2023 Ver.

데이터 리터러시

생성형 AI - 랭체인 (LLM) 활용

백엔드(풀스택) 엔지니어(자바,파이썬) 양성

문유정

010-8766-5119

zeein119@gmail.com

DATA LITERACY

데이터 문해력

정의

글을 읽고, 읽은 내용을 이해하는 능력이 리터러시(문해력)라면, 데이터를 읽고 그 안에 숨겨진 의미를 파악하는 데이터 해독 능력과 이를 전달하는 능력을 데이터리터러시(Data Literacy) 라고 한다.

KCA 한국방송통신전파진흥원 참조

데이터 문해력

데이터 읽기

“ 해석 ”

- 기술통계학 (Descriptive Statistics)
  : 대량 자료, 표, 그래프
- 추론통계 (Inferential Statistics)
  : 일부로 전체의 성질 파악 : 확률, 빈도, 추정, 검정, 분산분석
- 베이즈 통계
  : 과거의 데이터로 미래 예측, 주관적 확률 적용
- 육하원칙(5W1H)
  Who / What / When / Where
  How / Why ⟶ Insight 도출

데이터 쓰기

“ 설득 ”

- 행동 도출
- 연속적 스토리텔링
  시각화
  디자인씽킹(Design thinking)
  : 문제를 해결할 수 있는 아이디어를 선택
  : 공감, 정의, 아이디어 구상, 프로토타입 제작, 테스트
- 육하원칙(5W1H)
  Who / What / When / Where
  How / Why ⟶ Insight 도출

데이터

문자 데이터

“ 연산불가 ”

- 연산 적용시 의미 성립 불가
- 갯수측정(빈도측정, COUNT) 가능
- 숫자로 변환 (인코딩, Encoding)
- 텍스트 마이닝
  정보검색 (토큰화 , 형태소 분석)
  자연어처리 (NLP , 요약 , 품사태깅 , 감성 분석)
  정보 추출 (특징선택 , 특징추출 , 엔티티 식별)

숫자 데이터

“ 연산가능 ”

- 연산 적용시 의미 성립
- 집계함수(Aggregation Function) 가능
- 지표(Indicator, Index) / 스케일 변환
  단위가 다른 데이터는 직접 비교 불가 : 스케일 조정
  정규화 (MinMax Scaling , Normalization) : 0 ~ 1
  표준화 (Standardization) : -3 ~ 3

데이터 구조

행

“ 분석가 ”

- 저차원
- 객체 (Object) : 정보의 모음
  관측치 (Observation) / 표본 (Example) / 샘플 (Sample)
  데이터 분석가 선호 / 의사권자 비선호
- 열 → 행 : UnPivot / Melt

열

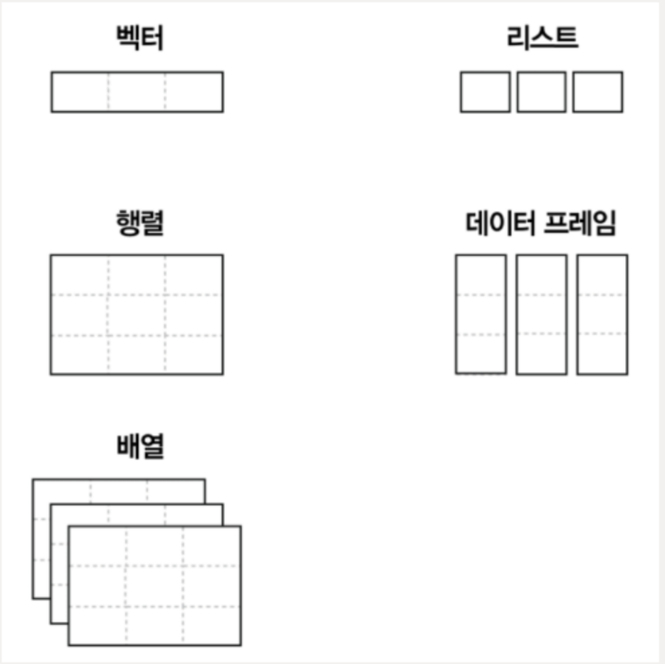
“ 의사결정자 ”

- 고차원 / 한화면
- 속성 (Property) / 변수 (Variable)
  특징 (Feature) / 차원 (Dimension)
  의사권자 선호 / 데이터 분석가 비선호
- 종속변수 : 연구대상, 목적변수, Output
  독립변수 : 조사자료, 설명변수, Input
  식별변수 : 행을 식별하게 하는 변수 : 가장 왼쪽에 위치
  측정(특징)변수 : 속성을 설명하는 변수
- 행 → 열 : Pivot / Cast

데이터 집합

- 스칼라 (Scala) : 0차원 , 원소 1개
- 벡터 (Vector) : 1차원 , 1개 이상 원소
- 행렬 (Matrix) : 2차원 , 행 / 열
- 배열 (Array , Tensor) : 3차원 이상
  행 / 열 / 차원(면)

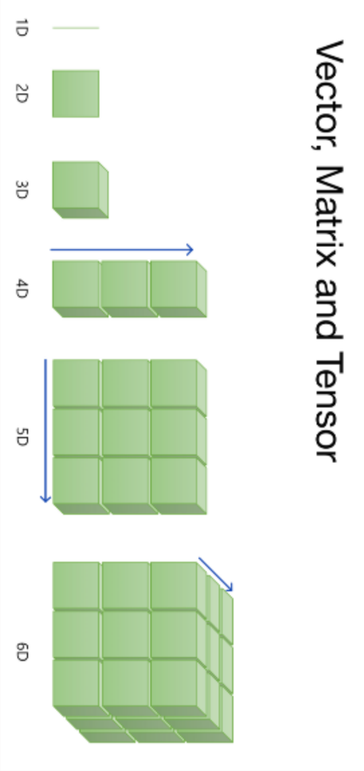
- 문자/숫자/bool 같은 종류 원소만 가능
- bool < 정수 < 실수 < 문자
  자료형이 큰 데이터 입력 - 기존 자료가 큰 자료형으로 변경
  자료형이 작은 데이터 입력 - 입력 자료만 큰 자료형으로 변경

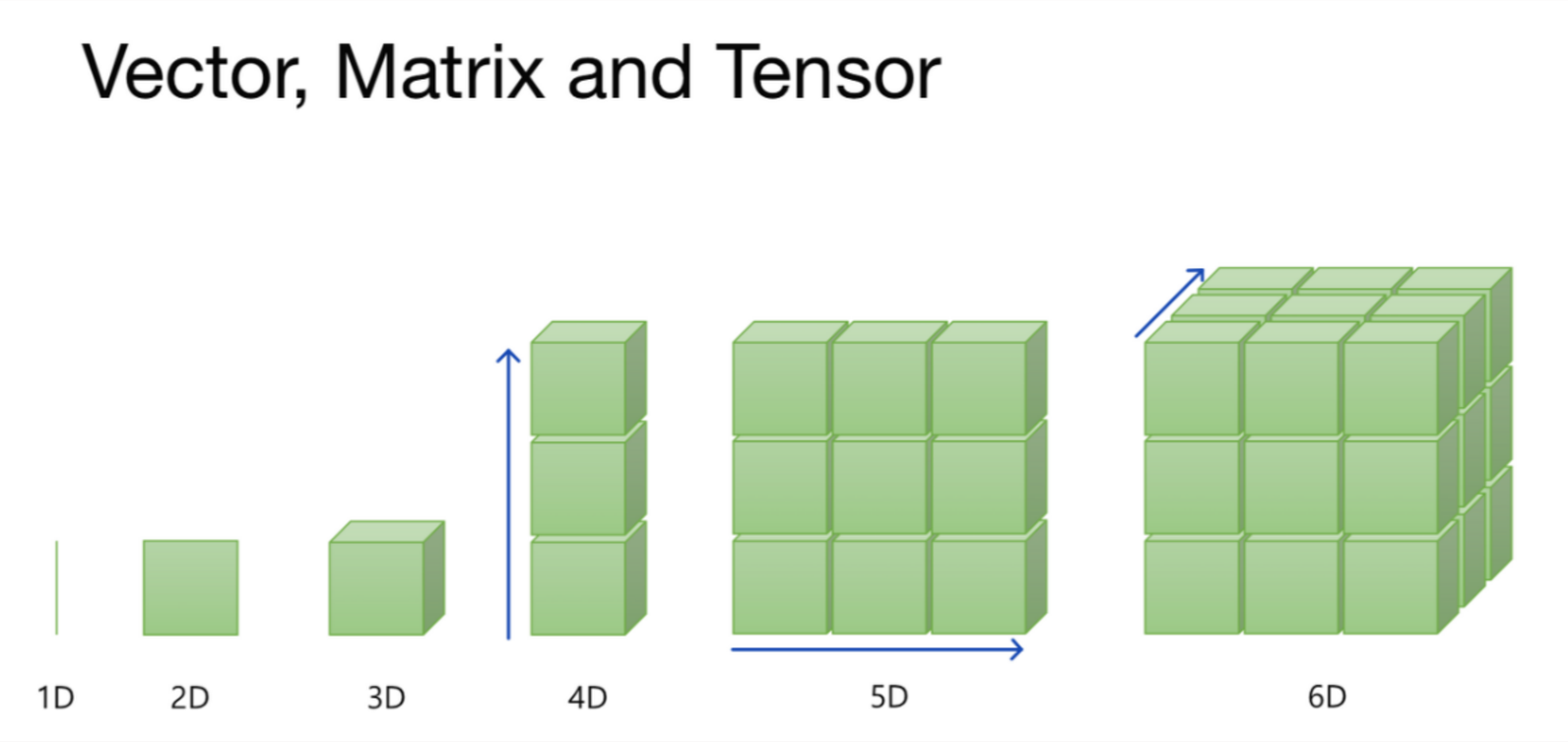


- 리스트 (List) : 1차원
  서로 다른 자료형 원소 가능
- 데이터프레임 (DataFrame) : 2차원
  서로 다른 자료형 열 가능
  같은 열은 자료형이 일치해야 함

데이터 집합

출처 : https://softkau19.tistory.com/6





데이터 관리

출처 : https://hooun.tistory.com/193

- 코드영역 : 프로그램 저장
- 데이터 영역 : 클래스 단위
- 힙(heap) 영역 : 객체 단위

  모든 스레드에서 참조 가능

  프로그래머가 수동 할당/해제

  넓은 공간 / 느린 처리 속도 / 메모리 단편화
- 스택(Stack) 영역 : 함수 단위 / 원시형(Primitive) 단위

  하나의 스레드에서만 참조 가능

  함수 호출시 자동 할당 / 함수 반환시 자동 해제

  작은 공간 / 빠른 처리 속도 / 메모리 부족

데이터 피라미드

출처 : Wikipedia

데이터 유형

정형
- RDB / CSV

반정형
- HTML / JSON
- WEB Log

비정형
- SNS / Email
- 이미지 / 동영상
- 음성 / IoT

데이터 크기

KB	MB	GB	TB	PB	EB	ZB	YB
Kilo	Mega	Giga	Tera	Peta	Exa	Zetta	Yotta
10^3	10^6	10^9	10^{12}	10^{15}	10^{18}	10^{21}	10^{24}
2^{10}	2^{20}	2^{30}	2^{40}	2^{50}	2^{60}	2^{70}	2^{80}

데이터 분석 선순환 구조

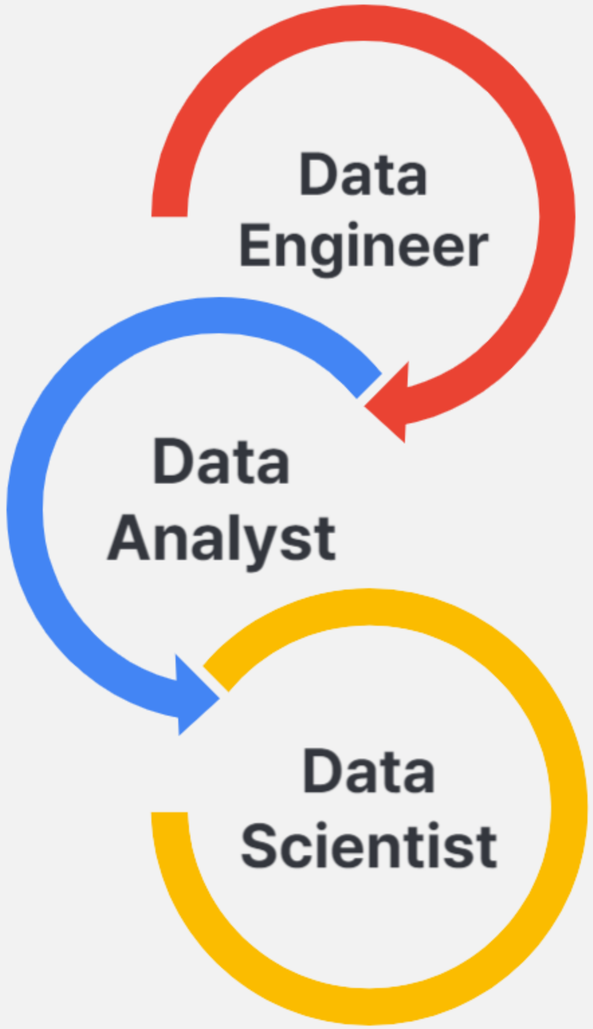
데이터 분석 도구

유형	특징
Rstudio	통계 컴퓨팅, 그래픽스를 위한 프로그래밍 언어인 R을 위한 오픈 소스 통합 개발 환경
Jupyter	Python 을 위한 오픈 소스 통합 개발 환경
Excel	2003 : 65,536 / 256 2007이후 : 1,048,576 / 16,384
SQL	통계적 분석과 데이터 마이닝 등에 사용되는 통계 분석 프로그램
Power BI	통계적 분석과 데이터 마이닝 등에 사용되는 통계 분석 프로그램
Tableau	데이터 시각화 프로그램
SPSS	통계적 분석과 데이터 마이닝 등에 사용되는 통계 분석 프로그램
SAS	통계 분석 프로그램. 빅데이터에 유용
STATA	통계 분석 프로그램. 시계열 분석
Zeppelin	데이터 분석 및 시각화 프로그램. 빅데이터 분석
Rapid Miner	GUI 기반 예측 분석 통합 플랫폼
Qlik Sense	데이터 분석 및 시각화 프로그램
Spark	메모리 DB 활용 빅데이터 분석 분산처리 시스템

기술동향

	2016	2018	2019	2020	2021	2022	2023
1	SQL	Python	Python	Python	Python	Data Visualization	Python and R
2	Hadoop	R	R	R	SQL	Python	Hadoop
3	Python	SQL	SQL	SQL	R	SQL/NoSQL	NoSQL
4	Java	Hadoop	Spark	Spark	Spark	Social Media Mining	Machine Learning
5	R	Spark	Hadoop	Hadoop	AWS	Statistics	Data Visualization
6	Hive	Java	Java	Java	Java	Natural Language Processing/Machine Learning	Probability and Statistics
7	MapReduce	SAS	Tableau	AWS	Tableau	Excel	Curiosity, Common Sense, and Communication Skills
8	NoSQL	Tableau	AWS	Tableau	Hadoop	High-Level Math	Innovation
9	Pig	Hive	SAS	SAS	TensorFlow	Teamwork	Data Intuition
10	SAS	Scala	Hive	Hive	Scala	Communication	Business Expertise

Data Scientist



Citizen Data Scientist
(CDS)

데이터 분석 절차

주제 / 목적

분석 목적 선정
요구사항 정리

한식, 양식, 중식

데이터 수집

데이터 스크래핑
데이터 허브
사내 보유 데이터
데이터 마켓

재료 구매

데이터 전처리

결측치
이상치
정규화 / 표준화
특징 선택
특징 추출

재료 손질

데이터 분석

확증적 데이터 분석
- 가설검정

탐색적 데이터 분석
- 데이터 시각화

머신러닝 / 딥러닝

레시피

시각화 / 시스템

패턴 분석
시스템 배포 / 서빙

플레이팅

가설검정 : 귀무가설이 잘못됐음을 주장

대립가설

“ 의심/차이 ”

- 나(연구자)의 주장.
  의심사항. 일반적이지 않은 사항
  차이(연관,효과,...)가 있다.
  p-value가 유의수준보다 작다.
  p-value : 대립가설이 틀릴 확률
  유의수준 : 가설 판단 기준 : 5% / 1% / 0.1%
  통계적으로 유의하다.
  대립가설 채택 : 분석에 사용
  대립가설 기각 : 분석에 사용하지 못함

귀무가설

“일반적/동일 ”

- 대립가설의 반대
  일반적인 사항
  차이(연관,효과,...)가 없다.
  p-value가 유의수준보다 크다.
  p-value : 귀무가설이 맞을 확률
  유의수준 : 가설 판단 기준 : 5% / 1% / 0.1%
  통계적으로 유의하지 않다.
  귀무가설 기각 : 대립가설 채택

전주의적 속성(Preattentive Attribute)

데이터 시스템

정규화

중복입력 배제
효율적 입력
데이터 관리성 증대

조인 증가로 인한 성능 저하

효율

비정규화

테이블 결합
편리한 검색
고정형 검색

데이터 일관성 고려 필요

결합

카디널리티

특정 데이터 집합의 유니크한 값의 개수

크면 중복도가 낮고 집계가 느림

중복

운영시스템

컴퓨터 + 데이터

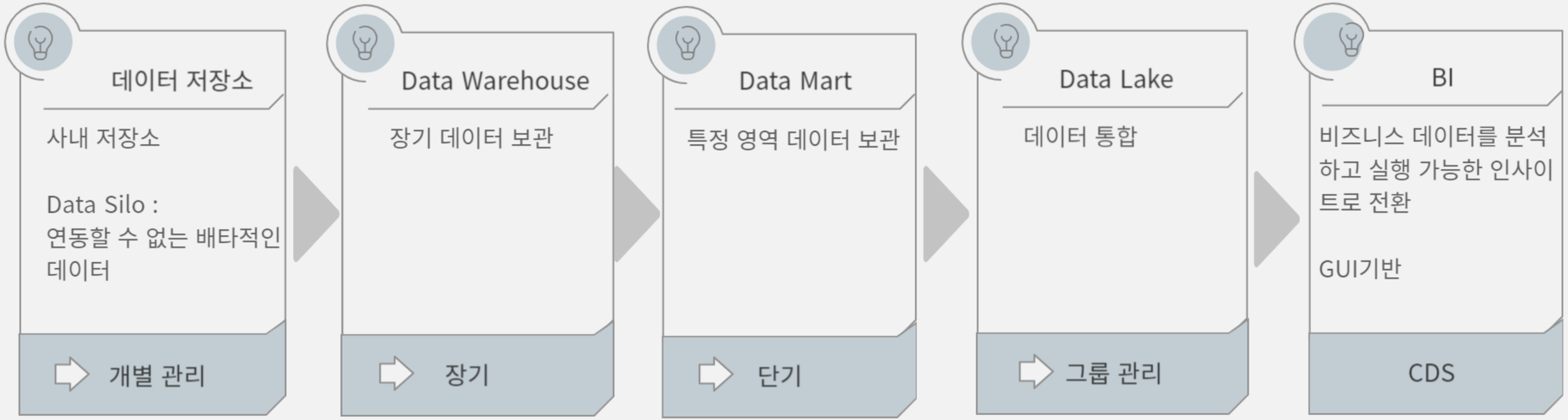
운영

정보시스템

ETL
DW
BI

이관

정보 시스템


데이터 저장소
사내 저장소
Data Silo :
연동할 수 없는 배타적인 데이터
개별 관리
Data Warehouse
장기 데이터 보관
장기
Data Mart
특정 영역 데이터 보관
단기
Data Lake
데이터 통합
그룹 관리
BI
비즈니스 데이터를 분석하고 실행 가능한 인사이트로 전환
GUI기반
CDS

머신러닝 개요

지도학습

	회귀	분류
선형회귀	O	X
정규화(Regularization) - Ridge , LASSO	O	X
k최근접 이웃 알고리즘 (K Nearest Neighbors)	O	O
결정트리 (Decision Tree)	O	O
랜덤 포레스트 (Random Forest)	O	O
서포트 벡터 머신 (Support Vector Machine)	O	O
서포트 벡터 머신 (Support Vector Machine) - 커널 기법	O	O
신경망 (Neural Network)	O	O
로지스틱 회귀 (Logistic Regression)	X	O
나이브 베이즈 분류 (Naive Bayes Classification)	X	O

비지도학습

	군집	차원축소
k평균 알고리즘 (K Means)	O	X
가우시안 혼합 모델 (Gaussian Mixture Model)	O	X
주성분분석 (Principal Component Analysis, PCA)	X	O
잠재 의미 분석 (Latent Semantic Analysis, LSA)	X	O
음수 미포함 행렬 분해 (Non-negative Matrix Factorization, NMF)	X	O
잠재 디리클레 할당 (Latent Dirichlet Allocation, LDA)	X	O
국소 선형 임베딩 (Local Linear Embedding, LLE)	X	O
t-분포 확률적 임베딩 (t-distributed Stochastic Neighbor Embedding, t-SNE)	X	O

종속변수 : 숫자(회귀) 문자(분류)

선형회귀 / 딥러닝

타겟(Target) : 정답. 종속변수

특성(Feature) : 자료. 독립변수

Y = aX + b :

Y : 종속변수. 예측값

X : 독립변수. 데이터

a : 계수(Coefficient) - 가중치(Weight)

b : 절편(Intercept) - 편향(Bias)

CNN

선형회귀 : 이미지 평탄화(Flatten)
필터(Filter) : 학습 대상 / 가중치 집합 / 특성
풀링(Pooling) : 정보 압축
이미지 > 필터 > 특성맵(Feature Map) > 풀링

RNN

이전 데이터가 다음 데이터에 영향 : 시간적/공간적 연속성
상태값 : 이전 시점의 출력값 / 이전 데이터의 특성 전달 / 임시 메모리 역할

GAN(Generative Adversarial Network)

모델끼리 경쟁

- 생성자(Generator) : 감별자를 속이지 못한 데이터 학습
- 감별자(Discriminator) : 생성자가 만든 가짜 데이터 학습
- 감별자 맞히는 확률 50% > 학습 종료

LLM(Large Language Model) I

자연어 처리(National Language Processing)
- 토큰(token) : 처리 기본 단위
- 인덱싱 : 토큰에 일련번호 부여
- 어휘 목록(Vocabulary) : 토큰 집합
- 임베딩(Embedding) : 문자를 특성이 담긴 숫자로 변환 / 인코딩(Encoding)

워드 임베딩 차원
- 트랜스포머(512차원) / BERT(768차원) / GPT3(1280차원)

전이학습 : 새로운 모델에게 학습된 가중치 전달

파인튜닝(Fine-Tunning) : 전달받은 가중치를 추가 학습으로 미세 조정하여 모델을 특화
- 입력 〉 전이학습 〉 파인튜닝 〉 출력

LLM(Large Language Model) II

인코더-디코더 모델 : 데이터 열 입력 / 데이터 열 출력
- 인코더 : 문장 -〉 숫자(문맥 정보 포함/컨텍스트 벡터) / 디코더 : 숫자(컨텍스트 벡터) -〉 문장
- 토큰별 인코더/디코더(RNN셀) 적용 / seq2seq (Sequence-to-Sequence) : 번역

어텐션 메커니즘 : 입력이 길어지면 과거 데이터의 상태값이 희석되어 영향력이 낮아지는 RNN단점 보완
- 디코더가 토큰별로 단어 생성할때 입력 문장 단어 전체를 다시 검토
- 디코더 토큰 상태값과 인코더 각 토큰별 상태값 사이의 유사도 평가

트랜스포머의 어텐션 : 각 단어 임베딩값 + 문장 내 순서 정보
- 인코더 셀프 어텐션 : 입력 문장 / 디코더 셀프 어텐션 : 생성 문장

BERT(Bidirectional Encoder Representations from Transformers) :
- 인코더 사용 / 문장 일부 예측 / 문장 전체 맥락 이해 / 33억개 단어
- 단어 임베딩 벡터 출력(문맥 정보 포함) / 컨텍스트 벡터는 미제공

GPT(Generative Pretrained Transformers) :
- 디코더 사용 / 다음 단어 예측 / 문장 생성 / 대화 / GPT3(570GB 텍스트)
- 순차적 텍스트 생성 / 다음 단어 출현 확률