8-6. 토픽 모델링

-문서 집단에 숨어있는 주제 찾기

-중심 단어를 위주로 추출

8-7. 문서 군집화

-비슷한 텍스트 구성의 문서를 군집화하는 것

-앞서 한 문서 분류와는 달리, 문서 군집화는 비지도학습 기반으로 동작한다.

-군집별 핵심 단어 찾기

Cluster\_centers 배열 내 인덱스 값이 필요하다.

8-8. 문서 유사도

1. 코사인 유사도

벡터와 벡터 간의 유사도를 비교할 때, 벡터의 크기보다는 벡터의 상호 방향성이 얼마나 유사한지에 기반함. 즉, 두 벡터 사이의 사잇각을 구해 얼마나 유사한지 수치로 적용함.

유사도 cosθ는 두 벡터의 내적을 총 벡터 크기의 합으로 나눈 것.

8-9. 한글 텍스트 처리

-띄어쓰기와 다양한 조사로 인해 nlp 처리에 어려움 발생.

-한글 형태소 패키지 KoNLPy