
파이썬 기반
빅데이터 처리
및 분석기술
Crayon

이번 시간에는

1	데이터 분석 방법의 이해와 Colab 활용 방법
2	Numpy 이해하기
3	Pandas 기초 실습
4	Pandas 심화 실습
5	빅데이터 시각화 (Matplotlib)
6	빅데이터 회귀분석
7	빅데이터 분류분석
8	빅데이터 군집분석

이번 시간에는

학습 목표

- 빅데이터의 개념을 설명할 수 있다.
- 복잡하고 거대한 데이터의 처리에 대한 절차와 방법에 대해서 설명할 수 있다.
- Colab을 이해하고 코드 수행에 활용할 수 있다.

이번 시간에는

1 데이터 분석 방법의 이해

- 빅데이터 개념
- 데이터 유형
- 데이터 분석 및 처리 과정

1. 빅데이터의 개념

1) 빅데이터란?

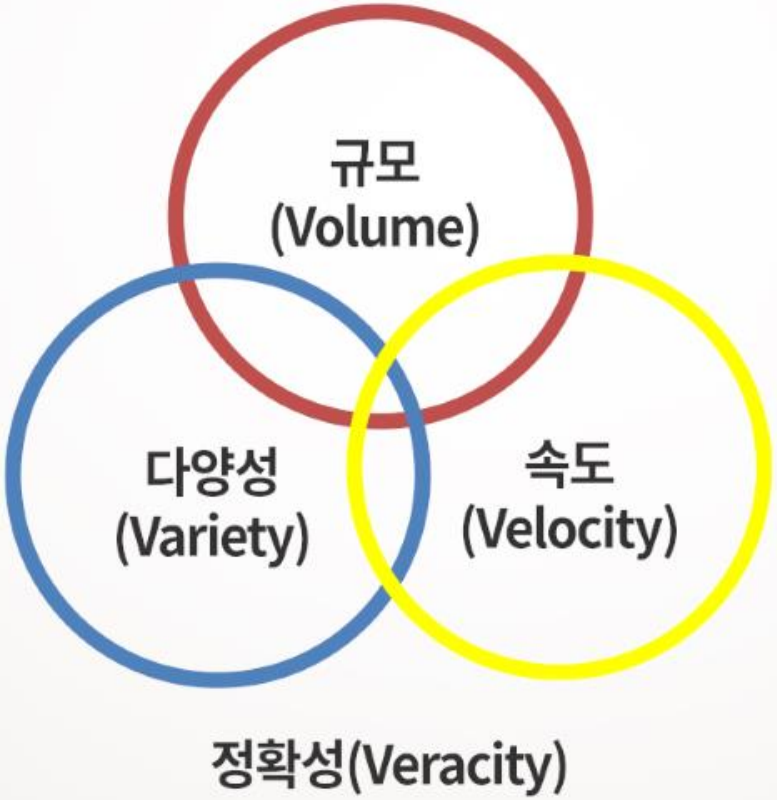
◆ 정의

- 기존 '관계형 데이터베이스 관리 시스템(RDBMS: Relational Database Management System)'과 같은 전통적 데이터 관리 기법으로 다루기 힘든 데이터
- 새로운 솔루션이 필요

1. 빅데이터의 개념

1) 빅데이터란?

◆ 특징



1. 빅데이터의 개념

1) 빅데이터란?

◆ 처리 단계

수집 → 선별 → 저장 → 검색 → 시각화

◆ 처리의 어려움

- 데이터의 다양성
- 단일화, 획일화된 처리 불가능

1. 빅데이터의 개념

2) 데이터 사이언스

◆ 요구 사항

빅데이터를 다루는 능력

+

컴퓨팅, 알고리즘, 머신러닝에 대한 구축 경험

+

R, 파이썬, 자바, 하둡 등의 도구 활용 능력

1. 빅데이터의 개념

2) 데이터 사이언스

◆ 파이썬 언어

- 데이터 과학용 라이브러리와 전문화된 SW 지원
- 데이터 과학에 잘 맞는 프로그래밍 언어

2. 데이터 유형

- 구조적 데이터
- 비구조적 데이터
- 자연어 데이터
- 기계 생성 데이터
- 그래프 기반 데이터
- 오디오·비디오·이미지 데이터
- 스트리밍 데이터

2. 데이터 유형

1) 구조적 데이터

◆ 정의

- 여러 개의 단순 데이터가 어떠한 구조를 가지고 모여서 이루어진 복합적인 데이터

◆ 단점

- 데이터 모델에 의존적임
- 고정된 필드를 가진 레코드에 기반함

예

DB용 테이블
엑셀 등

[출처01] Getty Images Bank

2. 데이터 유형

2) 비구조적 데이터

◆ 정의

• 데이터 모델에 맞지 않은 데이터

예

이메일 등

- 발신자, 제목, 본문과 같은 구조적 요소 포함
- 내용 자체는 비구조화적 요소로 이루어짐

[출처02] Getty Images Bank

2. 데이터 유형

3) 자연어

◆ 정의

- 인간이 일상적으로 사용하는 언어

◆ 단점

- 언어학에 대한 지식이 필요함
- 컴퓨터로 처리하기 까다로움

[출처03] Getty Images Bank

2. 데이터 유형

3) 자연어

◆ 처리 방법

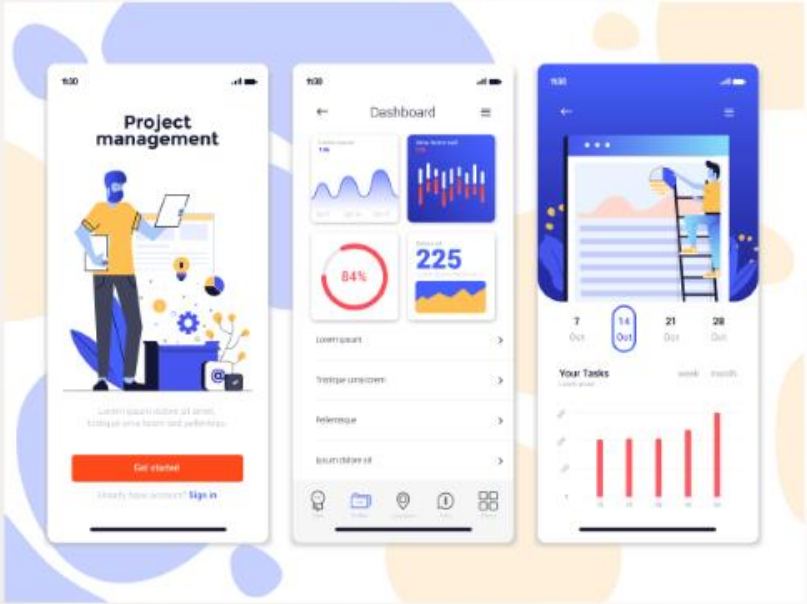
- 주제 파악, 요약문 작성, 텍스트 완성, 정서 분석 등 일부 성공
  - 특정 주제에 맞는 모델이 대다수
  - 일반화 모델 개발에 한계가 있음

2. 데이터 유형

4) 기계 생성 데이터

◆ 정의

- 사물인터넷을 통해 생성되는 데이터



[출처04] Getty Images Bank

예

웹서버 로그
네트워크 이벤트 로그
원격 검침 데이터 등

2. 데이터 유형

4) 기계 생성 데이터

◆ 특징

- 많은 양의 데이터가 빠른 속도로 생성됨
- 빠른 처리와 분석을 통한 활용 기술이 필요

2. 데이터 유형

5) 그래프 데이터

◆ 정의

- 수행하고자 하는 작업을 연산들의 의존 관계 및 선후 관계에 맞추어 **그래프 형식**으로 나타내는 것

◆ 표현 방법

- 그래프 이론의 노드, 에지, 가중치로 표현

예

SNS 친구 사이의 영향력 평가 등

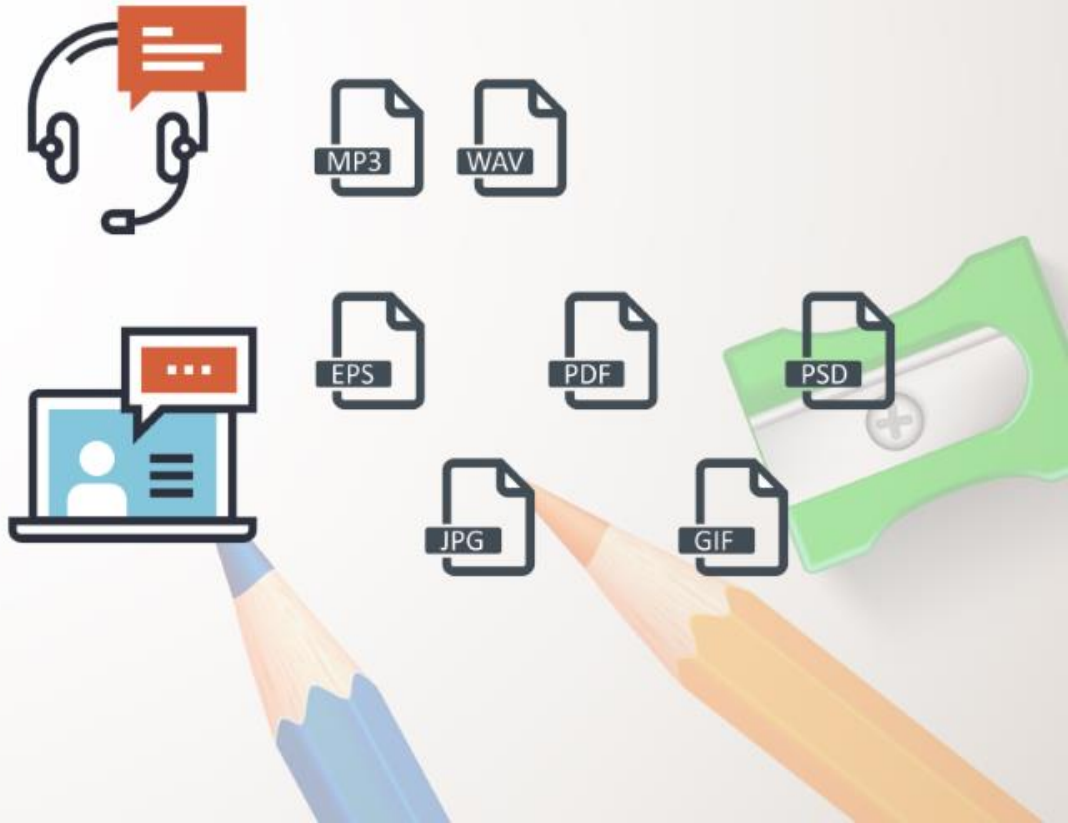
[출처05] Getty Images Bank

2. 데이터 유형

6) 오디오 · 이미지 데이터

◆ 특징

- 효과, 목소리 클립, 음악을 포함한 **소리들의 단편적인 모음**들
- 처리하기 까다로운 데이터 유형 중 하나



◆ 이미지 데이터의 경우

- 사물과 사람의 구분에 대한 처리가 쉽지 않음
- 딥러닝 학습 알고리즘 개발로 사진에서 동물과 사람의 구별 가능

[출처06] Getty Images Bank

2. 데이터 유형

7) 스트리밍 데이터

◆ 정의

- 일괄적으로 데이터 저장소에 저장되지 않고 **사건이 발생할 때마다** 시스템에 입력되는 데이터

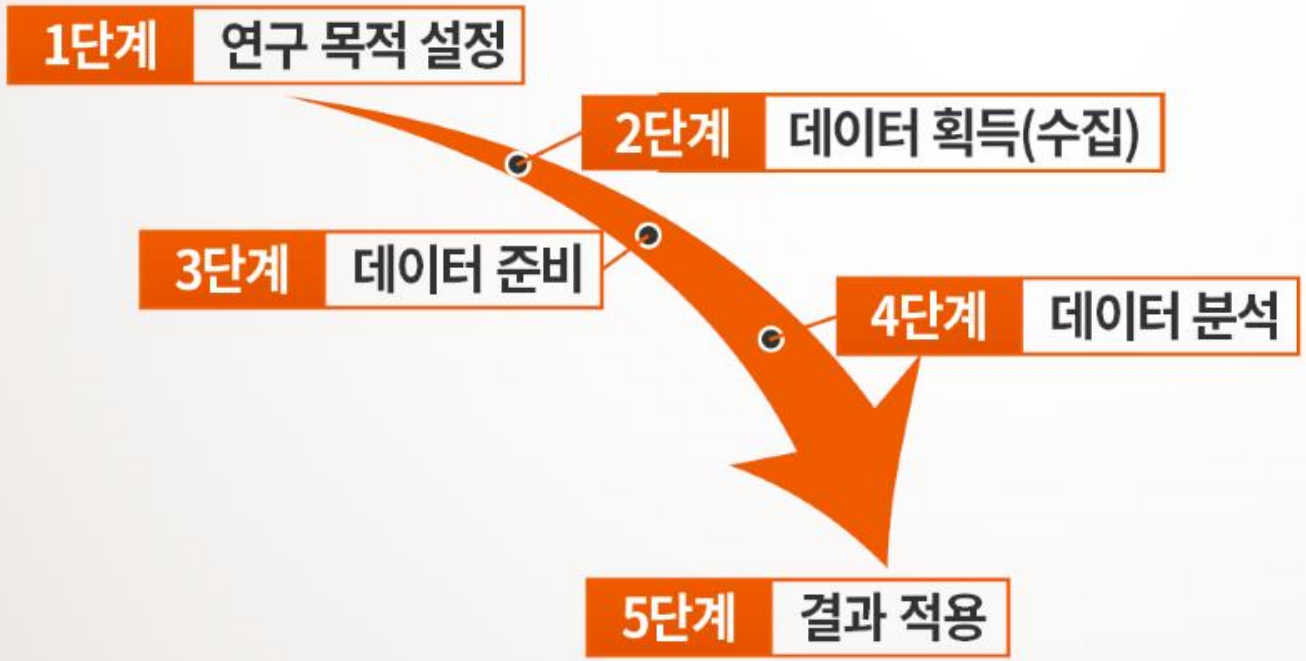
[출처07] 픽사베이

예

실시간 트렌드 데이터
운동 경기 생중계 데이터
주식 시장 데이터 등

3. 데이터 분석 및 처리 과정

1) 진행 단계



3. 데이터 분석 및 처리 과정

1) 진행 단계

◆ 1단계 : 연구 목적 설정

- 무엇을 조사할 것인가?
- 결과물로 어떤 이익을 낼 것인가?
- 어떤 데이터와 자원이 필요한가?
- 일정 및 산출물은 어떻게 되는가?

3. 데이터 분석 및 처리 과정

1) 진행 단계

◆ 2단계 : 데이터 획득(수집)

- 프로그램에서 사용할 데이터가 존재하는가?
- 데이터의 품질은 어느 정도인가?
- 데이터에 대한 접근이 가능한가?
  - 타사에서 얻을 수 있는가?
    엑셀 파일로 존재하는가?

3. 데이터 분석 및 처리 과정

1) 진행 단계

◆ 3단계 : 데이터 준비

- 데이터의 품질을 높여 데이터를 원활히 사용할 수 있게 해주는 단계
- 3개의 하위 단계

데이터

정제	데이터 출처로부터 잘못된 데이터를 제거하는 단계
통합	데이터 보충을 위해 여러 출처의 데이터를 결합
변환	의미나 내용을 바꾸지 않고 모양 및 포맷 변화

3. 데이터 분석 및 처리 과정

1) 진행 단계

◆ 4단계 : 데이터 분석

유형

기초 통계 분석	평균 분산 등 데이터 분포를 파악
클러스터링	비슷한 성격의 항목들을 그룹핑하기
연관 관계 분석	같이 자주 발생되는 패턴 찾기
분류	카테고리 중 어디에 속하는지 판별하기
예측	추가, 매출 등 수치를 예측하는 것

3. 데이터 분석 및 처리 과정

1) 진행 단계

◆ 4단계 : 데이터 분석

- 데이터 내에 존재하는 변수들의 상호 작용, 데이터의 분포, 이상점 존재 유무 등을 점검해야 함
- 평가 기준을 고려해야 함

3. 데이터 분석 및 처리 과정

1) 진행 단계

◆ 5단계 : 결과 적용

- 분석 모델을 실제 상황에 적용하는 단계
- 실제 적용 전 시뮬레이션을 수행
- 시각화 활용
  - 분석 결과를 쉽게 이해할 수 있도록 시각적으로 표현하고 전달하는 과정