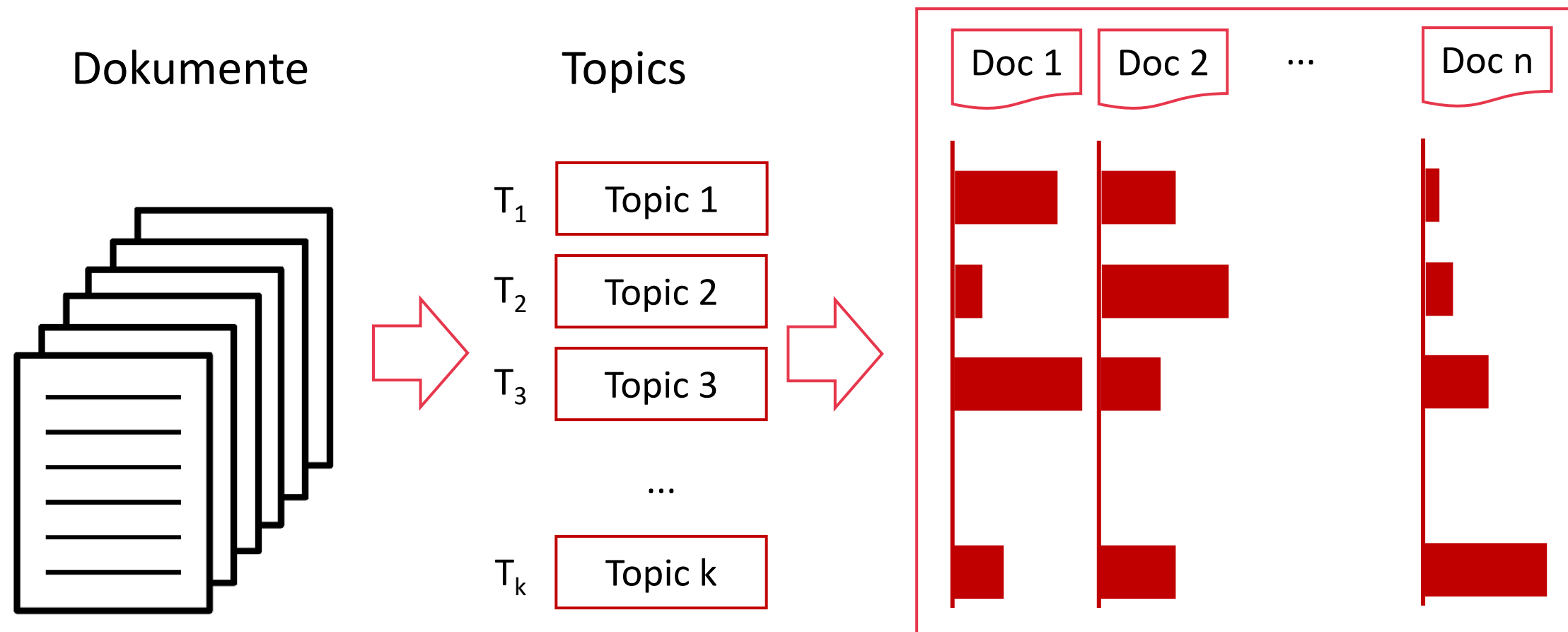


Topic Analyse

Gesucht: Versteckte/Latente Struktur

- 1) Welche Themen gibt es überhaupt?
- 2) Wie sind deren Verteilungen pro Dokument?



Motivation für Topic-Analyse (I)

Worum geht es in Texten?

- Identifikation von Themen
- Identifikation von verwandten Themen
- Zoom-In / Zoom-Out: spezifischere und breitere Themen
- Veränderung der Themen über die Zeit

Was könnten beispielhafte Fragestellungen in den Nachrichten, sozialen Netzwerken (Facebook, Twitter, ...), User Foren sein?

Wie könnte man Themen charakterisieren?

Was ist der Unterschied zu Clustering?

Motivation für Topic-Analyse (II)

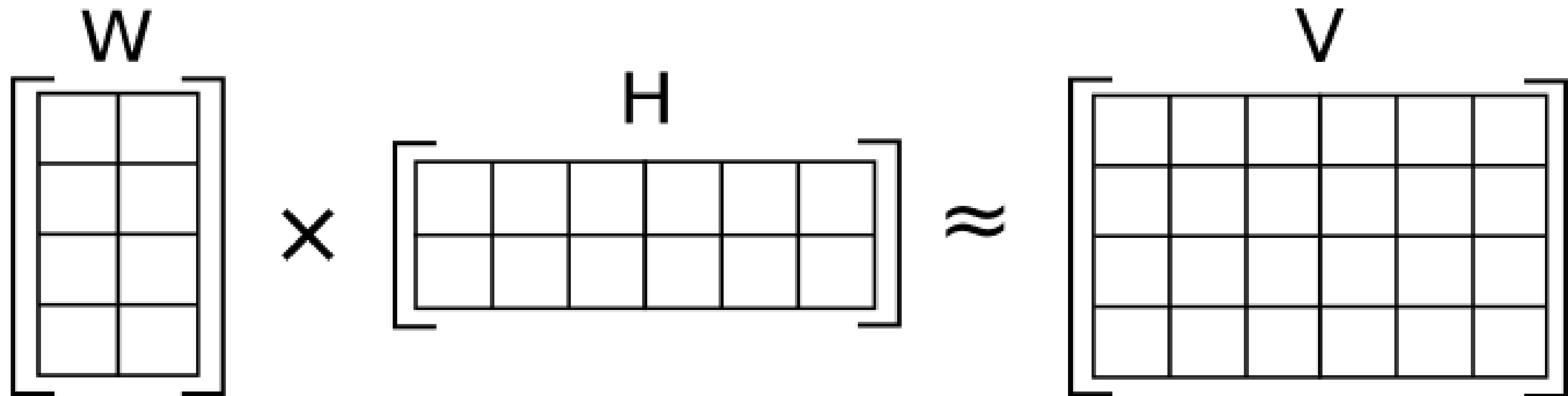
Topics unterstützen

- Information Retrieval (Suche)
- Klassifikation (zusätzliche Features)
- Datenexploration und -verständnis
- Recommendations
- Organisation großer Textdatenbestände

Anwendungsfälle

- Topics aus Job-Beschreibungen extrahieren und zu Bewerbern zuordnen
- Lesern relevante Themenlisten mit Artikeln präsentieren
- Kundenkommentare inhaltlich analysieren

Non-Negative Matrix Factorization (NMF)



Latent Dirichlet Allocation (LDA)

Versuche, diese beiden Ziele zu erreichen:

1. In jedem Dokument: Weise die Worte möglichst wenigen Topics zu (für hohe Aussagekraft)
2. In jedem Topic: Weise möglichst wenigen Termen eine hohe Wahrscheinlichkeit zu (spezifische Terme)

Beide Ziele sind gegensätzlich

- Wird ein Dokument d genau einem Topic T zugewiesen (#1 gut), so müssen zwangsläufig alle Worte w in dem Dokument eine Wahrscheinlichkeit $p(w|T) > 0$ für dieses Topic haben (#2 schlecht)
- Haben die Topics nur sehr wenige Worte (#2 gut), so muss ein Dokument aus vielen Topics bestehen, um alle Worte abzudecken (#1 schlecht)

Werden beide Ziele ausbalanciert, ergeben sich interessanterweise automatisch Gruppen häufig gemeinsam auftretender Terme.

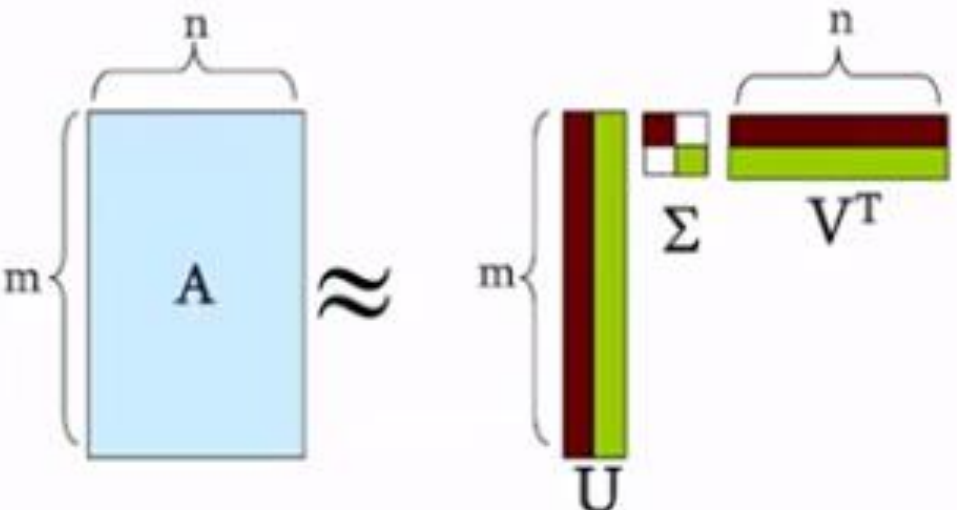
Singulärwert-Zerlegung

Wenn Rang r der Matrix klein ist, dann wird A zerlegt in

- schmale, hohe Matrix U
- flache, breite Matrix V

Dimensionalitätsreduktion auf $k < r$ Dimensionen:

- Reduziere Σ auf die größten k Singulärwerte

$$A \approx U \Sigma V^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i^T$$


The diagram illustrates the SVD decomposition $A \approx U \Sigma V^T$. Matrix A (blue, $m \times n$) is approximated by the product of matrix U (green, $m \times k$), matrix Σ (red, $k \times k$), and matrix V^T (green, $k \times n$). The dimensions m , n , and k are indicated by brackets. The matrix U is shown as a tall, narrow rectangle, Σ as a small square, and V^T as a wide, short rectangle.

SVD - Idee animiert

