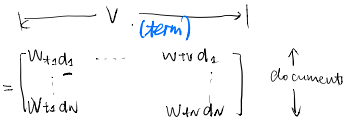
$w_{i(w),d} = f_{i(w)} \times IDF(i(w))$

Vector Notation

$vec(d) = \{ \cdots \; w_{i(w),d} \; \cdots \}$

- word document matrix A



- $W = USV^T$

PCA and LSA

- Suppose vec(D) zero mean
- Subtract mean from each doc vec. $vec(d_n)$ → so that the mean is zero
- Cov. matrix Σ of vec(D)

$$\Sigma = \frac{1}{N-1} W^T W = \frac{1}{N-1} (USV^T)^T (USV^T)$$

$$= \frac{1}{N-1} V S^T U^T U S V^T$$

$$= \frac{1}{N-1} V S^2 V^T$$

$$\Sigma = V \frac{S^2}{N-1} V^T = V \bar{S} V^T$$

$$\bar{S}_{ij} = \frac{S_{ij}^2}{N-1}$$

→ but don't get U if using PCA

in the form of eigenvector Decomposition (the decomp. is unique)

description of docs in terms of topics

Summary:

Document Vectors vec(D) with ☆ 0 mean vector

LSA and PCA give same orthonormal basis (V)

⇒ LSA, PCA : Same set of topics

D denotes eigenvalue matrix

$d_{ij} = \bar{S}_{ij} = \frac{S_{ij}^2}{N-1}$

Effect of subtracting Mean document vector from vec(D)

Normalize the vector the get the PCA = LSA