LSA (Latent Semantic Analysis) 잠재 의미 분석 – 위키피디아

문서의 집합과 그 문서들이 포함하는 용어간의 개념을 생성하여 그 사이의 관계를 분석하는 자연어 처리 기술

LSA는 유사한 단어들이 유사한 내용의 텍스트를 생성한다고 가정

단락 당 단어의 수를 포함하는 행렬(행이 고유한 단어, 열이 각 단락)이 텍스트를 구성

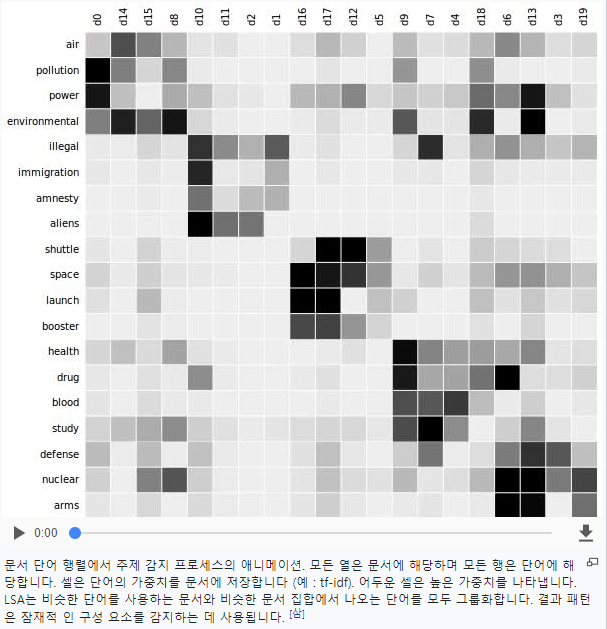
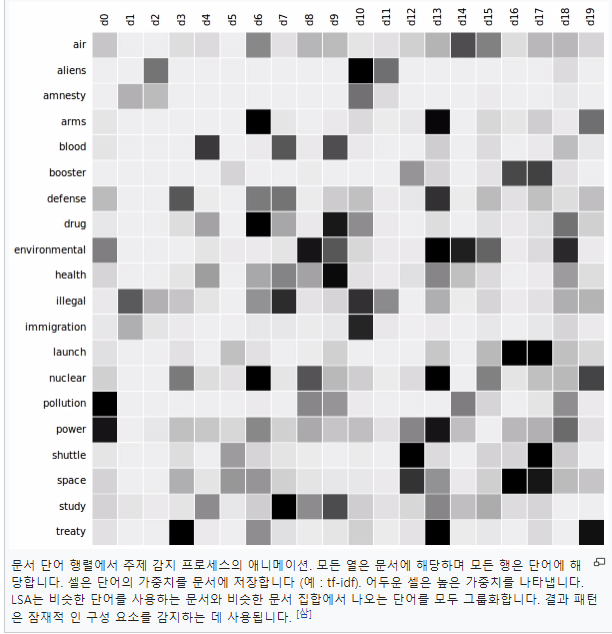
열(단락)이 비슷한 구조를 유지하면서 행(단어)을 줄이는 데에 SVD(Singular Value Decomposition)이라는 수학 기법을 사용

임의의 두 행(단어)으로 이루어진 두 벡터 사이 각에 코사인을 취하여 단어를 비교한다. 그 값이 1에 가까우면 유사한 단어들이고 0에 가까우면 유사하지 않은 단어들이다.

Occurrence Matrix 발생 행렬

문서에서의 용어 등장을 묘사하는 용어-문서 행렬(term-document matrix)을 사용. 행은 단어, 열은 문서인 희소행렬

tf-idf(term frequency – inverse document frequency) : 행렬의 원소의 가중치(weight)가 문서에서의 용어의 등장 수에 비례, 희귀하게 등장하는 용어는 상대적인 중요도를 반영하여 가중치를 부여



Rank Lowering 순위 낮추기

발생 행렬을 구성했다면, LSA는 용어-문서 행렬을 위한 낮은 순위 근사치(low-rank approximation)을 찾는다.

(1) 원래의 용어-문서 행렬은 계산하기에 너무 크다 -> 근사값을 취하여 줄이기

(2) 잡음이 너무 많아서 제거하기 위함: de-noisified matrix

(3) 그저 등장하는 단어의 나열 -> 각 문서와 관련 있는 모든 단어들, 주로 동의어의 집합

{(car), (truck), (flower)} --> {(1.3452 \* car + 0.2828 \* truck), (flower)}

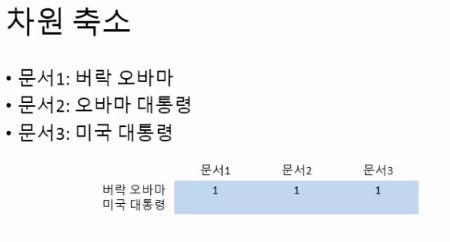
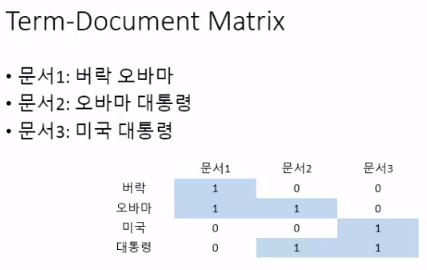
동의어 식별 문제를 완화: 비슷한 의미를 가진 용어들로 이루어진 차원을 병합하기 때문

다의어 문제를 완화: 올바른 방향을 가리키는 다차원 단어의 구성 요소가 유사한 의미를 공유하는 단어 구성 요소에 추가되기 때문

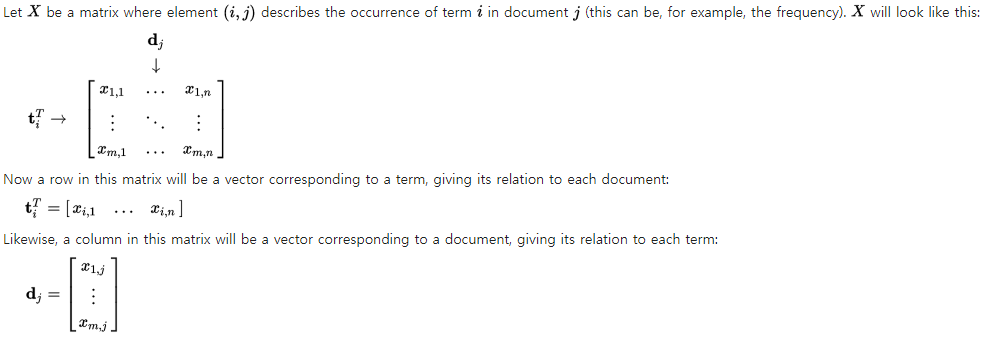
LSA의 개념 - 마인드스케일 강의

차원축소. 정보가 많이 손실되지 않으면서 의미 있는 자료로 가공(보기 쉽게)

1. 단어가 아닌 의미로 재배치
2. 동음이의어, 오탈자 처리
3. 문서에 존재하는 noise가 줄어 더 잘 분류됨



Derivation 파생



두 용어 벡터의 내적 tiTtp: 문서집합에서의 두 단어의 연관성을 나타냄

XXT는 이 모든 내적을 포함함. Element (i,p) of XXT = element (p,i) of XXT = tiTtp

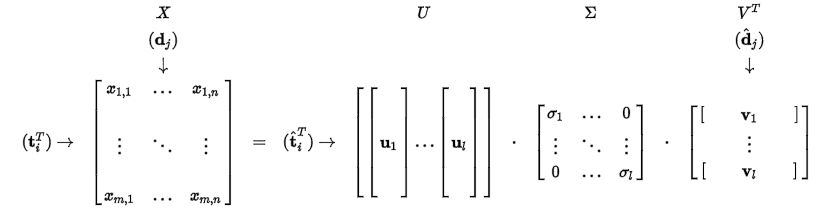
XTX는 문서 벡터의 내적을 포함함. (위와 마찬가지)

SVD(singular value decomposition)에 따라 X = U∑VT를 만족하는 집합 U와 V가 존재한다. (U와 V는 orthogonal matrix(UTU=UUT=I), ∑는 diagonal matrix(원소 (i,p)에 대해 i!=p일 경우 0의 값을 갖는다))



∑2가 diagonal이기 때문에 U는 XXT의 eigenvector에 무조건 포함이 되어야 한다. V 또한 XTX의 eigenvector에 무조건 포함이 되어야 한다. XXT와 XTX는 같은 값의 0이 아닌 eigenvalue를 갖는다.

(<https://en.wikipedia.org/wiki/Singular_value_decomposition#Relation_to_eigenvalue_decomposition> 참고)



뉴스 기사와 같은 형태는 LSA로 안정적인 패턴으로 클러스터링 하는 것이 좋음

SNS 분석과 같은 이슈 분석은 새로운 패턴을 찾아볼 수 있는 LDA(Latent Dirichlet Allocation) 알고리즘으로 클러스터링 하는 것이 좋다.

클러스터링(클러스터 분석): 주어진 데이터들의 특성을 고려해 데이터 집단(클러스터)을 정의하고 데이터 집단의 대표할 수 있는 대표점을 찾는 것으로 데이터 마이닝의 한 방법