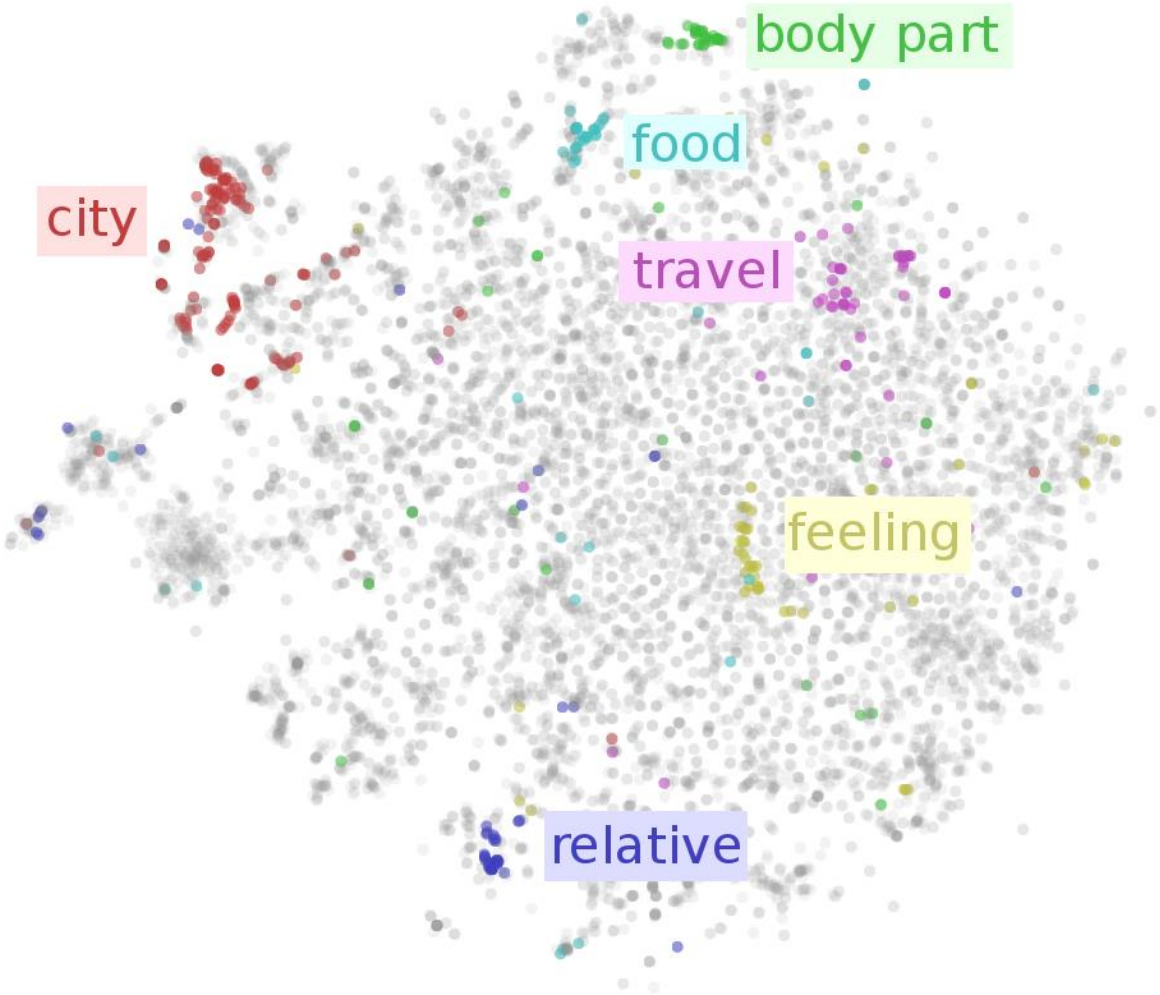
단어 임베딩



임베딩(Embedding)

컴퓨터가 자연어를 처리할 수 있게 하려면 자연어를 계산 가능한 형식인 임베딩으로 바꿔줘야 합니다.

■ 임베딩(embedding)

- 임베딩은 자연어를 숫자의 나열인 벡터로 바꾼 결과 혹은 그 일련의 과정 전체를 가리키는 용어입니다.
- 단어나 문장 각각을 벡터로 변환해 벡터 공간에 '끼워 넣는다(embed)'는 취지에서 임베딩이라는 이름이 붙었습니다.

■ 임베딩이 중요한 이유

- 임베딩에는 말뭉치(corpus)의 의미, 문법 정보가 응축돼 있습니다.
- 임베딩은 벡터이기 때문에 사칙연산이 가능하며, 단어/문서 관련도(relevance) 역시 계산할 수 있습니다.
- 최근 임베딩이 중요해진 이유는 전이 학습(transfer learning) 때문입니다.
- 전이 학습이란 특정 문제를 풀기 위해 학습한 모델을 다른 문제를 푸는 데 재사용하는 기법입니다.
  예컨대 대규모 말뭉치를 미리 학습(pre train)한 임베딩을 문서 분류 모델의 입력값으로 쓰고,
  해당 임베딩을 포함한 모델 전체를 문서 분류 과제를 잘할 수 있도록 업데이트(fine-tuning)하는 방식
- 대규모 말뭉치를 학습시켜 임베딩을 미리 만들고(pre train),
  이후 임베딩을 포함한 모델 전체를 문서 분류 과제에 맞게 업데이트합니다(fine-tuning).

출처 : https://bit.ly/2NLIAGV

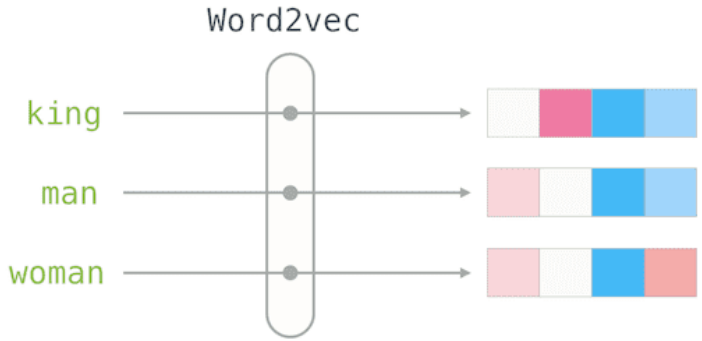
임베딩 종류

■ 단어 수준 임베딩

- NPLM : Neural Probabilistic Language Model
- Word2Vec
- FastText
- LSA(Latent Semantic Analysis, 잠재의미 분석)
- GloVe
- Swivel

■ 문장 수준 임베딩

- Doc2Vec(Document Embedding with Paragraph Vectors)
- LDA (Latent Dirichlet Allocation, 잠재 디리클레 할당)
- ELMo(Embeddings from Language Model)
- BERT(Bidirectional Encoder Representations from Transformers)



원핫 인코딩(one hot encoding)

고유 값에 해당하는 칼럼에만 1을 표시하고 나머지 칼럼에는 0을 표시하는 방법입니다.

Human-Readable

Pet
Cat
Dog
Turtle
Fish
Cat

Machine-Readable

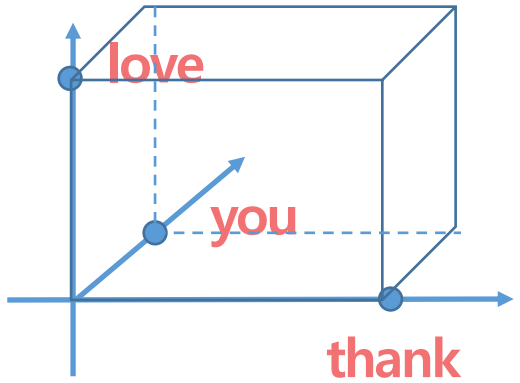
Cat	Dog	Turtle	Fish
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
1	0	0	0

- 단어를 범주형 변수로 변환
- 이진벡터로 표현
- 벡터의 각 차원이 단어 하나를 나타냄
- 모든 단어들의 유사도가 없음

"thank you"

"love you"

단어	임베딩
thank	[1, 0, 0]
you	[0, 1, 0]
love	[0, 0, 1]



원 핫 인코딩 실습

onehot_encoding.ipynb

희소표현 vs 밀집표현(분산표현)

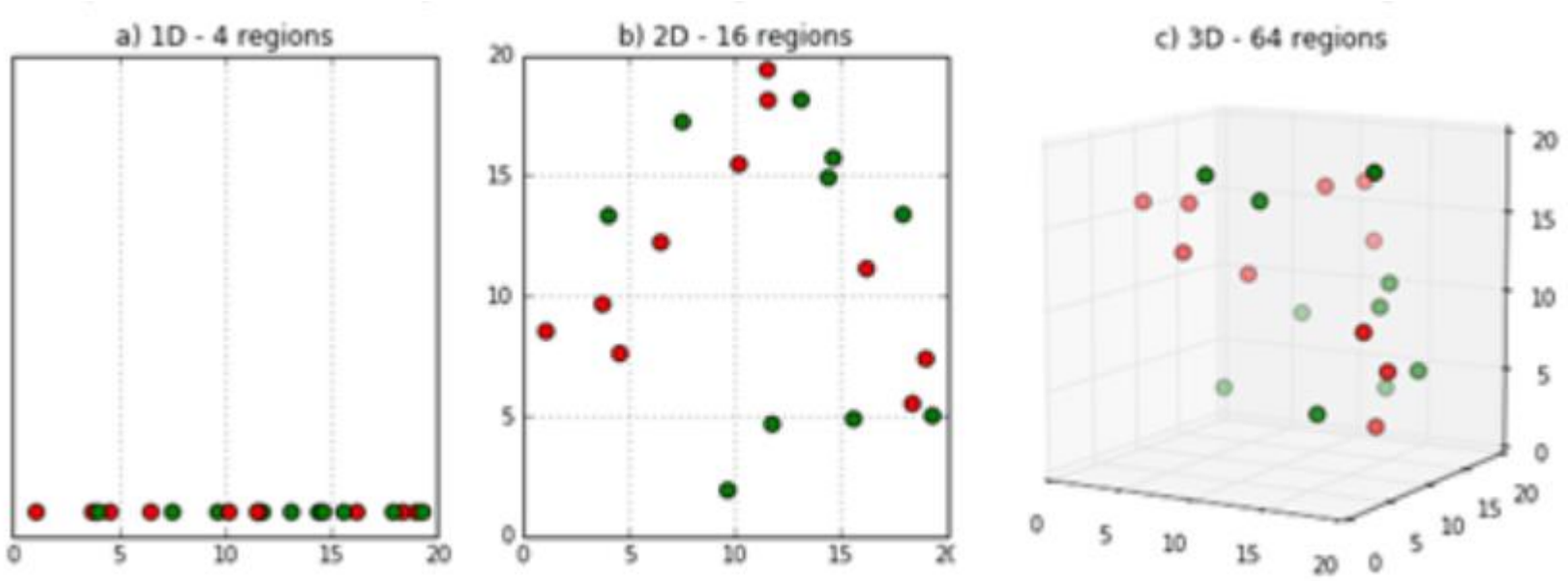
■ 희소표현(sparse representation)

■ 밀집표현(distributed representation)
분산표현(distributed representation)

https://dreamgonfly.github.io/blog/word2vec-explained/

차원의 저주(The curse of dimensionality)

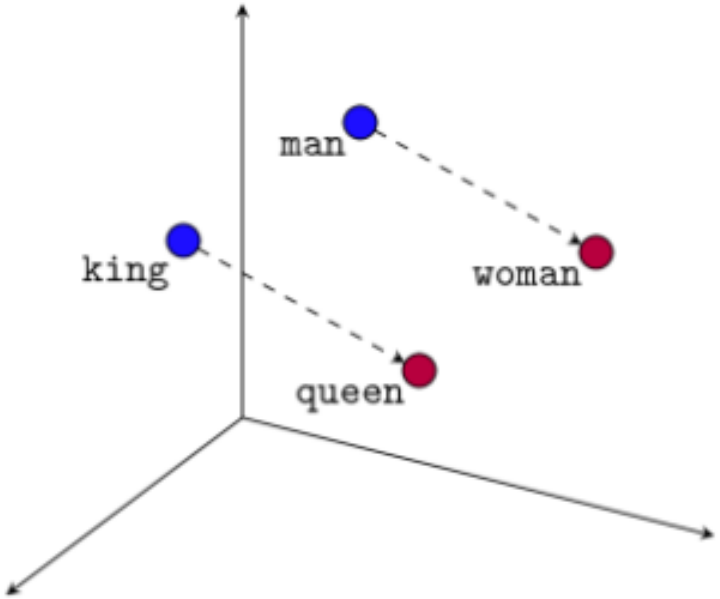
- 데이터 학습을 위해 차원이 증가하면서 학습데이터 수가 차원의 수보다 적어져 성능이 저하되는 현상
- 차원이 증가할 수록 개별 차원 내 학습할 데이터 수가 적어지는(sparse) 현상 발생
- 해결책 : 차원을 줄이거나 데이터를 많이 획득



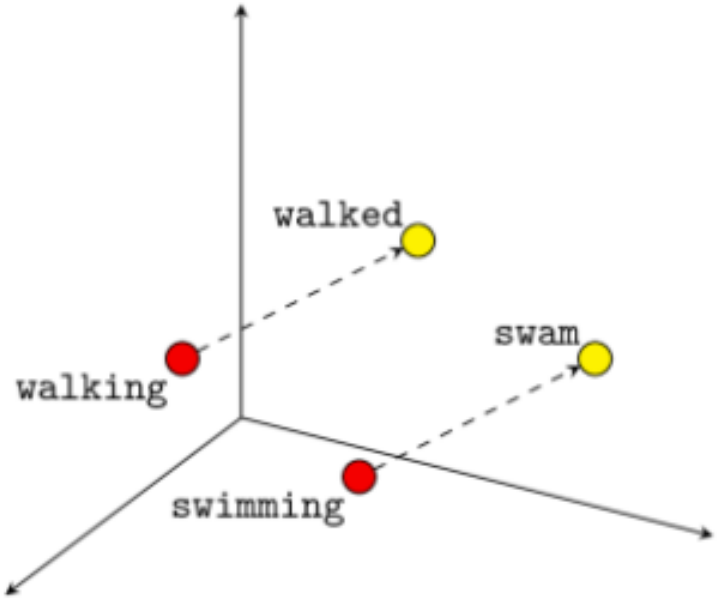
https://deepai.org/machine-learning-glossary-and-terms/curse-of-dimensionality

밀집표현(분산표현) 장점

임베딩 벡터의 차원을 축소하고 단어간 유사도를 계산할 수 있습니다



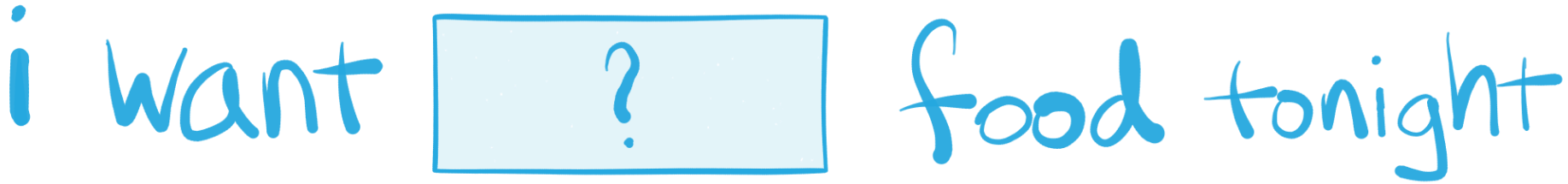
성별



시제

국가-수도

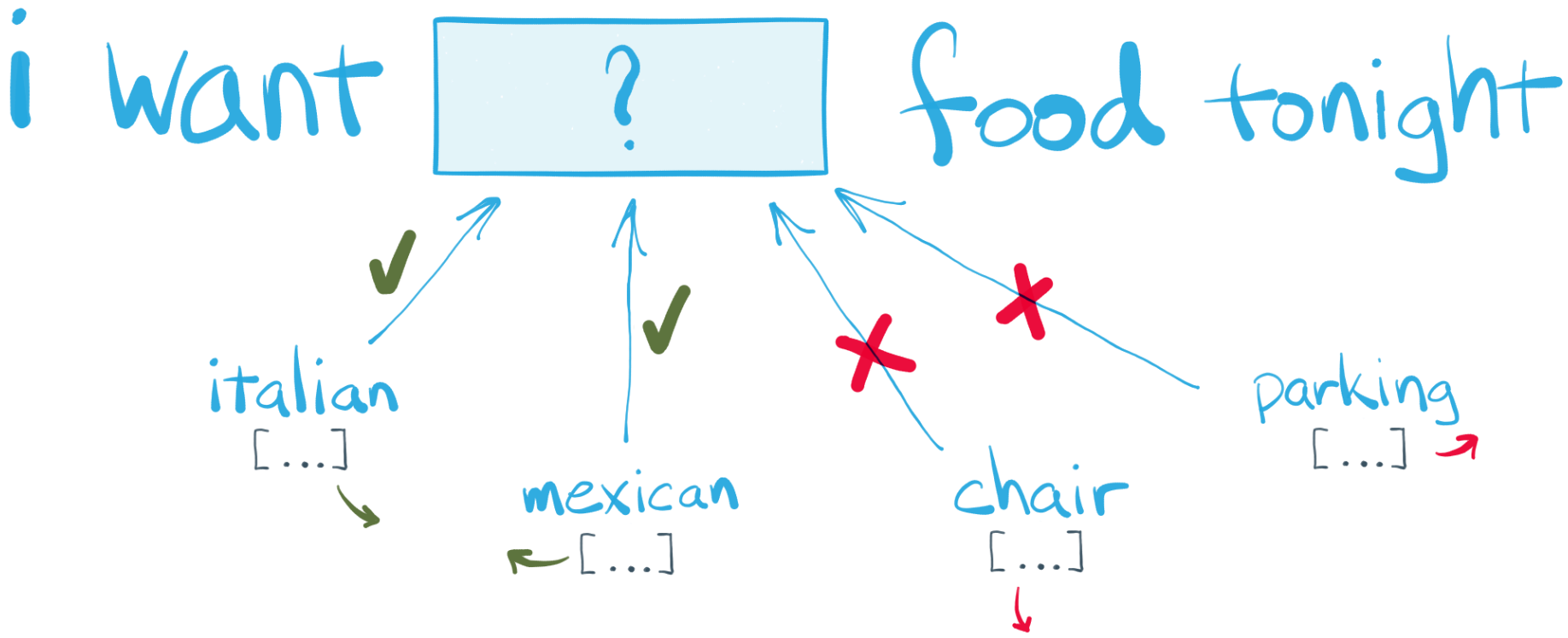
Word2Vec



https://deepai.org/machine-learning-glossary-and-terms/curse-of-dimensionality

Word2Vec

단어의 주변을 보면 그 단어를 안다. You shall know a word by the company it keeps. - 언어학자 J.R. Firth (1957)

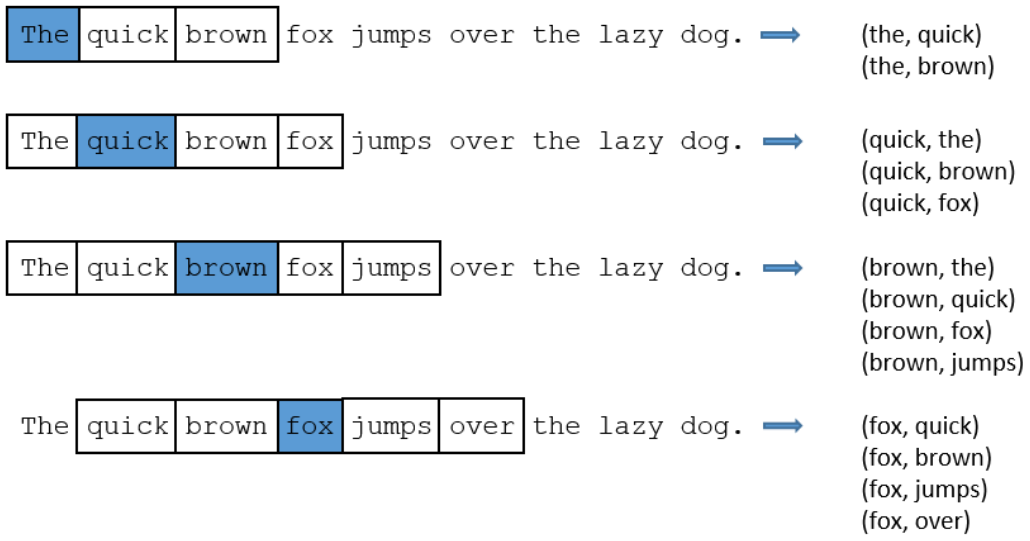


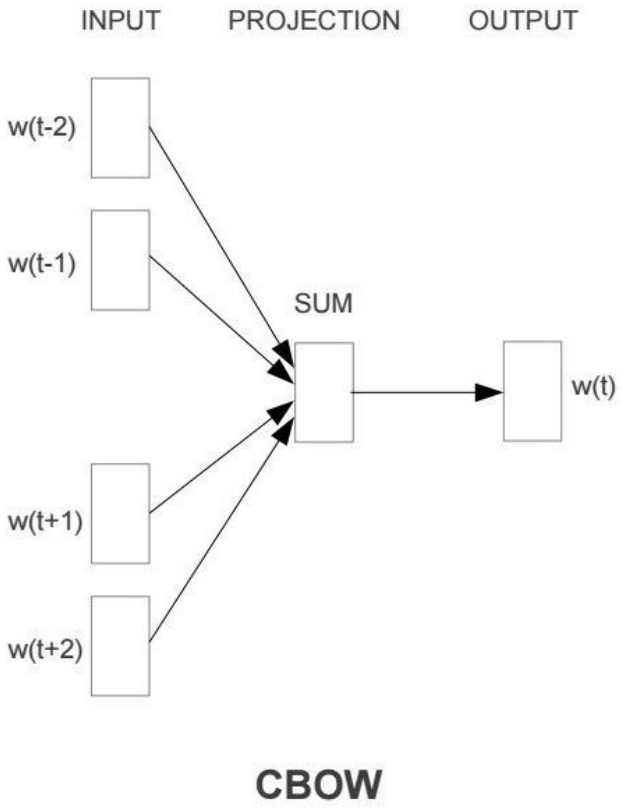
https://deepai.org/machine-learning-glossary-and-terms/curse-of-dimensionality

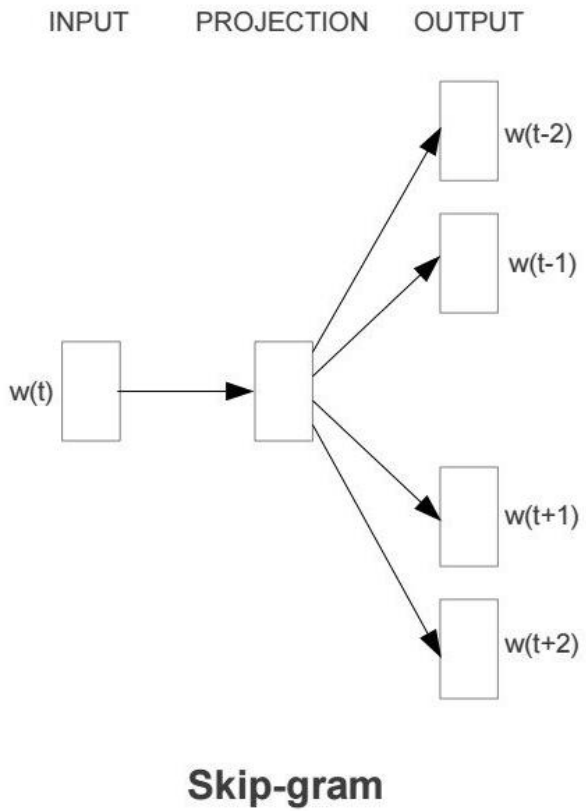
Word2Vec 알고리즘

맥락으로 단어를 예측하는 CBOW(continuous bag of words)과 단어로 맥락을 예측하는 skip-gram 모델이 있습니다.

- CBOW(Continuous Bag Of Words: 주변 단어들의 임베딩 벡터의 합을 이용하여 타깃 단어를 예측
- skip-gram : 타깃의 임베딩을 이용하여 주변 단어들을 예측







https://mubaris.com/posts/word2vec/

Word2Vec 실습

```
model = Word2Vec(sentences=sentences, size=100, window=4, min_count=2, sg=1)
model.save('nsmc.model')
```

- sentences : 학습 문장 데이터(입력)
- size : 단어 임베딩 벡터의 차원(크기)
- window : 주변 단어 윈도우의 크기
- min_count : 단어 최소 빈도 수 제한(학습에서 제외)
- sg : 0(CBOW 모델), 1(skip-gram 모델)

word2vec_model.ipynb

OOV (Out of Vocabulary)

생성된 코퍼스 데이터만 사용하여 학습한 임베딩 모델은 코퍼스 외부에 존재하는 단어에는 대응할 수 없는 문제가 있습니다.

"체크카드 만드는 법 알려줘"

"체카 만드는 법 알려줘"

FastText

FastText는 Facebook에서 만든 단어 임베딩 및 텍스트 분류 학습을 위한 라이브러리입니다.
FastText에서는 각 단어는 글자 단위 n-gram의 구성으로 취급하며, subword를 고려하여 학습합니다.

- Word2Vec는 OOV(Out Of Vocabulary)에 대해서는 벡터를 못 구하고, 빈도수가 낮은 단어에도 학습이 부족합니다.
- FastText는 이를 보완하기 위하여, 단어를 구성하는 subwords (substrings) 의 벡터의 합으로 단어 벡터를 표현합니다
- FastText는 새로운 단어에 대해서도 단어의 형태적 유사성을 고려한 적당한 word representation 을 얻도록 도와줍니다.

https://fasttext.cc/

Library for efficient text classification and representation learning

FastText 실습

fasttext_model.ipynb


Look Dick Look look at Jane
see Jane laugh and play
이봐 딕, 제인 좀 봐
즐겁게 놀고 있는 제인을 보라구

문장 토큰화(Tokenizing)

Look Dick Look look at Jane see Jane laugh and play

Look Dick Look see pretty Jane

look dick look look at jane see jane laugh and play

look dick look see pretty jane

단어사전(Vocab) 만들기

look dick look look at jane see jane laugh and play

look dick look see pretty jane

단어사전(Vocab)	
look	0
dick	1
at	2
jane	3
see	4
laugh	5
and	6
play	7
pretty	8

단어를 숫자로 치환

look dick look look at jane see jane laugh and play

0 1 0 0 2 3 4 3 5 6 7

단어사전(Vocab)	
look	0
dick	1
at	2
jane	3
see	4
laugh	5
and	6
play	7
pretty	8

[0, 1, 0 , 0, 2, 3, 4, 3, 5, 6]

입력 데이터 길이 맞추기

딥러닝 모델에 입력으로 사용하기 위해서는 길이를 같게 해 주어야 합니다.

Look Dick Look look at Jane see Jane laugh and play

Look Dick Look see pretty Jane

[0, 1, 0 , 0, 2, 3, 4, 3, 5, 6]
[0, 1, 0 , 4, 8, 3, 0, 0, 0, 0]

단어의 차원

너무 큰 차원의 데이터를 학습할 때, 차원의 저주(curse of dimensionality)에 빠질 수 있습니다.

look dick look look at jane see jane laugh and play

0 1 0 0 2 3 4 3 5 6 7

look	[**1**, 0, 0, 0, 0, 0, 0, 0, 0]
dick	[0, **1**, 0, 0, 0, 0, 0, 0, 0]
look	[**1**, 0, 0, 0, 0, 0, 0, 0, 0]
look	[**1**, 0, 0, 0, 0, 0, 0, 0, 0]
at	[0, 0, **1**, 0, 0, 0, 0, 0, 0]
jane	[0, 0, 0, **1**, 0, 0, 0, 0, 0]
see	[0, 0, 0, 0, **1**, 0, 0, 0, 0]
Jane	[0, 0, 0, **1**, 0, 0, 0, 0, 0]
laugh	[0, 0, 0, 0, 0, **1**, 0, 0, 0]
and	[0, 0, 0, 0, 0, 0, **1**, 0, 0]
play	[0, 0, 0, 0, 0, 0, 0, **1**, 0]

실제 Corpus에서는 단어가 많으므로,
단어의 차원이 1,000차원 이상이 됩니다.

[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0 0, 0, 0, 0, 0, 0, 0, 0 0, 0, 0, 0, 0, 0, 0,
0 0, 0, 0, 0, 0, 0, 0, 0.....
.....
.....
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

임베딩(Embedding)

Embedding은 큰 차원을 줄여주는 역할을 하며, Sparsity문제를 해결합니다.

```
[0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                                ..........
                                ..........
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```

```
[-0.01208562, -0.02042891, 0.00930309, -0.01072387, 0.00621265,
-0.03476477, 0.02996606, 0.02715156, -0.04755617, -0.03230421,
-0.00678114, 0.04859651, 0.03986677, 0.01091999, -0.03999164,
-0.01312722]
```

단어 임베딩의 한계

주변 단어를 통해 학습이 이루어지기 때문에 문맥을 고려하지 못하며, 동형어 , 다의어 등에서 성능이 안 좋습니다.

10과 4의 차는 6이다. → 차[8]
표준국어대사전

차가 식으니 어서 드세요.
표준국어대사전

결혼 10년 차에 내 집을 장만했다. → 차[4]
표준국어대사전

잠이 막 들려던 차에 전화가 왔다. → 차[4]
표준국어대사전

부모와 자식 간의 세대 차가 크다. → 차[8]
표준국어대사전

100에서 49를 빼면 그 차가 얼마인가? → 차[8]
표준국어대사전

그녀는 손님에게 대접할 차를 내왔다. → 차[2]
표준국어대사전

차[4] 次 ★

1. 의존명사 '번', '차례'의 뜻을 나타내는 말.
2. 의존명사 어떠한 일을 하던 기회나 순간.
3. 의존명사 수학 방정식 따위의 차수를 이르는 말.

유의어 번[4] 차례[2]

표준국어대사전

차[2] ★

1. 명사 차나무의 어린잎을 달이거나 우린 물.
2. 명사 식물의 잎이나 뿌리, 과실 따위를 달이거나 우리거나 하여 만든 마실 것을 통틀어 이르는 말. 인삼차, 생강차, ...
3. 명사 식물 [같은 말] 차나무(차나뭇과의 상록 활엽 관목).

유의어 차나무

표준국어대사전

차[6] 車

1. 명사 바퀴가 굴러서 나아가게 되어 있는, 사람이나 짐을 실어 옮기는 기관. 자동차, 기차, 전차, 우차, 마차 따위를 ...
2. 명사 화물을 '[1]'에 실어 그 분량을 세는 단위.
3. 명사 운동 '車' 자를 새긴 장기짝. 한편에 둘씩 모두 넷이 있고 일직선으로 가로나 세로로 몇 칸이든지 다닌다.

https://ko.dict.naver.com/#/search?query=%EC%B0%A8

Thank you

kgpark88@gmail.com