

데이터 수집 및 관리

Kyungsik Han

본 영상에서 배울 내용







- 데이터 관련 내용 학습
 - 수집
 - 종류
 - 형식
 - 타입
 - 행렬
 - 품질

데이터 수집

- 데이터 수집
 - 데이터 분석에 필요한 데이터를 확보하는 것
- 데이터
 - 이미 보유하고 있는 데이터
 - 새로 수집해야 할 데이터
- 데이터 수집에 시간과 비용 고려

The six Vs of big data

Big data is a collection of data from various sources, often characterized by what's become known as the 3Vs: *volume, variety and velocity*. Over time, other Vs have been added to descriptions of big data:

VOLUME	VARIETY	VELOCITY	VERACITY	VALUE	VARIABILITY
The amount of data from myriad sources.	The types of data: structured, semi-structured, unstructured.	The speed at which big data is generated.	The degree to which big data can be trusted.	The business value of the data collected.	The ways in which the big data can be used and formatted.
					

데이터 종류

구분	종류 예시
비즈니스, 웹 서비스 활동	매출 기록, 거래 내역, 웹 로그, 고객 정보
소셜 네트워크 활동	다양한 SNS 데이터
정부, 공공 기관	공공 인프라, 기상, 예산, 회의 자료 등
과학 기술 연구 활동	연구 실험 및 측정 데이터
각종 센서 장비	센서 데이터 (빛, 소리, 바람, 해양, 지질 등)
데이터 형태별	숫자, 텍스트, 오디오, 비디오 등

데이터

- 데이터 인스턴스(instance)와 속성(attribute) 모임
- 속성 혹은 특징(feature)은 데이터 인스턴스 특성을 나타냄
- 여러 개의 데이터 속성이 하나의 데이터 인스턴스를 만듦

Attributes (variable/feature)

Instances (sample/point/ case/object)	구분	키	몸무게	나이
	A	174cm	70kg	21세
	B	170cm	61kg	27세
	C	162cm	73kg	29세

데이터 형식

형식	종류 예시
정형 (structured)	<ul style="list-style-type: none">• 데이터의 포맷이 정해져 있는 데이터 (CSV)• 서식이 정해진 데이터• 통계표, 기업의 매출 기록
비정형 (unstructured)	<ul style="list-style-type: none">• 미리 정해진 포맷을 가지지 않는 데이터• 블로그, 트위터 데이터와 같이 임의의 문장으로 구성• 오디오, 비디오, 인기도 수치 데이터
반정형 (semi-structured)	<ul style="list-style-type: none">• 데이터 내부는 논리적 형식을 가지고 있으나 외형상으로는 데이터 포맷이 정형 데이터처럼 완전히 정의되어 있지 않음• 센서 데이터, 시간대별로 웹 사용자의 기록 등

데이터 타입

유형별 분류

형식	내용
문자형	이름, 주소, 텍스트 본문 등
숫자형	매출, 통계 수치, 센서 측정값
바이너리(binary)형	오디오, 비디오, 실행 파일 등 읽을 수 없는 파일 형태 (010101011110001100...)

속성별 분류

형식	내용
범주형 (categorical)	<ul style="list-style-type: none">클래스를 구분할 때 사용성별, 국가명, 요일, 사람 이름 등
순서형 (ordinal)	<ul style="list-style-type: none">순서가 의미를 가지는 데이터옷 사이즈, 달력 월/일 등
연속형 (continuous)	<ul style="list-style-type: none">숫자의 값이 어떤 의미를 가지는 연속적 데이터무게, 길이, 온도, 키, 몸무게 등

데이터 행렬 (matrix)

- 데이터의 속성이 숫자형만으로 구성될 경우 데이터를 다차원 행렬로 생각 가능
- 각 차원이 하나의 속성으로 표현
- 데이터는 $m \times n$ 행렬로 표기 가능하며 m 행은 데이터 인스턴스를, n 열은 데이터 속성을 나타냄

The diagram shows a data matrix with 4 columns and 4 rows. The first row is a header with blue background and white text. The subsequent three rows have light blue backgrounds. A bracket above the columns is labeled 'Attributes (variable/feature)' in red and blue. A bracket to the left of the rows is labeled 'Instances (sample/point/case/object)' in red and blue.

Attributes (variable/feature)			
구분	키	몸무게	나이
A	174cm	70kg	21세
B	170cm	61kg	27세
C	162cm	73kg	29세

Instances (sample/point/case/object)

데이터 품질 (quality)

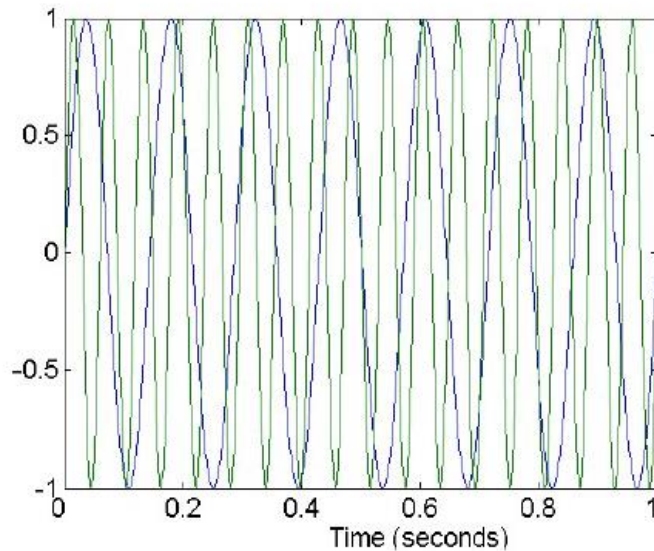
- 데이터 품질에 대한 문제는 어떤 것이 있을까?
- 데이터 품질 문제를 어떻게 발견할까?
- 데이터 품질 문제를 어떻게 해결할까?
- 데이터 품질에 대한 문제들
 - 결측치(missing value)
 - 노이즈(noise)와 이상치(outlier)
 - 중복 데이터

결측치 (Missing value)

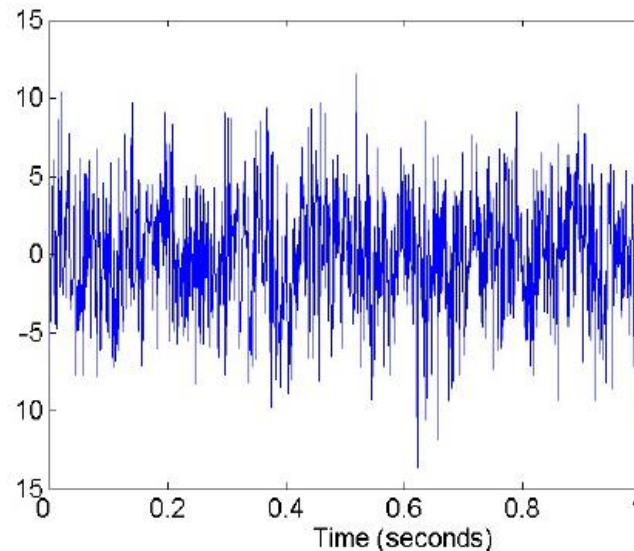
- 결측치의 원인
 - 정보가 모이지 않음
 - 데이터 속성이 모든 경우에 적용 가능하지 않음
- 결측치 문제를 해결하는 방법
 - 데이터 삭제하기
 - 결측치를 예상해서 채워 넣기
 - 분석 단계에서 결측치를 무시하기

노이즈 (Noise)

- 원본에서 조금씩 값이 변경되는 현상
 - 품질이 좋지 않은 전화 통화에서 사람의 목소리
 - 영상 통화



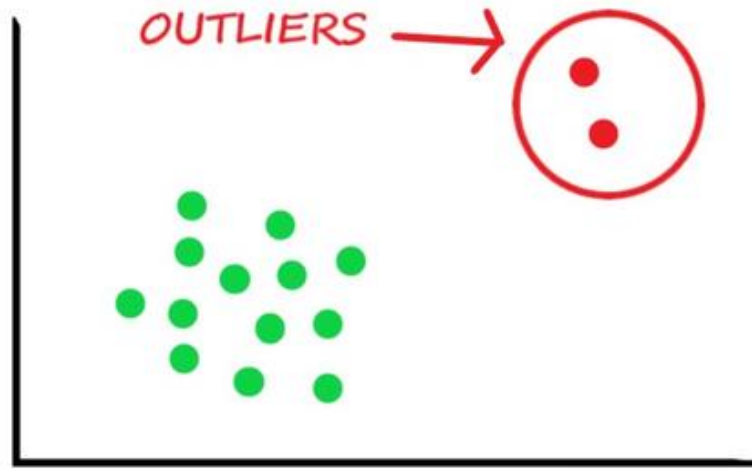
Two Sine Waves



Two Sine Waves + Noise

이상치 (Outlier)

- 데이터의 전체적인 패턴과 전혀 다른 양상을 보이는 데이터들
- 이상치를 고려하지 않고 분석하면 잘못된 결과를 초래할 수 있음



중복 데이터

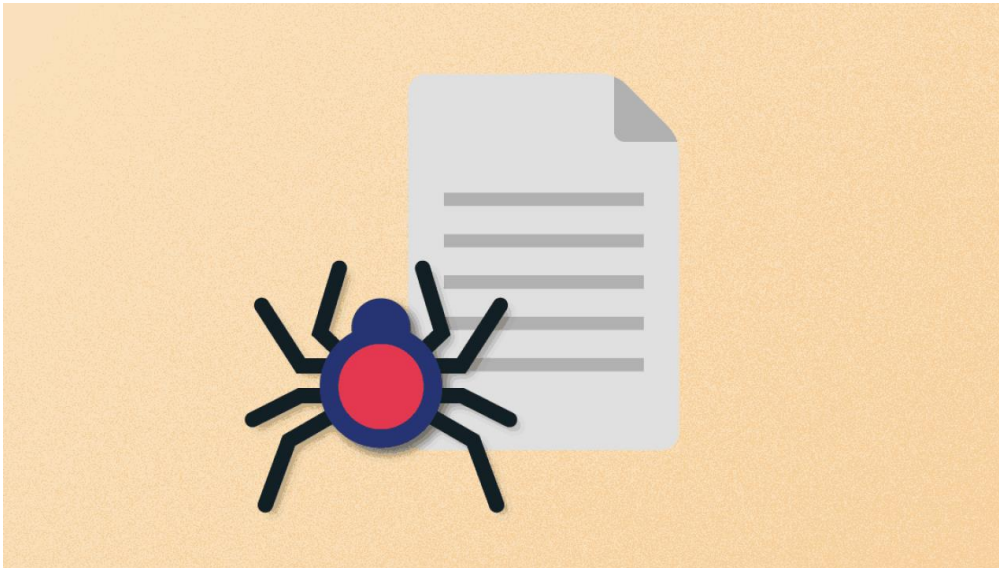
- 데이터셋은 중복되거나 거의 중복된 데이터를 포함할 수 있음
 - 여러 데이터를 합칠 때 발생할 가능성이 높음
 - 한 사람이 여러 이메일 주소를 가지는 경우
- 데이터 정제 (data cleaning)
 - 중복 데이터 문제를 해결하는 방법

데이터 수집 계획

- 고려 사항
 - 수집 가능 여부 (보유 기관이 정책 등)
 - 수집(업데이트) 주기 결정 (일회성, 한시간/하루/한달에 한번 등)
 - 획득 및 통신 비용 (무료, 유료, 통신 비용 등)
 - 센서 데이터의 활용 가능성 (데이터 포맷 변경 처리 작업 고려)
 - 정답 데이터 셋 확보 여부 (예측 모델의 정확도 측정)

데이터 수집 방법

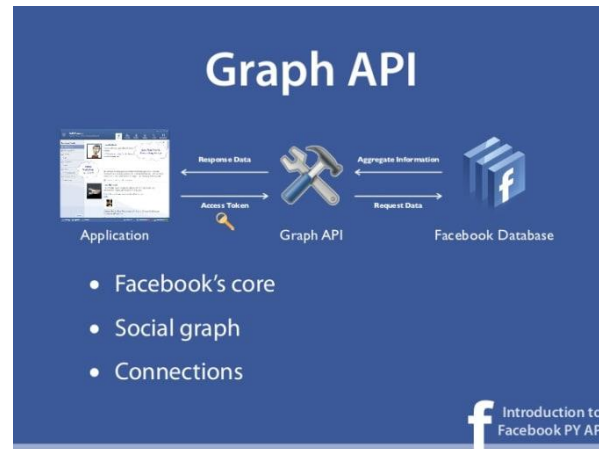
- 웹 사이트, 파일전송프로그램, 직접 전달, API (Application Program Interface)
- 데이터 크롤링(Crawling)
 - 프로그램을 자동화 하여 데이터를 수집



<https://juraganonline.org/apa-itu-web-crawler/>

API를 이용한 데이터 수집

- API (Application Program Interface)
 - URL 형태로 데이터 소재지 주소를 알려주고 데이터를 읽는 방법으로서 사용
- 대부분 큰 규모의 사이트에서 API 형태로 데이터를 제공
 - 네이버, 다음: 상위 검색어, 뉴스, 검색 결과, 지도 정보, 카페 정보 등
 - 소셜 네트워크: 자기 정보, 친구 정보, 특정 키워드가 있는 글 등
- 일반적인 API 방식은 데이터량을 제한
- API를 사용하려면 사용자를 등록하고 권한 허용을 위한 고유 키를 발급 받아야 함



데이터 마켓

- 데이터 거래 장터
- 데이터 수집 및 관리 비용 절감
- 데이터 마켓 예시
 - 캐글 (Kaggle): 데이터 분석 공모전 위주로 운영
 - data.go.kr: 우리나라 정부 공공 데이터
 - data.gov: 미국 정부 공공 데이터

kaggle™



데이터 저장 기술

구분	설명	비고
Relational Data Base	<ul style="list-style-type: none">관계형 데이터의 저장, 수정, 관리SQL 구분을 통하여 데이터베이스의 생성, 수정 및 검색 서비스 제공	<ul style="list-style-type: none">OracleMSSQLmySQL
NoSQL (Not-Only SQL)	<ul style="list-style-type: none">처리해야 할 데이터의 용량이 늘어나면서 확장성과 비용을 고려하여 등장하게 됨데이터 사이즈가 점점 커지는 문제를 해결하기 위해 key-value, document key-value, column 기반의 NoSQL 구조가 발전됨	<ul style="list-style-type: none">MongoDBCassandraHbase
분산파일 시스템	<ul style="list-style-type: none">분산된 서버 기반 파일 읽기/쓰기 등 연산을 지원하는 대규모 데이터 저장소수 TB ~ 수백 PB 이상의 데이터 저장 지원	<ul style="list-style-type: none">HDFS (Hadoop File System)Apache Spark

다음 영상에서 배울 내용

- 개발 환경 구축
 - Anaconda
 - Jupyter Notebook

수고하셨습니다.