

# Computer-Linguistische Anwendungen

CLA | B.Sc. | LMU



# Eigenschaften der SVD die wir Nutzen:

- Wichtig:
  - Jeder Singulärwert ist eine Angabe wie wichtig jede Dimension ist
- Wenn wir die weniger wichtigen Dimensionen auf 0 setzen, erhalten wir wichtige Informationen, aber werden überflüssige Details los
- Diese Details sind oft:
  - Noise - Reduzierte SVD Vektoren sind oft bessere Repräsentationen, weil sie weniger noise enthalten
  - Jene, welche Dinge unähnlicher machen, auch wenn sie ähnlicher sind - Reduzierte SVD Vektoren sind auch hier besser da sie Ähnlichkeiten besser abbilden.
- Analogie für “Weniger Details sind Besser”
  - Bild einer blauen Blume
  - Bild einer gelben Blume
  - Das Weglassen der Farb-Dimension erleichtert die Ähnlichkeit zu sehen



# Reduktion der Dimensionalität auf 2D

$U$	1	2	3	4	5
ship	-0.44	-0.30	0.00	0.00	0.00
boat	-0.13	-0.33	0.00	0.00	0.00
ocean	-0.48	-0.51	0.00	0.00	0.00
wood	-0.70	0.35	0.00	0.00	0.00
tree	-0.26	0.65	0.00	0.00	0.00

$\Sigma_2$	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	0.00	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

$V^T$	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$
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	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00

Wir setzen nur Singulärwerte in  $\Sigma$  auf 0. Dies sorgt dafür, dass die entsprechenden Dimensionen in  $U$  und  $V^T$  auch auf 0 gesetzt sind, wenn wir das Produkt von  $C = U \Sigma V^T$  bilden.



# Reduktion der Dimensionalität auf 2D

$U$	1	2	3	4	5
ship	-0.44	-0.30	0.00	0.00	0.00
boat	-0.13	-0.33	0.00	0.00	0.00
ocean	-0.48	-0.51	0.00	0.00	0.00
wood	-0.70	0.35	0.00	0.00	0.00
tree	-0.26	0.65	0.00	0.00	0.00

$\Sigma_2$	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	0.00	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

$V^T$	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$
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	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00

Wir setzen nur Singulärwerte in  $\Sigma$  auf 0. Dies sorgt dafür, dass die entsprechenden Dimensionen in  $U$  und  $V^T$  auch auf 0 gesetzt sind, wenn wir das Produkt von  $C = U \Sigma V^T$  bilden.



# Reduktion der Dimensionalität auf 2D

$C_2$	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$
ship	0.85	0.52	0.28	0.13	0.21	-0.08
boat	0.36	0.36	0.16	-0.20	-0.02	-0.18
ocean	1.01	0.72	0.36	-0.04	0.16	-0.21
wood	0.97	0.12	0.20	1.03	0.62	0.41
tree	0.12	-0.39	-0.08	0.90	0.41	0.49

$U$	1	2	3	4	5	$\Sigma_2$	1	2	3	4	5
ship	-0.44	-0.30	0.57	0.58	0.25	1	2.16	0.00	0.00	0.00	0.00
boat	-0.13	-0.33	-0.59	0.00	0.73	2	0.00	1.59	0.00	0.00	0.00
ocean	-0.48	-0.51	-0.37	0.00	-0.61	3	0.00	0.00	0.00	0.00	0.00
wood	-0.70	0.35	0.15	-0.58	0.16	4	0.00	0.00	0.00	0.00	0.00
tree	-0.26	0.65	-0.41	0.58	-0.09	5	0.00	0.00	0.00	0.00	0.00

$V^T$	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$
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

$$C_2 \sim C$$



# Original Matrix $C$ vs. reduzierte Matrix $C_2 = U \Sigma_2 V^T$

$C$	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$
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

$C_2$	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$
ship	0.85	0.52	0.28	0.13	0.21	-0.08
boat	0.36	0.36	0.16	-0.20	-0.02	-0.18
ocean	1.01	0.72	0.36	-0.04	0.16	-0.21
wood	0.97	0.12	0.20	1.03	0.62	0.41
tree	0.12	-0.39	-0.08	0.90	0.41	0.49

Wir können  $C_2$  als zweidimensionale Repräsentation der Matrix  $C$  betrachten. Hier haben wir eine Dimensionsreduktion auf 2D vorgenommen.



# Aufgabe

$C$	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$
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

Berechnen Sie die Ähnlichkeit zwischen  $d_2$  und  $d_3$  der Ursprungs-Matrix  $C$  und der reduzierten Matrix  $C_2$

$C_2$	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$
ship	0.85	0.52	0.28	0.13	0.21	-0.08
boat	0.36	0.36	0.16	-0.20	-0.02	-0.18
ocean	1.01	0.72	0.36	-0.04	0.16	-0.21
wood	0.97	0.12	0.20	1.03	0.62	0.41
tree	0.12	-0.39	-0.08	0.90	0.41	0.49



# Warum ist die reduzierte Matrix $C_2$ besser als $C$ ?

$C$	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$
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

$C_2$	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$
ship	0.85	0.52	0.28	0.13	0.21	-0.08
boat	0.36	0.36	0.16	-0.20	-0.02	-0.18
ocean	1.01	0.72	0.36	-0.04	0.16	-0.21
wood	0.97	0.12	0.20	1.03	0.62	0.41
tree	0.12	-0.39	-0.08	0.90	0.41	0.49

Ähnlichkeit von  $d_2$  und  $d_3$  in  $C$ :

0

Ähnlichkeit von  $d_2$  und  $d_3$  in  $C_2$ :

$0.52 * 0.28 + 0.36 * 0.16 + 0.72 * 0.36 + 0.12 * 0.20 + -0.39 * -0.08 \approx 0.52$   
(dot product, cosine)





# Word2vec: Lernen via Matrix Faktorisierung

- Sammeln und gewichten der Cooccurrence Matrix
- Berechnen der SVD der Cooccurrence Matrix
- Reduzieren des Raumes (Dimensionalitätsreduktion)
- Embeddings = Links-Singuläre Vektoren (Linke Matrix)



# Reduktion der Dimensionalität auf 2D

$U$	1	2	3	4	5
ship	-0.44	-0.30	0.00	0.00	0.00
boat	-0.13	-0.33	0.00	0.00	0.00
ocean	-0.48	-0.51	0.00	0.00	0.00
wood	-0.70	0.35	0.00	0.00	0.00
tree	-0.26	0.65	0.00	0.00	0.00

$\Sigma_2$	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	0.00	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

$V^T$	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$
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	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00

Wir setzen nur Singulärwerte in  $\Sigma$  auf 0. Dies sorgt dafür, dass die entsprechenden Dimensionen in  $U$  und  $V^T$  auch auf 0 gesetzt sind, wenn wir das Produkt von  $C = U \Sigma V^T$  bilden.

