Assignment 4

David Boerema (s3683869)
Marios Souroulla (s4765125)
Marius Captari (s4865928)
Max Valk (s3246922)

Group 3

October 19, 2021

A

a)

To construct vector q we simply increment each index that corresponds to a term in the search query by the amount of occurrences of that term in said query. Then, we apply the given formula to fold this query within the SVD space:

$$\hat{\vec{q}} = \vec{q}^T S^{-1} \tag{1}$$

b)

To calculate the cosine similarity, we first reduce the dimensionality of q, Dt_i (where i is the column of the current document) and S by `max_dimensions`. Then, we scale q and Dt_i by S and calculate the cosine similarity. In the end, we get the following scores:

```
C1: 0.997
C2: 0.895
C3: 0.997
C4: 0.979
C5: 0.846
M1: -0.176
M2: -0.163
M3: -0.157
M4: -0.043
```

This shows that all C documents have a good score and all M documents have a bad one, as expected. C1 and C3 score the best overall.

c)

Calculating the cosine similarity between two terms in our semantic space we proceed the same way as when calculating the similarity between a given query and a document. We start off by getting the rows representing the terms we want to compare through their respective index. We then reduce the dimension of both vectors by `max_dimensions`. It is now possible to scale both terms by S. Once both terms are scaled we can calculate the cosine similarity. By iterating through every pair of terms in our semantic space we get the cosine similarity for each pair. The matrix containing these results is represented in Table 1.

B

The goal of this exercise was to take the term by document matrix and calculate the Term Frequency - Inverse Document Frequency (TF-IDF) matrix and Log-Entropy (LE) matrix.

	computer	eps	graph	human	interface	minors	response	survey	system	time	trees	user
computer	1.000	0.816	0.233	0.796	0.864	0.245	0.979	0.733	0.908	0.979	0.202	0.999
eps	0.816	1.000	-0.372	0.999	0.996	-0.360	0.680	0.205	0.983	0.680	-0.401	0.837
graph	0.233	-0.372	1.000	-0.403	-0.288	1.000	0.428	0.833	-0.195	0.428	0.999	0.197
human	0.796	0.999	-0.403	1.000	0.992	-0.392	0.655	0.172	0.976	0.655	-0.432	0.818
interface	0.864	0.996	-0.288	0.992	1.000	-0.276	0.743	0.291	0.995	0.743	-0.318	0.882
minors	0.245	-0.360	1.000	-0.392	-0.276	1.000	0.439	0.839	-0.183	0.439	0.999	0.209
response	0.979	0.680	0.428	0.655	0.743	0.439	1.000	0.857	0.803	1.000	0.399	0.970
survey	0.733	0.205	0.833	0.172	0.291	0.839	0.857	1.000	0.381	0.857	0.814	0.707
system	0.908	0.983	-0.195	0.976	0.995	-0.183	0.803	0.381	1.000	0.803	-0.226	0.923
time	0.979	0.680	0.428	0.655	0.743	0.439	1.000	0.857	0.803	1.000	0.399	0.970
trees	0.202	-0.401	0.999	-0.432	-0.318	0.999	0.399	0.814	-0.226	0.399	1.000	0.166
user	0.999	0.837	0.197	0.818	0.882	0.209	0.970	0.707	0.923	0.970	0.166	1.000

Table 1: Matrix for cosine similarity between terms in the document.

a) TF-IDF

For this matrix we applied the formula below (2) which corresponds to the TF-IDF weighting scheme.

$$A_{ij} = f_{ij} \log \frac{n}{\sum_j x(f_{ij})} \tag{2}$$

Below is the resulting matrix:

$$A = \begin{bmatrix} 1.50 & 1.50 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.50 & 1.50 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.10 & 1.10 & 1.10 \\ 1.50 & 0.00 & 0.00 & 1.50 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 1.50 & 0.00 & 1.50 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.50 & 1.50 \\ 0.00 & 1.50 & 0.00 & 0.00 & 1.50 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.50 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.50 \\ 0.00 & 1.10 & 1.10 & 2.20 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.50 & 0.00 & 0.00 & 1.50 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.10 & 1.10 & 1.10 & 0.00 \\ 0.00 & 1.10 & 1.10 & 0.00 & 1.10 & 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}$$

b) Log-Entropy

For this matrix we used the log-entropy formula below (3) which corresponds to the LE weighting scheme

$$A_{ij} = \log 1 + f_{ij} \left[1 + \left(\sum_j \frac{p_{ij} \log p_{ij}}{\log n} \right) \right] \tag{3}$$

Below is the resulting matrix:

$$A = \begin{bmatrix} 0.48 & 0.48 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.47 & 0.47 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.35 & 0.35 & 0.35 \\ 0.47 & 0.00 & 0.00 & 0.47 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.47 & 0.00 & 0.47 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.46 & 0.46 \\ 0.00 & 0.48 & 0.00 & 0.00 & 0.48 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.48 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.48 \\ 0.00 & 0.38 & 0.38 & 0.60 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.48 & 0.00 & 0.00 & 0.48 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.47 & 0.47 & 0.47 & 0.00 \\ 0.00 & 0.37 & 0.37 & 0.00 & 0.37 & 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}$$