Euron 15주차 예습과제_개념정리_한송희

Chapter8 텍스트 분석

06. 토픽 모델링(Topic Modeling)-20 뉴스 그룹

토픽 모델링: 문서 집합에 숨어 있는 주제를 찾아내는 것

→ 머신러닝 기반 토픽 모델링을 사용해 숨어 있는 중요 주제를 효과적으로 찾아낼 수 있음

-LSA(Latent Semantic Analysis)

-LDA(Latent Dirichlet Allocation): e.g. 8개 토픽별로 1000개의 word 피처가 해당 토픽별로 연관도 값을 가지고 있음

07. 문서 군집화 소개와 실습

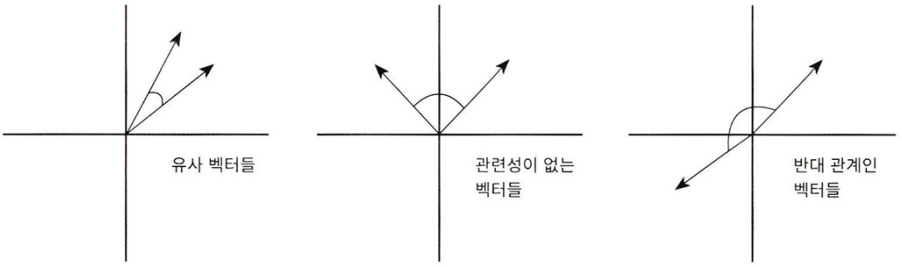
문서군집화: 비슷한 텍스트 구성의 문서를 군집화하는 것. 문서 군집화는 동일한 군집에 속하는 문서를 같은 카테고리 소속으로 분류할 수 있으므로 앞으로 소개한 텍스트 분류 기반의 문서 분류와 유사 but 문서 군집화는 학습 데이터 세트가 필요 없는 비지도 학습 기반으로 동작함

08. 문서 유사도

-코사인 유사도: 두 벡터 사이의 사잇각을 구해서 얼마나 유사한지 수치로 적용하는 것

두 벡터 사잇각

두 벡터의 사잇각에 따라서 상호 관계는 다음과 같이 유사하거나 관련이 없거나 아예 반대 관계가 될 수 있습니다.



두 벡터 A와 B의 코사인 값은 다음 식으로 구할 수 있습니다(고등학교 때 배웠겠지만, 아마 기억의 흔적이 없을 것입니다). 두 벡터 A와 B의 내적 값은 두 벡터의 크기를 곱한 값의 코사인 각도 값을 곱한 것입니다.

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

따라서 유사도 $cos\theta$는 다음과 같이 두 벡터의 내적을 총 벡터 크기의 합으로 나눈 것입니다(즉, 내적 결과를 총 벡터 크기로 정규화(L2 Norm)한 것입니다).

$$\text{similarity} = cos\theta = \frac{A \cdot B}{\| A \| \| B \|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

문서를 피처 벡터화 변환하면 차원이 매우 많은 희소 행렬이 되기 쉬움. 이러한 희소 행렬 기반에서 문서와 문서 벡터 간의 크기에 기반한 유사도 지표(유클리드 거리)는 정확도가 떨어지기 쉬움.+문서가 매우 긴 경우 단어의 빈도수도 더 많을 것이기 때문에 이러한 빈도수에만 기반해서는 공정한 비교를 할 수 없다

09. 한글 텍스트 처리-네이버 영화 평점 감성 분석

한글 NLP처리의 어려움: 일반적으로 한글 언어 처리는 영어보다 어려움

-띄어쓰기: 띄어쓰기를 잘못하면 의미가 왜곡되어 전달됨

-다양한 조사: 경우의 수가 많아 어근 추출 등의 전처리시 제거가 어려움

KoNLPy: 대표적인 파이썬 한글 형태소 패키지. 기존 엔진은 유지한채 파이썬 기반에서 인터페이스를 제공함