

한국어임베딩

4-4~4-6

전찬웅

# 목차

1. 잠재 의미 분석

2. GloVe

3. Swivel

# 잠재 의미 분석

차원 축소 방법의 일종인 특이값 분해를 수행해  
데이터의 차원 수를 줄여 계산의 효율성을 키우고 잠재  
의미를 이끌어내기 위한 방법론

# PPMI 행렬

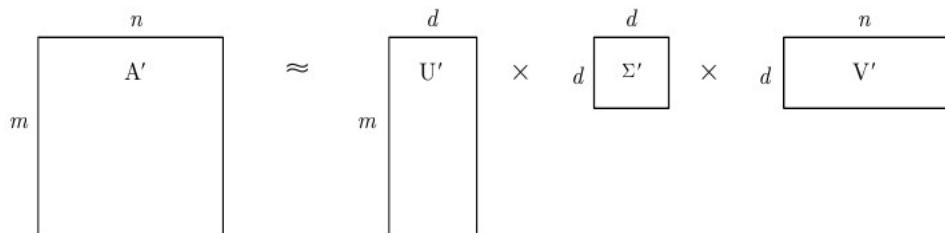
PMI(A, B)  $\longrightarrow$  PPMI(A, B)  $\longrightarrow$  SPMI(A, B)

$$\text{PMI}(A, B) = \log \frac{P(A, B)}{P(A) \times P(B)}$$

$$\text{PPMI}(A, B) = \max(\text{PMI}(A, B), 0)$$

$$\text{SPMI}(A, B) = \text{PMI}(A, B) - \log k$$

# Truncated SVD

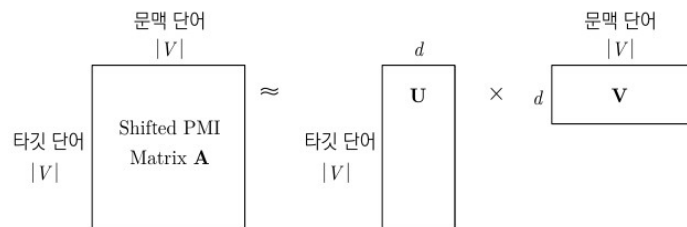


The diagram illustrates the Truncated SVD decomposition of a matrix  $A'$ . On the left is a square matrix  $A'$  with dimensions  $m$  (height) and  $n$  (width). To its right is an approximation symbol  $\approx$ . Further right is the product of three matrices: a tall rectangular matrix  $U'$  with dimensions  $m$  and  $d$ , followed by a small square matrix  $\Sigma'$  with dimensions  $d$  and  $d$ , and finally a wide rectangular matrix  $V'$  with dimensions  $d$  and  $n$ . Multiplication symbols  $\times$  are placed between  $U'$  and  $\Sigma'$ , and between  $\Sigma'$  and  $V'$ .

그림 4-13 truncated SVD

- $U'$ 는 단어 임베딩,  $V'$ 는 문서 임베딩에 대응
- $N, m$ 개의 단어, 문서벡터들이  $d$ 차원만으로 표현가능
- 효과: 단어 문맥간의 내재적인 의미 보존하며 입력데이터의 노이즈, 희소성을 줄일 수 있다

# SGNS



$$\mathbf{A}_{ij}^{\text{SGNS}} = \mathbf{U}_i \cdot \mathbf{V}_j = \text{PMI}(i, j) - \log k$$

- $k$ 는 skip-gram 모델의 네거티브 샘플 수 의미
- 단어의 의미가 내적 값으로 나타나고, 관련성이 높을수록 내적 값이 크게 나타난다.

# GloVe

수식 4-20 GloVe의 목적함수

$$\mathcal{J} = \sum_{i,j=1}^{|V|} f(\mathbf{A}_{ij})(\mathbf{U}_i \cdot \mathbf{V}_j + \mathbf{b}_i + \mathbf{b}_j - \log \mathbf{A}_{ij})^2$$

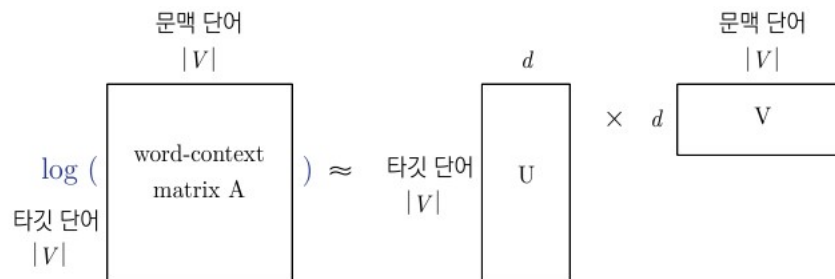
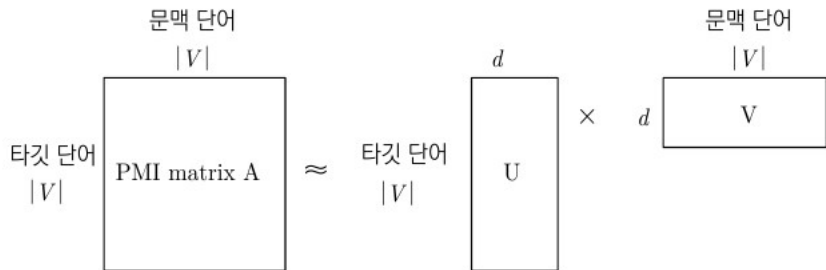


그림 4-15 그림으로 이해하는 GloVe

# Swivel

Swivel의 목적함수 1 (말뭉치에 동시 등장한 케이스가 한 건이라도 있는 경우)

$$\mathcal{J} = \frac{1}{2} f(x_{ij}) (\mathbf{U}_i \cdot \mathbf{V}_j - \text{PMI}(i, j))^2$$



Swivel의 목적함수 2 (말뭉치에 동시 등장한 케이스가 한 건도 없는 경우)

$$\mathcal{J} = \log [1 + \exp(\mathbf{U}_i \cdot \mathbf{V}_j - \log |D| + \log \mathbf{A}_{i*} + \log \mathbf{A}_{*j})]$$