빅데이터시스템

강의 과정에 대한 이해

education.drone@gmail.com
by sEOK-iLL, jEONG

목 차

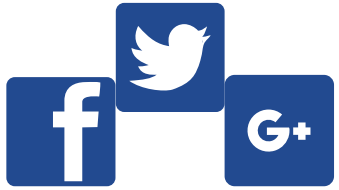
1. 빅데이터

BigData

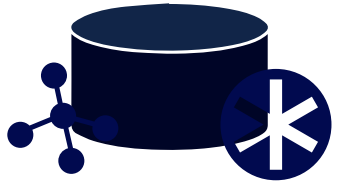
- 빅데이터의 의의
  - 기존의 DataBase 를 넘어서는 대량의 정보를 찾고 결과를 분석하는 기술
  - 정해진 규칙만으로 존재하는 데이터가 아닌 비정형으로 수시로 가치창출

소비자 행태에 대한 정보수집

개인형 맞춤정보 제공 필요



소셜 플랫폼 사용 증가

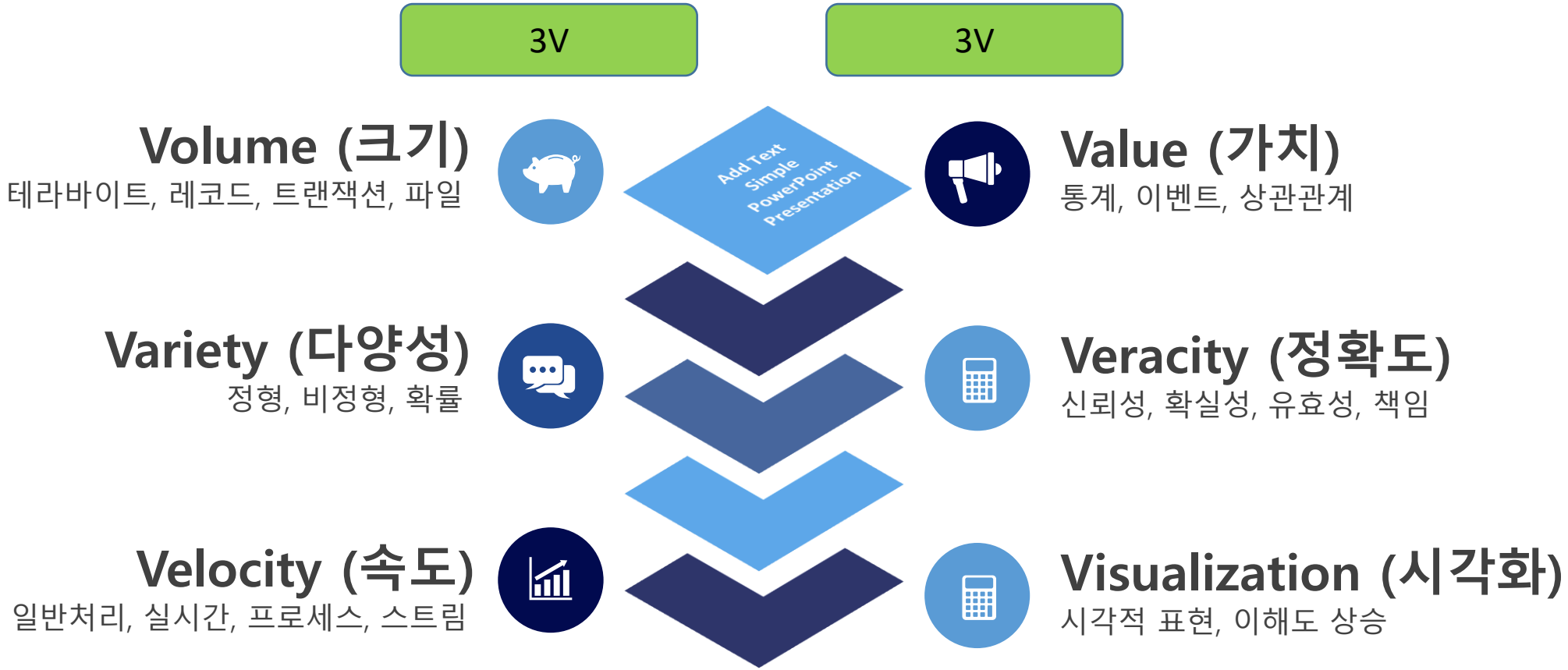


비정형 데이터 폭증

1. 빅데이터

BigData 구성요소 (6V)

※ 기업에게는, 빅데이터가 생산성향상, 비용절감, 고객서비스개선, 수익증대, 혁신, 빠른시장대응 등의 경영상 중요정보 제공



2. 빅데이터시스템

빅데이터시스템

- 시스템의 필요성
  - 인터넷을 통해 기하급수적으로 데이터 생산(하루에 25억 기가바이트)
  - 많은 데이터를 활용하여 수집하고, 분석하여 비즈니스 가치를 생성할 필요
  - 데이터를 중심으로 모든 생활 기반 사업들이 4차 산업혁명을 주도
  - 기업들은 데이터전략을 만들고 이를 활용하기 위한 전문가 양성에 박차
- 시스템의 종류

구 분	구글	하둡생태계
대용량분산처리	Map/Reduce(2004)	Hadoop(2006)
배치	Sawzall(2005)	Pig, Hive(2008)
키/밸류 엔진	BigTable(2006)	HBase(2008)
온라인쿼리	Dremel(2010)	Impala(2012)
대용량파일시스템	-	Hadoop 2.0(2013)

출처:빅데이터시스템(제이펍, 황세규)

2. 빅데이터시스템

하둡(Hadoop: High-Availability Distributed Object-Oriented Platform)

- 하둡이란

  - 빅데이터 처리를 위해 최적화된 플랫폼 제공
  - Java 언어로 개발
  - 클러스터에서 사용할 수 있는 분산파일시스템과 분산처리시스템을 제공
  - 아파치 소프트웨어 재단의 오픈 소스 프레임워크
  - 하둡생태계를 통해 다양한 기능 확장

- 장단점

장 점	단 점
▪ 오픈소스로 라이선스에 대한 비용 부담이 적음 ▪ 시스템을 중단하지 않고, 장비의 추가가 용이(Scale Out) ▪ 일부 장비에 장애가 발생하더라도 전체 시스템 사용성에 영향이 적음(Fault tolerance) ▪ 저렴한 구축 비용과 비용대비 빠른 데이터 처리 ▪ 오프라인 배치 프로세싱에 최적화	▪ HDFS에 저장된 데이터를 변경 불가 ▪ 실시간 데이터 분석 같이 신속하게 처리해야 하는 작업에는 부적합 ▪ 너무 많은 버전과 부실한 서포트 ▪ 설정의 어려움

출처:위키독스(https://wikidocs.net/22654)

2. 빅데이터시스템

MongoDB

※ NoSQL(Not Only SQL) : 관계형 데이터베이스가 아닌 다른 형태의 데이터 저장 기술
※ Document Oriented : 키식별자와 밸류 형식의 데이터 저장 방식을 사용

▪ 몽고DB란

- 빅데이터 처리를 위해 NoSQL 타입의 도큐먼트지향 데이터베이스 시스템
- Json 형태로 데이터를 저장
- 스키마를 자주 변경 해야 하는 대량의 데이터를 처리하는 데 강한 특징
- 뉴욕시에 기반을 둔 10gen(현재는 MongoDB)에서 만든 자유-오픈 소스 DB
- 응답속도와 안정성 측면에서 우수

▪ 장단점

장 점	단 점
▪ 사용방법이 쉬움 ▪ 스키마 리스 구조로 데이터 모델 변경 및 필드 확장이 용이하고 다양한 형태의 데이터를 저장 ▪ 쿼리 프로세싱이 단순화 되어 있어 대용량 데이터 처리 성능이 향상 ▪ 대용량 데이터 저장이 가능	▪ 많은 인텍스 사용 시 메모리 가용성 확보 필요 ▪ 데이터 중복에 의해 데이터 일관성이 저하되고 용량이 증가 ▪ 트랜잭션 지원이 RDBMS 대비 미약 ▪ 스키마에 대한 메타 데이터가 없어 스키마 확보를 위해 전체 도큐먼트를 조사할 필요가 있음

출처:네이버블로그(https://blog.naver.com/classmethodkr/222822485338)

3. 우리가 할 일

RoadMap

Hadoop설치

- ✓ VM 셋업
- ✓ JDK
- ✓ Python
- ✓ Hadoop Engine
- ✓ Spark Engine
- ✓ Zeppelin

빅데이터분석

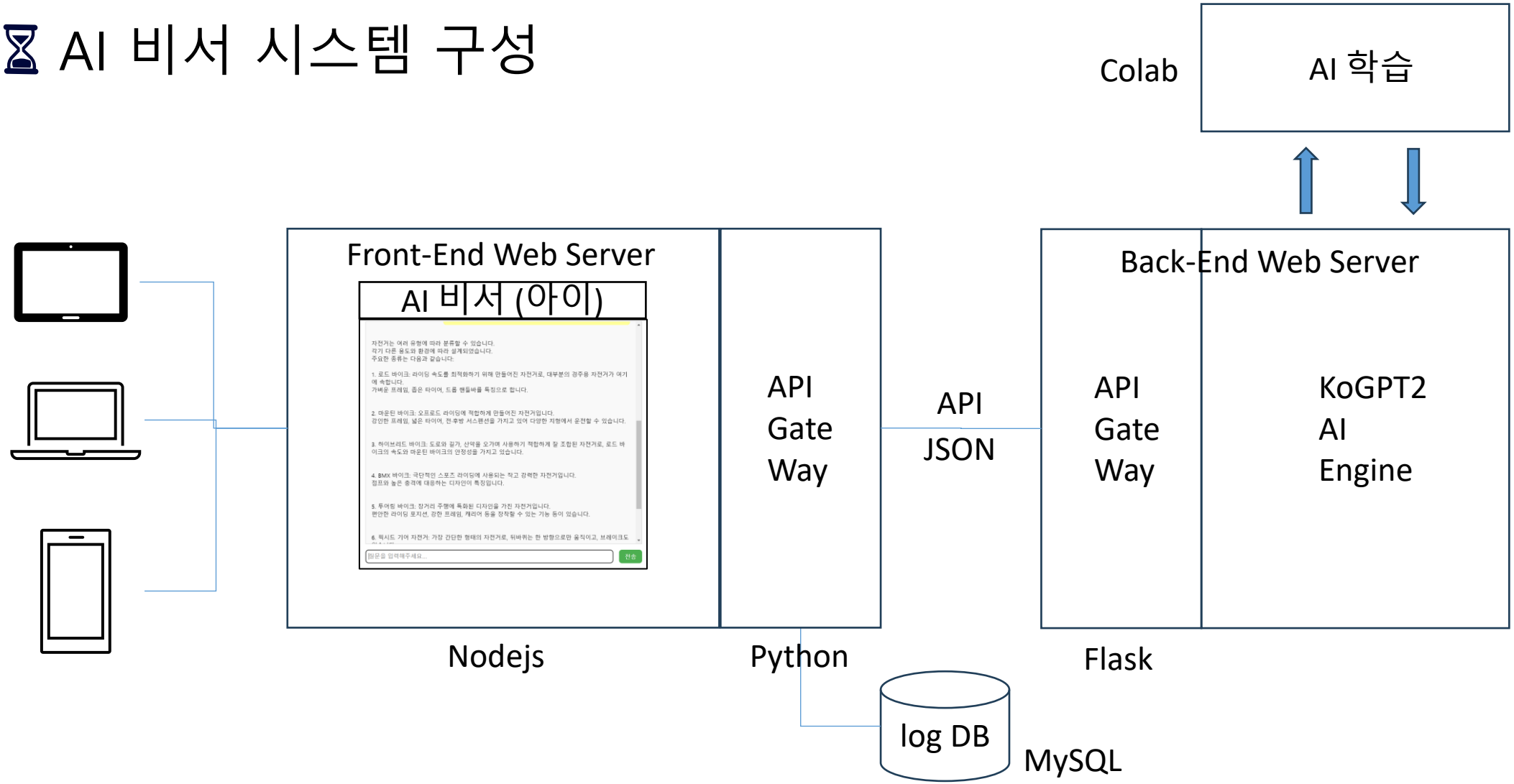
- ✓ 빅데이터 산업의 이해
- ✓ 파이썬 프로그래밍
- ✓ 크롤링
- ✓ 통계분석
- ✓ 텍스트빈도분석
- ✓ 지리정보분석
- ✓ 회귀분석/분류분석
- ✓ 텍스트마이닝

AI 비서학습

- ✓ 챗봇 데이터 수집
- ✓ Flask 웹서버
- ✓ Nodejs API 연동
- ✓ KoGPT2 환경구성
- ✓ Colab을 이용한 학습
- ✓ 말풍선생성기 활용
- ✓ MySQL
- ✓ 챗봇 비서 만들기

3. 우리가 할 일

AI 비서 시스템 구성



4. 전달 사항

학습 목표

- 빅데이터를 구성하는 시스템과 빅데이터 분석에 대한 이해와 실습을 통한 정확한 문제의 정의와 창의적 능력을 도출
- 데이터 분석의 기초를 이해하고 이를 활용하여 서비스를 구축하는 데 필요한 요건과 해결방안을 학습하는 능력 배양
- 데이터분석에 필요한 서비스 플랫폼 구축에 따른 원리에 대하여 기본적 이해와 관련 솔루션을 활용한 운용 기술 습득

학습 내용

- 하둡설치 (VM 셋업, JDK, Python, Hadoop Engine, Spark Engine, Zeppelin)
- 빅데이터분석 (빅데이터 산업의 이해, 파이썬 프로그래밍, 크롤링, 통계분석, 텍스트빈도분석 등)
- AI 비서 학습 (챗봇 데이터 수집, Flask 웹서버, Nodejs API 연동, KoGPT2 환경구성, Colab 등)

4. 전달 사항

교재

주교재	- PowerPoint 로 만든 pdf 자료 - 데이터 과학 기반의 파이썬 빅데이터 분석 (이지영 지음, 한빛아카데미)
부교재	- 필요 시, 영상 공유

4. 전달 사항

학사 일정

- 3/6 : 1학기 개강
- 3/2 ~ 3/3 : 신입생 오리엔테이션
- 4/24 ~ 4/28 (8주차) : 중간고사
- 5/4 : JEIU 축제
- 6/12 ~ 6/16 (15주차) : 보강주
- 6/17 : 개교기념일
- 6/19 ~ 6/23 (16주차) : 기말고사
- 6/26 : 하계방학
- 9/4 : 2학기 개강
- 9/4 ~ 9/15 : 수강신청 정정기간
- 10/23 ~ 10/27(8주차) : 중간고사
- 10/27 : JEIU 체육대회
- 12/11 ~ 12/15(15주차) : 보강주
- 12/18 ~ 12/22(16주차) : 기말고사
- 12/25 : 동계방학

4. 전달 사항

수업 방식

- 주간 : 대면 수업이 원칙 (비대면 수업 불가, 보강도 대면)
- 전일제 : 비대면 수업으로 진행 하되, 중간평가, 기말평가 중 1회 이상 대면수업(시험)
  그외 대면 수업 필요 시, 학생 의견 반영

수업 진행

- 대면 : 일반적인 강의, 실습

- 비대면 : Zoom 회의(저번주 복습, 수업내용 개략 설명) + 동영상 강의 + 실습(집 PC 활용)
  ※ 출석인정 : 동영상시청 (Zoom실시간 수업 참여)
  ※ 질문사항은 Zoom 회의나 오픈 채팅방 이용
  ※ 영상은 수업시간 50% 이상

4. 전달 사항

강의 일정 (1/3)

일자 (월 / 일)	교육 내용	반 구분				비 고
		A / B 반		C 반		
		대면	비대면	대면	비대면	
9 / 8	- 본 강의에 대한 오리엔테이션 - 4차 산업혁명과 데이터과학 * 4차산업혁명의 이해, 4차 산업혁명을 실현하는 데이터 과학 등	√			O	1주차
9 / 15	- 빅데이터의 이해와 활용 * 빅데이터의 이해, 빅데이터의 활용 - 하둡(Hadoop 설치) 과 Mongo DB * 빅데이터시스템에 대한 기초 사항을 숙지하고 환경 설정	√			O	2주차
9 / 22	- 데이터 과학 기반의 빅데이터 분석 * 빅데이터 산업의 이해, 빅데이터 분석 방법과 접근법 등	√			O	3주차
9 / 29 (보강주순연)	- 추석연휴	√			O	4주차 휴무보강
10 / 6	- 데이터 분석을 위한 파이썬 프로그래밍 * 변수와 객체, 자료형과 연산자, 조건문과 반복문, 함수, 파일처리 - 파이썬 플라스크(Flask)를 활용한 웹서버 구축	√			O	5주차

4. 전달 사항

강의 일정 (2/3)

일자 (월 / 일)	교육 내용	반 구분				비 고
		A / B 반		C 반		
		대면	비대면	대면	비대면	
10 / 13	- 파이썬 크롤링 – API 이용 * 네이버 API를 이용한 크롤링, 공공데이터 API 기반 크롤링	√			O	6주차
10 / 20	- 파이썬 크롤링 – 라이브러리 이용 * 정적 웹 페이지 크롤링, 동적 웹 페이지 크롤링 - Nodejs 를 활용한 웹서버와 플라스크 연동 API 응용	√			O	7주차
10 / 27	- 중간 평가	√		√		8주차
11 / 3	- 빅데이터 분석을 위한 인공지능 챗봇 비서 구현 * KoGPT2 를 활용한 챗봇 환경 구축	√			O	9주차
11 / 10	- 통계 분석 : 기술통계분석과 그래프, 상관분석과 히트맵 - Colab 을 활용한 인공지능 챗봇 비서 학습 데이터 수집 및 학습	√			O	10주차
11 / 17	- 텍스트 빈도 분석 : 영문분석/한글분석 및 워드클라우드 - 말풍선 생성기 활용	√			O	11주차

4. 전달 사항

강의 일정 (3/3)

일자 (월 / 일)	교육 내용	반 구분				비 고
		A / B 반		C 반		
		대면	비대면	대면	비대면	
11 / 24	- 지리정보 분석 * 주소데이터분석 및 지오맵, 행정구역별 데이터분석 및 블록맵 - 인공지능 챗봇 비서 강화 학습	√			O	12주차
12 / 1	- 회귀분석 * 선형 회귀분석 및 그래프, 회귀분석 및 그래프 - 인공지능 챗봇 비서 질의/응답 자동화를 위한 DB 환경 구축	√			O	13주차
12 / 8	- 분류분석 * 로지스틱 회귀분석, 결정 트리 분석 - 인공지능 챗봇 비서 활용-I	√			O	14주차
12 / 15 (보강주)	- 텍스트 마이닝 * 감성분석 및 토픽 모델링, 감성분석 및 바차트 - 인공지능 챗봇 비서 활용-II	√			O	15주차
12 / 22	- 기말 평가	√		√		16주차

4. 전달 사항

수업 도구

- 교과목 카톡방 : 공지사항, Zoom URL 안내, 출석, 질의응답 등
- Zoom : 비대면 출석 체크, 수업
- LMS : 교과목 공지사항, 자료 업로드/다운로드, 동영상 시청, 과제 제출 등

출석 인정

- 대면수업주차는 실제 출석, 비대면수업주차는 동영상 시청(zoom실시간수업)

과제

- 과제의 종류에 따라, 다음 대면 수업 시, 과제 발표 또는 제출 시점까지 제출
- 실습과제 내용 제출 시기 : 기말고사 이전까지 (기말고사 이후 제출 시, 미인정)

4. 전달 사항

평가항목 및 배점

출석	중간평가	기말평가	기타(과제/태도)
20	30	30	20

- 1회 결석 시 -5점,
  4회 결석이면 F

- 대면수업 주차 :
  실제 출석 및 지각 체크

- 비대면수업 주차 :
  학기말에 주차별 동영상
  시청 여부

- 강의실에서 대면평가

- 과제에 따라, 수시점수부여
- 실습과제는 기말고사전,
  제출한 과제 확인 후,
  점수 부여

참고 자료

- 자바와 파이썬으로 만드는 빅데이터시스템(제이펍, 황세규)
- 위키독스(https://wikidocs.net/22654)
- 네이버블로그(https://blog.naver.com/classmethodkr/222822485338)
- 데이터분석과 인공지능 활용 (NOSVOS, 데이터분석과인공지능활용편찬위원회 편)

QnA