Latent Semantic Indexing (LIS)

Motivating Example

- In search application, we use a word-document vector for efficient indexing and retrieval.

	w_1	w_2	w_3	-----	w_m →
d_1	2.9				
d_2	3.4				
⋮					
d_n ↓					

A word-document vector is a n x m vector Every cell indicates "importance" of the word in the document.

→ TF-IDF
(Term Frequency - Invers Document Frequency)

D_1 = Dog says Bow

D_2 = Cat says Mew

D_3 = Bow Bow Mew Mew

	Dog	says	Bow	Cat	Mew
D_1	1	1	1	0	0
D_2	0	1	0	1	1
D_3	0	0	2	0	2

Query : "Cat Mew"

Query Vector: [0 0 0 1 1]

Use similarity measures: Cosine!

Cosin (q_1, d_1) Cosin (q_1, d_2) Cosin (q_1, d_3)

Rank 3 Rank 1 Rank 2

We want to reduce dimensions of word-document vector.

- Number of documents & words are very high.
- Without reducing dimensions, most of the feature (dimension, word count) for a document will be zero. (Sparse)

① Reduce dimension to avoid sparcity

② We want to relate word-word & doc-doc.

Bank → Financial Inst.

Bank → River Bank

	w_1 = Bank	w_2 = money	w_3 = River
d_1	1	1	0
d_2	1	0	1

P_1 = Bank money

P_2 = Bank River

	good	nice	movie
D_1	1	0	1
D_2	0	1	1

D_1 = good movie

D_2 = Nice movie

q = "Nice movie"

Latent Semantic Indexing

An m x n dimension matrix can be divided it into 3 matrices

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

Orthogonal

Diagonal

Singular Value Decomposition (SVD)

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

C = word document vector $(m \times n)$

m = no. of docs

n = no. of words

U = $m \times r$ matrix (doc (doc) matrix)

V^T = $r \times n$ matrix (word (word) matrix)

Σ = Diagonal matrix ($\sqrt{\text{Eigen values}}$) (Singular Values)

(Ref: Diagonalization)

$$S A S^{-1} = D$$

↑

Eigen Values

SVD Example

$$C = \begin{bmatrix} 5 & 5 \\ -1 & 7 \end{bmatrix}$$

Find U, Σ, V^T

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

$$C^T C = (V \Sigma^T U^T) \cdot (U \Sigma V^T)$$

$$\boxed{C^T C = V \Sigma^T \Sigma V^T} \quad \text{—— (1)}$$

$$CV = U\Sigma \quad \text{—— (2)}$$

① Find C^TC

$$C^TC = \begin{bmatrix} 5 & -1 \\ 5 & 7 \end{bmatrix} \begin{bmatrix} 5 & 5 \\ -1 & 7 \end{bmatrix}$$

$$= \begin{bmatrix} 26 & 18 \\ 18 & 74 \end{bmatrix}$$

② Find Eigen values & vectors of C^TC,

$$\lambda_1 = 20, \quad \lambda_2 = 80$$

$$V_1 = \begin{bmatrix} -3/\sqrt{10} \\ 1/\sqrt{10} \end{bmatrix} \qquad V_2 = \begin{bmatrix} 1/\sqrt{10} \\ 3/\sqrt{10} \end{bmatrix}$$

③ Define V & Σ

Σ = square root of Eigen value, diagonal matrix

$$\Sigma = \begin{bmatrix} 2\sqrt{5} & 0 \\ 0 & 4\sqrt{5} \end{bmatrix}$$

$$V = \begin{bmatrix} -3/\sqrt{10} & 1/\sqrt{10} \\ 1/\sqrt{10} & 3/\sqrt{10} \end{bmatrix}$$

How to find U ?

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

$$= U\Sigma\Sigma^T U^T$$

$$CC^T = \begin{bmatrix} 5 & 5 \\ -1 & 7 \end{bmatrix} \begin{bmatrix} 5 & -1 \\ 5 & 7 \end{bmatrix}$$

$$= \begin{bmatrix} 50 & 30 \\ 30 & 50 \end{bmatrix}$$

$$\lambda_1 = 20, \qquad \lambda_2 = 80$$

$$u_1 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} \qquad u_2 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

$$U = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

$$\begin{bmatrix} 5 & 5 \\ -1 & 7 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \cdot \begin{bmatrix} 2\sqrt{5} & 0 \\ 0 & 4\sqrt{5} \end{bmatrix} \cdot \begin{bmatrix} -\sqrt{3}/10 & 1/\sqrt{10} \\ 1/\sqrt{10} & \sqrt{3}/10 \end{bmatrix}$$

Word document matrix A

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

$r \times n$

word-word matrix

doc-doc matrix

$m \times r$

singular value ($\sqrt{\text{Eigen values}}$)

"Ignore low singular values"

First r singular values non zero