[7] 문서 군집화 소개와 실습

문서 군집화?
- 비슷한 텍스트구성의 문서를 군집화하는 것.
- 텍스트 기반의 문서 분류와 유사하지만, 문서 군집화는 텍스트 기반의 문서 분류와는 다르게 학습 데이터 세트가 필요없는 비지도 학습 기반으로 동작함.

[8] 문서 유사도

코사인 유사도?
문서와 문서 간의 유사도 비교는 일반적으로 코사인 유사도를 사용함.
코사인 유사도는 벡터와 벡터 간의 유사도를 비교할 때 벡터의 크기보다느 벡터의 상호 방향성이 얼마나 유사한지에 기반함. 즉, 코사인 유사도는 두 벡터 사이의 사잇각을 구해서 얼마나 유사한지 수치로 적용한 거임.

두 벡터 사잇각?
두 벡터의 사잇각에 따라서 상호 관계는 유사하거나 관련이 없거나 아예반대 관계가 될 수 있음.
두 벡터 A와 B의 내적 값 : 두 벡터의 크기를 곱한 값의 코사인 각도 값을 곱한 것.
따라서, 유사도 cosθ는 두 벡터의 내적을 총 벡터 크기의 합으로 나눈 것.

<코사인 유사도가 문서의 유사도 비교에 가장 많이 사용되는 이유>
1. 문서를 피처 벡터화 변환하면 차원이 매우 많은 희소 행렬이 되기 쉬움. 이러한 희소 행렬 기반에서 문서와 문서 벡터간의 크기에 기반한 유사도 지표는 정확도가 떨어지기 쉬움.

2. 문서가 매우 긴 경우 단어의 빈도수도 더 많을 것이기 때문에 이러한 빈도수에만 기반해서는 공정한 비교를 할 수 없음.

[9] 한글 텍스트 처리 - 네이버 평점 감성 분석

* 한글 NLP 처리의 어려움.
이유 : 띄어쓰기와 다양한 조사 때문.

* KoNLPy
- 파이썬의 대표적인 한글 형태소 패키지