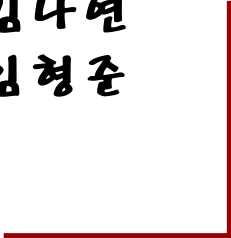




NLP Project

강유정	김근호	김나연
김정현	박진우	임형준



Index

1. NLP?

- Text Classification
- Text Similarity
- NLG
- Machine Comprehension

2. Text Vectorization

- One-hot Encoding
- Count-base Method
- Predictive Method
- Glove

NLP?

Text
Classification

Text Similarity

NLG

Machine
Comprehension

Text Classification



1. 스팸메일 자동분류하기

스팸으로 의심되는 메일을 스팸메일함으로 자동분류하여
스팸없는 깨끗한 네이버 메일을 이용하실 수 있습니다.

스팸 자동 분류

스팸 자동 분류 항목

☐ 내 네이버 메일 주소로 받은 메일이 있으면 스팸으로 자동 분류 ?

☐ 발신사명, 참조에 내 메일 주소가 있으면 스팸으로 자동 분류 ?

☐ 보낸 사람의 주소가 주소록에 있으면 스팸으로 자동 분류

☒ 카페메일함을 사용하여 카페 사정 메일을 차단 ?

☒ 언어별 스팸 분류 설정 사용 ?

☐ 모든 외국어 메일을 스팸으로 분류 ☒ 스팸으로 분류할 언어 직접 선택

☒ 영어 ☐ 일본어 ☐ 러시아어 ☐ 중국어

스팸 자동여동 ☒ 사용함 ☐ 사용 안 함

[사용함으로 설정하면 발송처의 스팸으로 의심되는 메일이 자동으로 스팸메일함으로 이동됩니다.
차단된 메일은 스팸메일함에서 스팸자동여동으로 표시되며, 언어별 스팸메일함으로 이동되면 해당 메일 통수도 변경됩니다.]

Supervised Learning

- ✓ Naïve Bayes Classifier
- ✓ SVM
- ✓ Neural Network

Unsupervised Learning

- ✓ K-means Clustering
- ✓ Hierarchical Clustering

Text Similarity

"이 노래 누가 만들었어?"

"지금 나오는 노래의 작곡가는 누구야?"

Similar??

- ✓ Jaccard Similarity

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

- ✓ Cosine Similarity

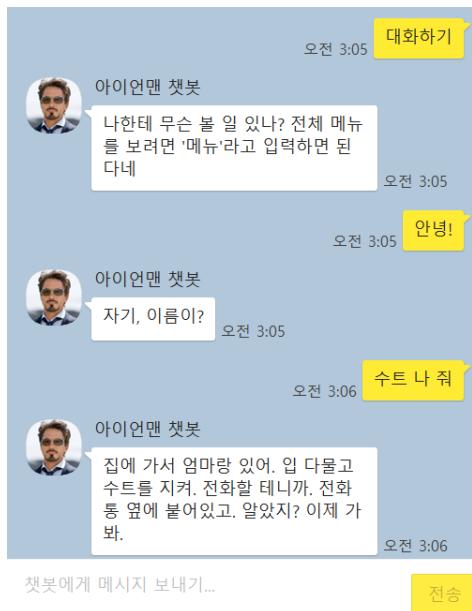
$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

- ✓ Euclidean Distance

$$d(\mathbf{q}, \mathbf{p}) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

- ✓ Mantattan Similarity

$$d = \sum_{i=1}^n |x_i - y_i|$$



입력된 문장의 이해(NLU)



연관된 단어 벡터 추출



문장 구조 정하기



문장 내에서 단어 배열



문장 생성

Machine Comprehension

Task 1: Single Supporting Fact

Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A:office

✓ bAbI

Task 2: Two Supporting Facts

John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground

✓ SQuAD

출처: <https://yerevann.github.io/>



What is the mustache made of?

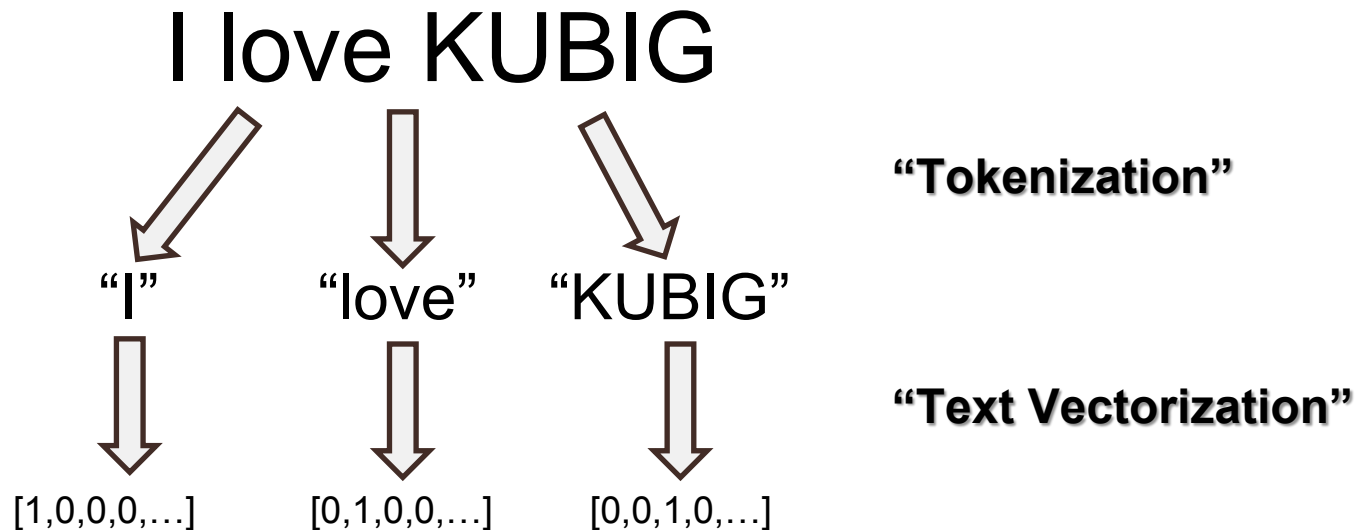
AI System

bananas

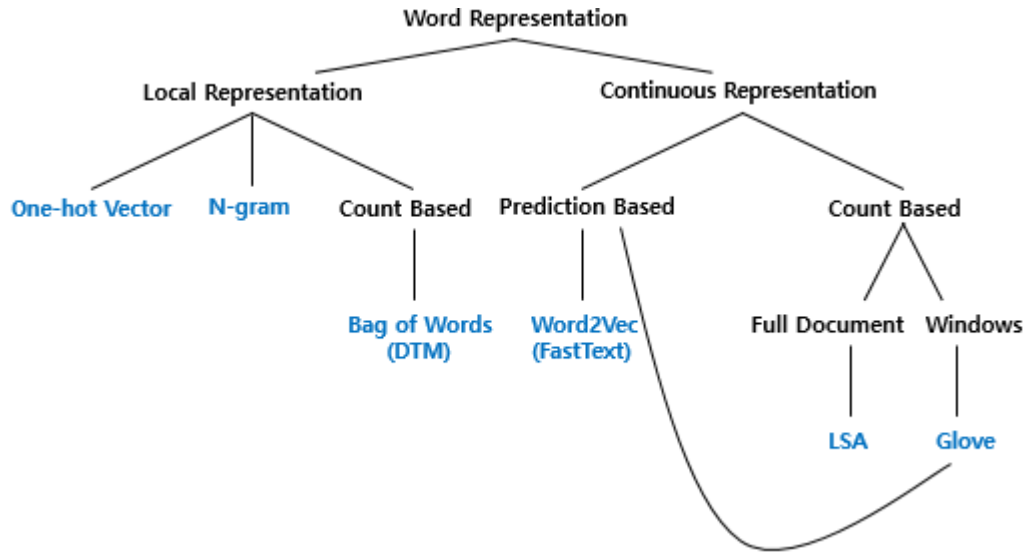
✓ VQA

출처: <https://visualqa.org/challenge.html>

Text Vectorization



Text Vectorization



출처: <https://wikidocs.net/31767>

Local Representation

- 단어 그 자체만 가지고 표현
- 대부분이 0인 vector
- 단어의 연관성, 의미 표현 불가

Continuous Representation

- 주변 단어 참고 후 표현
- 단어의 연관성, 의미 표현 가능

One-hot Encoding

King Love Queen



word	one-hot
King	[1,0,0]
Love	[0,1,0]
Queen	[0,0,1]

- ✓ 간단하고 직관적인 계산 방법
- ✓ 벡터의 크기가 커지는 위험
- ✓ 비효율적인 저장 방식
- ✓ 단어의 의미 / 특성 표현 불가

출처: <https://blog.naver.com/timtaeil/221335952229>

Count-based Method

Example corpus:

- ♥ I like playing tennis
- ♥ I like sweets
- ♥ I enjoy skiing

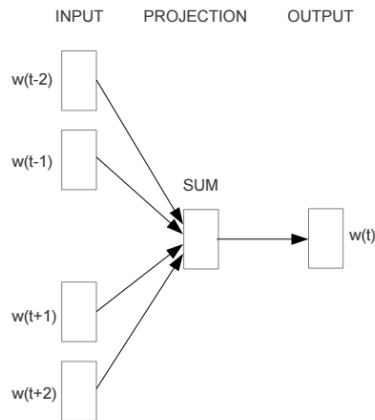
	I	like	enjoy	playing	tennis	sweets	skiing
I	0	2	1	0	0	0	0
like	2	0	0	1	0	1	0
enjoy	1	0	0	0	0	0	1
playing	0	1	0	0	1	0	0
tennis	0	0	0	1	0	0	0
sweets	0	1	0	0	0	0	0
skiing	0	0	1	0	0	0	0

출처: <https://badootech.badoo.com/>

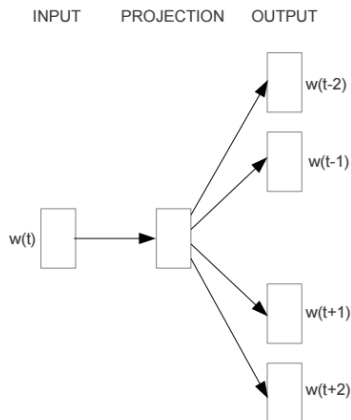
- ✓ Co-occurrence Matrix 이용
- ✓ SVD 등을 활용해 벡터 생성
→ 차원 축소 가능
- ✓ 한 번만 계산으로 벡터 생성 가능
- ✓ 주변 단어를 고려한 벡터 생성

Predictive Method

창욱은 냉장고에서 ____ 꺼내서 먹었다. ____ ____ 음식을 ____ ____.



CBOW

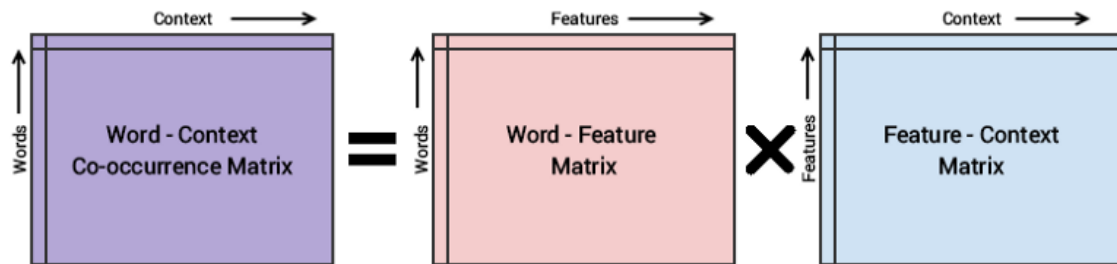


Skip-gram

- ✓ 특정 문맥에서 나올 단어를 예측
- ✓ CBOW / Skip-gram 존재
- ✓ 단어 간의 유사도 측정에 강점
- ✓ 복잡한 특징까지도 잘 파악

출처: <https://towardsdatascience.com>

Glove



- ✓ Count-based Method와 predictive method 동시에 활용
- ✓ 각각의 단점 보완(통계 정보 반영, 유추 작업도 가능)
- ✓ 벡터 사이의 내적이 co-occurrence probability가 되도록 모델링

Q&A