08 텍스트 분석

07 문서 군집화

- Document Clustering

- 텍스트 분류 기반 문서 분류 : 결정 카테고리 값을 가진 학습데이터세트 필요

- 문서 군집화 : 비지도학습 기반

08 문서 유사도

- 코사인 유사도

$$AB = \|A\|\|B\|\cos\theta$$

$$\text{similarity} = \cos\theta = \frac{AB}{\|A\|\|B\|} = \frac{\Sigma AB}{\sqrt{\Sigma A^2}\sqrt{\Sigma B^2}}$$

09 한글 텍스트 처리

- 띄어쓰기, 조사로 인해 라틴어보다 어려움

- KoNLP