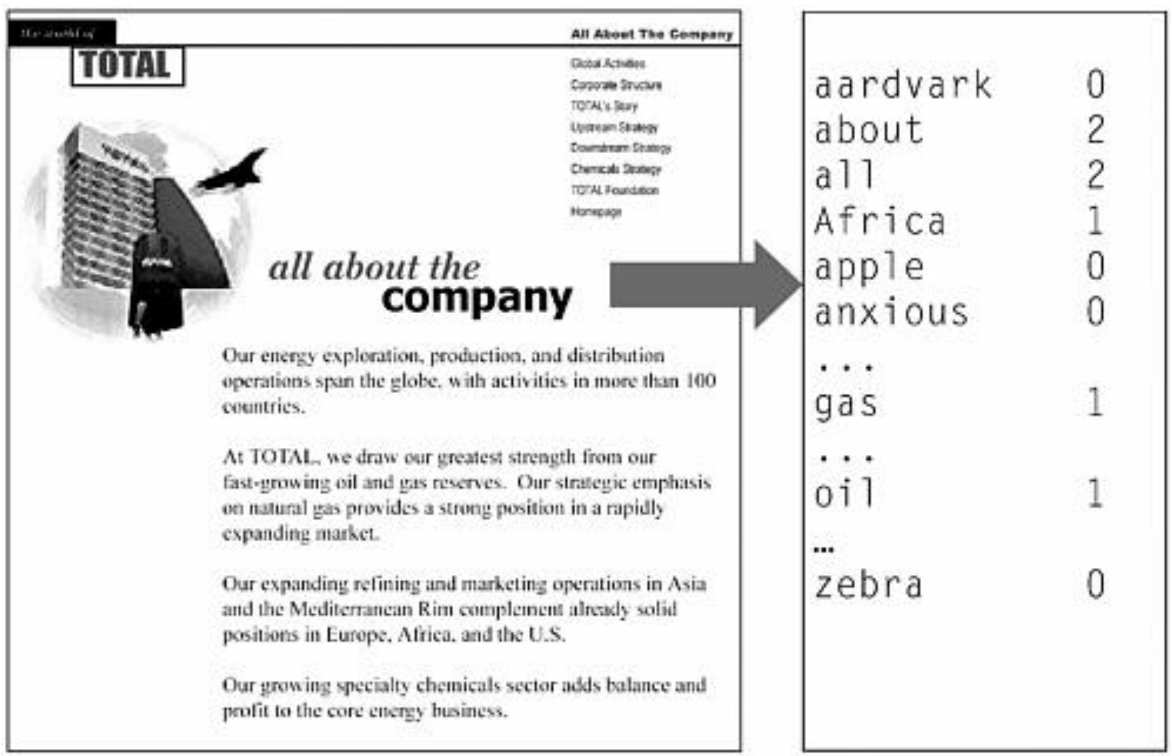
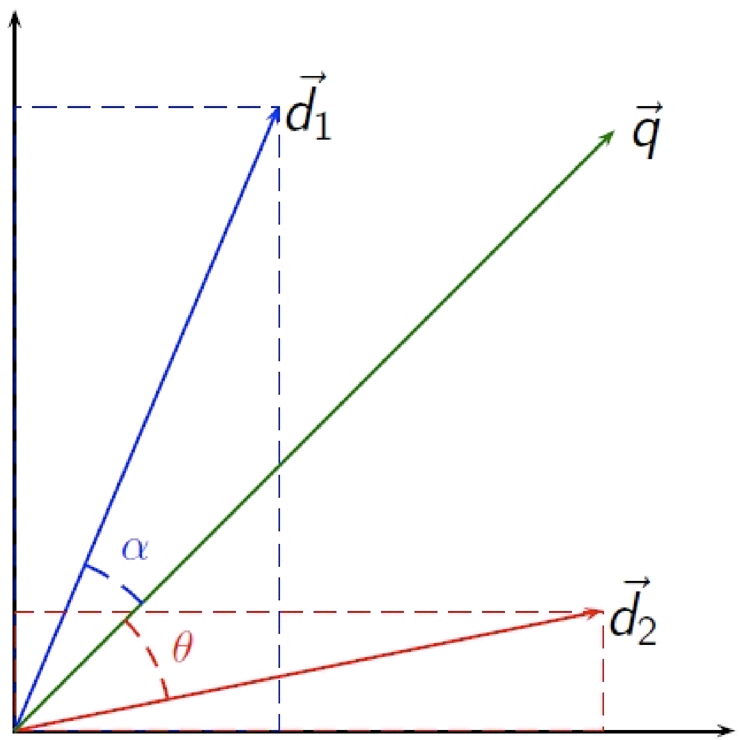
Salton's Vector Space Model (Prior to 1988)

- Represent each document by a high-dimensional vector in the space of words



Query

- Compute the similarity between *queries(q)* and *documents(d)*



$$\cos(\boldsymbol{q}, \boldsymbol{d}) = \frac{\boldsymbol{q}^T \boldsymbol{d}}{\|\boldsymbol{q}\| \|\boldsymbol{d}\|}$$

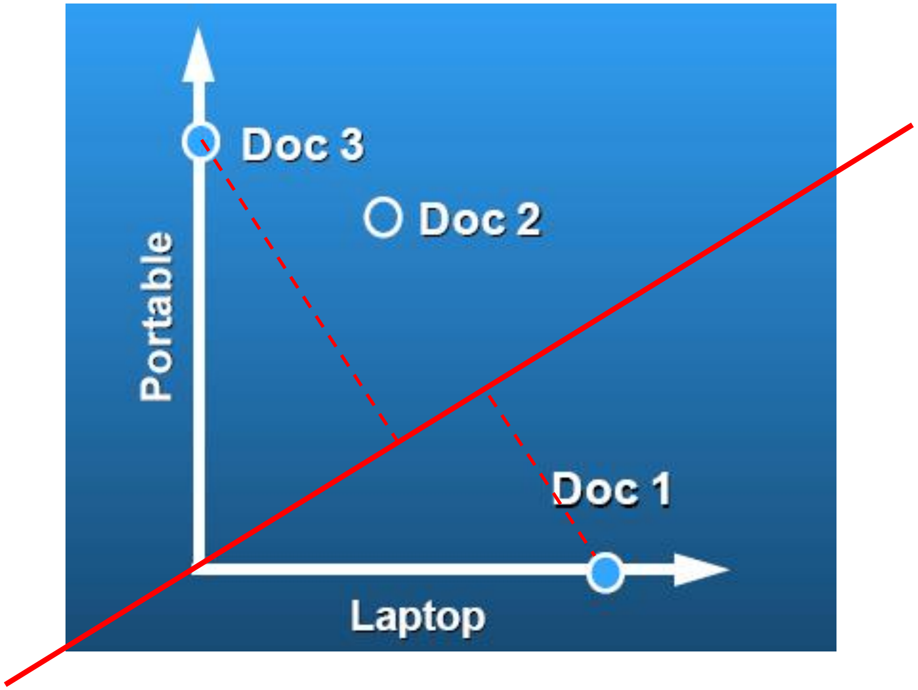
Simple, intuitive

Fast to compute, because both they are sparse

Retrieval Methods

- Rank documents according to similarity with query
- Term weighting schemes, for example, TF-IDF

Problem of Vector Space Model



Possible Solution: Principle Component Analysis

Problem of PCA

- Main steps for computing PCs:
  - Form the covariance matrix S.
  - Computes the first d eigenvectors $\{\boldsymbol{a}_i\}_{i=1}^{d}$.
  - The transformation A is given by

$$A = [\boldsymbol{a}_1, \cdots \boldsymbol{a}_d]$$

$$S = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \in R^{m \times m}$$

We have the computational problem if m is very large.

Singular Value Decomposition (SVD)

- For an arbitrary matrix $X \in \mathcal{R}^{m\times n}$ there exists a factorization as follows:

$$X = U\Sigma V$$

- where

$$U \in \mathcal{R}^{m\times m}, V \in \mathcal{R}^{n\times n}, UU^T = U^T U = I, VV^T = V^T V = I$$

$$\text{diagonal matrix } \Sigma \in \mathcal{R}^{m\times n}$$

- If $rank(X) = d$

$$U \in \mathcal{R}^{m\times d}, V \in \mathcal{R}^{d\times n}, U^T U = I, VV^T = I$$

$$\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \cdots \sigma_d) \quad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d > 0$$

SVD: Low-rank Approximation

- SVD can be used to compute optimal **low-rank approximations**.
- Approximation problem:

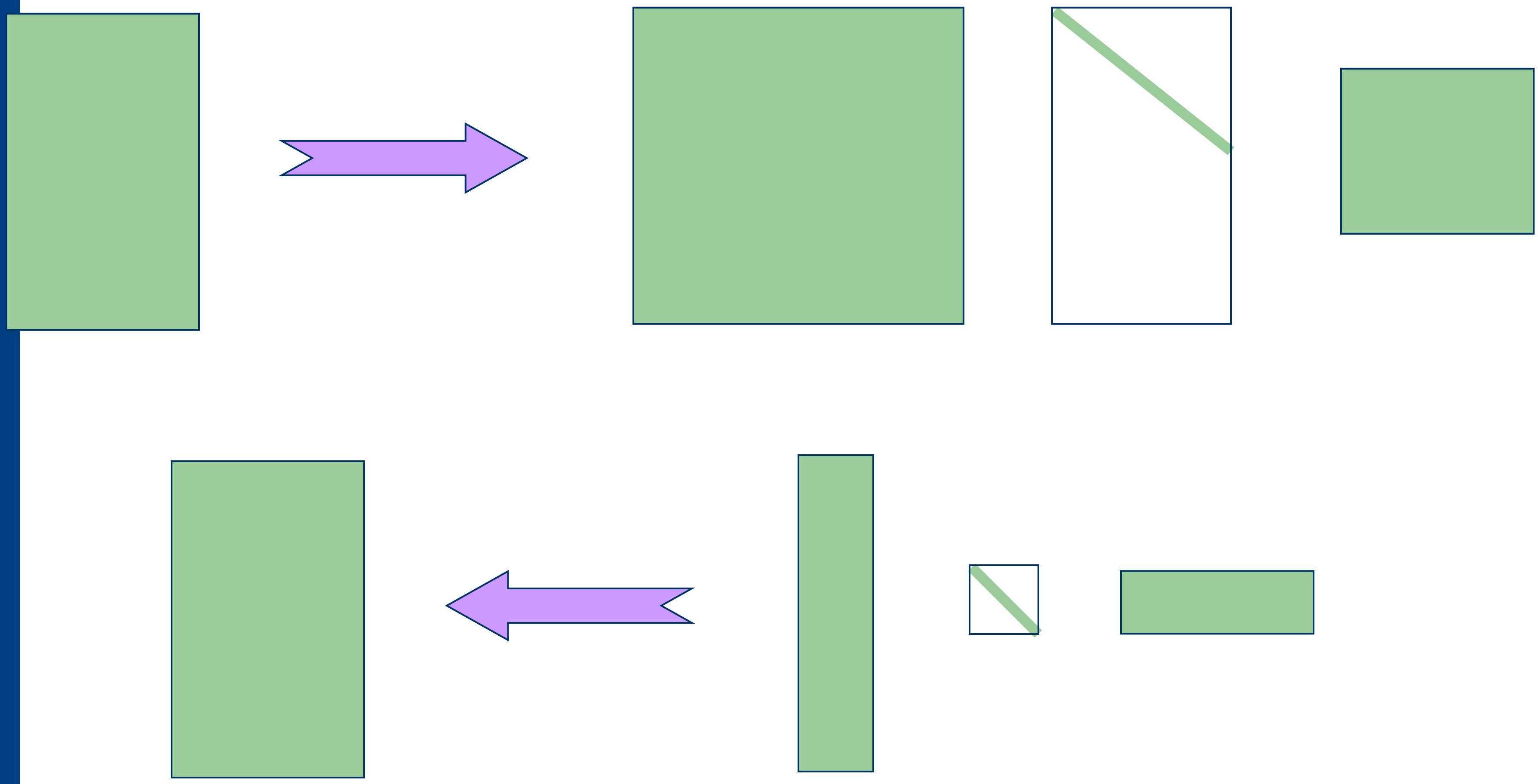
$$X^* = \underset{rank(\tilde{X})=k}{\operatorname{argmin}} \left\| X - \tilde{X} \right\|_F^2$$

- Solution via SVD

$$X^* = U\mathrm{diag}(\sigma_1, \cdots, \sigma_k, \underbrace{0, \cdots, 0}_{\text{set small singular values to zero}})V$$

C. Eckart, G. Young, The approximation of a matrix by another of lower rank. Psychometrika, 1, 211-218, 1936.

Low rank approximation by SVD



SVD Solution of PCA

$$\overline{\boldsymbol{x}} = \mathbf{0}$$

$$S = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \qquad \frac{1}{n}XX^T \qquad XX^T \in R^{m\times m} \qquad n \ll m$$

$$XX^T = U\Sigma V^T V\Sigma^T U^T = U(\Sigma\Sigma^T)U^T$$

$$X^T X = V\Sigma^T U^T U\Sigma V^T = V(\Sigma^T\Sigma)V^T$$

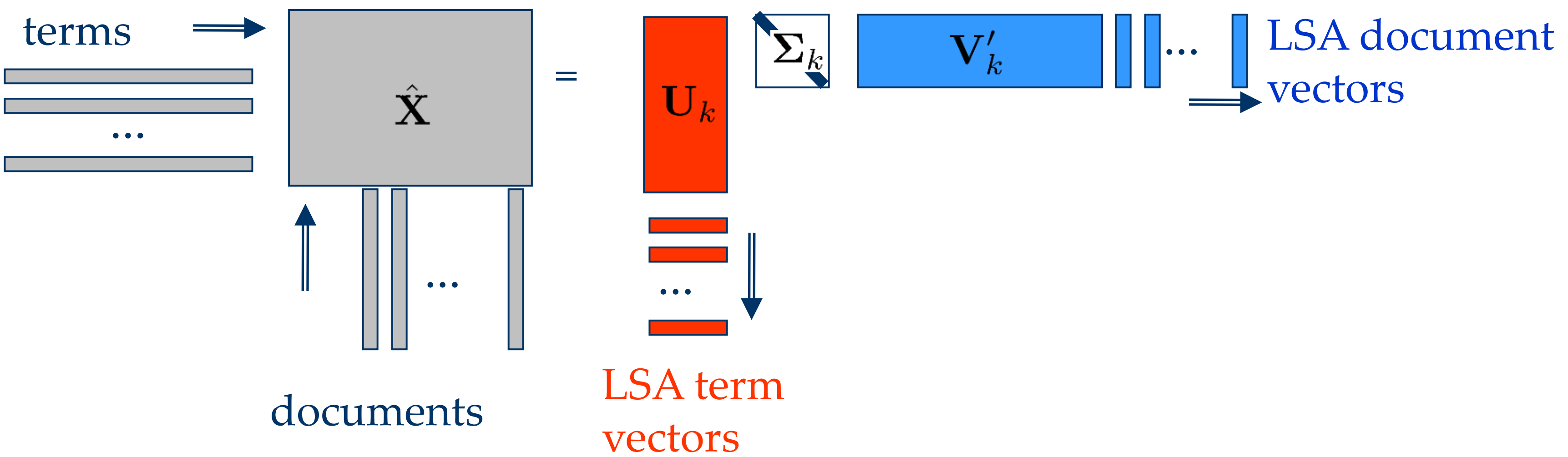
$$X = U\Sigma V^T$$

$$XV = U\Sigma V^T V = U\Sigma$$

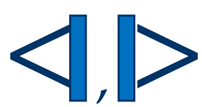
$$XV\Sigma^{-1} = U$$

Latent Semantic Analysis (Indexing)

- The Latent Semantic Analysis via SVD can be summarized as follows:



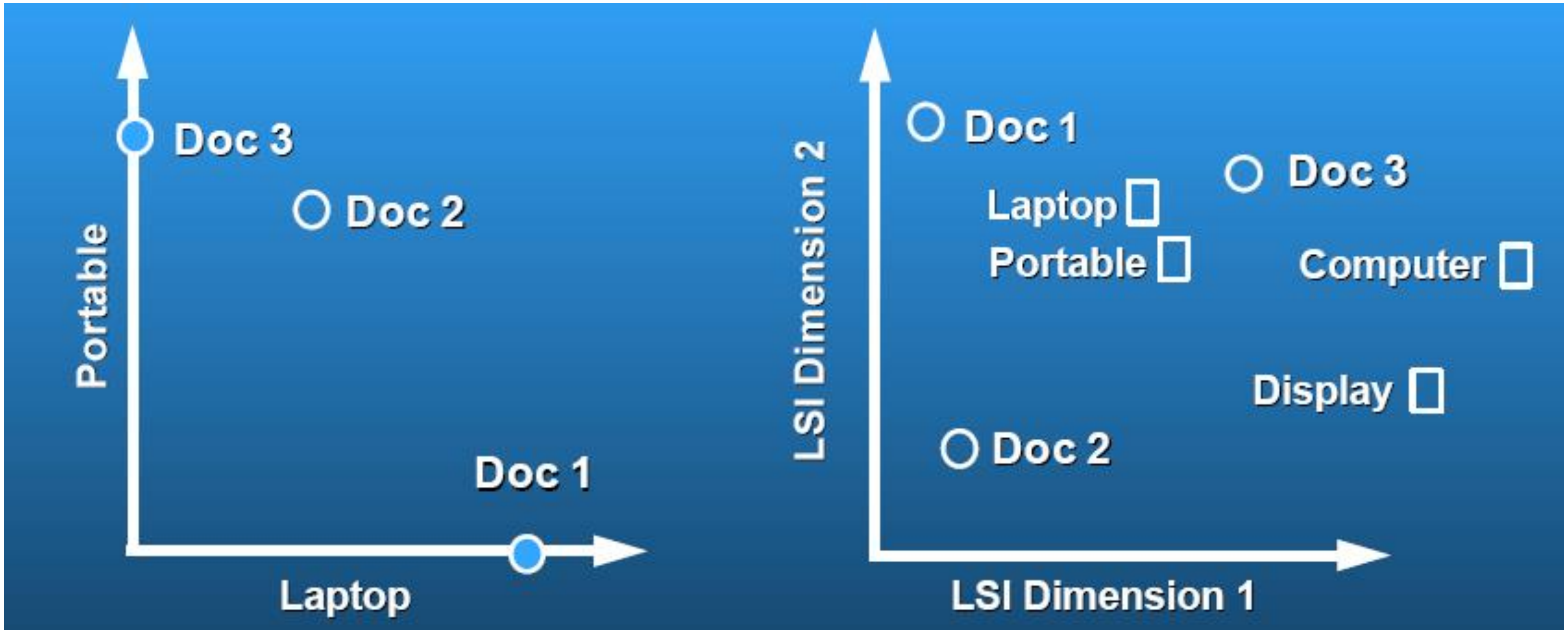
- Document **similarity**



- $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle = \langle \Sigma_k \boldsymbol{v}_i, \Sigma_k \boldsymbol{v}_j \rangle$

Latent Semantic Analysis

- **Latent semantic space**: illustrating example



Matrix Factorization

Deng Cai (蔡登)

College of Computer Science
Zhejiang University

dengcai@gmail.com

What Is Matrix Factorization?

$$X \in \mathcal{R}^{m\times n}$$

$$UV = X \qquad U \in \mathcal{R}^{m\times k}, V \in \mathcal{R}^{k\times n}$$

- Is this factorization unique?

$$\Sigma \in \mathcal{R}^{k\times k} \qquad U\Sigma\Sigma^{-1}V = X$$

$$U\Sigma V = X$$

  - Every column of U and every row of V are normalized

- Does this factorization always exist?

$$UV = \tilde{X} \approx X \qquad \|X - UV\|_F^2$$

Why Matrix Factorization?

$$X = UV$$

$$\begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ x_{13} & x_{23} & \cdots & x_{n3} \\ \vdots & \vdots & & \vdots \\ x_{1m} & x_{2m} & \cdots & x_{nm} \end{bmatrix} = \begin{bmatrix} u_{11} & \cdots & u_{k1} \\ u_{12} & \cdots & u_{k2} \\ u_{13} & \cdots & u_{k3} \\ \vdots & & \vdots \\ u_{1m} & \cdots & u_{km} \end{bmatrix} \times \begin{bmatrix} v_{11} & v_{21} & \cdots & v_{n1} \\ \vdots & \vdots & & \vdots \\ v_{1k} & v_{2k} & \cdots & v_{nk} \end{bmatrix}$$

$$\begin{bmatrix} \\ \boldsymbol{x}_i \\ \\ \end{bmatrix} = v_{i1} \begin{bmatrix} \\ \boldsymbol{u}_1 \\ \\ \end{bmatrix} + v_{i2} \begin{bmatrix} \\ \boldsymbol{u}_2 \\ \\ \end{bmatrix} + \cdots + v_{ik} \begin{bmatrix} \\ \boldsymbol{u}_k \\ \\ \end{bmatrix}$$

- Each column vector of X can be represented as a linear combination of column vectors of U
- Each column vector of V can be regarded as a low dimensional representation of corresponding column vector of X

Relation to Dimensionality Reduction

$$X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_n] = UV = U[\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots \boldsymbol{v}_n]$$

$$\boldsymbol{x}_i = U\boldsymbol{v}_i \qquad \boldsymbol{x}_i \in \mathcal{R}^m, \boldsymbol{v}_i \in \mathcal{R}^k$$

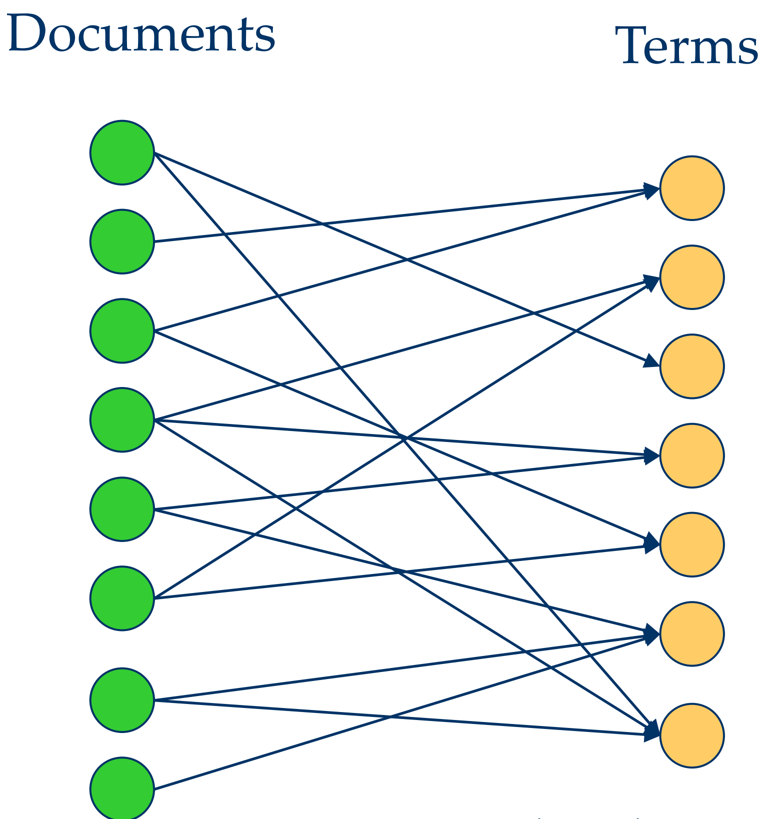
- If there is a matrix $A \in \mathcal{R}^{k\times m}$ which satisfies:

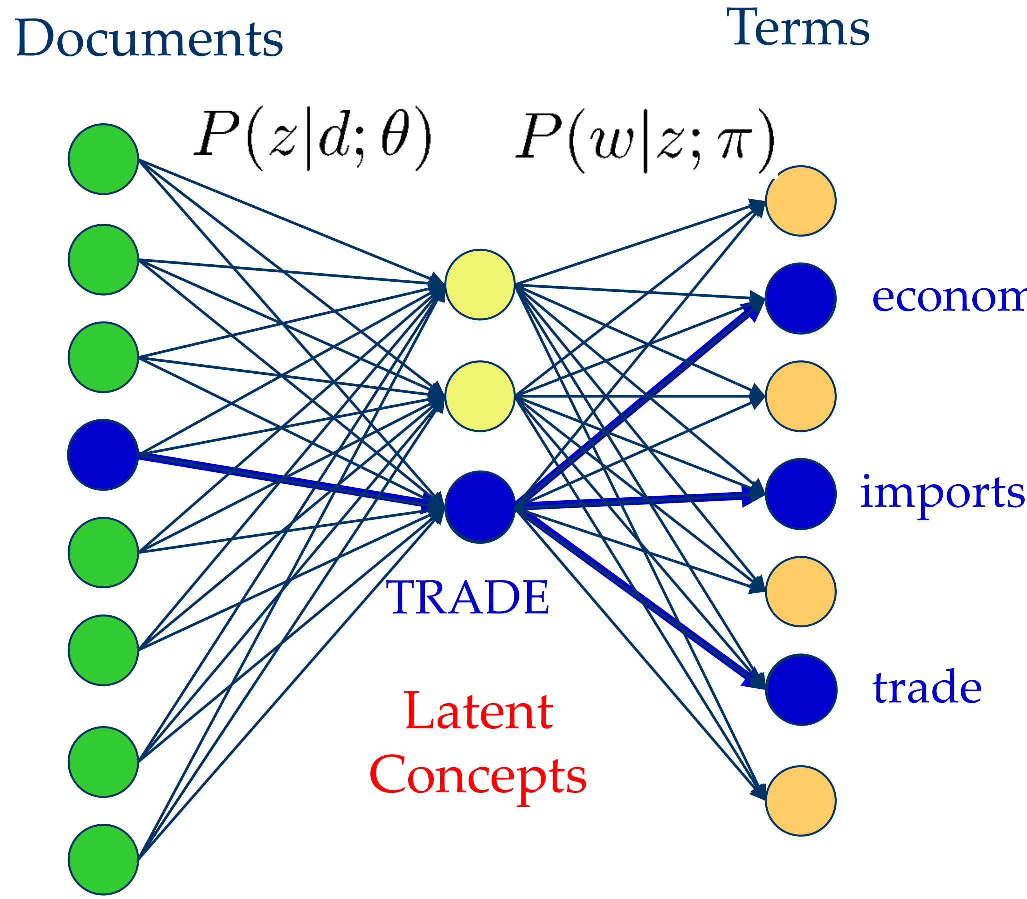
$$AU = I$$

$$A\boldsymbol{x}_i = \boldsymbol{v}_i$$

- In DR, we learn the transformation matrix
- In MF, we learn the basis matrix

Relation to Topic Modeling





$$P(w|d) = \frac{n(d,w)}{\sum_{w'} n(d,w')}$$

$$X = \begin{bmatrix} P(w_1|d_1) & \cdots & P(w_1|d_n) \\ \vdots & \ddots & \vdots \\ P(w_m|d_1) & \cdots & P(w_m|d_n) \end{bmatrix}$$

$$\hat{P}(w|d) = \sum_z P(w|z)P(z|d)$$

Relation to Topic Modeling

$$P(w|d) = \frac{n(d,w)}{\sum_{w'} n(d,w')}$$

$$\hat{\boldsymbol{P}}(\boldsymbol{w}|\boldsymbol{d}) = \sum_{\boldsymbol{z}} \boldsymbol{P}(\boldsymbol{w}|\boldsymbol{z})\boldsymbol{P}(\boldsymbol{z}|\boldsymbol{d})$$

$$X = \begin{bmatrix} P(w_1|d_1) & \cdots & P(w_1|d_n) \\ \vdots & \ddots & \vdots \\ P(w_m|d_1) & \cdots & P(w_m|d_n) \end{bmatrix}$$

$$U = \begin{bmatrix} \hat{P}(w_1|z_1) & \cdots & \hat{P}(w_1|z_k) \\ \vdots & \ddots & \vdots \\ \hat{P}(w_m|z_1) & \cdots & \hat{P}(w_m|z_k) \end{bmatrix}$$

$$\hat{X} = \begin{bmatrix} \hat{P}(w_1|d_1) & \cdots & \hat{P}(w_1|d_n) \\ \vdots & \ddots & \vdots \\ \hat{P}(w_m|d_1) & \cdots & \hat{P}(w_m|d_n) \end{bmatrix}$$

$$V = \begin{bmatrix} \hat{P}(z_1|d_1) & \cdots & \hat{P}(z_k|d_1) \\ \vdots & \ddots & \vdots \\ \hat{P}(z_1|d_n) & \cdots & \hat{P}(z_k|d_n) \end{bmatrix}$$

$$X \approx \hat{X} = UV^T$$

Nonnegative Matrix Factorization

$$X \in \mathcal{R}^{m \times n}$$

$$U \in \mathcal{R}^{m \times k}, \qquad V \in \mathcal{R}^{k \times n}$$

$$UV = \tilde{X} \approx X$$

$$u_{ij} \geq 0, v_{ij} \geq 0$$

- Low rank assumption (*k* hidden factors)
- Nonnegative assumption

Non-negative Matrix Factorization

$$X \cong \hat{X} = UV^{T}, u_{ij} \geq 0, v_{ij} \geq 0$$

- Two cost functions
  - Euclidean distance

$$||A - B||^2 = \Sigma_{ij}(A_{ij} - B_{ij})^2$$

  - Divergence

$$D(A||B) = \Sigma_{ij}(A_{ij} \log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij})$$

Optimization Problems

- *Minimize* $||X - UV^T||^2$ *with respect to* U *and* V, *subject to the constraints* $U, V \geq 0$.

- *Minimize* $D(X||UV^T)$ *with respect to* U *and* V, *subject to the constraints* $U, V \geq 0$.

NMF Optimization (Euclidean Distance)

$$\min\left\| |X - UV^T| \right\|^2, s.t. u_{ij} \geq 0, v_{ij} \geq 0$$

$$J = \left\| |X - UV^T| \right\|^2 = \mathrm{tr}\big((X - UV^T)^T (X - UV^T)\big)$$

$$= \mathrm{tr}(X^T X - X^T U V^T - V U^T X + V U^T U V^T)$$

Γ, same size as U

Φ, same size as V

$$\mathcal{L} = \mathrm{tr}(X^T X) - 2\mathrm{tr}(X^T U V^T) + \mathrm{tr}(V U^T U V^T) + \mathrm{tr}(\Gamma U^T) + \mathrm{tr}(\Phi V^T)$$

$$\frac{\partial \mathcal{L}}{\partial U} = -2XV + 2UV^T V + \Gamma \qquad (UV^T V)_{ik} u_{ik} - (XV)_{ik} u_{ik} = 0$$

$$\boxed{u_{ik} \leftarrow \frac{(XV)_{ik}}{(UV^T V)_{ik}} u_{ik}}$$

$$\frac{\partial \mathcal{L}}{\partial V} = -2X^T U + 2VU^T U + \Phi \qquad (VU^T U)_{jk} v_{jk} - (X^T U)_{jk} v_{jk} = 0$$

$$\boxed{v_{jk} \leftarrow \frac{(X^T U)_{jk}}{(VU^T U)_{jk}} v_{jk}}$$

Multiplicative Update Rules

- *The Euclidean distance* $\|X - UV^T\|^2$ *is nonincreasing under the update rules*

$$u_{ik} \leftarrow \frac{(XV)_{ik}}{(UV^TV)_{ik}} u_{ik} \qquad v_{jk} \leftarrow \frac{(X^TU)_{jk}}{(VU^TU)_{jk}} v_{jk}$$

The Euclidean distance is invariant under these updates if and only if U *and* V *are at a stationary point of the distance.*

NMF vs PLSA

$X \cong \hat{X} = UV^T, u_{ij} \geq 0, v_{ij} \geq 0$

$$D(A||B) = \sum_{ij}\left(A_{ij}\log\frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij}\right) \quad = \sum_{ij}(A_{ij}\log A_{ij} - A_{ij} - A_{ij}\log B_{ij} + B_{ij})$$

$$\max \sum_{ij}(A_{ij}\log B_{ij} - B_{ij})$$

$$X = [n(d_i, w_j)] \times diag\left(\frac{1}{l(d_i)}\right) \qquad U = [p(w_j|z_k)] \qquad V^T = [p(z_k|d_i)]$$

$$\max \sum_i \frac{1}{l(d_i)} \sum_j n(d_i, w_j) \log \sum_k p(w_j|z_k)p(z_k|d_i) - n$$

$$l(\theta, \pi; \mathbf{N}) = \sum_{d,w} n(d,w)\log(\sum_z P(w|z;\theta)P(z|d;\pi))$$

Sparse Coding

$$X \approx \hat{X} = UV^T$$

$$\begin{bmatrix} \\ \boldsymbol{x}_i \\ \\ \end{bmatrix} = v_{i1} \begin{bmatrix} \\ \boldsymbol{u}_1 \\ \\ \end{bmatrix} + v_{i2} \begin{bmatrix} \\ \boldsymbol{u}_2 \\ \\ \end{bmatrix} + \cdots + v_{ik} \begin{bmatrix} \\ \boldsymbol{u}_k \\ \\ \end{bmatrix}$$

$$\begin{aligned} &\text{minimize}_{U,V} \quad \left|\left|X - UV^T\right|\right|_F^2 + \lambda f(V) \\ &\text{subject to} \quad \Sigma_i u_{i,k}^2 \leq c, \forall k = 1, \dots, K. \end{aligned}$$

- Represent input vectors approximately as a weighted linear combination of a small number of "basis vectors."

Matrix Factorization: Summary

$$X \in \mathcal{R}^{m \times n}$$

$$U \in \mathcal{R}^{m \times k}, \qquad V \in \mathcal{R}^{k \times n}$$

$$UV = \tilde{X} \approx X$$

- Low rank assumption (*k* hidden factors)
  - SVD
- Nonnegative assumption
  - NMF
- Sparseness assumption
  - Sparse Coding