

# Computer-Linguistische Anwendungen

CLA | B.Sc. | LMU



# Input Matrix



Slides recycled from B. Roth / H. Schuetze

# Input Matrix: Outline

- Erinnerung:

Wir möchten die **Embedding Parameter** durch Matrix **Faktorisierung** lernen

- Dafür brauchen wir eine **Input Matrix** für die Matrix **Faktorisierung**
- Wir schauen uns an wie eine Input Matrix erzeugt wird (**PPMI**, **Cooccurrence**)
- Diese Methode kommt aus der **Information Retrieval**
  - Demnach betrachten wir ein Beispiel aus diesem Feld



# Vektor Repräsentationen: Wörter vs. Dokumente

- Statistische NLP & Deep Learning:

**Embeddings** als Model für Wort-Ähnlichkeit

- In **Information Retrieval**:

Vektor **Repräsentationen** als Model für Anfrage-Dokument Ähnlichkeit

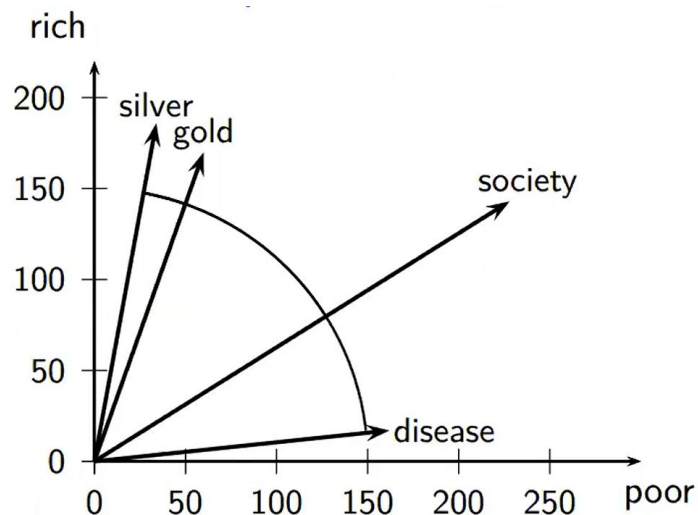
- Einfache **Suchmaschine**:

- BenutzerIn gibt Anfrage ein
- Anfrage wird in **Anfrage-Vektor** transformiert
- Dokumente werden in **Dokument-Vektoren** transformiert
- Ähnlichkeit des Anfrage und Dokument Vektors werden verglichen
- Dokument(e) mit höchster Ähnlichkeit werden zurückgegeben

tf-idf



# Basis für WordSpace: Cooccurrence → Ähnlichkeit



Die Ähnlichkeit zwischen zwei Worten ist der **Kosinus** des **Winkels** zwischen den beiden Vektoren


Kleiner Winkel: **silver** und **gold** sind ähnlich

Mittlerer Winkel: *silver* und *society* sind **nicht sehr ähnlich**

Großer Winkel: *silver* und *disease* sind noch weniger ähnlich




# Dokumente geordnet nach Ähnlichkeit zur Anfrage

automobile prices 

[All](#) [News](#) [Shopping](#) [Images](#) [Maps](#) [More](#) [Settings](#) [Tools](#)

About 69,500,000 results (0.41 seconds)

## Automobile aus Deutschland - 2,4 Mio. Gebrauchte- & Neuwagen

 [www.autoscout24.de/auto/mobile](https://www.autoscout24.de/auto/mobile)

4.3 ★★★★★ rating for autoscout24.de

Jetzt schnell, einfach & unkompliziert Autos aller Marken in Ihrer Nähe finden.

Europaweite Angebote · Alle Fahrzeugdetails · Kostenlos verkaufen · Ausgezeichneter Service

Modelle: VW Turan, Kia Sportage, BMW X1, Audi A3

AutoScout24 Neuwagen

from €8,000.00

verschiedene Modelle

Neuwagen

from €10K

verschiedene Modelle

Fabrikneue Autos

from €12.5K

verschiedene Modelle

## Kelley Blue Book - New and Used Car Price Values, Expert Car Reviews

<https://www.kbb.com/>

Check KBB car price values when buying and selling new or used vehicles. Recognized by consumers and the automotive industry since 1926.

[Resale Value](#) · [Used Car Prices](#) · [New Cars](#) · [Motorcycles](#)

[ ... ]

## NADAguides: New Car Prices and Used Car Book Values

<https://www.nadaguides.com/>

Research the latest new car prices, deals, used car values, specs and more. NADA Guides is the leader in accurate vehicle pricing and vehicle information.

[New Car Prices & Used Car ...](#) · [Motorcycles](#) · [RV Prices and Values](#) · [Trucks](#)

[ ... ]

[ ... ]



# Wörter geordnet nach Ähnlichkeit zum Anfrage-Wort

silver

1.000 silver, 0.865 bronze, 0.842 gold, 0.836 medal, 0.826, medals, 0.761 relay, 0.740 medalist,  
0.737 coins ...



# Setup für Cooccurrence **Count Matrix**

		$w_2$				
		rich	poor	silver	society	disease
$w_1$	rich					
	poor					
	silver					
	society					
	disease					





# Cooccurrence Count Matrix

		$w_2$				
		rich	poor	silver	society	disease
$w_1$	rich	$CC(w_1, w_2)$	$CC(w_1, w_2)$	$CC(w_1, w_2)$	$CC(w_1, w_2)$	$CC(w_1, w_2)$
	poor	$CC(w_1, w_2)$	$CC(w_1, w_2)$	$CC(w_1, w_2)$	$CC(w_1, w_2)$	$CC(w_1, w_2)$
	silver	$CC(w_1, w_2)$	$CC(w_1, w_2)$	$CC(w_1, w_2)$	$CC(w_1, w_2)$	$CC(w_1, w_2)$
	society	$CC(w_1, w_2)$	$CC(w_1, w_2)$	$CC(w_1, w_2)$	$CC(w_1, w_2)$	$CC(w_1, w_2)$
	disease	$CC(w_1, w_2)$	$CC(w_1, w_2)$	$CC(w_1, w_2)$	$CC(w_1, w_2)$	$CC(w_1, w_2)$



# PPMI der Cooccurrence Count Matrix

PMI: pointwise mutual information

$$\text{PMI}(w, c) = \log \frac{P(wc)}{P(w)P(c)}$$

PPMI =

positive pointwise mutual information

$$\text{PPMI}(w, c) = \max(0, \text{PMI}(w, c))$$

More generally (with offset  $k$ ):

$$\text{PPMI}(w, c) = \max(0, \text{PMI}(w, c) - k)$$

		$w_2$				
		rich	poor	silver	society	disease
$w_1$	rich	$\text{PPMI}(w_1, w_2)$	$\text{PPMI}(w_1, w_2)$	$\text{PPMI}(w_1, w_2)$	$\text{PPMI}(w_1, w_2)$	$\text{PPMI}(w_1, w_2)$
	poor	$\text{PPMI}(w_1, w_2)$	$\text{PPMI}(w_1, w_2)$	$\text{PPMI}(w_1, w_2)$	$\text{PPMI}(w_1, w_2)$	$\text{PPMI}(w_1, w_2)$
	silver	$\text{PPMI}(w_1, w_2)$	$\text{PPMI}(w_1, w_2)$	$\text{PPMI}(w_1, w_2)$	$\text{PPMI}(w_1, w_2)$	$\text{PPMI}(w_1, w_2)$
	society	$\text{PPMI}(w_1, w_2)$	$\text{PPMI}(w_1, w_2)$	$\text{PPMI}(w_1, w_2)$	$\text{PPMI}(w_1, w_2)$	$\text{PPMI}(w_1, w_2)$
	disease	$\text{PPMI}(w_1, w_2)$	$\text{PPMI}(w_1, w_2)$	$\text{PPMI}(w_1, w_2)$	$\text{PPMI}(w_1, w_2)$	$\text{PPMI}(w_1, w_2)$

Diese Matrix wird verwendet für die Matrix Faktorisierung und wird für die Word Embeddings verwendet.

# Anwendungsbeispiel: Wort-Dokument Matrix (Information Retrieval)

	doc 1	doc 2	doc 3	doc 4	doc 5	query
anthony	5.25	3.18	0.0	0.0	0.0	0.35
brutus	1.21	6.10	0.0	1.0	0.0	0.0
caesar	8.59	2.54	0.0	1.51	0.25	0.0
calpurnia	0.0	1.54	0.0	0.0	0.0	0.0
cleopatra	2.85	0.0	0.0	0.0	0.0	0.0
mercy	1.51	0.0	1.90	0.12	5.25	0.88
worser	1.37	0.0	0.11	4.15	0.25	0
...						

... als nächstes: Matrix Faktorisierung