LSA (Latent Semantic Analysis) 잠재 의미 분석

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

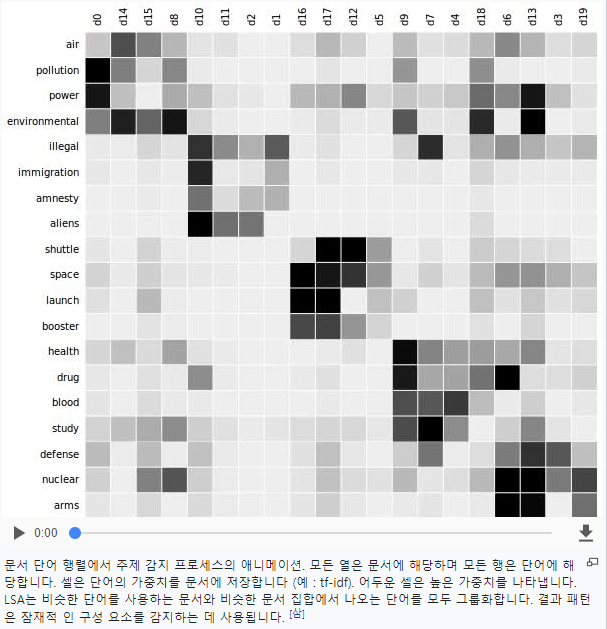
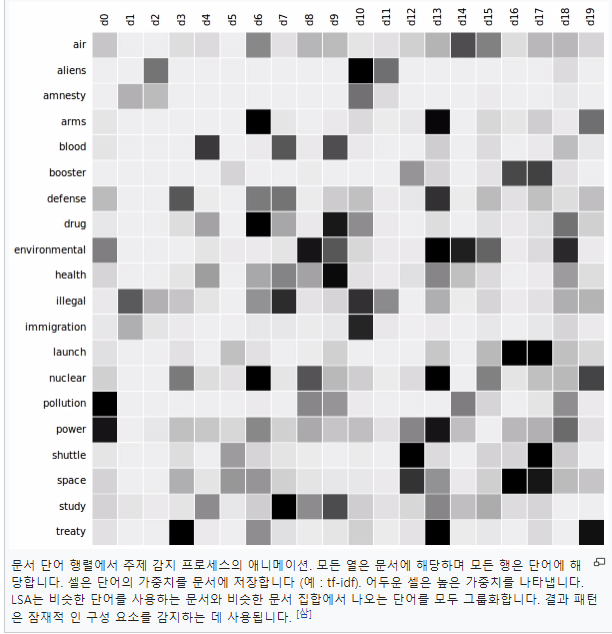
열(단락)이 비슷한 구조를 유지하면서 행(단어)을 줄이는 데에 SVD(Singular Value Decomposition)이라는 수학 기법을 사용

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

Derivation 파생

