Martyna Stasiak id. 156071

INFORMATION RETRIEVAL – SHORT EXERCISES II – VECTOR SPACE MODEL AND LATENT SEMANTIC INDEXING

I. Consider a set of terms **T** = {t_1, t_2, t_3, t_4} and the following collection of two documents: **D1** = {t_1 t_2 t_1 t_2 t_3} and **D2** = {t_4 t_2 t_2 t_3}. Consider query **Q** = {t_1 t_4}. Represent D1, D2, and Q using TF (normalized Bag-Of-Words).

TF	t_1	t_2	t_3	t_4
D1	2/2	2/2	1/2	0
D2	0	2/2	1/2	1/2
Q	1/1	0	0	1/1

max = 2

Compute IDFs for all four terms (note that only D1 and D2 are included in the collection).

	t_1	t_2	t_3	t_4
IDF	log2	log1=0	log1=0	log 2

$idf_i = \log(\frac{N}{df_i})$

II. Consider the below term-document matrix **C** for the bag-of-words representation of five documents **D1-D5** in the space of six terms t_1-t_6. Using the SVD factorization method, matrix **C** has been decomposed into matrices **K**, **S**, and $\mathbf{D}^T$ given below. The rank of **C** is 4 (4 ≤ min{6,5}), so 4 concepts (semantic dimensions) were discovered.

C =

	D1	D2	D3	D4	D5
t_1	5	5	0	0	1
t_2	4	5	1	1	0
t_3	5	4	1	1	0
t_4	0	0	4	4	4
t_5	0	0	5	5	5
t_6	1	1	4	4	4

K = (term × concept)

	c1	c2	c3	c4
t_1	-0.27	0.55	-0.78	0
t_2	-0.29	0.47	0.44	-0.71
t_3	-0.29	0.47	0.44	0.71
t_4	-0.45	-0.29	-0.01	0
t_5	-0.56	-0.36	-0.02	0
t_6	-0.50	-0.18	-0.05	0

S =

13.74	0	0	0
0	10.88	0	0
0	0	1.36	0
0	0	0	1

D^T = (concept × document)

	D1	D2	D3	D4	D5
c1	-0.32	-0.32	-0.52	-0.52	-0.5
c2	0.63	0.63	-0.25	-0.25	-0.29
c3	-0.02	-0.02	0.41	0.41	-0.82
c4	0.71	-0.71	0	0	0

Answer the following questions:

- What is the informativeness value of the most important concept? Answer: 13.74
- Based on the informativeness values of all concepts, which seems the most obvious value for the reduced number of dimensions k? Answer: k = 2
- What is the (numerical value of the) mapping of term t_6 to the most important (informative) concept? Answer: -0.50
- What is the vector representing document D3 in the space of four discovered concepts? Answer: [-0.52 , -0.25 , 0.41 , 0]