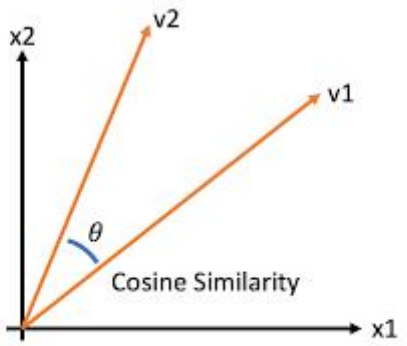
preview 15 ch8(2) parkboyeong

문서군집화 : 비슷한 텍스트 구성의 문서를 군집화 하는 것.
동일 카테고리 소속으로 분류할 수 있고, 이는 문서분류와 유사
비지도 학습

문서유사도-문서와 문서간의
코사인(cosine)유사도=두 벡터사이의 cosine value




벡터화->희소행렬기반에서 문서간의 유사도 지표는 정확도가 떨어진다.
빈도수뿐만 아니라 문서크기까지 고려해야 한다

한글 텍스트 처리
한글자연어처리의 어려움과 그 원인은 '띄어쓰기', '다양한조사' 때문이다.
KoNLPy 한글 형태소 패키지. 말뭉치를 형태소 어근단위로 쪼개고 각 형태소에 품사 태깅을 부착하는 작업이다
re 공백으로 변환
(네이버영화리뷰)sns분석에 적합한 twitter클래스 morphs() 형태소로 토큰화해 list반환

https://wikidocs.net/92961