

CNU 데이터 분석 교육



14th lecture
"Movie review analysis"

2022 - 11 - 30

지난 시간?

1. Convolution Neural Network

오늘은 무엇을?

1. 네이버 영화 리뷰를 학습하여 긍정/부정 리뷰 판별하기
 - NLP(Natural Language Processing), 자연어처리
 - 언어를 인식 및 처리할 수 있는 분석 방법.

텍스트 마이닝?

텍스트 마이닝?

1. 텍스트 데이터 (비정형)에서 가치와 의미가 있는 정보를 찾아내는(Mining) 기술
2. 언어는 그 표현의 형태가 매우 다양하고 복잡하여 일괄된 규칙으로 규정하기 힘들고, 끊임없이 변화한다
3. 이를 자연언어처리 (Natural Language Processing) 기술로 해결한다.

1. 문서분류

000 총류

- 010 도서학, 서지학
- 020 문헌정보학
- 030 백과사전
- 040 강연집, 수필집, 연설문집
- 050 일반연속간행물
- 060 일반학회, 단체, 협회, 기관
- 070 신문, 언론, 저널리즘
- 080 일반전집, 총서
- 090 향토자료

100 철학

- 110 형이상학
- 120 인식론, 인과론, 인간학
- 130 철학의 체계
- 140 경 학
- 150 동양철학, 사상
- 160 서양철학
- 170 논 리 학
- 180 심 리 학
- 190 윤리학, 도덕철학

200 종교

- 210 비교종교
- 220 불 교
- 230 기 독 교
- 240 도 교
- 250 천 도 교
- 260 신 도
- 270 힌두교, 브라만교
- 280 이슬람교(회교)
- 290 기타 제종교

300 사회과학

- 310 통 계 학
- 320 경 제 학
- 330 사회학, 사회문제
- 340 정 치 학
- 350 행 정 학
- 360 법 학
- 370 교 육 학
- 380 풍속, 예절, 민속학
- 390 국방, 군사학

400 자연과학

- 410 수 학
- 420 물 리 학
- 430 화 학
- 440 천 문 학
- 450 지 학
- 460 광 물 학
- 470 생명과학
- 480 식 물 학
- 490 동 물 학

500 기술과학

- 510 의 학
- 520 농업, 농학
- 530 공학, 공업일반, 토목공학, 환경공학
- 540 건축공학
- 550 기계공학
- 560 전기공학, 전자공학
- 570 화학공학
- 580 제 조 업
- 590 생활과학

600 예술

- 610 건 축 물
- 620 조각, 조형예술
- 630 공예, 장식미술
- 640 서 예
- 650 회화, 도화
- 660 사진예술
- 670 음 악
- 680 공연예술, 매체예술
- 690 오락, 스포츠

700 언어

- 710 한 국 어
- 720 중 국 어
- 730 일본어, 기타아시아제어
- 740 영 어
- 750 독 일 어
- 760 프랑스어
- 770 스페인어, 포르투갈어
- 780 이탈리아어
- 790 기타제어

800 문학

- 810 한국문학
- 820 중국문학
- 830 일본문학, 기타아시아문학
- 840 영미문학
- 850 독일문학
- 860 프랑스문학
- 870 스페인, 포르투갈문학
- 880 이탈리아문학
- 890 기타제문학

900 역사

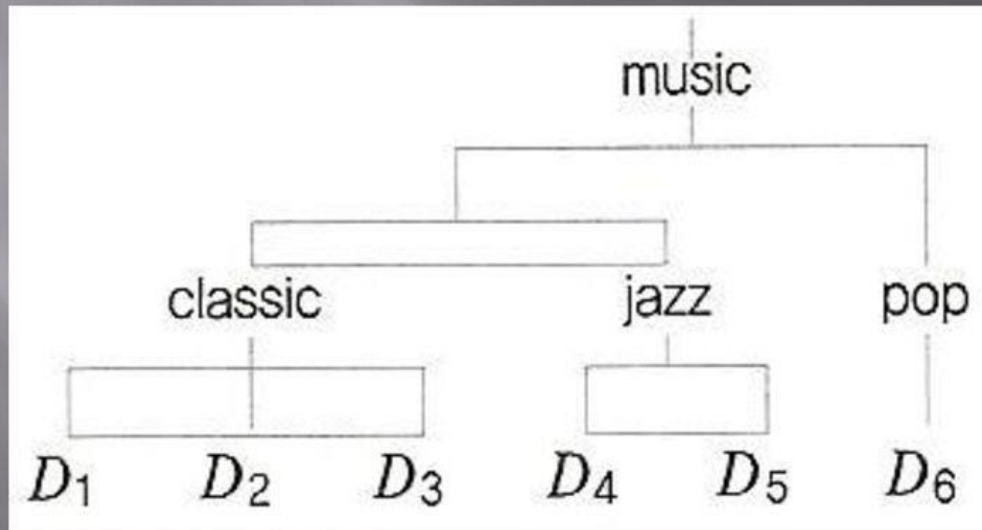
- 910 아 시 아
- 920 유 럽
- 930 아프리카
- 940 북아메리카
- 950 남아메리카
- 960 오세아니아
- 970 양극지방
- 980 지 리
- 990 전 기

2. 문서군집

$$similarity = \cos(\Theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

문서 클러스터링 알고리즘 (계층적 클러스터링)

- 트리 구조를 형성하여 클러스터링을 하는 모델
그림은 D1~D6문서들을 계층적 클러스터링한 예.



3. 정보추출

수학을 모른다면 인생을 살아가는데 아무런 지장이 없지만, 수학을 아는 것은 인생에 있어서 무한한 재미를 가져다 줄 수 있다. 수학은 본질을 자연의 본질을 파악하기 위한 하나의 언어이며, 수학을 잘 하면 머리가 잘 돌아간다. 학창 시절 수학을 잘하는 학생들은 못하는 사람들에게겐 동경의 대상이 되기도 하였으며, 가끔 수학을 모르는 학생이 잘하는 이성친구에게 다가가 물어보다가 커플이 되는 경우도 있다. 이렇듯 수학은 인생에 재미를 더하며, 인생을 풍요롭게 만들지만, 수학을 어려워하는 자에게는 한 없는 고통이 될 수도 있다. 수학은 세상을 바꾸는데 있어서 필수적인 학문이며, 완벽한 논리를 추구하는 수학의 성질을 통해 자연의 숨어있는 규칙과 우주의 신비를 알아갈 수 있다. 하지만 수학은 넘모 힘들고 어렵기 때문에 잘하기가 쉽지 않다. 수학 잘하는 사람들 진짜 신기해 천재야 천재 수학은 크게 위상수학, 대수학, 해석학으로 나눌 수 있으며 현 시대에는 분야를 가리지 않고 융복합적인 연구가 활발하게 진행되고 있다.

3. 정보추출

수학을 모른다면 인생을 살아가는데 아무런 지장이 없지만, 수학을 아는 것은 인생에 있어서 무한한 재미를 가져다 줄 수 있다. **수학은 본질을 자연의 본질을 파악하기 위한 하나의 언어이며**, 수학을 잘 하면 머리가 잘 돌아간다. 학창 시절 수학을 잘하는 학생들은 못하는 사람들에게겐 동경의 대상이 되기도 하였으며, 가끔 수학을 모르는 학생이 잘하는 이성친구에게 다가가 물어보다가 커플이 되는 경우도 있다. 이렇듯 수학은 인생에 재미를 더하며, 인생을 풍요롭게 만들지만, 수학을 어려워하는 자에게는 한 없는 고통이 될 수도 있다. 수학은 세상을 바꾸는데 있어서 필수적인 학문이며, 완벽한 논리를 추구하는 **수학의 성질을 통해 자연의 숨어있는 규칙과 우주의 신비를 알아갈 수 있다.** 하지만 수학은 넘모 힘들고 어렵기 때문에 잘하기가 쉽지 않다. 수학 잘하는 사람들 진짜 신기해 천재야 천재 수학은 크게 위상수학, 대수학, 해석학으로 나눌 수 있으며 현 시대에는 분야를 가리지 않고 융복합적인 연구가 활발하게 진행되고 있다.

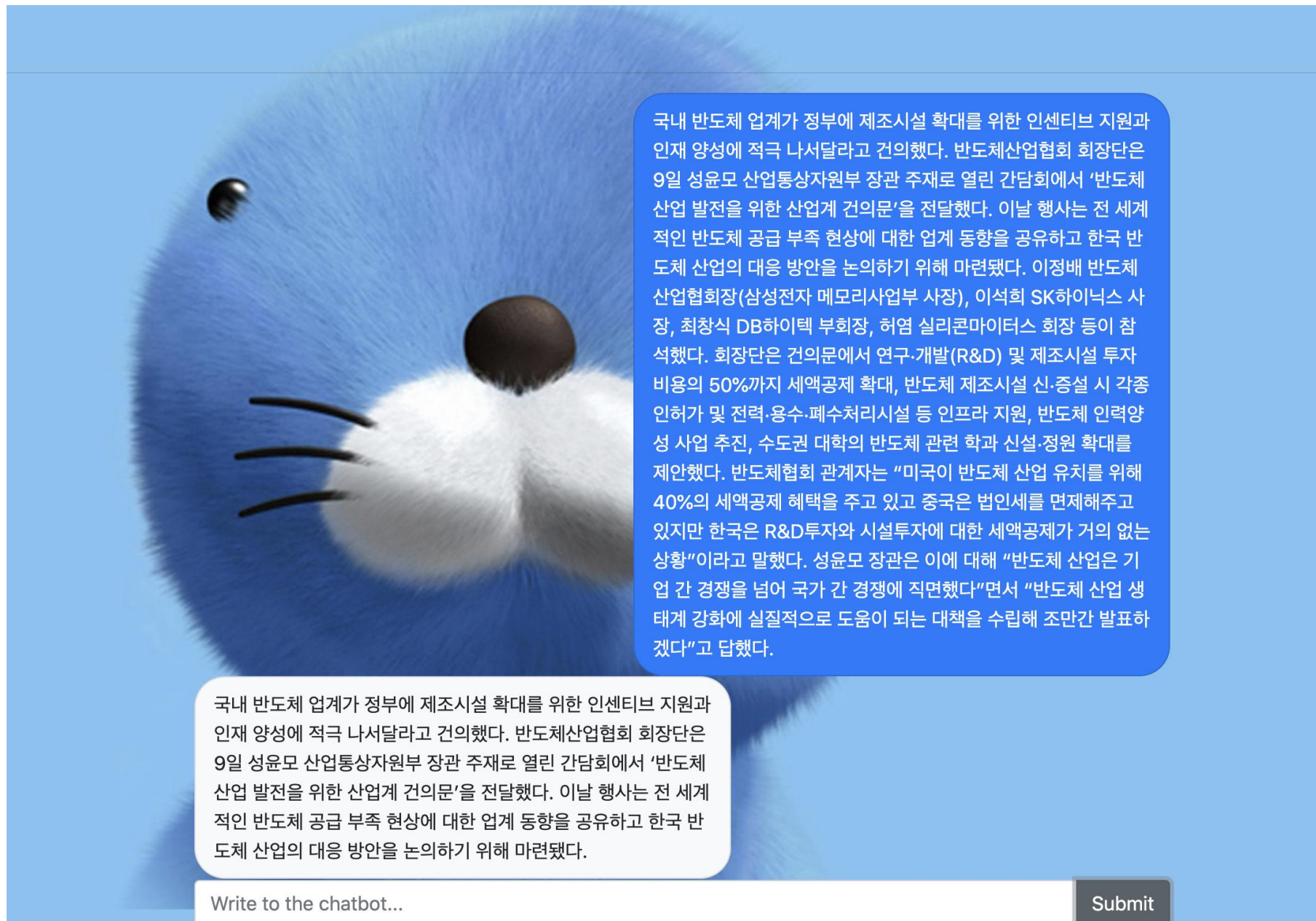
3. 정보추출

수학을 모른다면 인생을 살아가는데 아무런 지장이 없지만, 수학을 아는 것은 인생에 있어서 무한한 재미를 가져다 줄 수 있다. 수학은 본질을 자연의 본질을 파악하기 위한 하나의 언어이며, 수학을 잘 하면 머리가 잘 돌아간다. 학창 시절 수학을 잘하는 학생들은 못하는 사람들에게겐 동경의 대상이 되기도 하였으며, 가끔 수학을 모르는 학생이 잘하는 이성친구에게 다가가 물어보다가 커플이 되는 경우도 있다. 이렇듯 수학은 인생에 재미를 더하며, 인생을 풍요롭게 만들지만, **수학을 어려워하는 자에게는 한 없는 고통이 될 수도 있다.** 수학은 세상을 바꾸는데 있어서 필수적인 학문이며, 완벽한 논리를 추구하는 수학의 성질을 통해 자연의 숨어있는 규칙과 우주의 신비를 알아갈 수 있다. 하지만 **수학은 넘모 힘들고 어렵기 때무네 잘하기가 쉽지 않다.** 수학 잘하는 사람들 진짜 신기해 천재야 천재 수학은 크게 위상수학, 대수학, 해석학으로 나눌 수 있으며 현 시대에는 분야를 가리지 않고 융복합적인 연구가 활발하게 진행되고 있다.

3. 정보추출

수학을 모른다면 인생을 살아가는데 아무런 지장이 없지만, 수학을 아는 것은 인생에 있어서 무한한 재미를 가져다 줄 수 있다. 수학은 본질을 자연의 본질을 파악하기 위한 하나의 언어이며, 수학을 잘 하면 머리가 잘 돌아간다. 학창 시절 수학을 잘하는 학생들은 못하는 사람들에게겐 동경의 대상이 되기도 하였으며, 가끔 수학을 모르는 학생이 잘하는 이성친구에게 다가가 물어보다가 커플이 되는 경우도 있다. 이렇듯 수학은 인생에 재미를 더하며, 인생을 풍요롭게 만들지만, 수학을 어려워하는 자에게는 한 없는 고통이 될 수도 있다. 수학은 세상을 바꾸는데 있어서 필수적인 학문이며, 완벽한 논리를 추구하는 수학의 성질을 통해 자연의 숨어있는 규칙과 우주의 신비를 알아갈 수 있다. 하지만 수학은 넘모 힘들고 어렵기 때문에 잘하기가 쉽지 않다. 수학 잘하는 사람들 진짜 신기해 천재야 천재 수학은 크게 위상수학, 대수학, 해석학으로 나눌 수 있으며 현 시대에는 분야를 가리지 않고 융복합적인 연구가 활발하게 진행되고 있다.

4. 문서 요약



국내 반도체 업계가 정부에 제조시설 확대를 위한 인센티브 지원과 인재 양성에 적극 나서달라고 건의했다. 반도체산업협회 회장단은 9일 성윤모 산업통상자원부 장관 주재로 열린 간담회에서 '반도체 산업 발전을 위한 산업계 건의문'을 전달했다. 이날 행사는 전 세계적인 반도체 공급 부족 현상에 대한 업계 동향을 공유하고 한국 반도체 산업의 대응 방안을 논의하기 위해 마련됐다. 이정배 반도체 산업협회장(삼성전자 메모리사업부 사장), 이석희 SK하이닉스 사장, 최창식 DB하이텍 부회장, 허엽 실리콘마이터스 회장 등이 참석했다. 회장단은 건의문에서 연구·개발(R&D) 및 제조시설 투자 비용의 50%까지 세액공제 확대, 반도체 제조시설 신·증설 시 각종 인허가 및 전력·용수·폐수처리시설 등 인프라 지원, 반도체 인력 양성 사업 추진, 수도권 대학의 반도체 관련 학과 신설·정원 확대를 제안했다. 반도체협회 관계자는 "미국이 반도체 산업 유치를 위해 40%의 세액공제 혜택을 주고 있고 중국은 법인세를 면제해주고 있지만 한국은 R&D투자와 시설투자에 대한 세액공제가 거의 없는 상황"이라고 말했다. 성윤모 장관은 이에 대해 "반도체 산업은 기업 간 경쟁을 넘어 국가 간 경쟁에 직면했다"면서 "반도체 산업 생태계 강화에 실질적으로 도움이 되는 대책을 수립해 조만간 발표하겠다"고 답했다.

국내 반도체 업계가 정부에 제조시설 확대를 위한 인센티브 지원과 인재 양성에 적극 나서달라고 건의했다. 반도체산업협회 회장단은 9일 성윤모 산업통상자원부 장관 주재로 열린 간담회에서 '반도체 산업 발전을 위한 산업계 건의문'을 전달했다. 이날 행사는 전 세계적인 반도체 공급 부족 현상에 대한 업계 동향을 공유하고 한국 반도체 산업의 대응 방안을 논의하기 위해 마련됐다.

Write to the chatbot...

Submit

공학석사 학위논문

가사의 감정 분석과 구조 분석을 이용한
노래간 유사도 측정

Popular Music Similarity Evaluation
using Emotion and Structure Analysis on Lyrics

2016 년 2월

서울대학교 대학원
컴퓨터공학부
이재환

[Aidea] ⑩ “지금 내 감정과 잘 어울리는 노래는”...AI 추천 음악으로 힐링

윤영주 기자 | 입력 2021.11.20 14:57 | 댓글 1 | 좋아요 0

스마트인재개발원 'SOULFUL' 팀, 감정 기반 음악 추천 서비스 제안
비지도 학습·딥러닝 기술 등 활용...사용자 위치의 날씨 및 감정 분석
이모티콘으로 감정을 표현하면, 지금 기분에 어울리는 맞춤 곡 재생
스트리밍 서비스 비롯한 챗봇·음악 치료 등 다양한 산업에 접목 가능
"늦은 나이는 없다, 스마트인재개발원 경험은 내 인생의 터닝 포인트"

<http://www.aitimes.com/news/articleView.html?idxno=141511>

f

응용

사랑 고백

아티스트명	곡명	가사	장르
SS501	Snow Prince	girl 이른 햇살에 비치는 너의 싱그런 모습 내 맘을 가져간걸까 i wanna b...	국내 댄스/일렉
이달의 소녀	Perfect Love	oh baby hold on 보자마자 놀랐잖아 넌 내 이상형 짜릿하게 스친 순간 부...	국내 댄스/일렉
김하온 (HAON)	LOVE ! DANCE ! (Prod. BOYCOLD)	i don't wanna waste my time 어제 꿔던 꿈이 생각나 넌 별안간...	국내 힙합
박지훈	Let's Love	baby good morning what a beautiful day 아침 날씨...	국내 팝/어쿠스틱
수지 (SUZY)	I LOVE YOU BOY	그댈 보기 위해 먼 길을 건너왔네요 힘든 길인걸 알면서도 그저 걸었죠 i fly f...	국내 발라드, OST/BGM

자유롭게 날아

아티스트명	곡명	가사	장르
세훈&찬열	What a life	시스템 종료를 click 오늘 미세먼지는 free 차 키 챙겨서 나가 오랜만에 도...	국내 힙합
곽진언	자유롭게	자기가 자유롭게 쉽잖은 세상이지만 알잖아 나는 언제나 네 편인 걸 그러니 자...	국내 발라드
YB (윤도현밴드)	나는 나비	내 모습이 보이지 않아 앞길도 보이지 않아 나는 아주 작은 애벌레 살이 터져 ..	국내 락/메탈
볼빨간사춘기	여행	저 오늘 떠나요 공항으로 핸드폰 꺼 놔요 제발 날 찾진 말아줘 시끄럽게 소리를 질...	국내 인디, 국내 락/메탈
태연 (TAEYEON)	I (feat. 버벌진트)	빛을 쏘는 sky 그 아래 선 아이 i 꿈꾸듯이 fly my life is a be.	국내 락/메탈

<https://mail.so/musicdata/posts/7251b9>

요즘 정치권은 연일 청년, 청년, 청년
그리고 그들을 잡기 위한 커뮤니티 정치학..

2022 대선, 정치권이 부르는 '청년'은 누구인가

[대선 D-100] 20대 대선 결정하는 그들...
2030 MZ세대, 키플레이어 떠올라

"기득권 두 당이 '남초 커뮤니티' 담론에 목숨 거는 동안..."

바람과 검증, 대립과 혐오...대선 정국 움직이는 커뮤니티 정치학

[2030 정치 선 넘기] '웹코'에서 대선 후보들이 찾지 못하는 것

"커뮤니티 한줌론 외치고 이해 못한 사람들 안타깝다"

여야 대선주자, 청년세대 표심 향한 구애 나서...방식은 각각

- 커뮤니티 관련 언론 기사 제목 -

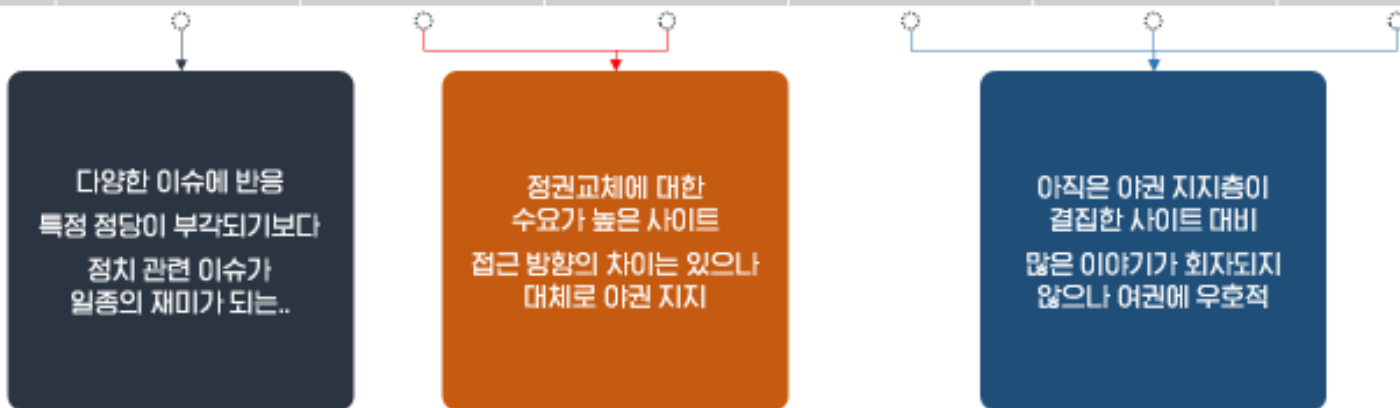
<http://bigdata.emforce.co.kr/index.php/2021120901/>

커뮤니티 사이트별 주요 연관어 해석

디시인사이드	MLB파크	에펠키리아
<p>문재영 발언 검찰총장 정치 풍자 놀이터로 대선이 가십거리로 소모</p>	<p>정권교체 가능성이 가장 높은 후보를 옹호하는, 아물따 정권교체!</p>	<p>페미니즘에 대한 백래시 (반발 심리)를 바탕으로 결집</p>
판지일보	보배드림	클리앙
<p>검찰개혁 문재인 대통령의 검찰개혁을 완수할 계승자 기대</p>	<p>여론조사 일베 드루킹 수세 모드로 전환한 사이 진보</p>	<p>건창개혁 여권 주자에 높은 관심 유지 적극적 정치 참여 독려</p>

세부 분석 기준 ③ 커뮤니티 사이트별 방문자들의 정치 성향 정리

	디시인사이드	MLB파크	에펠폰코리아	판지일보	보배드림	클리앙
방문 규모	고	저	고	저	저	저
정치 언급	고	고	고	고	고	고
관여도	초 고관여	고관여	초 고관여	고관여		
민감도	그룹 ①		그룹 ②	그룹 ③		
정치 성향	Gosship	무조건 정권교체!	페미니즘 반발!	현 정부 계승!	Shy 진보!	여권 적극 지지



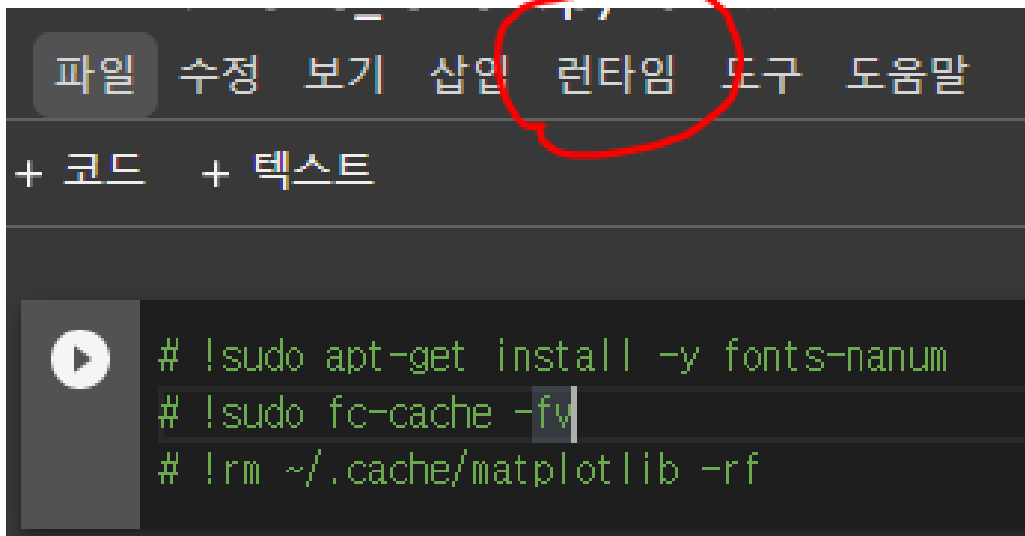
이런거 하려면
데이터 전처리가 필수!

Colab 한글폰트설치

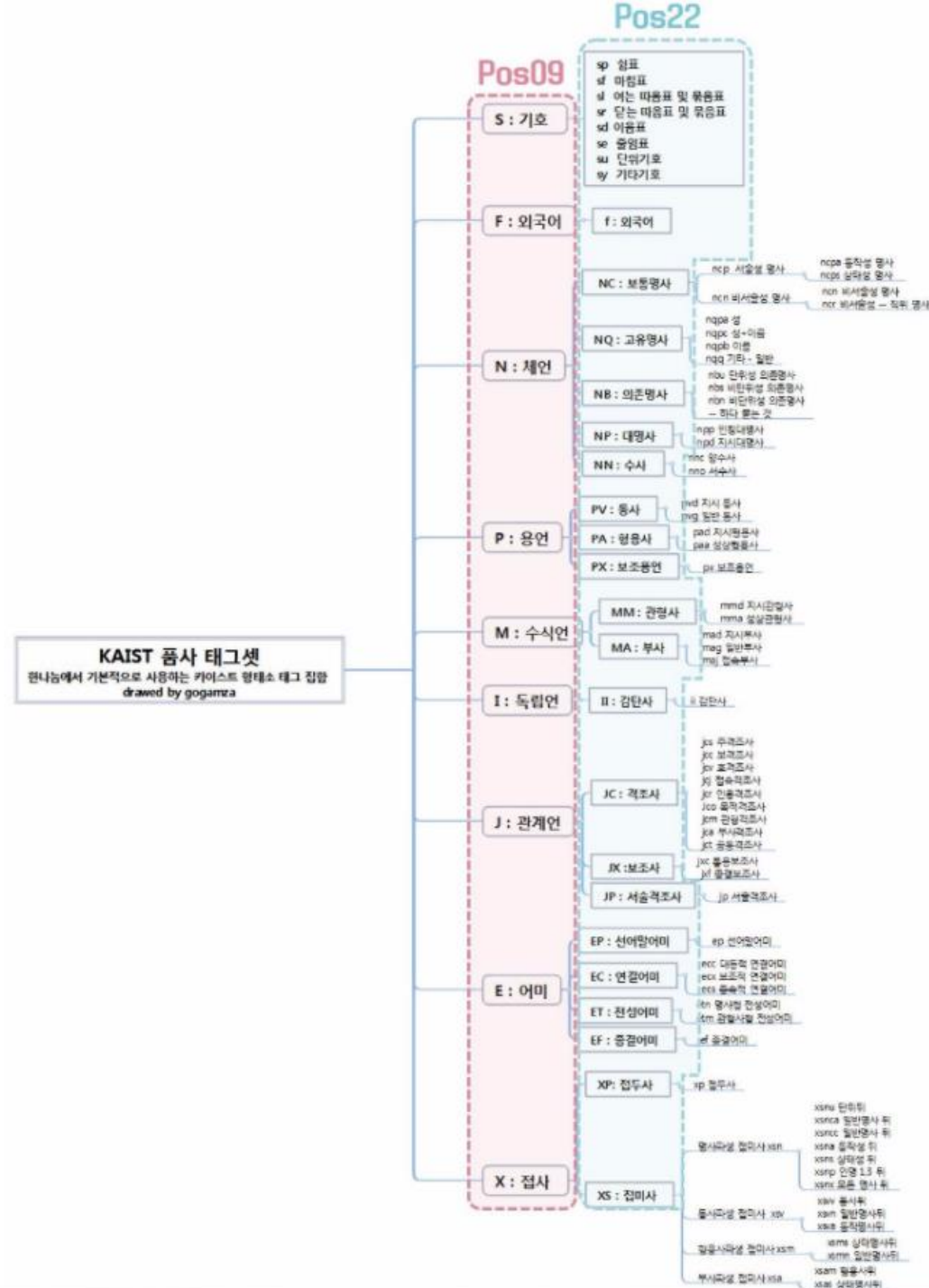
1. 아래 코드를 실행한다

```
!sudo apt-get install -y fonts-nanum  
!sudo fc-cache -fv  
!rm ~/.cache/matplotlib -rf
```

2. 코드 비활성화 후 런타임 다시 시작



형태소 분석기?



6B7%9C-%ED%91%9C%ED%98%84%EC%8B%9D-Regex

수학과 최명수

형태소 분석기?

실질의미유무	대분류(5언 + 기타)	세종 품사 태그	
		태그	설명
실질형태소	체언	NNG	일반 명사
		NNP	고유 명사
		NNB	의존 명사
		NR	수사
		NP	대명사
	용언	VV	동사
		VA	형용사
		VX	보조 용언
		VCP	긍정 지정사
		VCN	부정 지정사
		수식언	MM
	MAG		일반 부사
	MAJ		접속 부사
	독립언	IC	감탄사
형식형태소	관계언	JKS	주격 조사
		JKC	보격 조사
		JKG	관형격 조사
		JKO	목적격 조사
		JKB	부사격 조사
		JKV	호격 조사
		JKQ	인용격 조사
		JX	보조사
		JC	접속 조사
	선어말 어미	EP	선어말 어미
	어말 어미	EF	종결 어미
		EC	연결 어미
		ETN	명사형 전성 어미
		ETM	관형형 전성 어미
	접두사	XPN	체언 접두사

토큰화 하기 (Tokenize)

1. 엄마 학교 다녀왔습니다

엄마 : 1 학교 : 2 다녀왔습니다 : 3

→ [1, 2, 3]

2. 엄마 저 학교 다녀왔습니다

엄마 : 1 저 : 2 학교 : 3 다녀왔습니다 : 4

→ [1, 2, 3, 4]

토큰화 하기 (Tokenize)

1. 엄마 학교 다녀왔습니다

엄마 : 1 학교 : 2 다녀왔습니다 : 3

→ [1, 2, 3]

2. 엄마 저 학교 다녀왔습니다

엄마 : 1 학교 : 2 다녀왔습니다 : 3 저 : 4

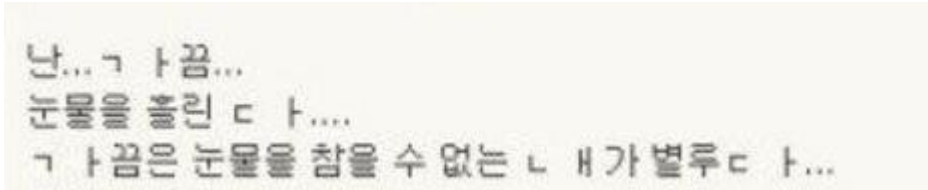
→ [1, 4, 2, 3]

토큰화 하기 (Tokenize)

1. 다음을 숫자에 대응되는 단어를 보고 문장을 완성해보시오

없는 : 1 나는 : 2 흘린다 : 3 눈물을 : 4
가끔 : 5 내가 : 6 별로다 : 7 참을 수 : 8

[2, 5, 4, 3, 5, 4, 8, 1, 6, 7] : ???



토큰화 하기 (Tokenize)

1. 물디브 가서 모히또 한잔 하자

물디브 : 1 가서 : 2 모히또 : 3 한잔 : 4 하자 : 5

→ [1, 2, 3, 4, 5]

토큰화 하기 (Tokenize)

1. 몰디브 : 1 가서 : 2 모히또 : 3 한잔 : 4 하자 : 5

→ [1, 2, 3, 4, 5] : 몰디브 가서 모히또 한잔 하자

→ [3, 2, 1, 4, 5] : 모히또 가서 몰디브 한잔 하자

→ [3, 4, 1, 2, 5] : 모히또 한잔 몰디브 가서 하자

토큰화 하기 (Tokenize)

1. ★고객님의 신용점수를 올릴수있는 절호의 기회! 대박 상품!★

2. 요청하신 고객님의 신용점수를 알려드립니다.

고객님의 : 1 신용점수를 : 2 올릴수있는 : 3 절호의 : 4 기회 : 5
★ : 6 ! : 7 대박 : 8 상품 : 9 요청하신 : 10 알려드립니다 : 11

1. : [6, 1, 2, 3, 4, 5, 6, 8, 9, 7, 6] - 11개
2. : [10, 1, 2, 11] - 4개

-Tokenize 한 문장의 길이가 다른데 어찌지



토큰화 하기 (Tokenize)

1. ★고객님의 신용점수를 올릴수있는 절호의 기회! 대박 상품!★

2. 요청하신 고객님의 신용점수를 알려드립니다.

고객님의 : 1 신용점수를 : 2 올릴수있는 : 3 절호의 : 4 기회 : 5
★ : 6 ! : 7 대박 : 8 상품 : 9 요청하신 : 10 알려드립니다 : 11

1. : [6, 1, 2, 3, 4, 5, 6, 8, 9, 7, 6] - 11개
2. : [0, 0, 0, 0, 0, 0, 0, 10, 1, 2, 11] - 11개

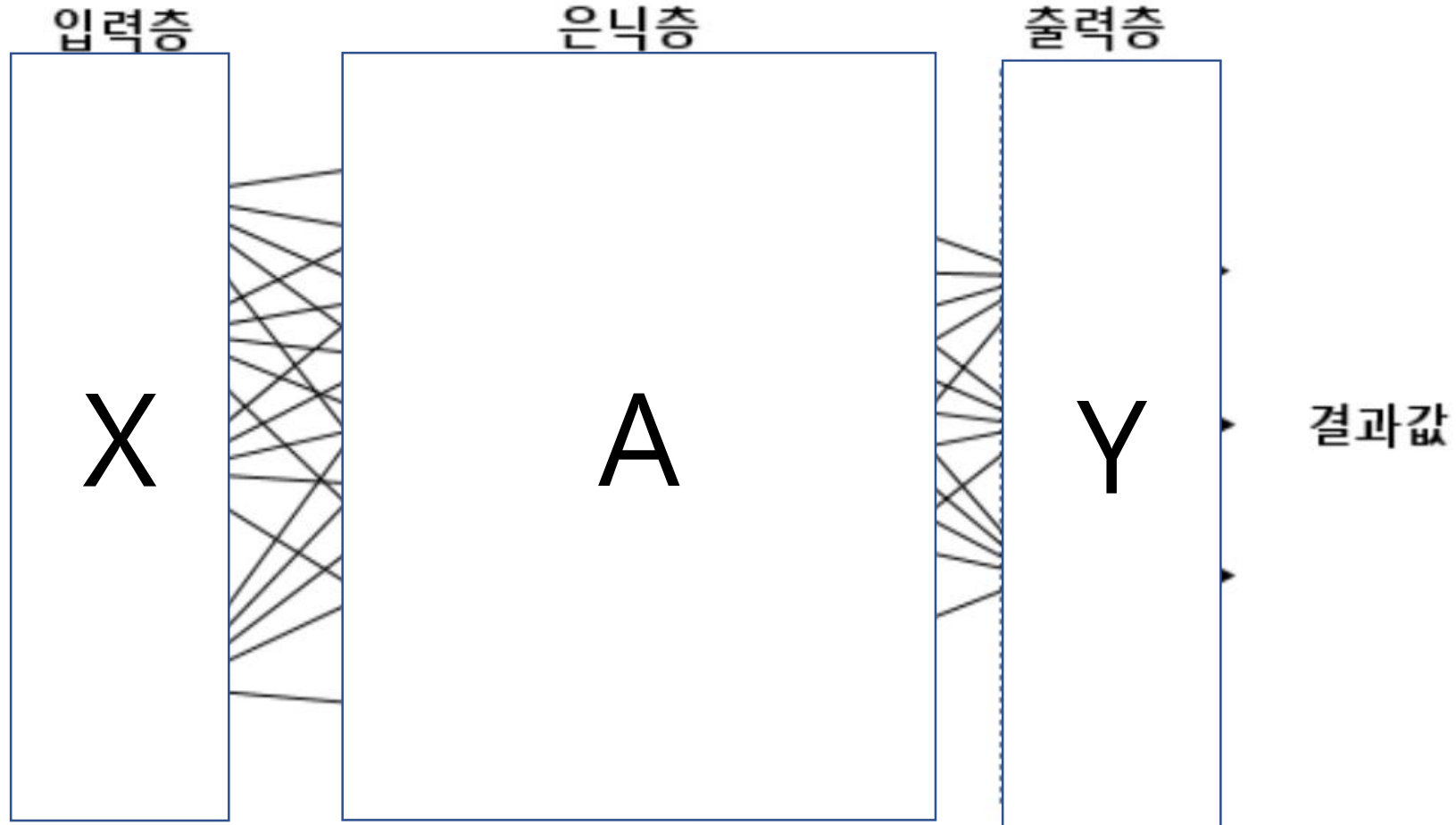


-Sequence Padding

즉, Tokenize와 Padding으로 단어들의 흐름을 벡터화한다!

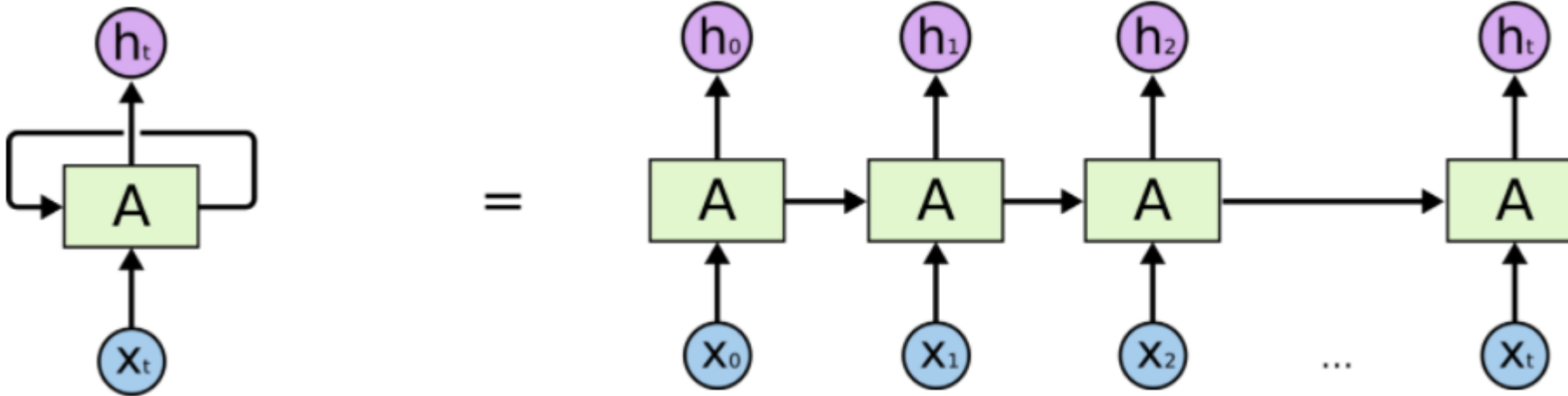
RNN (Recurrent Neural Network)

기존 인공신경망



RNN (Recurrent Neural Network)

순환신경망



<https://medium.com/humanscape-tech/rnn-recurrent-neural-network-%EC%88%9C%ED%99%98%EC%8B%A0%EA%B2%BD%EB%A7%9D-%EC%9D%84-%EC%9D%B4%ED%95%B4%ED%95%B4%EB%B3%B4%EC%9E%90-1697a5472af2>

RNN (Recurrent Neural Network)

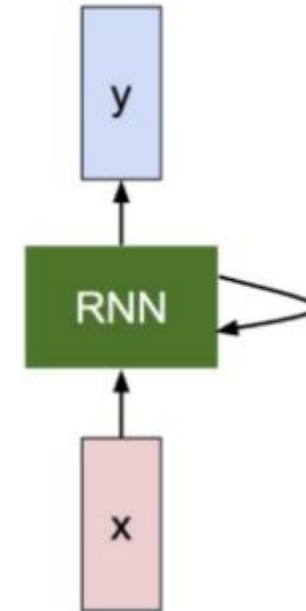
순환신경망

Recurrent Neural Network

We can process a sequence of vectors \mathbf{x} by applying a recurrence formula at every time step:

$$\boxed{h_t} = \boxed{f_W}(\boxed{h_{t-1}}, \boxed{x_t})$$

new state / some function with parameters W old state input vector at some time step



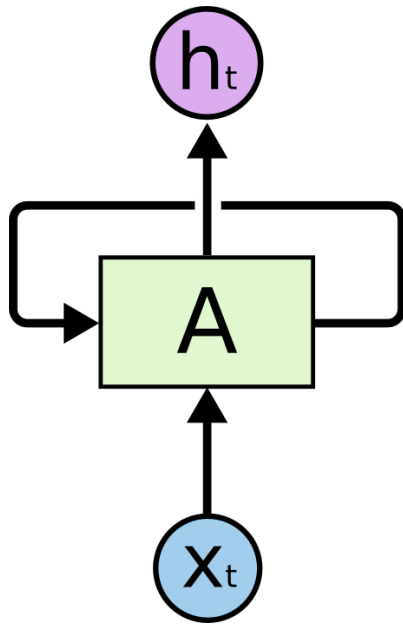
Fei-Fei Li & Andrej Karpathy & Justin Johnson

Lecture 10 - 15

8 Feb 2016

LSTM(Long Term Short Memory)

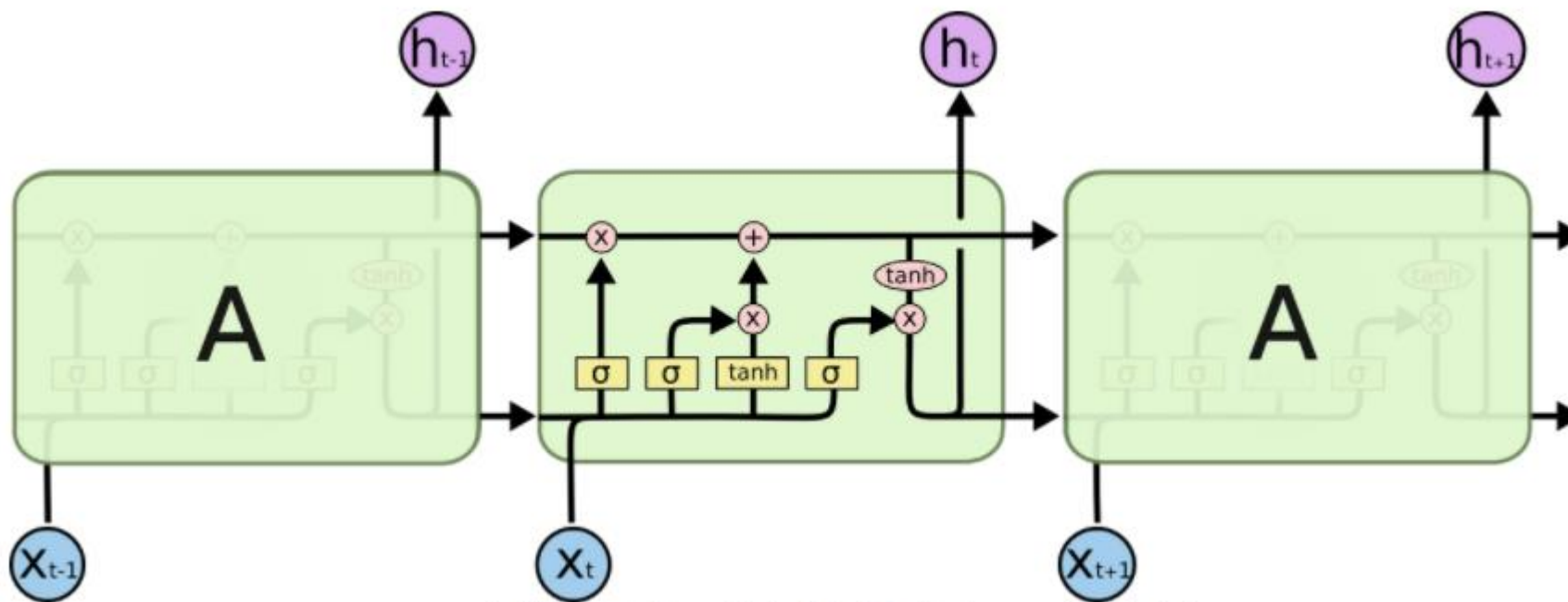
1. RNN(Recurrent Neural Network) 계열의 모델
2. 우리는 과거로부터 얻은 정보를 통해 판단하여 지금에 행동한다!
3. 이처럼 RNN은 이전 단계에서 얻은 정보가 지속되도록 한다



<https://dgkim5360.tistory.com/entry/understanding-long-short-term-memory-lstm-kr>

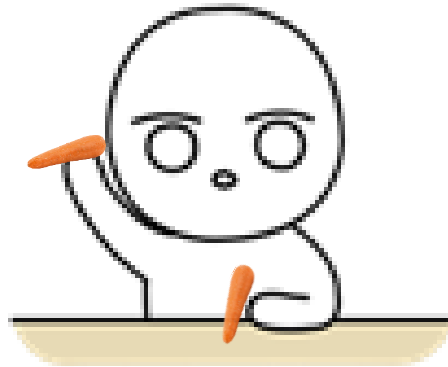
LSTM(Long Term Short Memory)

1. 기존 RNN의 긴 기간의 의존성(Long-term dependencies)을 극복

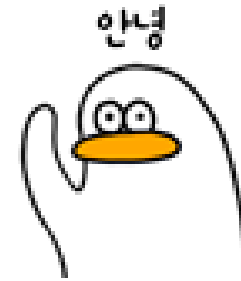


<https://dgkim5360.tistory.com/entry/understanding-long-short-term-memory-lstm-kr>

질문있습니다



끼
티



안녕!