자연어 처리가 뭔가요?

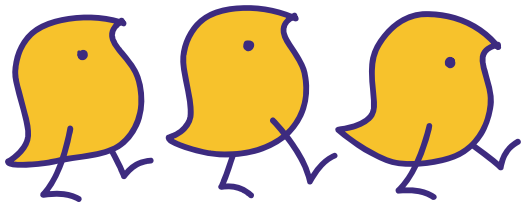
왜 회사들에서 관심가지나요?

SEnE 이재석

목차

Language Conference 2019

1. 오늘의 결론
2. 자연어처리 기본적 이해
    1. 기본적이론
    2. 챗봇
    3. 텍스트분석
    4. 워드임베딩
    5. 한글처리문제
3. 자연어처리 기법
    1. 단어주머니
    2. 이산표현
    3. 분산표현
    4. 임베딩
4. Word2Vec



1. 오늘의 결론

LangCon 2019

단어

차원

임베딩

말귀

분산표현

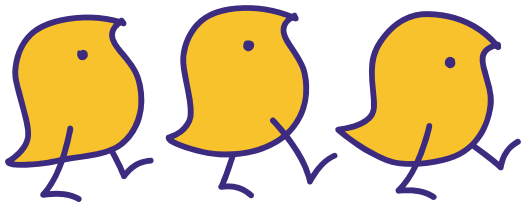
크롤링

클라우드

DB

정규표현식

도메인



2. 자연어처리 기본적 이해

LangCon 2019

자연 언어란?

- 정보 전달의 수단
- 인간 고유의 능력
- 인공 언어에 대응되는 개념
- 특정 집단에서 사용되는 모국어의 집합

⇒ 우리가 일상 생활 속에서 사용하는 언어(말)

자연 언어 처리란?

Language Conference 2019

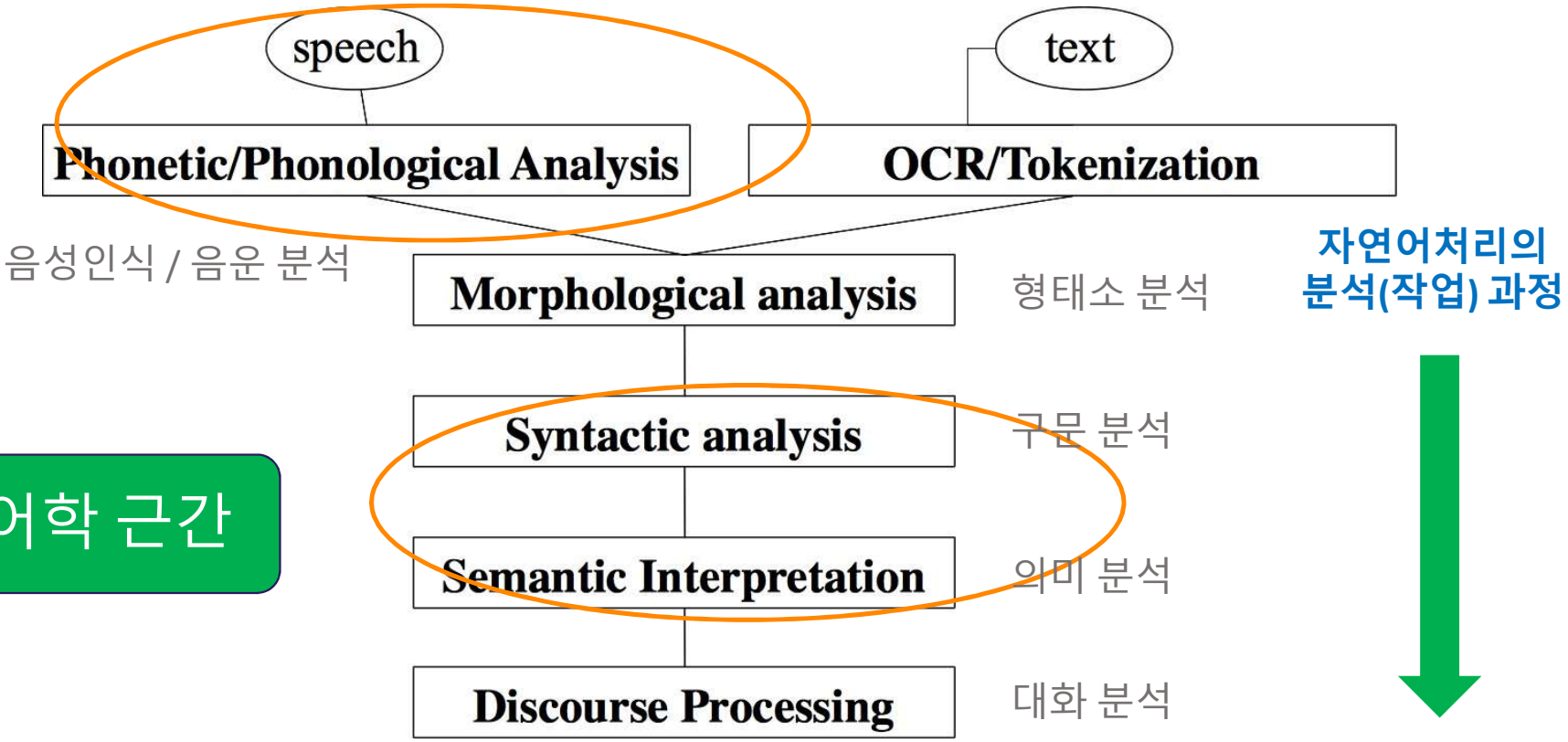
- 컴퓨터가 단어를 잘 분석(이해)할 수 있게 하도록
- 단어를 "표현" 하는 다양한 방법

- Representations of text
- NLP : Natural Language Processing

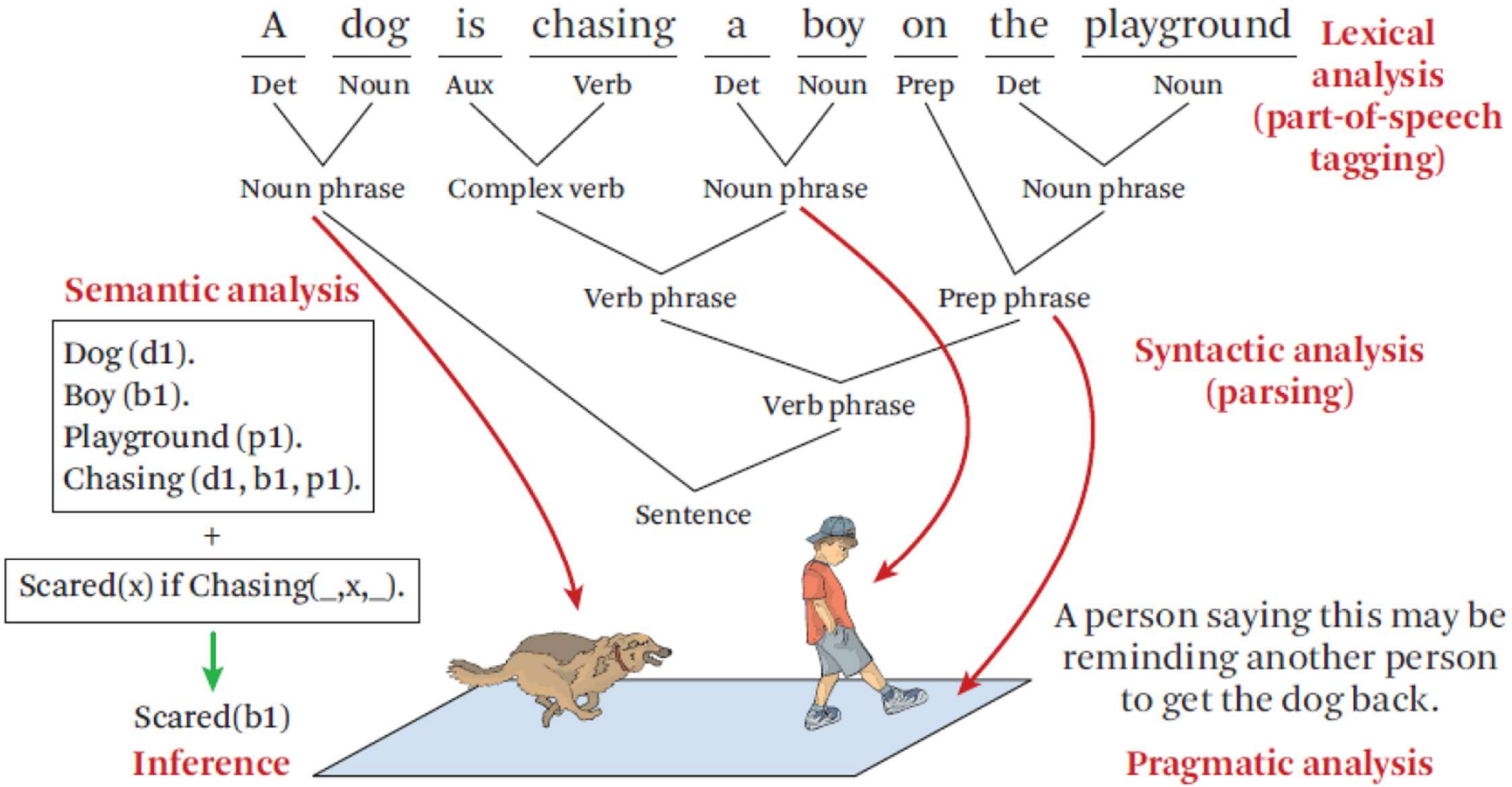
- CS224n: Natural Language Processing with Deep Learning

자연 언어 처리 분석 과정

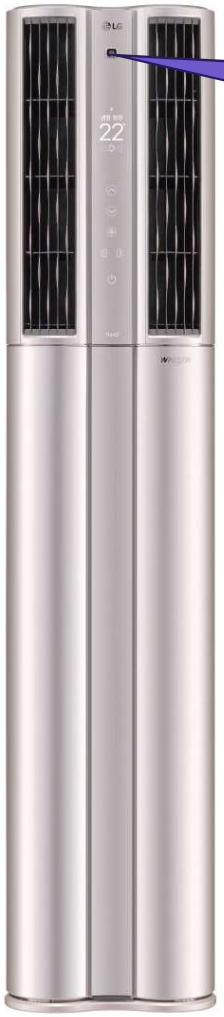
Language Conference 2019



언어학 근간


A dog is chasing a boy on the playground
Det Noun Aux Verb Det Noun Prep Det Noun
Lexical analysis (part-of-speech tagging)
Noun phrase
Complex verb
Noun phrase
Noun phrase
Verb phrase
Prep phrase
Verb phrase
Sentence
Syntactic analysis (parsing)
Semantic analysis
Dog (d1).
Boy (b1).
Playground (p1).
Chasing (d1, b1, p1).
+
Scared(x) if Chasing(_,x,_).
Scared(b1)
Inference
A person saying this may be reminding another person to get the dog back.
Pragmatic analysis (speech act)

자연 언어 분석 과정 예



에어컨을 가동할까요?

너무 더운 거 같지 않니?

몇 단계의 자연어처리 과정이 필요

1. 어떤 단어로 구성 : '너무(부사) 덥-(형용사)+-ㄴ(어미) 거(의존명사) 같-(형용사)+지(어미) 않-(보조용언)+-니(어미)+?(문장부호)'
2. 문장의 주어와 목적어 : '너는' 주어 생략
3. 의미 분석 : '덥'이 '기온이 높다'라는 것을 파악
4. 숨겨진 의도(**추론**) : 긍정 내지 부정 질문이 아닌, **요구(가동)**라는 것을 파악

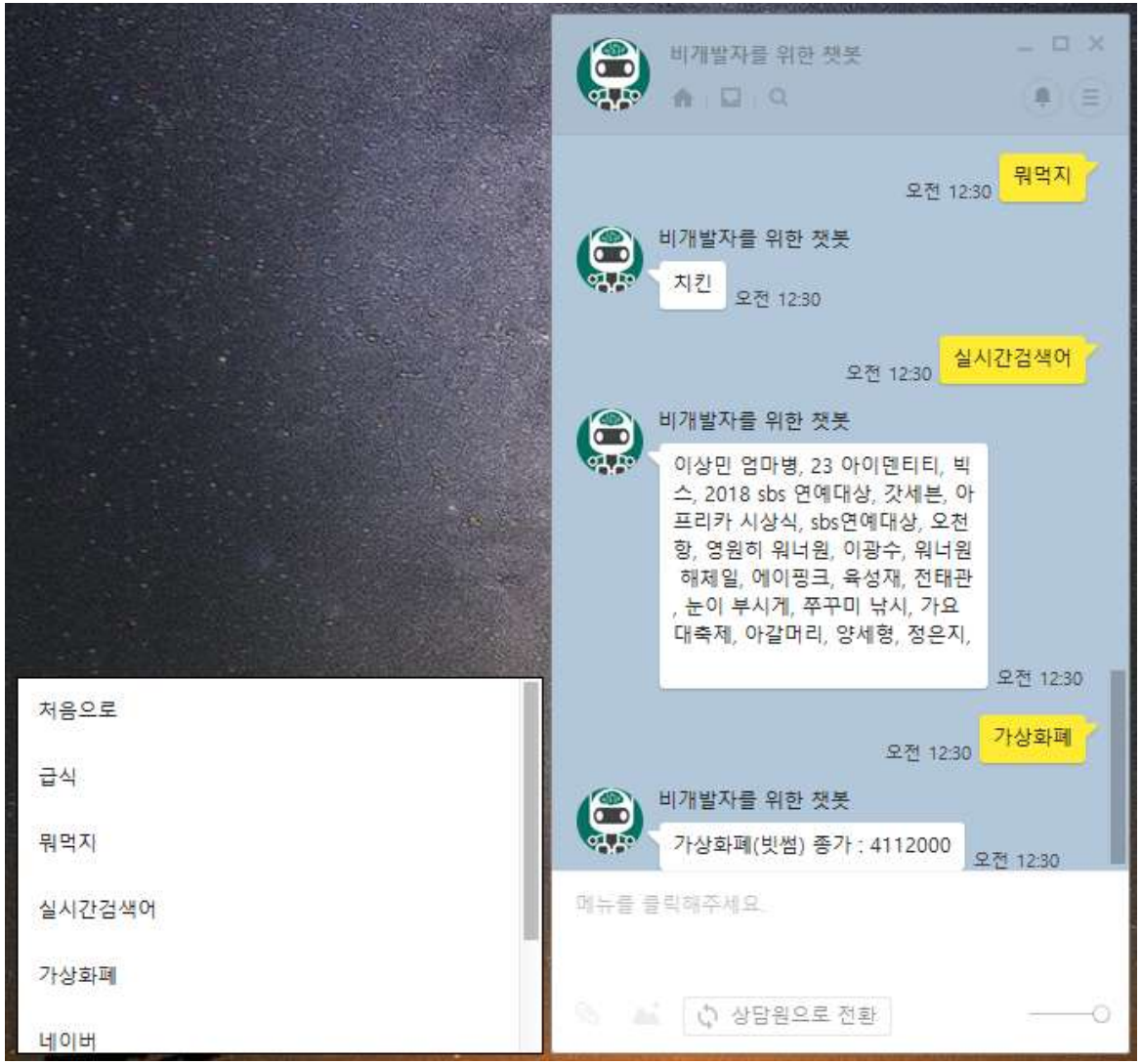
	1분과	2분과	튜토리얼	튜토리얼
1시 ~ 1시 40분	이재석 자연어 처리가 뭔가요? 왜 회사들에서 관심을 가지나요? (초급)	서범석 뉴스를 이용한 주식시장 예측(with kaggle)((python)	이광춘 유튜브 댓글 텍스트 분석(R) 조교: 김건희	박조은 청와대 국민청원 데이터로 파이썬 자연어처리 입문하기(python) 조교 : 박신홍, 김강민
쉬는 시간				
1시 50분 ~ 2시 30분	우종하 사례중심으로 본 감성챗봇의 미래	고동현 문장 속 단어		
쉬는 시간				
2시 40분~ 3시 20분	박혜웅 DIY 챗봇(python)	남내현 · 장정우 너의 목소리가 들려 (광화문 1번가 분석) (python)		

자연 언어 처리 연구 분야

- 음성 인식
- 맞춤법 검사
- 내용 요약
- 번역
- 사용자의 감정 분석
- 텍스트 분류 작업(스팸 메일 분류, 뉴스 기사 카테고리 분류)
- 질의 응답 시스템
- **챗봇**

챗봇 만들기 쉬워요

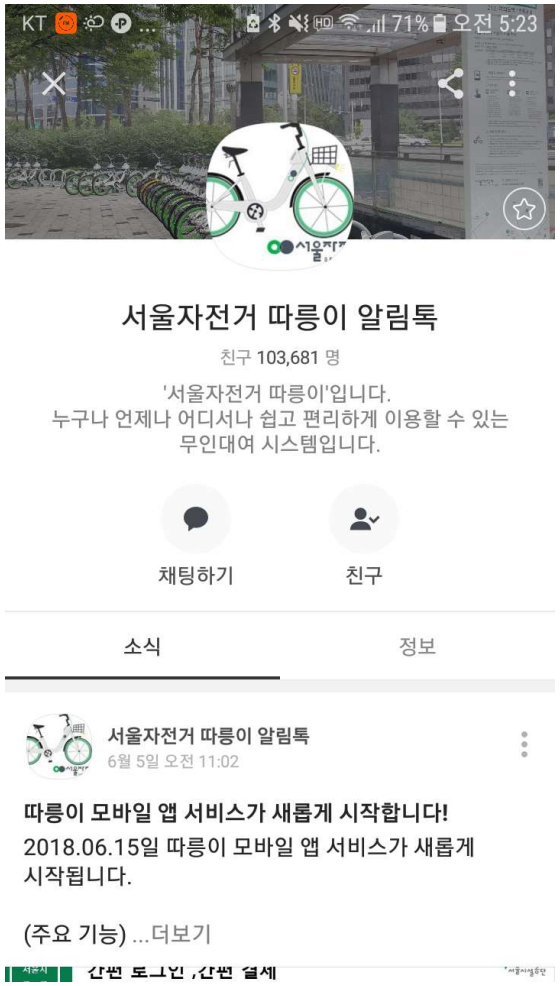
Language Conference 2019

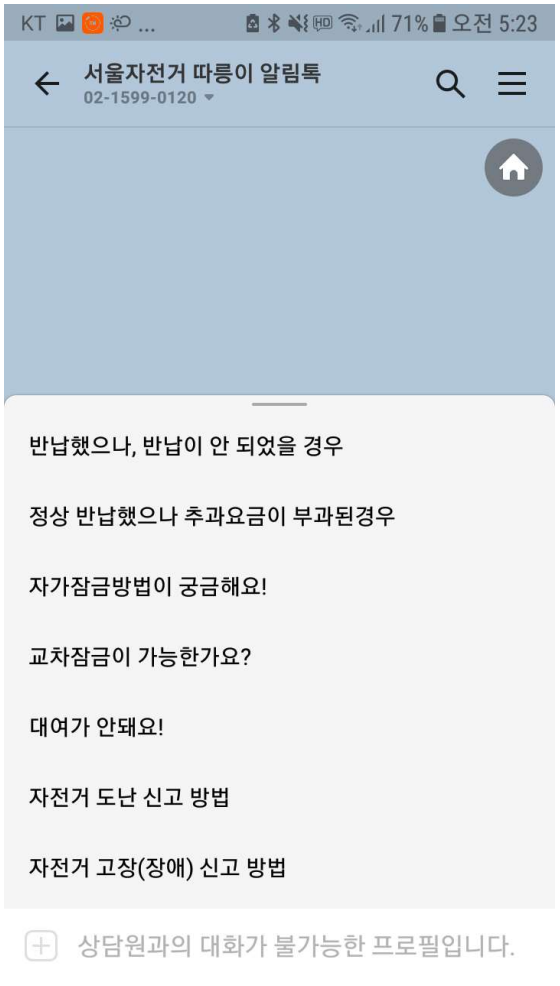


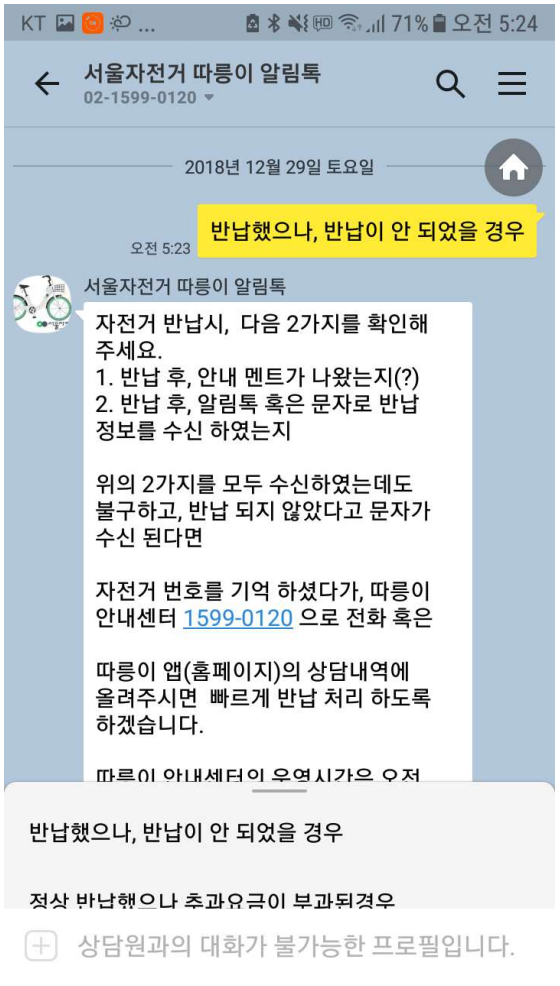
http://bitly.kr/at0zN

서울자전거 따릉이 알림톡

Language Conference 2019







챗봇이란?

"단순히 사람대신, 답변해주는 프로그램"

"사람의 말을 이해하고, 응답하는 가상비서"

"사용자 인터페이스 : 말"

"불고기맛피자는 뭐야?"

"불고기피자는 한국의 맛 불고기에
부드러운 치즈를 넣어 만든 피자입니다!"

인공지능 스피커

Language Conference 2019

" 불고기맛피자 를 주문해줘 ! "
" 불고기피자 가 먹고싶어 ! "

사람의 말을 얼마나 잘 알아듣고
똑똑한 대답을 하느냐가 중요

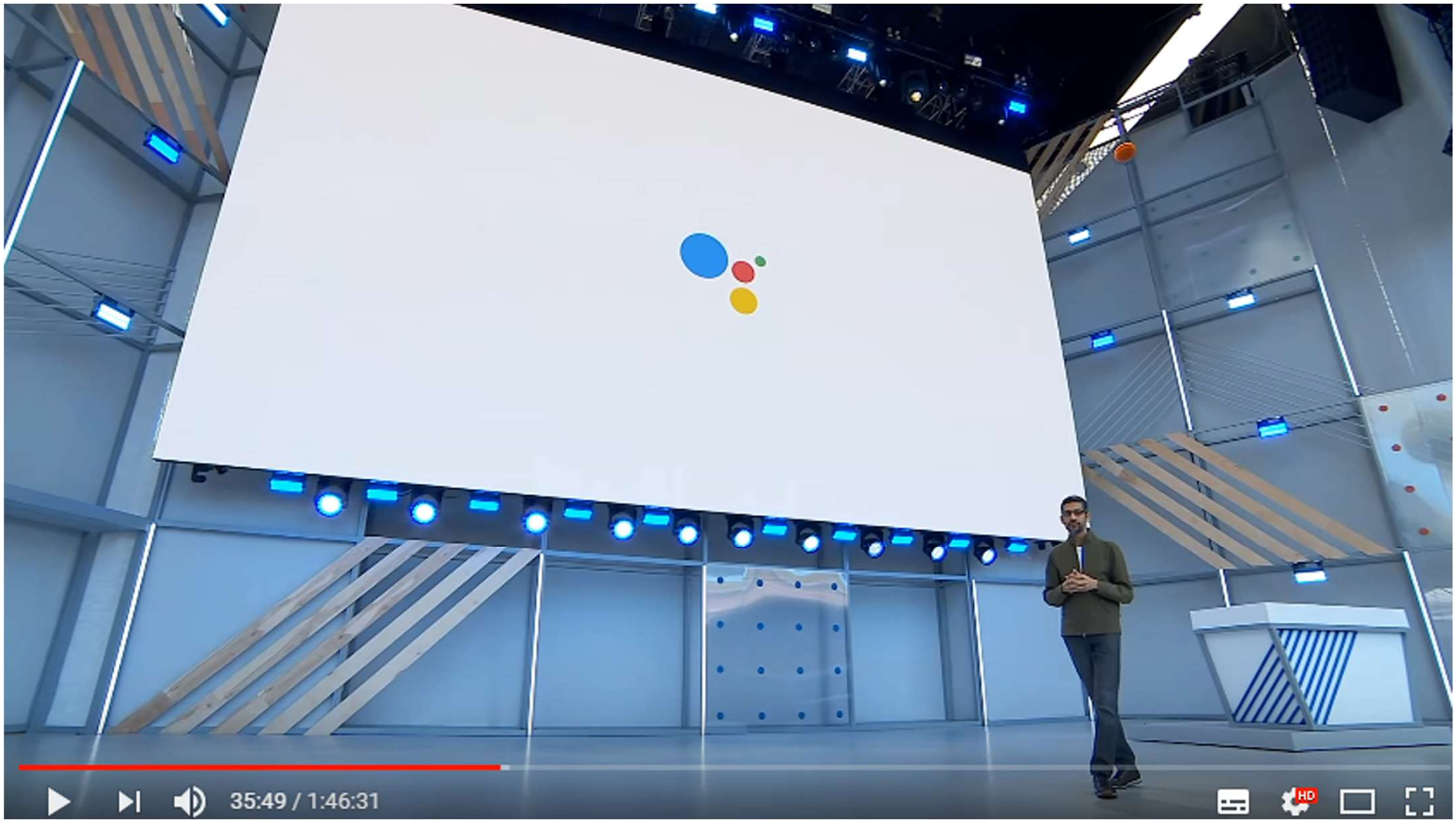
사람도 알아듣기 힘든 말!

Language Conference 2019

문디 가스나야 만다꼬 질질짜면서 방구석에 쌔리 공기가 있노
니가 그카이 가가 그카지 니가 안그카믄 가가 만다꼬 그칼끼고
금마 끌베이 가뜬데 엥가이해라 니속만 디비진다 아이가
우짜든 가네 단디 정리하고 그런너마 재끼뿌고 이자뿌라 영파이다
금마 아이라도 까리하고 혼빵가는 아들 천지빼까리다 고다꼬 솩쌤이 추잡꾸로 그게 뭐꼬
글그치게 스리 내한테 함 자피바라 고마 쎄리 마 똥방디를 콱 추차삘라니깐
아 마 쫌 인자 고마 질질짜라 엥가이 했으니까네

Google Duplex : 가상비서의 결정판!

Language Conference 2019



인공지능 스피커의 현주소

- 음악 재생, 검색
- 날씨 뉴스 확인 – 90%
- 질문하기
- 전화
- 버스 도착 알리미

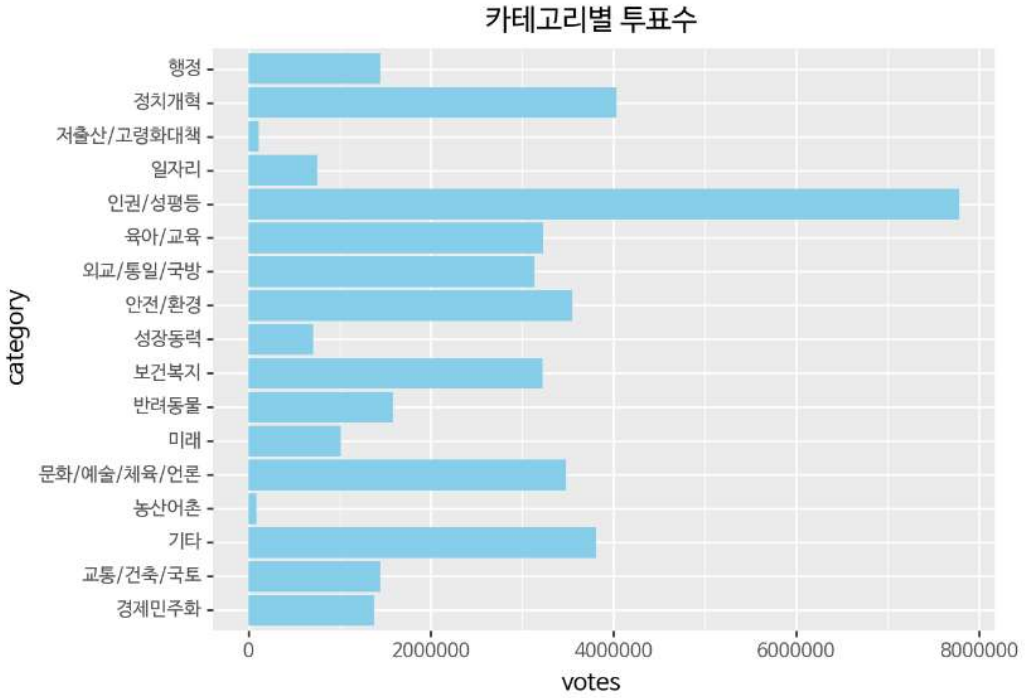
- 동영상 재생(유투브) – 우위 점유!
- 스마트홈 기기 제어(IoT) : (앱 설치 – 불편 – **신기한 쓰레기**) ↔ 음성으로 제어 가능하면?
- 중요한 것은 월등히 더 편한 use case 가 무엇인지?

	1분과	2분과	튜토리얼	튜토리얼
1시 ~ 1시 40분	이재석 자연어 처리가 뭔가요? 왜 회사들에서 관심을 가지나요? (초급)	서범석 뉴스를 이용한 주식시장 예측(with kaggle)((python)	이광춘 유튜브 댓글 텍스트 분석(R) 조교: 김건희	박조은 청와대 국민청원 데이터로 파이썬 자연어처리 입문하기(python) 조교 : 박신홍, 김강민
쉬는 시간				
1시 50분 ~ 2시 30분	우종하 사례중심으로 본 감성챗봇의 미래	고동현 문장 속 단어		
	쉬는 시간			
2시 40분~ 3시 20분	박혜웅 DIY 챗봇(python)	남내현 · 장정우 너의 목소리가 들려 (광화문 1번가 분석) (python)		

	1 분과	2 분과	튜토리얼	튜토리얼
3 시 30 분~ 4 시 10 분	고재선 트랜스퍼 러닝과 텍스트 문서의 분류 (python)	정건용 · 신민철 메세지(기사)와 메신저(기사 전달자)를 활용한 가짜뉴스 판별기 제작기(python)	강창훈 LUIS 머신러닝 자연어처리 기반 챗봇 개발 및 서비스하기 조교 : 김도경	최태균 텐서플로우로 시작하는 텍스트 분류(python) 조교: 김성근
쉬는 시간				
4 시 20 분~ 5 시	심상진 한국어 의존성 분석 이론 및 동향	홍지민 · 곽현석 인터넷 방송 크롤링을 통한 방송 하이라이트 예측(python)		
쉬는 시간				
5 시 10 분~ 5 시 50 분	김준혁 R 에서 텍스트 분석과 RcppMeCab(R)	김준민 리뷰에는 이미 별점이 있는데, 또 별점을 학습해서 뭘 하나요? (python)		

텍스트 분석 및 시각화(1)

Language Conference 2019



텍스트 분석 및 시각화(2) - 머신러닝

- 사람들이 무슨 생각으로 살아가는지 알아보는 방법들?
- 여론조사 : 설문지, 전화, 직접 방문(but, 미리 정답에 가까운 질문들 작성, 시간과 돈 제약)
- 구매 기록 : 싫어하는 것, 좋아하는 것 (but, 왜 샀는지?)

- 텍스트 분석 - sns 글 작성, 생각이나 느낌, 감정분석
- 글쓰는 사람만(특이한 사람만!)
- 답이 없을 수 있음

텍스트 분석 사용예

Language Conference 2019

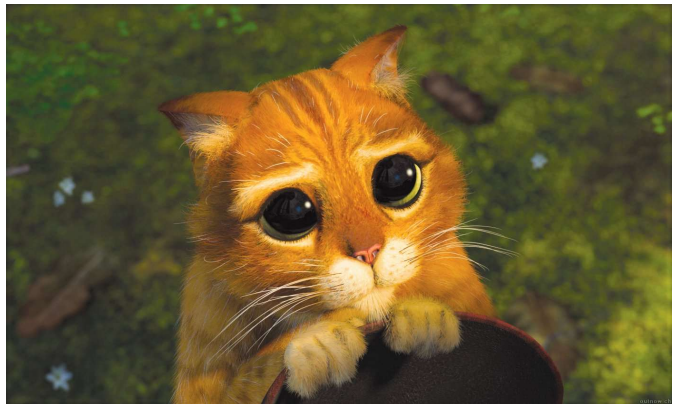
수 많은 데이터 속에서 가치를 찾아내는 텍스트 분석을 원하시나요?

	1분과	2분과	튜토리얼	튜토리얼
1시 ~ 1시 40분	이재석 자연어 처리가 뭔가요? 왜 회사들에서 관심을 가지나요? (초급)	서범석 뉴스를 이용한 주식시장 예측(with kaggle)((python)	이광춘 유튜브 댓글 텍스트 분석(R) 조교: 김건희	박조은 청와대 국민청원 데이터로 파이썬 자연어처리 입문하기(python) 조교 : 박신홍, 김강민
쉬는 시간				
1시 50분 ~ 2시 30분	우종하 사례중심으로 본 감성챗봇의 미래	고동현 문장 속 단어		
쉬는 시간				
2시 40분~ 3시 20분	박혜웅 DIY 챗봇(python)	남내현 · 장정우 너의 목소리가 들려 (광화문1번가 분석) (python)		

	1 분과	2 분과	튜토리얼	튜토리얼
3 시 30 분~ 4 시 10 분	고재선 트랜스퍼 러닝과 텍스트 문서의 분류 (python)	정건용 · 신민철 메세지(기사)와 메신저(기사 전달자)를 활용한 가짜뉴스 판별기 제작기(python)	강창훈 LUIS 머신러닝 자연어처리 기반 챗봇 개발 및 서비스하기 조교 : 김도경	최태균 텐서플로우로 시작하는 텍스트 분류(python) 조교: 김성근
쉬는 시간				
4 시 20 분~ 5 시	심상진 한국어 의존성 분석 이론 및 동향	홍지민 · 곽현석 인터넷 방송 크롤링을 통한 방송 하이라이트 예측(python)		
쉬는 시간				
5 시 10 분~ 5 시 50 분	김준혁 R 에서 텍스트 분석과 RcppMeCab(R)	김준민 리뷰에는 이미 별점이 있는데, 또 별점을 학습해서 뭘 하나요? (python)		

왜 잘 만든 Word Embeddings 필요한가?

Language Conference 2019

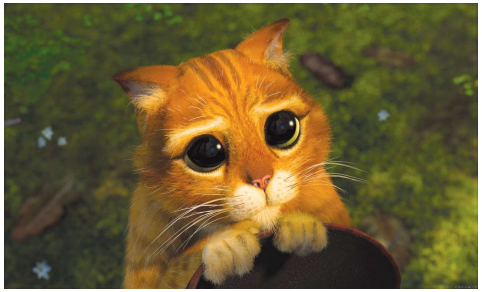


고양이
야옹이

- 단어를 "표현" 하는 다양한 방법 : "고양이" 어떻게 표현할까요?
- 전통적 방법 : 고양이 = Index[537], 야옹이 = Index[654]

단어를 "표현" 하는 다양한 방법

Language Conference 2019



- 고양이 = Index[1]
- 고양이 = [0, 1, 0, 0, 0]
- 고양이 = [0.0, 1.0, 0.3, 0.2, 0.9]

- 피자 = Index[4]
- 피자 = [0, 0, 0, 1, 0]
- 피자 = [0.5, 0.0, 1.0, 1.0, 0.0]

자연 언어 처리 기술 수준

Language Conference 2019

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

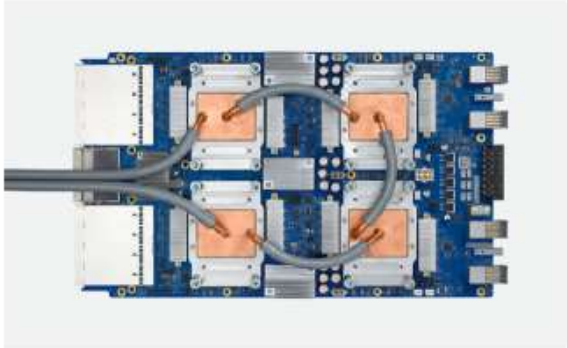
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Dec 13, 2018	BERT finetune baseline (ensemble) *Anonymous*	**83.536**	**86.096**
2 Nov 25, 2018	PAML+BERT (ensemble) *PINGAN GammaLab*	83.435	85.992
3 Dec 16, 2018	PAML+BERT-single (single model) *PINGAN GammaLab*	82.577	85.603
4 Nov 16, 2018	AoA + DA + BERT (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	82.374	85.310
5 Dec 12, 2018	BERT finetune baseline (single model) *Anonymous*	82.126	84.820
5 Dec 10, 2018	Candi-Net+BERT (ensemble) *42Maru NLP Team*	82.126	84.624

BERT

https://rajpurkar.github.io/SQuAD-explorer/

자연 언어 처리 기술 수준

Language Conference 2019



TPU 3.0

4일 소요(16개)

Tesla V100

5 1/3일 소요(64개)

약 42일 소요(8개)

RTX 2080 Ti

8 1/2일 소요(64개)

약 68일 소요(8개)

- 구글 클라우드를 통해 TPU v2와 v3(베타)의 사용이 가능하며,
- 최초 1계정 당 $300 무료 크래딧 제공
- 1대 1시간 당 각각 $4.5, $8.0가 청구됨(TPU 3.0 16개 4일 - **약 1,400만원**)

그런데, 처리할 언어가?

한글이라면

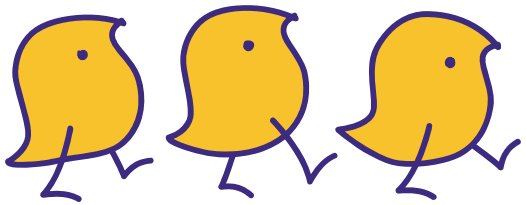
한글, 왜 이렇게 어려운가요?

<확률에 기반한 언어 모델링 관점에서 …>

- 교착어 - heljosa
- 언어별 특성 - 굉장히 자유로운 어순, 주어 생략 가능
- 잘못된 띄어쓰기 – 아버지 가방에 들어가신다(의미적 중의성)

- 모호한 문장 – 거만한 친구의 남편
- 복합명사
- 긴명사 – 니코틴아마이드 아데닌 다이뉴클레오타이드가 뭐야?
- 짧은 순간 발생했다 사라지는 비속어와 유행어(신조어) – 엌츜

	1 분과	2 분과	튜토리얼	튜토리얼
3 시 30 분~ 4 시 10 분	고재선 트랜스퍼 러닝과 텍스트 문서의 분류 (python)	정건용 · 신민철 메세지(기사)와 메신저(기사 전달자)를 활용한 가짜뉴스 판별기 제작기(python)	강창훈 LUIS 머신러닝 자연어처리 기반 챗봇 개발 및 서비스하기 조교 : 김도경	최태균 텐서플로우로 시작하는 텍스트 분류(python) 조교: 김성근
쉬는 시간				
4 시 20 분~ 5 시	심상진 한국어 의존성 분석 이론 및 동향	홍지민 · 곽현석 인터넷 방송 크롤링을 통한 방송 하이라이트 예측(python)		
쉬는 시간				
5 시 10 분~ 5 시 50 분	김준혁 R 에서 텍스트 분석과 RcppMeCab(R)	김준민 리뷰에는 이미 별점이 있는데, 또 별점을 학습해서 뭘 하나요? (python)		

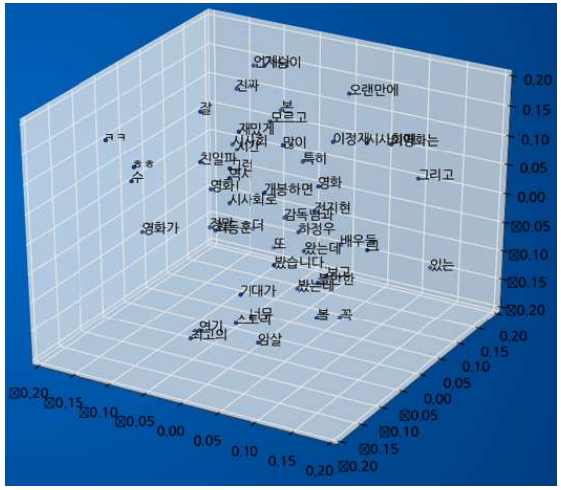


3. 자연어처리 기법

LangCon 2019

자연어처리 기법 - 두가지 방식

- Bag-of-words : 단어 주머니(카운트 기반)
- Word embedding : 워드 임베딩(벡터 기반)



Bag-of-words

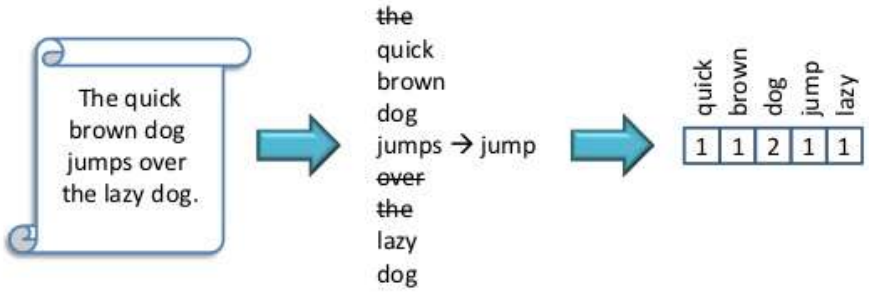
Language Conference 2019

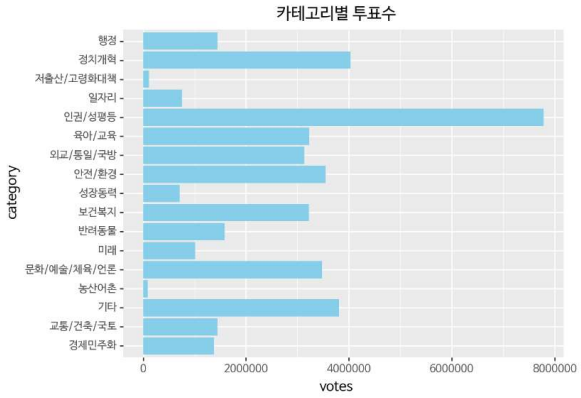
PART I • PART II

INTRO • BOW • ML • EVAL • APP

Bags of words

• Tokenize • Remove stop words • Lemmatize • Compute weights





TF-IDF

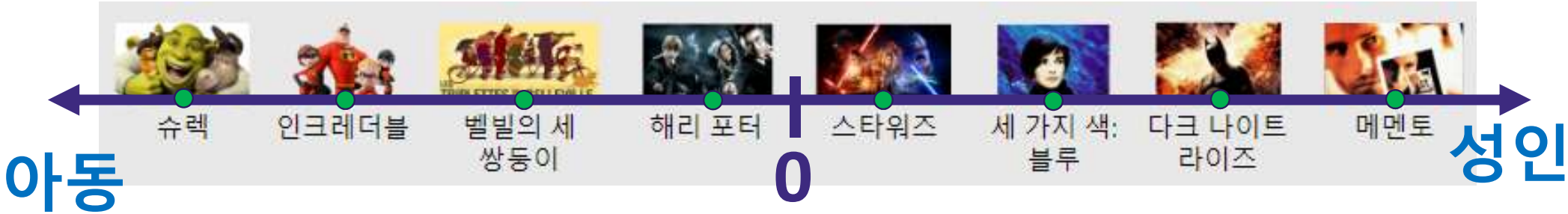
- term frequency–inverse document frequency
- TF : 문서 내 단어 개수 / 문서 내 모든 단어 수
- IDF : log(문서 전체의 수 / 단어를 포함한 문서의 수)
- TF-IDF : TF×IDF

TF-IDF

- term frequency–inverse document frequency
- TF : 문서 내 단어 개수 / 문서 내 모든 단어 수
- IDF : log(문서 전체의 수 / 단어를 포함한 문서의 수)
- TF-IDF : TF×IDF

임베딩(embedding)

- 단어를 벡터화하는 방법론
- Word2Vec, Glove, FastText
- 정보를 보존하면서, 단어 벡터를 만듬
- 의미론 : 분포 의미(의미 유사성)
- 하나의 단어를 벡터 공간상의 하나의 점으로 맵핑해주는 기법



이산 표현

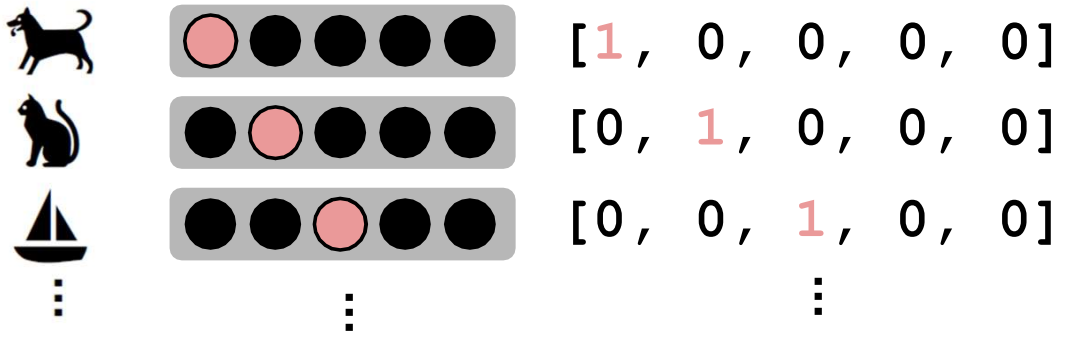
- 단어를 표현하는 가장 쉬운 방법은? 이산 표현 방법

$$w^{가} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, w^{가가(假家)} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, w^{가가대소} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \ldots w^{행자} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

- 단어에 대한 벡터 표현
- one-hot-vector : 고양이 = [0, 0, 0, ---, 1, ---,0]
- 전체 사전의 크기 |V|

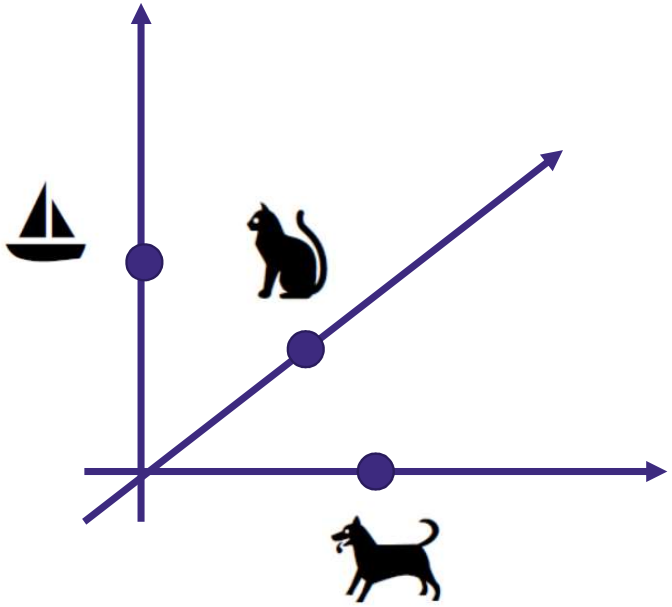
단어의 이산표현

1개 뉴론의 작동으로 1개의 개념을 나타냄



벡터형태로 나타내서 **one-hot vector**

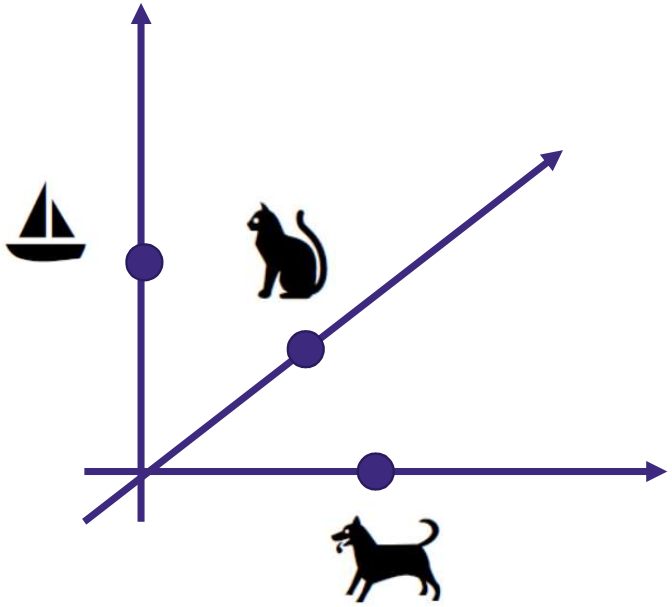
이산 표현 유사도



단어	원핫인코딩
개	[1, 0, 0]
고양이	[0, 1, 0]
배	[0, 0, 1]

- 단어의 유사성이 없다(45도).
- 이것만 가지고는 단어의 의미를 전혀 알 수 없다.
- 단어의 의미를 파악하는 벡터를 갖고 싶다

이산 표현 유사도

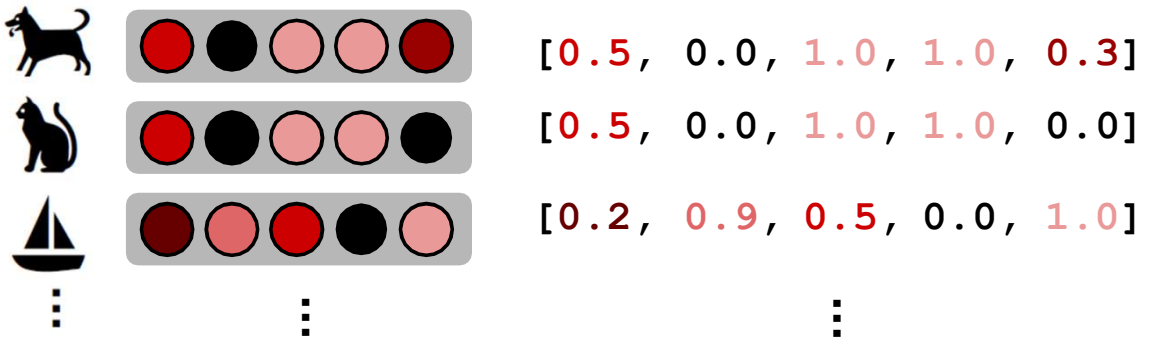


단어	원핫인코딩
개	[1, 0, 0]
고양이	[0, 1, 0]
배	[0, 0, 1]

- 단어의 유사성이 없다(45도).
- 이것만 가지고는 단어의 의미를 전혀 알 수 없다.
- 단어의 의미를 파악하는 벡터를 갖고 싶다

분산 표현

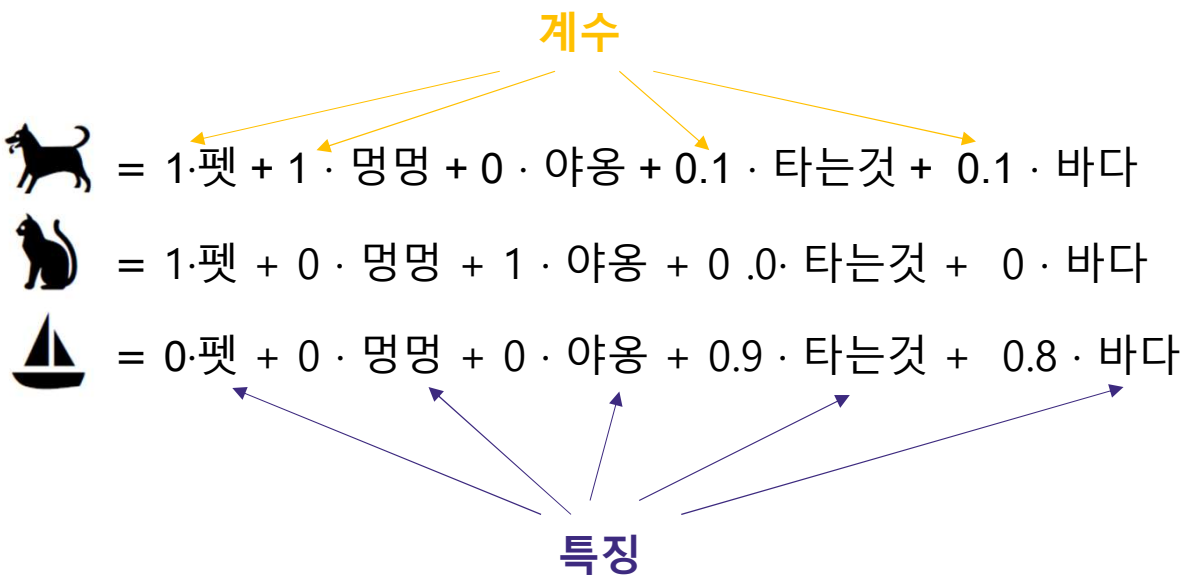
복수 뉴론의 작동으로 1개의 개념을 나타냄.



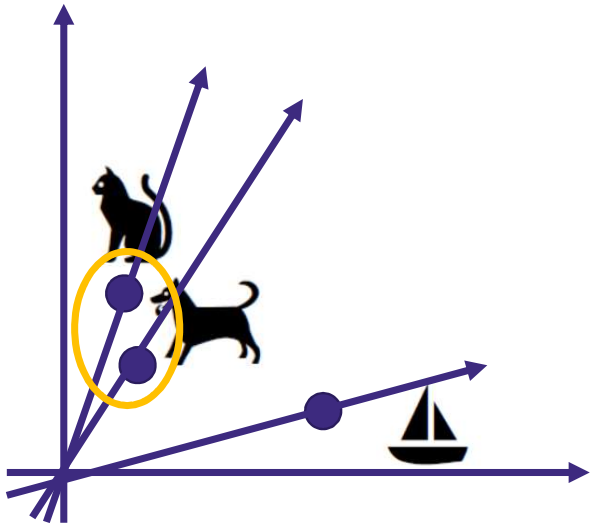
힌톤 교수(1986년) : 뉴런이 어떻게 개념을 나타내는가를 설명하기 위해 분산 표현 사용

분산 표현

개념을 **특징의 조합**으로 나타냄.

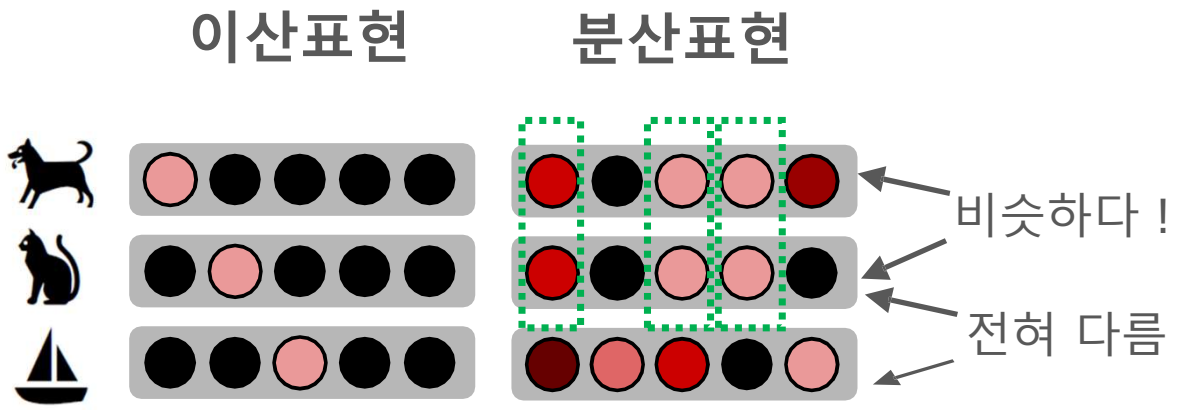


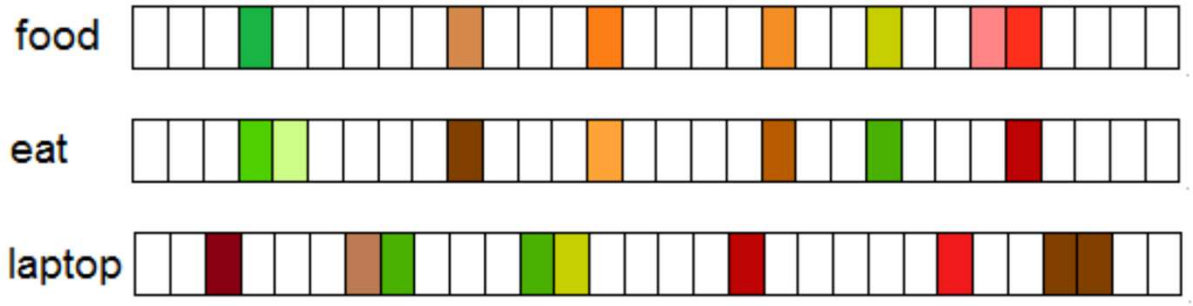
분산 표현 유사도



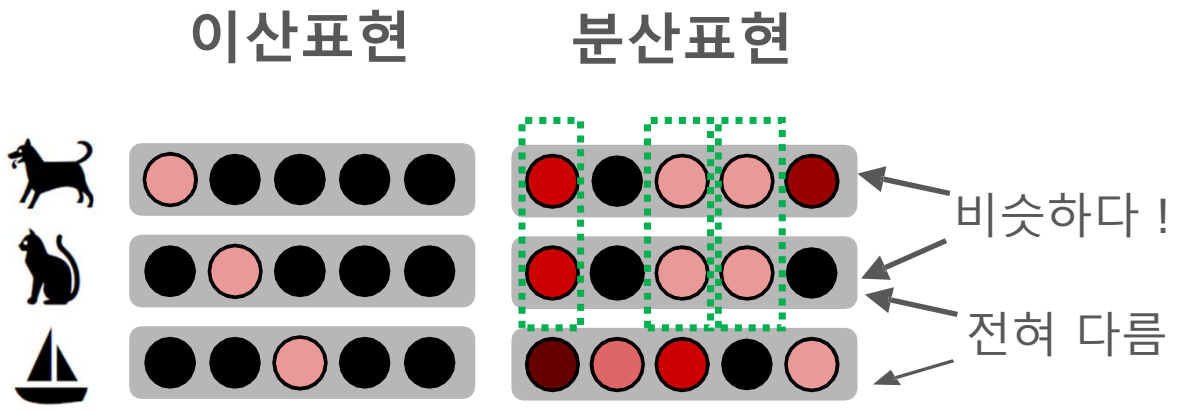
단어	원핫인코딩	임베딩
개	[1, 0, 0]	[1,2]
고양이	[0, 1, 0]	[1,3]
배	[0, 0, 1]	[5,1]

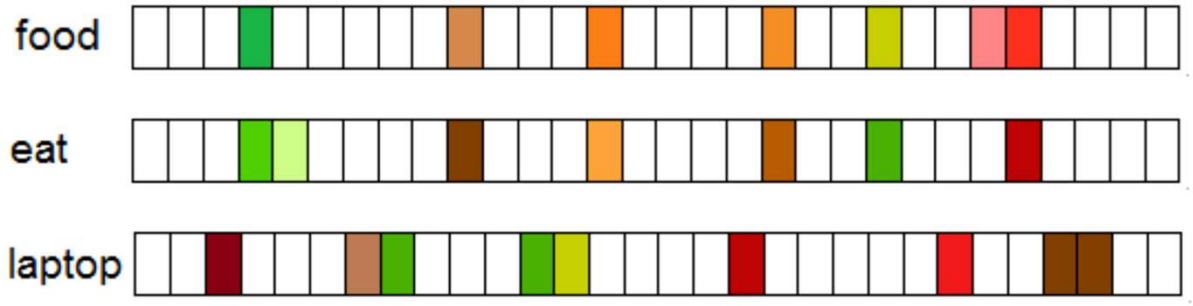
개념의 유사





개념의 유사





임베딩 – 영화 추천

- 100만명 사용자, 50만편의 영화 중 각 사용자가 본 영화 목록

- 어떤 영화가 서로 비슷한지 파악하는 방법이 필요
- 이를 위해 비슷한 영화가 서로 인접하도록 만들어진 저차원 공간에 영화를 임베딩할 수 있다.

- 임베딩 학습을 위해, 학습 데이터를 표현하는 방법

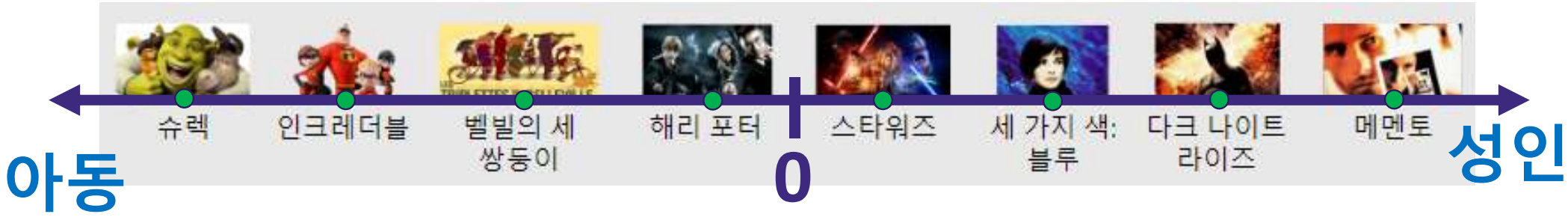
임베딩 – 영화 추천

- 100만명 사용자, 50만편의 영화 중 각 사용자가 본 영화 목록

- 어떤 영화가 서로 비슷한지 파악하는 방법이 필요
- 이를 위해 비슷한 영화가 서로 인접하도록 만들어진 저차원 공간에 영화를 임베딩할 수 있다.

- 임베딩 학습을 위해, 학습 데이터를 표현하는 방법

임베딩 – 1차원 수직선에 영화 정렬



- 1차원 수직선에 배열하여
- 서로 가장 가까운 관련성이 있는 영화가 서로 가장 가까이에 있도록
- 영화가 아동과 성인 중 어느 쪽을 대상으로 하는지 파악

임베딩 – 2차원 공간에서 영화 정렬

Language Conference 2019

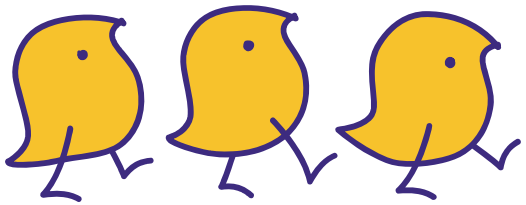
- 2차원 임베딩을 통해 (2가지 특성)
- **아동과 성인** 쪽 어느 쪽을 대상으로 하는지 여부와 **블록버스터 영화와 예술 영화** 중에서

- 어느 쪽에 가까운지 여부에 따라 서로 가깝고, 따라서 비슷한 것으로 추론되는 영화 간의 거리를 정의

- 결국 중요한 것은 **주어진 차원**에서
- 특정 영화의 값 이라기보다는
- 임베딩 공간에서 **영화들 간의 거리값**
- **숫자 벡터**로 표현

임베딩 – 2차원 공간에서 영화 정렬

	1 분과	2 분과	튜토리얼	튜토리얼
3 시 30 분~ 4 시 10 분	고재선 트랜스퍼 러닝과 텍스트 문서의 분류 (python)	정건용 · 신민철 메세지(기사)와 메신저(기사 전달자)를 활용한 가짜뉴스 판별기 제작기(python)	강창훈 LUIS 머신러닝 자연어처리 기반 챗봇 개발 및 서비스하기 조교 : 김도경	최태균 텐서플로우로 시작하는 텍스트 분류(python) 조교: 김성근
쉬는 시간				
4 시 20 분~ 5 시	심상진 한국어 의존성 분석 이론 및 동향	홍지민 · 곽현석 인터넷 방송 크롤링을 통한 방송 하이라이트 예측(python)		
쉬는 시간				
5 시 10 분~ 5 시 50 분	김준혁 R 에서 텍스트 분석과 RcppMeCab(R)	김준민 리뷰에는 이미 별점이 있는데, 또 별점을 학습해서 뭘 하나요? (python)		



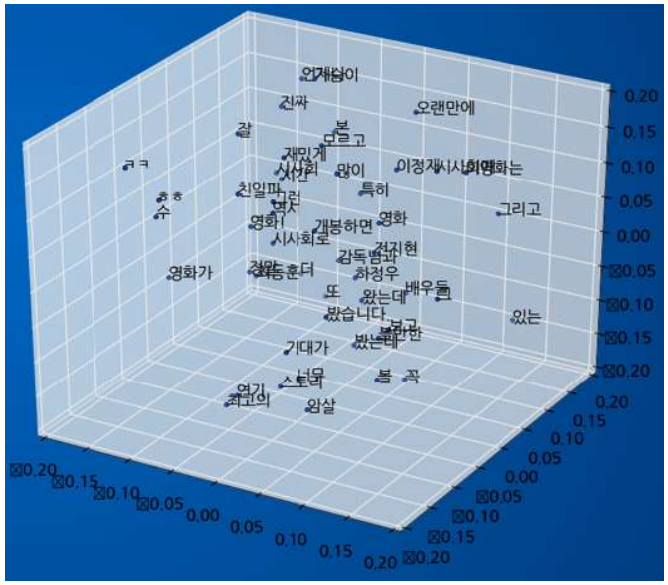
4. Word2Vec

LangCon 2019

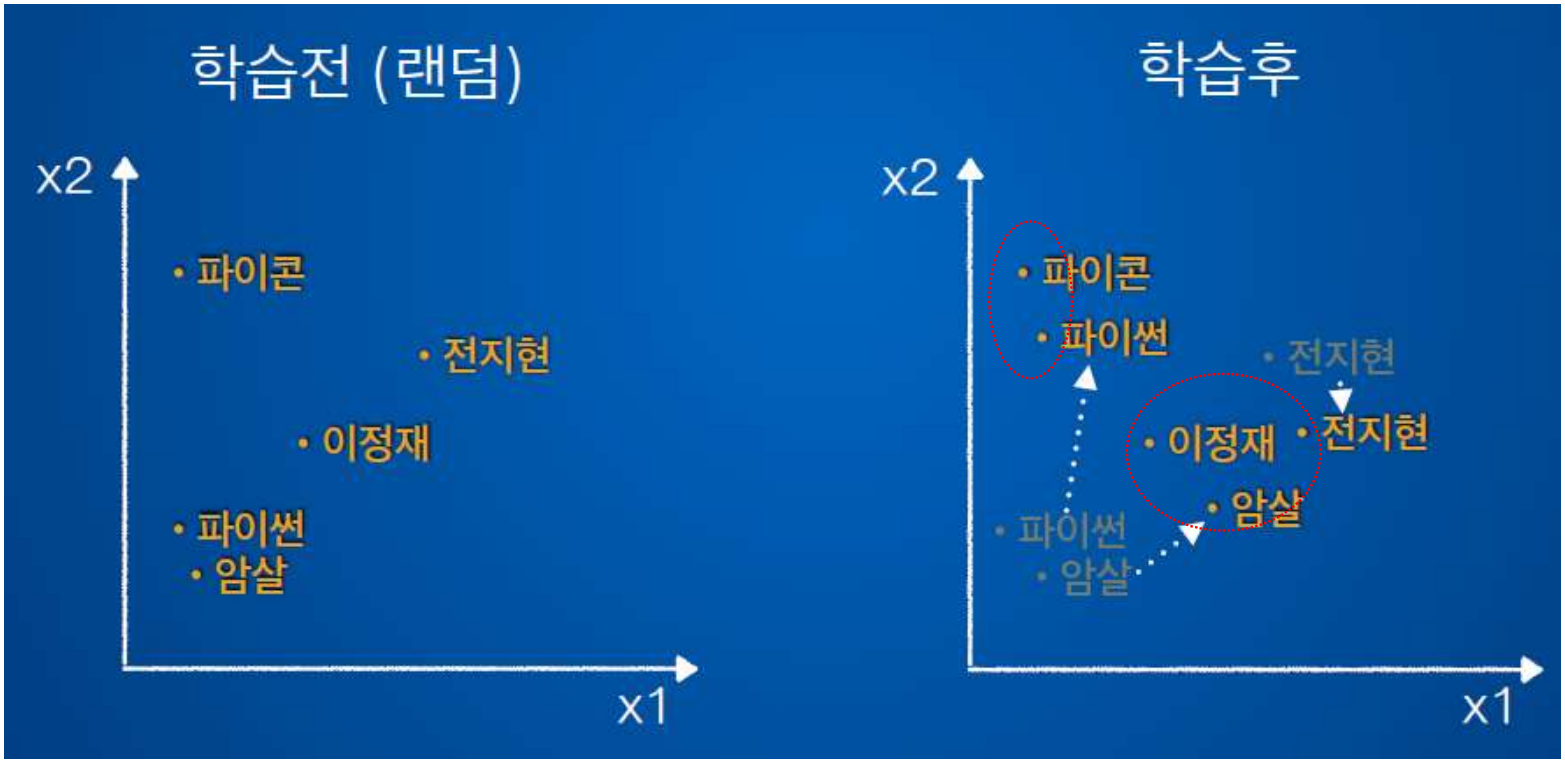
Word2Vec 이란

Language Conference 2019

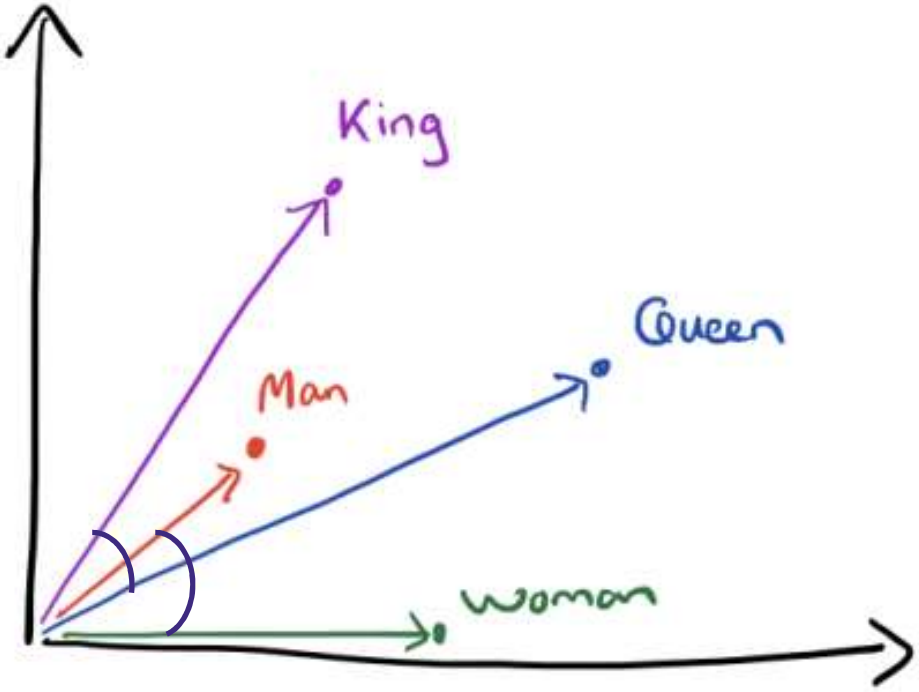
- **단어**를 **벡터**로 **바꿔주는** 알고리즘
- 2013년 구글에서 미콜로프 발표
- 2003년 bengio의 NNLM(Neural Network Language Model) 기반
- 자연어 처리 딥러닝(히든레이어 1개)
- Word를 100 ~ 300차원의 한정(차원이 적은)된 Vector로 효율적으로 표현.
- 학습 속도와 성능을 비약적으로 발전
- 단어의 **의미**를 알아내는 작업
- Word Embedding : Word2Vec, Glove



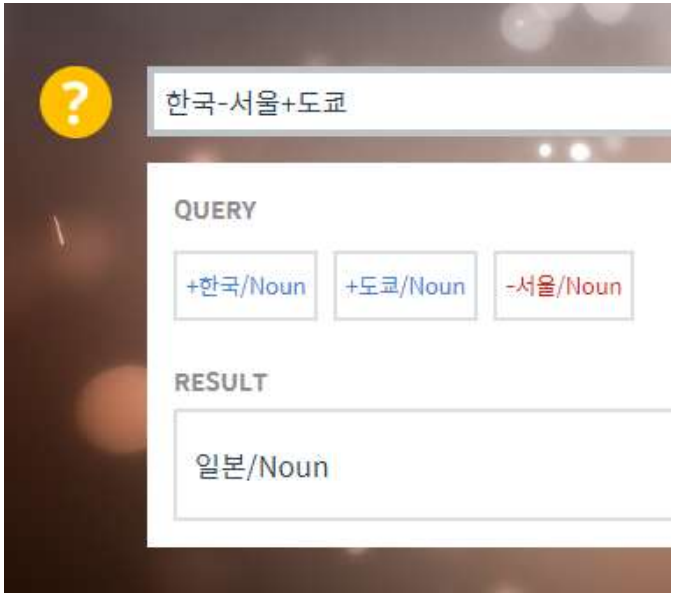
Word2Vec 으로 학습하면



Word2Vec 활용 예



Word Vectors



http://w.elnn.kr/

감사합니다

이제 자연어처리에 관심갖고 하나씩 즐겨봅시다

SEnE 이재석