Chapter. 04

토픽 모델링

잠재 의미 분석 (LSA)

FASTCAMPUS
ONLINE
금융공학/퀀트

강사. 장순용

| 키포인트

- 토픽 모델링.
- 특이값 분해 (SVD).
- 특이값 분해와 차원 축소.

I 잠재 의미 분석 (Latent Semantic Analysis, LSA)

- 여러 문서를 분석하여 몇 개의 공통된 토픽 (topic)을 추출해 냄.

  ⇔ TF IDF 행렬을 가지고 SVD 분해 하여 주요 성분을 추출해 내는 것.

- TF IDF 행렬 $\boldsymbol{M}$의 크기가 다음과 같다고 전제한다.

$$Size(\boldsymbol{M}) = m \times n$$

$$m = \text{문서의 개수}$$

$$n = \text{단어의 개수}$$

- 하나의 토픽 벡터는 길이가 m이며 행렬 $\boldsymbol{M}$의 주성분이다.

- 특이값 크기 순서로 2~5개의 토픽을 추출하여 해석한다.

I 잠재 의미 분석 (Latent Semantic Analysis, LSA)

- 다음과 같은 주요 활용 분야가 있다.

⇒ 문서의 군집화.

⇒ 문서 사이의 관계 분석.

⇒ 서치엔진의 페이지 인덱싱.

ㅣ특이값 분해 (SVD)

- 행렬 $\boldsymbol{M}$을 $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^t$와 같은 형태로 분해한다.

$\Rightarrow$ 행렬의 크기는 다음과 같다:

$$Size(\boldsymbol{M}) = m \times n$$

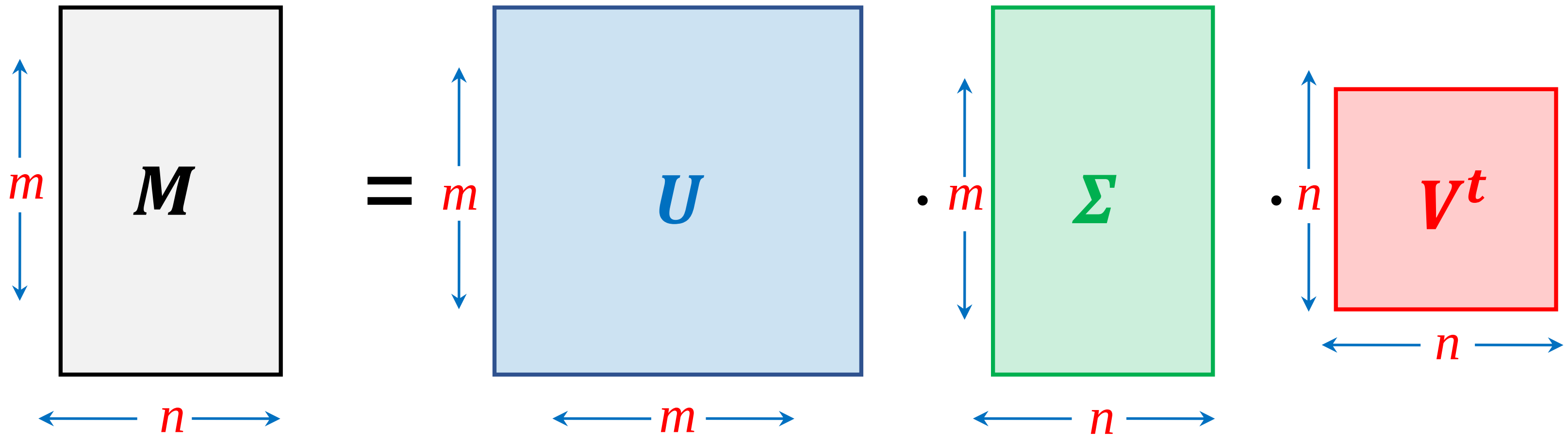
$$Size(\boldsymbol{U}) = m \times m$$

$$Size(\boldsymbol{\Sigma}) = m \times n$$

$$Size(\boldsymbol{V}) = n \times n$$

I 특이값 분해 (SVD)

- 행렬 $\boldsymbol{M}$을 $\boldsymbol{M} = \boldsymbol{U\Sigma V^t}$와 같은 형태로 분해한다.



I 특이값 분해 (SVD)

- 행렬 $\boldsymbol{M}$을 $\boldsymbol{M} = \boldsymbol{U\Sigma V^t}$와 같은 형태로 분해한다.

$\Rightarrow$ $\boldsymbol{\Sigma}$의 대각 원소가 바로 "특이값" 이다.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & 0 \\ 0 & \sigma_2 & & 0 & 0 \\ \vdots & & \ddots & \vdots & \\ 0 & 0 & \cdots & \sigma_m & 0 \end{bmatrix}$$

$\Rightarrow$ 이 특이값들은 양수이며 대→소의 순서로 정렬되어 있는 것이 원칙이다.

I 특이값 분해 (SVD)

- 행렬 $\boldsymbol{M}$을 $\boldsymbol{M} = \boldsymbol{U\Sigma V^t}$와 같은 형태로 분해한다.

$\Rightarrow$ $\boldsymbol{U}$의 개개 컬럼을 벡터로 가져온 것은 "왼쪽 특이벡터" 이다.

$\Rightarrow$ $\boldsymbol{V}$의 개개 컬럼을 벡터로 가져온 것은 "오른쪽 특이벡터" 이다.

$$\boldsymbol{U} = \begin{bmatrix} \uparrow & \cdots & \uparrow \\ \boldsymbol{u_1} & \cdots & \boldsymbol{u_m} \\ \downarrow & \cdots & \downarrow \end{bmatrix} \qquad \boldsymbol{V} = \begin{bmatrix} \uparrow & \cdots & \uparrow \\ \boldsymbol{v_1} & \cdots & \boldsymbol{v_n} \\ \downarrow & \cdots & \downarrow \end{bmatrix}$$

$\Rightarrow$ 특이벡터와 특이값 사이에는 다음과 같은 관계가 성립된다.

$$\boldsymbol{M}\,\boldsymbol{v_i} = \sigma_i\,\boldsymbol{u_i}$$

$\Rightarrow$ 특이벡터 사이에는 다음과 같은 직교 관계가 성립된다.

$$\boldsymbol{v_i} \cdot \boldsymbol{v_j} = \delta_{ij} \quad \Leftrightarrow \quad \boldsymbol{VV^t} = \boldsymbol{V^tV} = \boldsymbol{I}$$

$$\boldsymbol{u_i} \cdot \boldsymbol{u_j} = \delta_{ij} \quad \Leftrightarrow \quad \boldsymbol{UU^t} = \boldsymbol{U^tU} = \boldsymbol{I}$$

I 특이값 분해 (SVD)를 통한 차원 축소와 토픽 벡터

- r = 토픽의 가짓수라 할 때, 다음과 같은 차원축소가 가능하다.

$Size(\boldsymbol{M}) = m \times n$

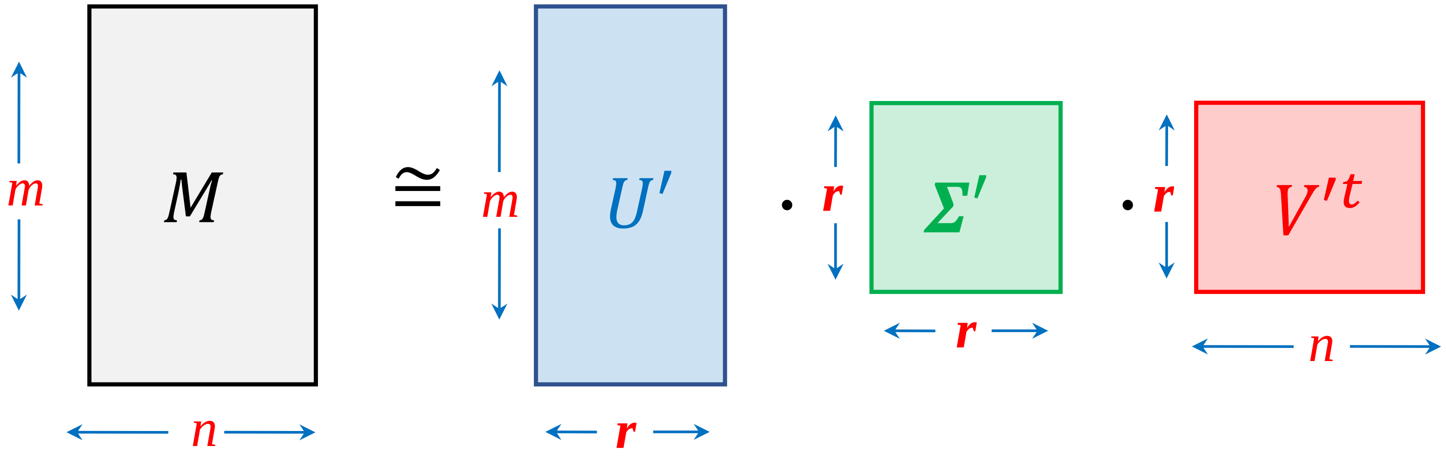
$Size(\boldsymbol{U}) = m \times m$ $\rightarrow m \times r$로 크기 축소.

$Size(\boldsymbol{\Sigma}) = m \times n$ $\rightarrow r \times r$ 로 크기 축소.

$Size(\boldsymbol{V}) = n \times n$ $\rightarrow n \times r$ 로 크기 축소.

I 특이값 분해 (SVD)를 통한 차원 축소와 토픽 벡터

- r = 토픽의 가짓수라 할 때, 다음과 같은 차원축소가 가능하다.



$\Rightarrow$ 차원 축소 $n \rightarrow r$

$\Rightarrow$ U'의 컬럼들이 토픽 벡터.

I 끝.

감사합니다.