\text{cosSim}(D_3, D_4) &= \frac{3.39794 \cdot 1.69897 + 60 \cdot 2}{60.7214 \cdot 2.62421} \approx 0.7893
\end{aligned}$$

Therefore we have the ordering $D_2 > D_3 > D_1$, which means that document D_4 is most similar to D_2.