임베딩(Embedding)

임베딩: 문자를 기계가 이해할 수 있는 숫자로 바꾼 결과 혹은 그 과정

원 핫 인코딩(One-Hot Encoding)

표현하고 싶은 단어의 인덱스에 1의 갑을 부여하고, 다른 인덱스에는 0을 부여하는 벡터 표현 방식

이렇게 표현한 벡터를 원 핫 벡터

단어	단어 인덱스	원-핫 벡터
you	0	[1, 0, 0, 0, 0, 0, 0]
say	1	[0, 1, 0, 0, 0, 0, 0]
goodbye	2	[0, 0, 1, 0, 0, 0, 0]
and	3	[0, 0, 0, 1, 0, 0, 0]
I	4	[0, 0, 0, 0, 1, 0, 0]
say	5	[0, 0, 0, 0, 0, 1, 0]
hello	6	[0, 0, 0, 0, 0, 0, 1]

단점

1. 공간 낭비(희소 벡터, 하나만 1이고 나머지는 전부 0)
2. 단어의 유사도를 표현하지 못함

희소 벡터(Sparse Vector) : 원소 대부분이 0인 벡터를 희소벡터라고 한다. 차원 감소의 핵심은 희소베터의 중요한 축을 찾아내어 더 적은 차원으로 다시 표현하는 것이다. 차원 감소의 결과로 원래 희소벡터의 원소 대부분이 0이 아닌 값으로 구성된 '밀집 벡터(Dense Vector)'로 변환된다.

Word2Vec

원 핫 인코딩방식의 단점을 보안

단어의 의미를 반영해 다차원 공간에 추론 기반으로 벡터화

뉴럴 네트워크 언어 모델(Neural Net Language Model, NNLM) 기반

분포가설에 기반하여 맥락을 이용해 중심 단어를 추론하는 작업

1. CBOW 모델: 주변 단어로부터 중심단어를 추측하는 신경망
2. Skip-Gram: 중심단어로부터 주변 단어를 추측하는 신경망