Assignment 6

Exercise 1 (IIR 18) [3 P.]

Given the singular value decomposition of the matrix C from the lecture.

C	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
wood	1	0	0	1	1	0
tree	0	0	0	1	0	1

Σ	1	2	3	4	5
1	2.16	0.00	0.00	0.00	0.00
2	0.00	1.59	0.00	0.00	0.00
3	0.00	0.00	1.28	0.00	0.00
4	0.00	0.00	0.00	1.00	0.00
5	0.00	0.00	0.00	0.00	0.39

U	1	2	3	4	5
1	0.44	-0.30	-0.57	0.58	-0.25
2	0.13	-0.33	0.59	0.00	-0.73
3	0.48	-0.51	0.37	0.00	0.61
4	0.70	0.35	-0.15	-0.58	-0.16
5	0.26	0.65	0.41	0.58	0.09

V	1	2	3	4	5
1	0.75	-0.29	-0.28	0.00	0.53
2	0.28	-0.53	0.75	0.00	-0.29
3	0.20	-0.19	-0.45	0.58	-0.63
4	0.45	0.63	0.20	0.00	-0.19
5	0.33	0.22	-0.12	0.58	-0.41
6	0.12	0.41	0.33	0.58	0.22

1. Calculate the reduced matrix C_3. That is the term-document matrix C reduced to 3 dimensions (see slides).

2. Compare the rankings of the query "ship ocean" for the matrices C and C_3: Rank the documents after relevance.

Solution

a)

To calculate C_3 we need the matrix Σ with the first three single values (other values are "zeroed out"):

Σ_3	1	2	3	4	5
1	2.16	0.00	0.00	0.00	0.00
2	0.00	1.59	0.00	0.00	0.00
3	0.00	0.00	1.28	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00

Furthermore V has to be transposed since the formula for singular value decomposition is $C = U\Sigma V^T$:

V^T	1	2	3	4	5	6
1	0.75	0.28	0.20	0.45	0.33	0.12
2	-0.29	-0.53	-0.19	0.63	0.22	0.41
3	-0.28	0.75	-0.45	0.20	-0.12	0.33
4	0.00	0.00	0.58	0.00	0.58	0.58
5	0.53	-0.29	-0.63	-0.19	-0.41	0.22

Now we can multiply U with Σ_3 and get the following matrix A:

A	1	2	3	4	5
1	0.950	-0.477	-0.730	0.000	0.000
2	0.281	-0.525	0.755	0.000	0.000
3	1.037	-0.811	0.474	0.000	0.000
4	1.512	0.557	-0.192	0.000	0.000
5	0.562	1.034	0.525	0.000	0.000

After multiplying A by matrix V^T we finally get the matrix C_3:

C_3	d_1	d_2	d_3	d_4	d_5	d_6
ship	1.055	-0.029	0.609	-0.019	0.296	-0.322
boat	0.152	0.923	-0.184	-0.053	-0.113	0.068
ocean	0.880	1.076	0.148	0.051	0.107	-0.052
wood	1.026	-0.016	0.283	0.993	0.645	0.346
tree	-0.025	0.003	-0.320	1.009	0.350	0.665

Remark:
The false matrix C_3 if multiplied with V instead of V^T:

C_3	d_1	d_2	d_3	d_4	d_5
ship	0.433	0.116	-0.295	-0.423	1.102
boat	0.215	0.053	-0.812	0.438	-0.174
ocean	0.645	0.039	-1.112	0.275	0.486
wood	1.252	-0.697	0.081	-0.111	0.761
tree	0.816	-0.811	0.382	0.305	-0.333

b)

C : d_1, d_2, d_3 or d_1, d_3, d_2

C_3 : d_1, d_2, d_3, d_5, d_4, d_6

Computation:

1. d_1: 1.055 + 0.88 = 1.935
2. d_2: (-0.029) + 1.076 = 1.047
3. d_3: 0.609 + 0.148 = 0.757
4. d_5: 0.296 + 0.107 = 0.403
5. d_4: (-0.019) + 0.051 = 0.032
6. d_6: (-0.322) + (-0.052) = -0.374

Exercise 2 (IIR 19) [3 P.]

The shingle representations of three documents are as follows: $d_3 = (0,0,1,0,0,0,1)^T$, $d_4 = (0,0,1,0,0,0,0)^T$, $d_5 = (1,1,1,0,1,1,1)^T$

We will use sketches of size 2. The two elements of a sketch are defined by the permutations $(2 \times n + 2)$ mod 7 and $(4 \times n + 1)$ mod 7. Based on this setup, what are the estimates of the three Jaccard coefficients $J(d_3, d_4)$, $J(d_3, d_5)$, and $J(d_4, d_5)$? Use the kind of table introduced in class to visualize the permutations and to calculate the final sketches.

Solution

	d_3	d_4	d_5
s_1	0	0	1
s_2	0	0	1
s_3	1	1	1
s_4	0	0	0
s_5	0	0	1
s_6	0	0	1
s_7	1	0	1

Hash functions for the permutation:
$h(x) = (2x + 2) \mod 7$
$g(x) = (4x + 1) \mod 7$

	d_3 slot		d_4 slot		d_5 slot	
		∞		∞		∞
		∞		∞		∞
$h(1) = 4$	–	∞	–	∞	4	4
$g(1) = 5$	–	∞	–	∞	5	5
$h(2) = 6$	–	∞	–	∞	6	4
$g(2) = 2$	–	∞	–	∞	2	2
$h(3) = 1$	1	1	1	1	1	1
$g(3) = 6$	6	6	6	6	6	2
$h(4) = 3$	–	1	–	1	–	1
$g(4) = 3$	–	6	–	6	–	2
$h(5) = 5$	–	1	–	1	5	1
$g(5) = 0$	–	6	–	6	0	0
$h(6) = 0$	–	1	–	1	0	0
$g(6) = 4$	–	6	–	6	4	0
$h(7) = 2$	2	1	–	1	2	0
$g(7) = 1$	1	1	–	6	1	0

Final sketches: $d_3 = (1,1)$, $d_4 = (1,6)$, $d_5 = (0,0)$

$$J(d_3, d_4) = \frac{1+0}{2} = 1/2$$
$$J(d_3, d_5) = \frac{0+0}{2} = 0$$
$$J(d_4, d_5) = \frac{0+0}{2} = 0$$