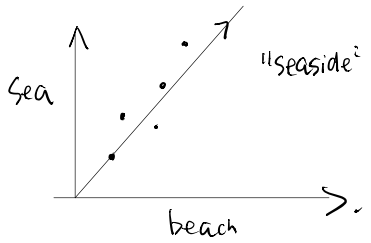
Latent Semantic Analysis. (LSA)

- Discover Latent topics
- Dimension Reduction.
- Represent words in terms of topics

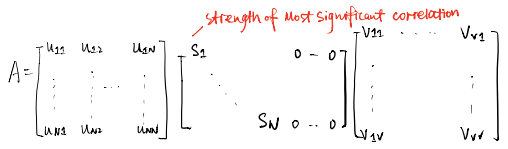
Latent Semantic Analysis



beach is often with sea

- 'word document' matrix A $=$ $\begin{bmatrix} w_{t_1}d_1 & \cdots & w_{tv}d_1 \\ \vdots & & \vdots \\ V_{t_1}d_N & & w_{tv}d_N \end{bmatrix}$ ← V → ; ↑ documents ↓
- Singular Value Decomposition (SVD) to A.
  "similar to eigenvector Decomposition"

$$A = USV^T.$$



U_{NxN}, V_{VV} are orthogonal matrices

$U \cdot U^T = I$.

$V \cdot V^T = I$

$S_1 \cdots S_N$: singular values

$S_1 \geq S_2 \cdots \geq S_N$.

Consider V

Columns: V dimensional. unit vectors.

orthogonal to each other.

$V = \{v_1, \cdots v_V\}$

$\Rightarrow$ orthonormal basis.

$\begin{bmatrix} V_{11} & \cdots & V_{1V} \\ V_{V1} & & V_{VV} \end{bmatrix}$

V_i : a document vector. $\Rightarrow$ Semantic class.
[topic]

Importance of the class : indicated by S_n.

$V_i = \begin{bmatrix} V_{1i} \\ \vdots \\ V_{ji} \\ \vdots \\ V_{Vi} \end{bmatrix}$ TF-IDF weight of the j^{th} term.

if V_{ji} is Large

$\Rightarrow$ this term in vocab. is significant

Consider. U

- $A v_n = U S V^T \cdot v_n = S_{nn} u_n$ Combination of documents

the nth topic

slides 15 20

topic-based representation.

vec(d) · v_n: magnitude of the component in direction of v_n.

$$top(d) = \begin{bmatrix} vec(d) \cdot v_1 \\ \vdots \\ vec(d) \cdot v_n \\ \hline vec(d) \cdot v_V \end{bmatrix} \xrightarrow{\text{truncate.}} \approx top(d) = V^T(n) \cdot vec(d)$$

reduced Dimension