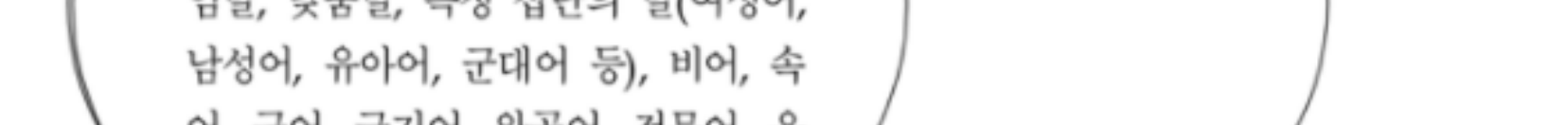
원-핫 인코딩(one-hot encoding)

텍스트를 유의미한 숫자(벡터)로 바꾸는 가장 손쉬운 방법론은 바로 '원-핫 인코딩(one-hot encoding)'이다. 이는 N개의 단어를 각각 N차원의 벡터로 표현하는 방식이다. 단어가 포함되는 자리엔 1을 넣고 나머지에는 0을 넣는다. 사전이 [인간, 펭귄, 문어, 사람]이라면 "인간"을 표현하는 벡터는 [1, 0, 0, 0]이 되는 식이다. 단어 하나에 인덱스 정수를 할당한다는 점에서 '단어 주머니(bag of words, BoW)'라 부르기도 한다.



[그림1] 원-핫인코딩은 한개의 요소만 1이고 나머지는 0인 N차원의 벡터로 표현된다

이 방식은 나름대로 좋은 성능을 내고, 지금까지도 사용하는사람들이 많지만 아주 큰 단점이 있다. 바로 컴퓨터가 단어의 의미 또는 개념 차이를 전혀 담지 못한다는 것이다. 예를 들어, '과학'과 '공학'의 관계는 '과학'과 '수박'의 관계와 차이가 없다.

[그림 2] 유의어와 관련어*1

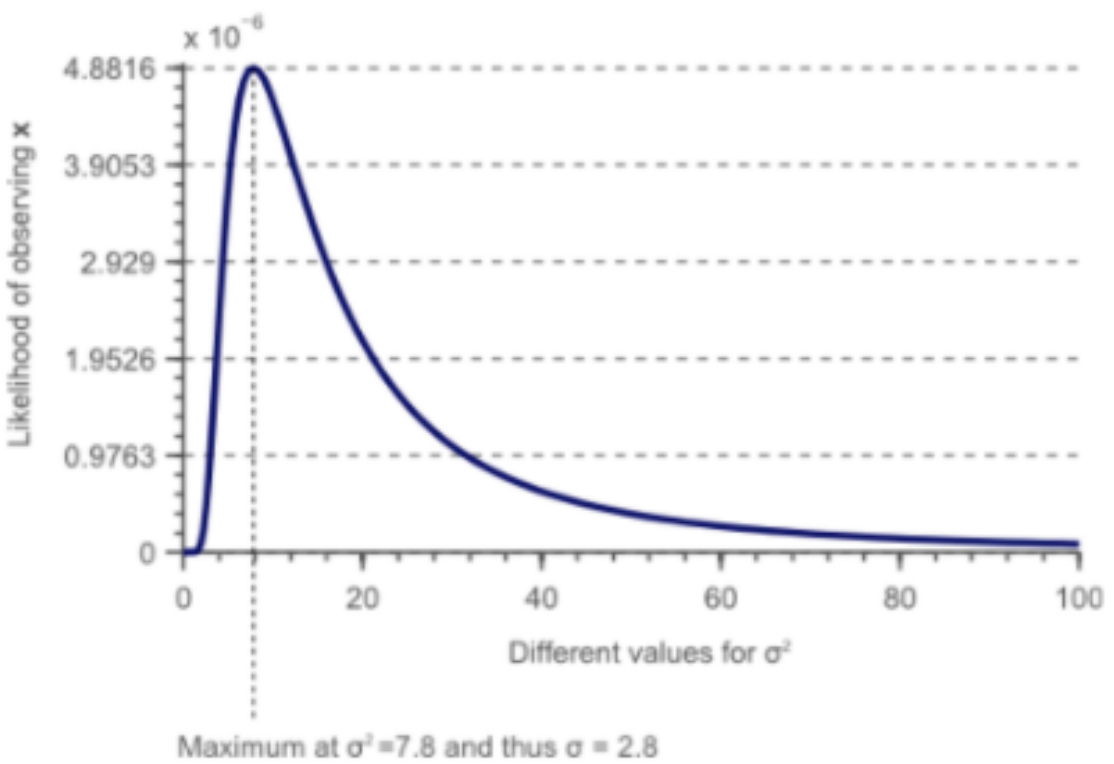
수학적으로 보자면, 원-핫 벡터들은 딱 하나의 요소만 1이고 나머지는 모두 0인 희소 벡터(sparse vector) 형태를 띤다. 이런 경우 두 단어 벡터의 내적(inner product)은 0으로 직교(orthogonal)를 이룬다. 이는 단어 간 존재하는 유의어, 반의어와 같은 특정한 관계나 의미를 전혀 담지 못한 채 서로 독립적(independent)으로만 존재한다는 것을 의미한다.

'차원의 저주(curse of dimensionality)' 문제도 발생한다. 하나의 단어를 표현하기 위해 말뭉치(corpus)에 존재하는 수만큼의 차원을 가지게 되면 계산 복잡성이 기하급수적으로 늘어난다. 예를 들어,40만 개의 고유의 언어 데이터 셋을 활용해 원-핫 인코딩 배열을 구성한다면 그 차원 수는 40만에 이르게 된다.

단어 활용 상황	차원(단위 : 만)
음성	2
PTB	5
대사전	50
구글 웹크롤 말뭉치(1TB)	1300

[표 1] 단어 수가 많아질수록 차원의 크기는 기하급수적으로 증가한다.*2

하지만 차원 수가 일정 수준을 넘어서면 분류기(classifier)의 성능은 되려 0으로 수렴한다는 점에서 봤을 때 제아무리 뛰어난 성능을 가진 컴퓨터라도 이런 고차원(high dimensionality)의 벡터를 학습하기는 어렵고 성능이 떨어지기 마련이다.



[그림 4] 특징 수가 일정 수준을 넘어서면 분류 성능은 오히려 낮아진다