숭실대학교 Soongsil University

빅데이터 분석 01

최대선
소프트웨어학부
숭실대학교

숭실대학교 Soongsil University

내용

- 강의 개요
- 4차 산업 혁명과 데이터 과학
- 파이썬 리뷰

강의 개요

숭실대학교 Soongsil University

목표

- 데이터 과학 및 빅데이터 분석 이론 습득
- 파이썬을 이용한 데이터 분석 능력 개발
- 실제 데이터를 활용한 프로젝트 수행 경험

과정

주차	내용	주차	내용
1	강의소개, 데이터과학, 파이썬 리뷰	9	군집분석
2	빅데이터 활용과 이해, 파이썬 라이브러리	10	지리정보 분석
3	크롤링	11	텀 프로젝트 제안 발표
4	통계분석1	12	텍스트 마이닝1
5	통계분석2	13	텍스트 마이닝2
6	회귀예측	14	딥러닝
7	분류	15	텀 프로젝트 발표
8 (4/26)	중간고사		

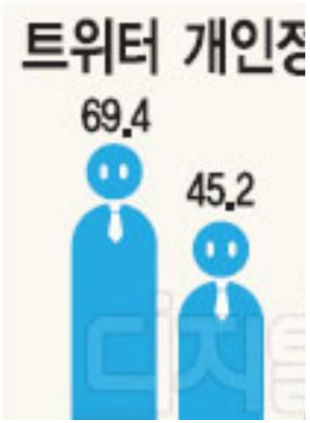
데이터 수집

- 공개 데이터
- API 활용
- 페이지 스크레이핑

SNS 개인정보 노출 심각

페북 이름 • 학교 등 조합 이용자 45% 식별...피싱 등 악용소지 커

이준기 기자 | 입력: 2013-10-13 20:00



페이스북과 트위터 등 소셜네트워크서비스(SNS)를 통한 개인정보 노출이 심각하다는 연구결과가 나왔다. SNS 사용자의 이름과 성별, 학교 정보 등이 그대로 노출돼 있어 개인 신상정보 조합만을 통해 이용자 절반 가량을 식별해낼 수 있는 것으로 확인됐다.

한국전자통신연구원(ETRI) 사이버보안연구단은 13일 빅데이터 개인정보 분석기술을 개발, 페이스북 657만개와 트위터 277만개 등 SNS 이용자 계정 934만개를 대상으로 개인정보 노출현황을 분석한 연구결과를 발표했다.

연구결과에 따르면 페이스북의 경우 성별(92%), 고등학교(47%), 혈액형(40%), 관심사(19%), 좋아하는 음악(14%) 등의 순으로 개인 신상정보

데이터 가공

- 파이썬 라이브러리
  - Pandas, numpy

통계분석

- 1 변수
  - Nominal : 명목형
  - Numeric : 수치형
- 2 변수 간 관계 분석

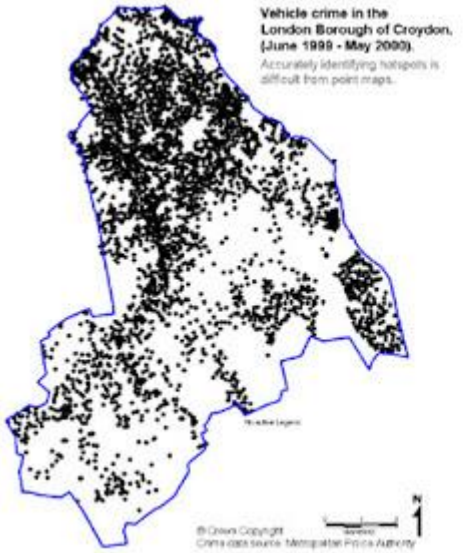
		종속변수	
		연속형 (명목형 중 순서)	비연속형 (수치형 중 이산형)
독립변수	연속	상관분석 (cor) 회귀분석 (lm)	
	비연속	평균치 비교 (t.test, aov)	도수, chisq

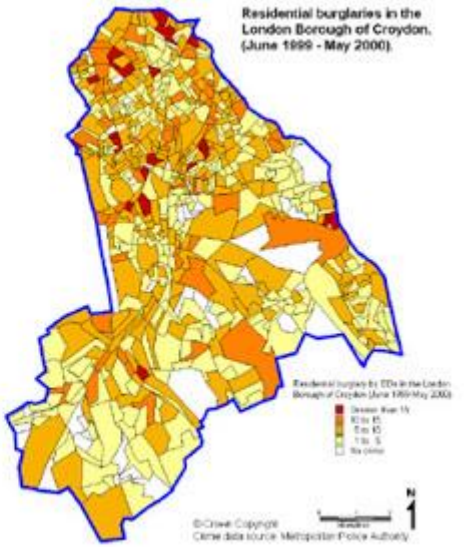
텍스트 분석

- 빈도 분석
- 임베딩
- 감성 분석

지리정보

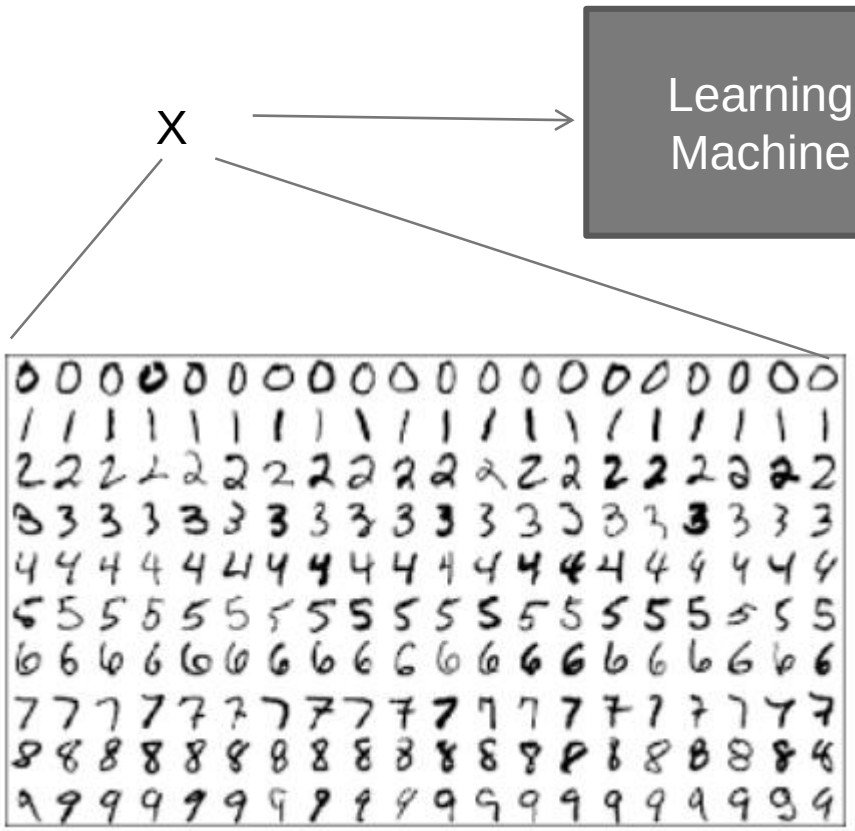
- 위치와 매출의 상관관계 등





분류, 회귀 예측

- Train

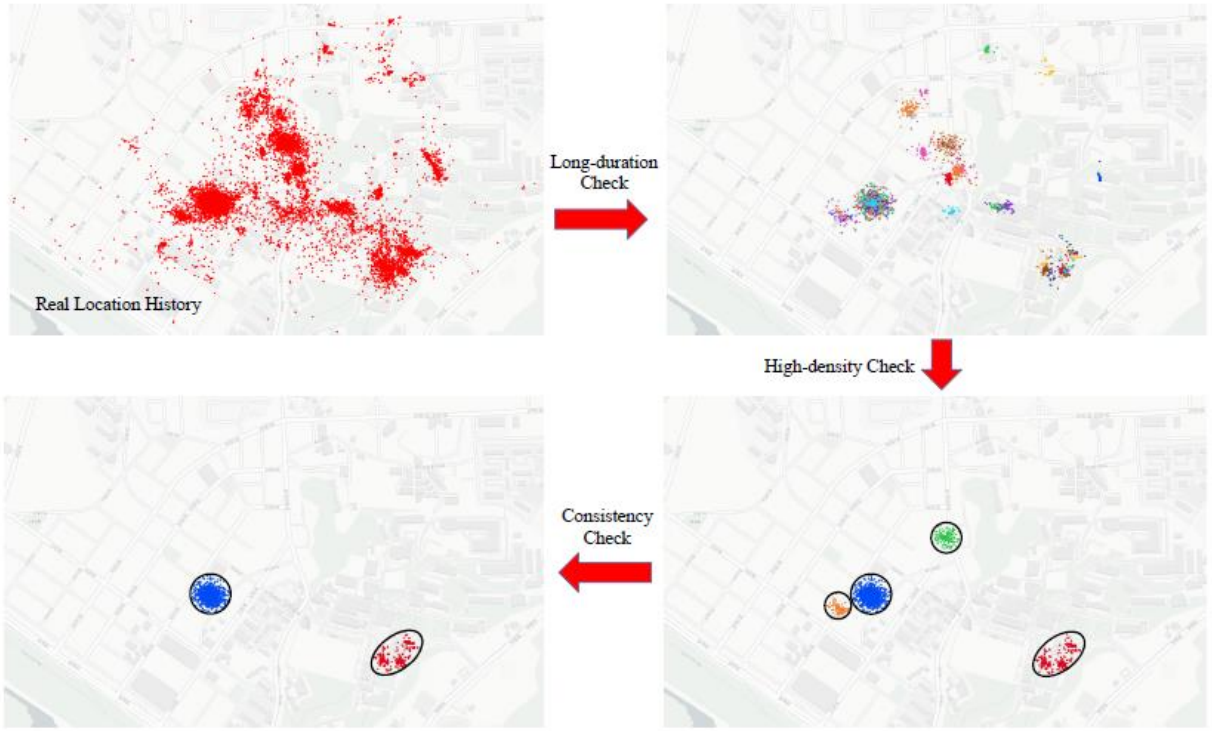


클래스 (Classification)
ex) A~ Z ,
신용불량 or not

수치 값 (Regression)
ex) 내년도 매출,
사기 가능성, ..

군집 분석

- Unsupervised learning



숭실대학교 Soongsil University

수업 방법

- 수업 : 이론, 실습, 과제 풀이
- 과제 : 수업 중 수행한 과제 + 숙제
- 퀴즈 : 2회 정도, 기존 강의 복습 차원
- 중간고사 : 실습 과제와 유사
- 텀 프로젝트
  - 개인 별로 주제를 정하고 데이터를 찾아서 분석
  - 분석 결과 = findings

평가

- 중간고사 30
- 텀 프로젝트 30
- 출석 10
- 과제 20
- 퀴즈 등 기타 10

공지 및 Q/A

숭실대학교 Soongsil University

공지

- 주로 오픈카톡방
  - https://open.kakao.com/o/gQo7rq7e
  - 반드시 실명으로 참여 바람
- 스마트캠퍼스

Q/A

- 공유할 만한 질문 : 오픈카톡방에 올리기
- 개인적 질문 : 메일로 sunchoi@ssu.ac.kr
- 스마트캠퍼스 질문은 잘 안봄

숭실대학교 Soongsil University

4차 산업혁명과 데이터 과학

01. 4차 산업혁명의 이해

- 1차 산업혁명 : 증기기관을 기반으로 한 기계화 혁명
- 2차 산업혁명 : 전기를 사용한 대량 생산 혁명
- 3차 산업혁명 : 컴퓨터와 인터넷이 보급, 지식 정보 혁명
- 4차 산업혁명 : 지능 정보 기술 혁명
  - 초연결, 초지능, 초융합

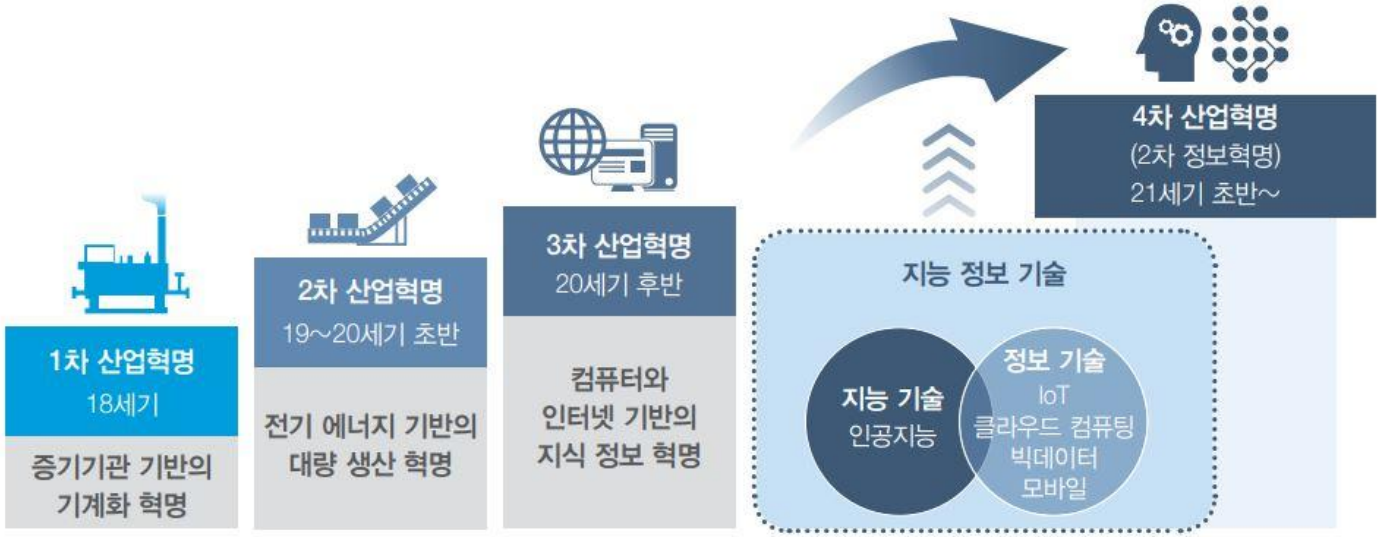


그림 1-1 산업혁명의 흐름(출처: 미래창조과학부 블로그)

01. 4차 산업혁명의 이해

초연결

- 사물과 공간, 인터넷의 상호의존성 증폭, 제품과 서비스의 연결성이 무한 확장
- 대표적 기술 : 사물인터넷, 5세대 통신(5G)

사물인터넷

- IoT : 언제나, 어디서나, 어느 것과도 연결될 수 있는 새로운 통신 환경
- RFID태그를 읽는 센서 네트워크(USN)에서 시작
- 사물과 사물 간 통신을 의미하는 M2M으로 발전
- 사람, 업무, 데이터까지 모든 것이 연결되어 상호 통신하는 만물인터넷 (IoE)으로 발전할 전망

01. 4차 산업혁명의 이해

5G

- **초고속**
  - 최대 20Gpbs , 일상적으로는 100Mbps
  - 1만배 이상 더 많은 트래픽을 수용
- **초연결**
  - 평방 킬로 미터 당 1백만 개의 기기 사용 가능
  - 배터리 하나로 10년 간 구동 가능한 고에너지 효율
- **초저지연**
  - 1ms 이하의 낮은 지연시간
  - 이동 간 제로 중단을 실현하는 고안정성

표 1-1 5G의 특징

특징	설명
초고속	초광대역 무선통신(eMBB)enhanced Mobile BroadBand • 유선과 무선의 차이가 없는 **대용량** 및 **고속**의 데이터 이용 환경 제공 • 4G에 비해 최대 20배 더 빠른 20Gbps까지 구현 가능(일상적으로는 100Mbps 보장 목표) • 모바일로 8K 콘텐츠 송수신 가능
초연결	대규모 사물통신(mMTC)massive Machine Type Communication • 산업 또는 일반인에게 IoT 사용 환경 제공 • 현재보다 최대 500배 더 많은 기기와 **고밀집** 연결 가능 • **고에너지 효율** • 스마트폰의 인터넷, PC의 인터넷을 넘어 진정한 IoT 가능
초저지연	고신뢰/초저지연 통신(uRLLC)ultra-Reliable Low-Latency Communication • 최대 1ms까지의 **낮은 지연성**과 **고안정성**을 목표로 데이터 통신 서비스의 품질(QoS)을 제공

01. 4차 산업혁명의 이해

- 인공지능(AI)
  - 1950년) 앨런 튜링의 튜링 머신 – 이미테이션 게임
  - 1956년) 다트머스대학교의 하계 컨퍼런스에서 Artificial Intelligence라는 용어가 처음 사용 (1차 전성기)
  - 1970년대) 컴퓨터의 계산 기능과 논리 체계의 한계로 인공지능 이론 구현에 실패 (1차 인공지능 겨울)
  - 1980년대) 신경망 다층 퍼셉트론 개발 (2차 전성기)
  - 신경망의 성능을 높이기 위한 학습 데이터 부족 및 계산 능력의 한계에 도달 (2차 인공지능 겨울)
    - Back propagation의 한계
  - 2000년대) 메모리, CPU, GPU 등의 하드웨어 성능 향상으로 신경망 연구가 다시 활발해짐
    - Back propagation 해결
  - 딥러닝의 성능 향상이 가속, 구글 딥마인드의 알파고가 바둑대회에서 우승(3차 전성기)
  - 2023년 생성 AI 붐 : ChatGPT 등

01. 4차 산업혁명의 이해

▶ 초지능

- 초지능 – 하이퍼 인텔리전스
  - 인간의 지능과 인공지능이 협력하여 더 스마트한 서비스를 제공
  - 인공지능 기능을 추가하여 사물을 더 스마트하게 만드는 사물의 지능화

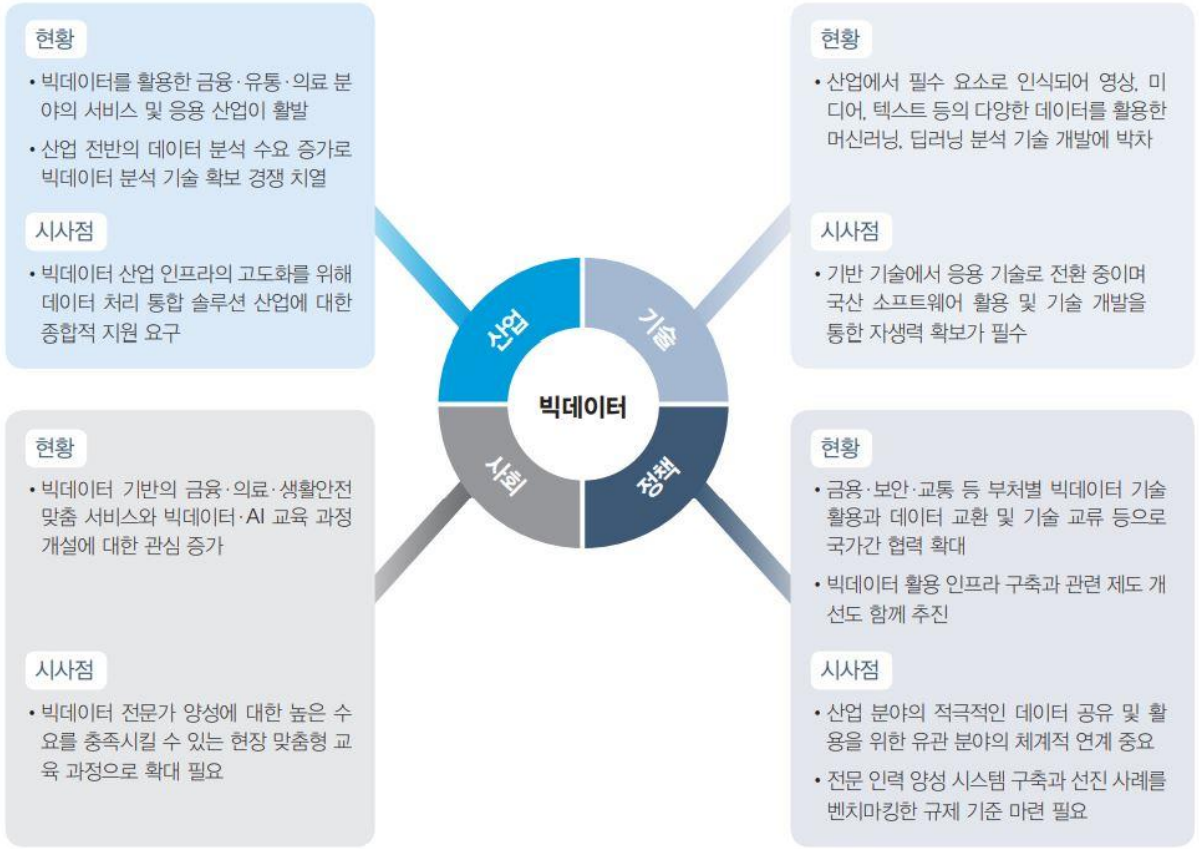
- 초지능 – 슈퍼 인텔리전스
  - 특이점 : 인공지능의 지능이 인간을 넘어섬
  - AGI (Artificial General Intelligence) : 모든 분야의 범용 인공지능

- 하이퍼 인텔리전스 -> 슈퍼 인텔리전스로 진화

01. 4차 산업혁명의 이해

▶ 빅데이터

- 디지털 환경에서 발생하는 모든 데이터를 의미
- 4차 산업 전 분야에서 분석, 활용
- 인공지능의 소스



01. 4차 산업혁명의 이해

▶ 초융합

- **디지털 트랜스포메이션**
  - 디지털 기술을 활용하여 기존 산업의 운영 및 생산의 효율성과 경쟁력을 높이는 프로세스의 변화를 의미
  - 기업은 디지털 기술을 활용하여 다양한 산업 분야에서 지속적인 혁신을 추진, 특히 제조업에 주목
  - 기존 비즈니스 모델뿐만 아니라 고객의 경험을 변화시키고 추가 수익 흐름을 창출

02. 4차 산업혁명을 실현하는 데이터 과학

숭실대학교 Soongsil University

▶ 데이터 과학

- 정형, 비정형 다양한 데이터로부터 지식과 인사이트를 추출하는데 과학적 방법론, 프로세스, 알고리즘, 시스템을 동원하는 융합분야 (위키)

▶ 데이터 과학과 IoT + 빅데이터 + AI

- IoT 를 구성하는 센서와 기기의 노드 : 감각 및 행동 기관
- 빅데이터 : 외부 센싱 데이터, 내부 처리 결과 데이터
- 인터넷/4G/5G : 인지된 자극과 명령을 전달하는 신경계
- AI : 인지된 자극을 처리하고 분석하여 명령을 내리는 두뇌
- IoT와 빅데이터, AI가 함께 선순환하며 발전하고 진화해야 함

03. 4차 산업혁명 서비스 사례

▶ 자율주행차

- 인지-판단-제어라는 3가지 단계로 동작
- 도로 환경에서 빅데이터를 수집하여 상황을 인지하고 판단한 뒤 신속하게 제어
- 핵심 기반 기술
  - 센서 (카메라, 라이다, 레이더) – 주변 환경의 빅데이터 수집 및 분석
  - 차량 부품의 빅데이터 수집 및 분석 (AI)
  - 차량 제어
- 자율주행차 5레벨

레벨	설명
레벨 0	자동화 기능이 미적용된 상태
레벨 1	운전자 보조주행: 운전자가 속도 또는 방향을 통제
레벨 2(현재)	부분적 자율주행: 차간 거리 및 속도 유지 등이 가능하지만 운전자가 주행에 적극 개입해야 하는 상태
레벨 3	조건부 자율주행: 자율주행 시스템을 운행하지만 비상시 몇 초 안에 운전자가 개입해야 하는 상태
레벨 4	고수준 자율주행: 비상시 차량이 일정 시간은 자체 대응하는 상태로 운전자가 차량 내에서 책을 읽어도 되는 수준
레벨 5	완전 자율주행: 어떠한 도로 환경에서도 무인 자율주행이 가능한 상태

03. 4차 산업혁명 서비스 사례

- 커넥티드 카
  - 정보통신기술과 자동차를 연결시킨 것으로 양방향 인터넷 및 모바일 서비스가 가능한 차량
    - V2X : V2V(Vehicle to Vehicle), V2I(Vehicle to Infrastructure), V2N (Vehicle to Nomadic Device), V2P(Vehicle to Pedestrian)
  - 차량과 도시의 모든 곳이 연결되어 스스로 위험을 감지하고 다른 자동차와의 거리나 속도를 제어하며 운전할 수 있음
  - 스스로 고장을 진단하여 필요한 조치를 취함
  - 인포테인먼트 : 영화 스트리밍 서비스나 실시간 날씨 및 뉴스 검색, 소셜 네트워크 서비스 등 다양한 운전자 맞춤형 서비스를 제공

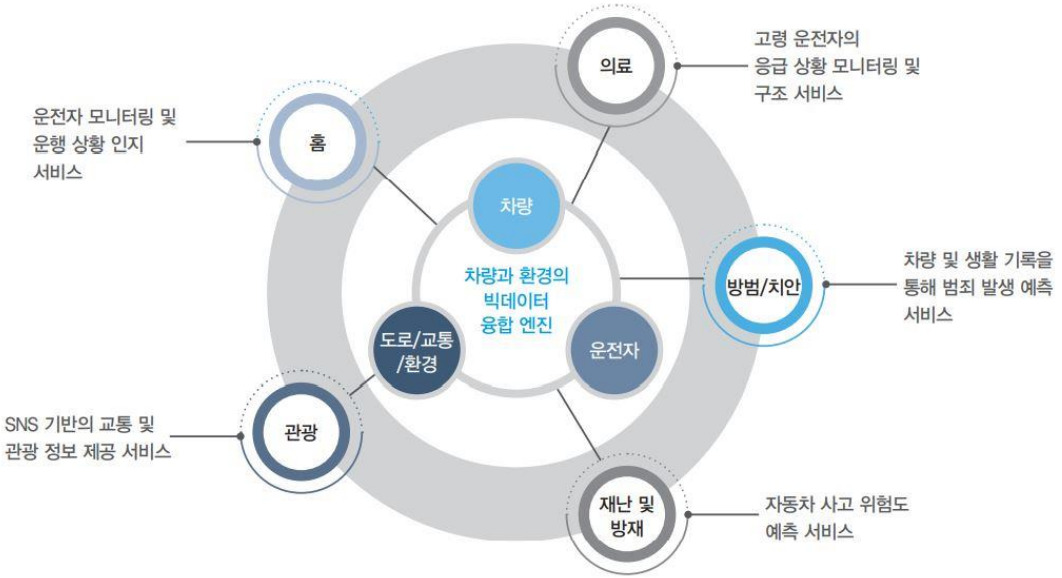


그림 1-7 커넥티드 카의 운전자 맞춤 서비스 개념도(출처: ETRI IoT 추진 계획)

03. 4차 산업혁명 서비스 사례

▶ 스마트 시티

- 도시 구성원과 시설 기관들이 네트워킹이 가능하도록 인터넷과 IoT 등의 통신 인프라가 갖춰진 것
- 빅데이터 및 AI와 융합하여 보다 편리하고 안전한 생활 및 업무 환경을 구현하는 4차 산업혁명 시대의 진화된 도시

▶ 핵심 기반 기술

- IoT와 AI, 빅데이터 분석, AR/VR/MR, 건강/교통/교육/기기제어 등의 요소 기술

표 1-4 스마트 시티의 구성 요소

구성 요소	기능
스마트 홈/사무실	• 주거, 사무실, 학교, 편의시설 등이 상호 유기적으로 연결되어 사용자 요구를 예측해서 해결 • 개인별 편의성 극대화 • 재택근무 등 업무 환경의 제한 완화
스마트 시설 관리	• 발전, 교량, 환경 등 사회 기간 시설의 실시간 관제를 통해 에너지 절약 및 운영 효율화
스마트 교통	• 교통 시설이나 도로 상황의 실시간 지능형 관제를 통해 시간 단축 및 운영 효율화 • 개인별 이동 상황에 따른 맞춤형 교통 편의 제공
스마트 교육	• 학생별 학습 수준에 따라 맞춤형 교육을 제공하는 AI 기반의 튜터링 시스템 보급
스마트 치안	• 빅데이터 분석을 기반으로 범죄, 테러, 사고 등의 징후 예측 및 예방 • 유사시 효과적인 구조 조치를 통해 안전한 생활 환경 구축
스마트 환경	• 신재생 및 청정 에너지 기술, 생활 환경의 위생 상태 측정 및 관리, 자원 재활용, 환경오염의 측정/예방/처리가 융합되어 쾌적하고 청결하며 안전한 생활 환경 구축
스마트 문화/여가	• 문화, 콘텐츠, 스포츠와 VR, AR, MR 기술이 융합하여 개인 맞춤형의 건강, 재미, 지식을 제공하는 복합적 오락, 운동, 문화 체험 환경 제공

03. 4차 산업혁명 서비스 사례

▶ 스마트 헬스 케어

- 개인의 건강에 대한 의료 정보, 기기, 시스템, 플랫폼을 다루는 산업 분야
- 건강 관련 서비스와 의료 IT가 융합된 종합 의료 서비스
- 고령화와 의료비 지출 증가라는 사회적 요인과 AI, 빅데이터, IoT, 5G 등의 기술 발전에 따라 지속적으로 성장
- 핵심 기반 기술
  - 종합 건강 정보 빅데이터 구축, 분야별 지식베이스 구축
  - 웨어러블, 원격 헬스 모니터링 : 건강 빅데이터에 대한 실시간 수집 및 분석, 진단 및 처방

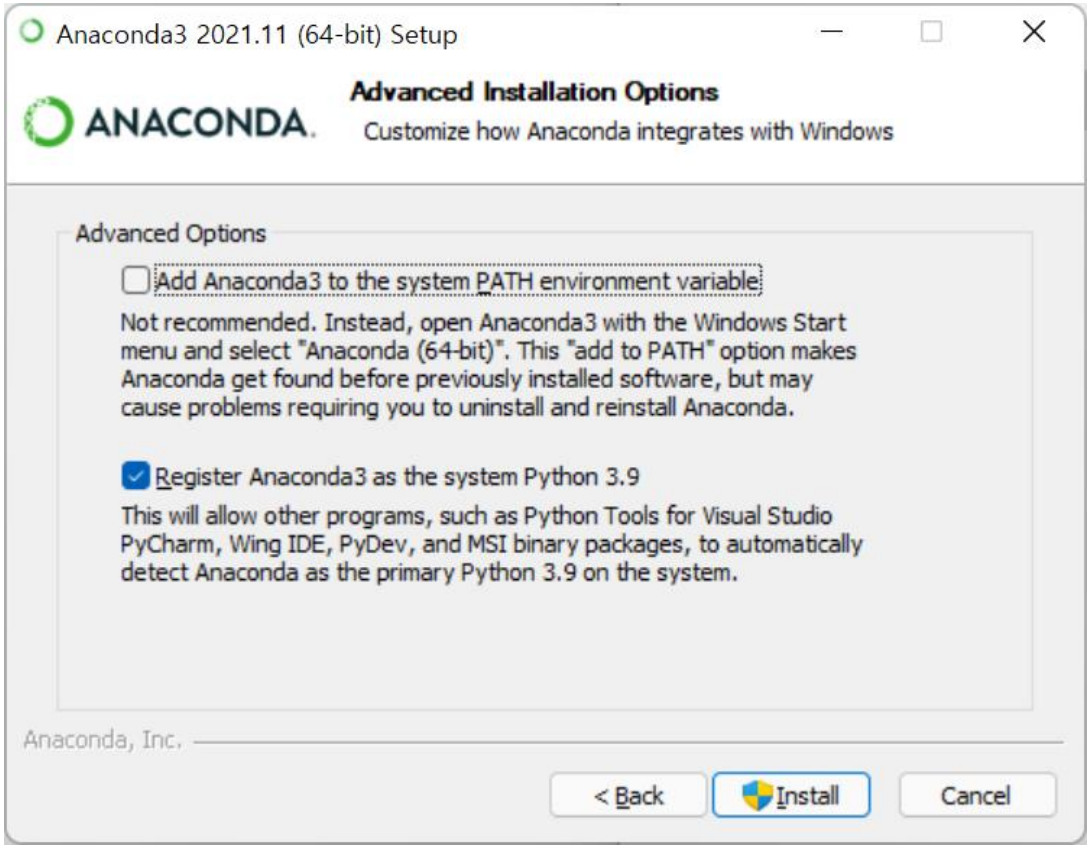
표 1-5 헬스 케어 서비스와 ICT 융합의 발전 과정

구분	Tele-헬스 케어	e-헬스 케어	u-헬스 케어	스마트 헬스 케어
시기	1999년대 중반	2000년대 초반	2000년대 후반	2010년 이후
핵심 서비스	병원 내 치료	치료, 의료 정보 제공	e-헬스 케어 + 원격 의료, 만성 질환자 관리로 질병 예방	u-헬스 케어 + 운동 및 식사량 등의 건강 생활 관리, 복지, 안전
공급자	병원	병원	병원, ICT 기업	병원, ICT 기업, 보험사, 헬스 케어 서비스 기업
주요 이용자	의료인	의료인, 환자	의료인, 환자, 일반인	의료인, 환자, 일반인
핵심 ICT 기술		초고속 인터넷	무선 인터넷	스마트 기기, 앱, AI, 빅데이터

파이썬 리뷰

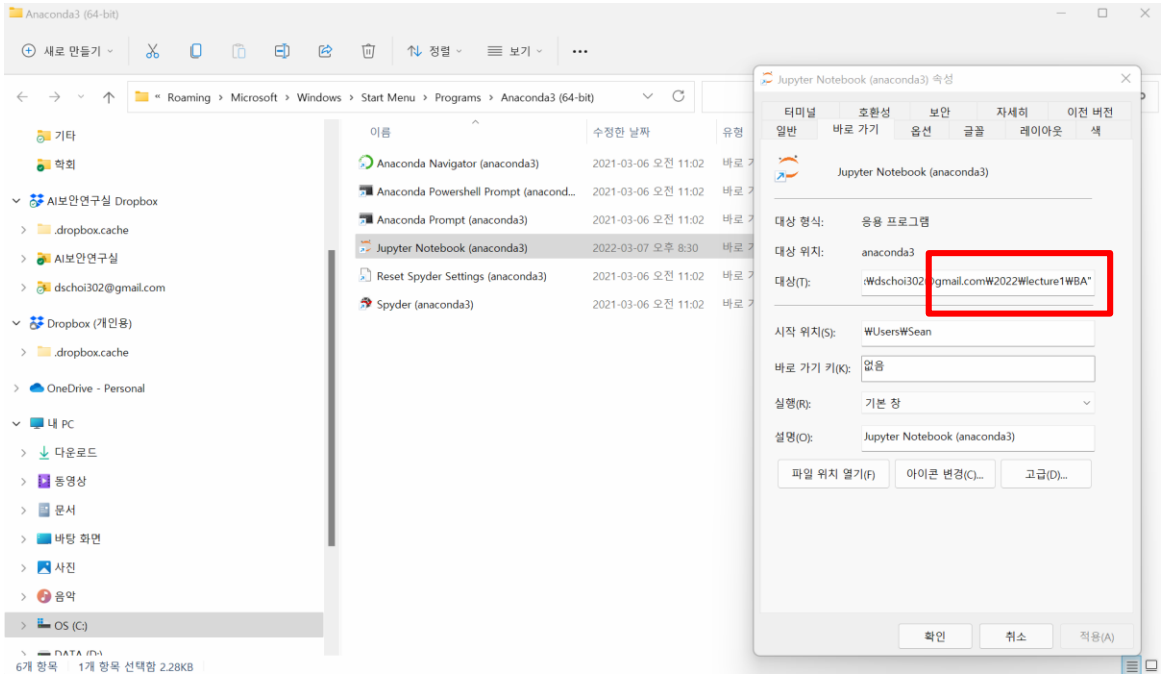
ANACONDA 설치

- www.anaconda.com/products/individual 설치
  - 본인의 운영체제에 맞게



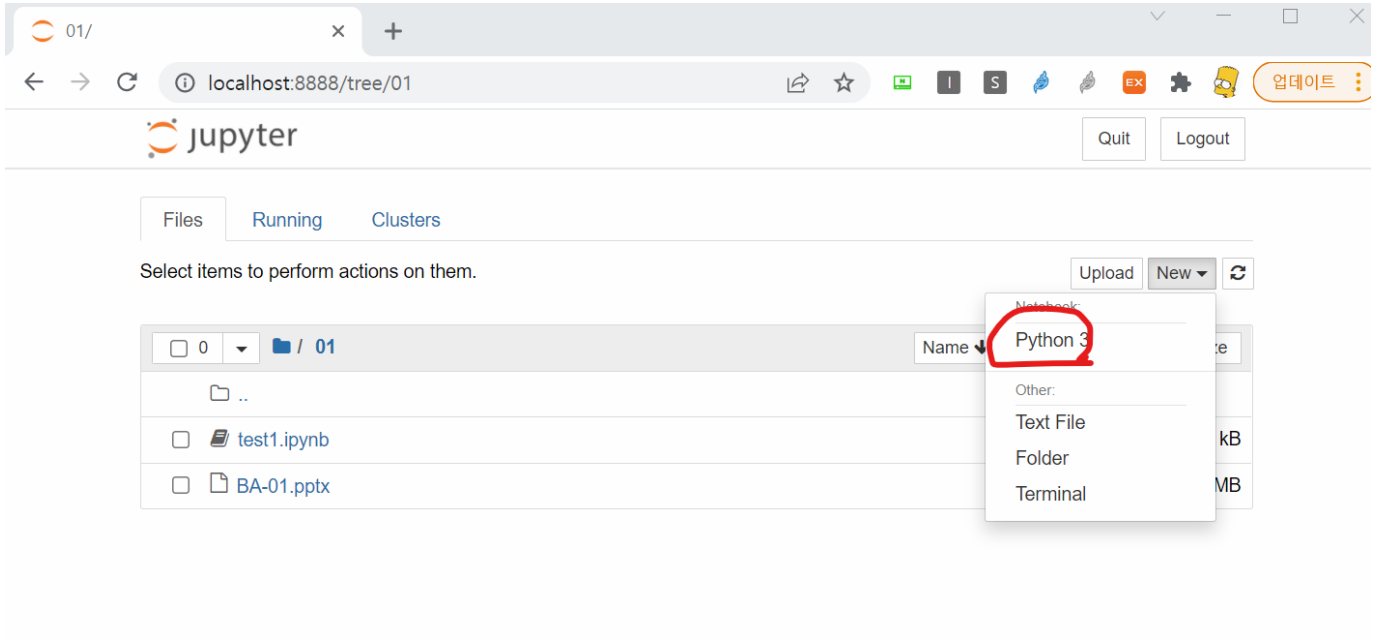
JUPYTER NOTEBOOK 실행

- 강의 폴더 만들기
  - BA
- 강의 폴더에서 시작하기
  - 시작 메뉴에서 검색
  - 파일위치 열기
  - 속성 변경

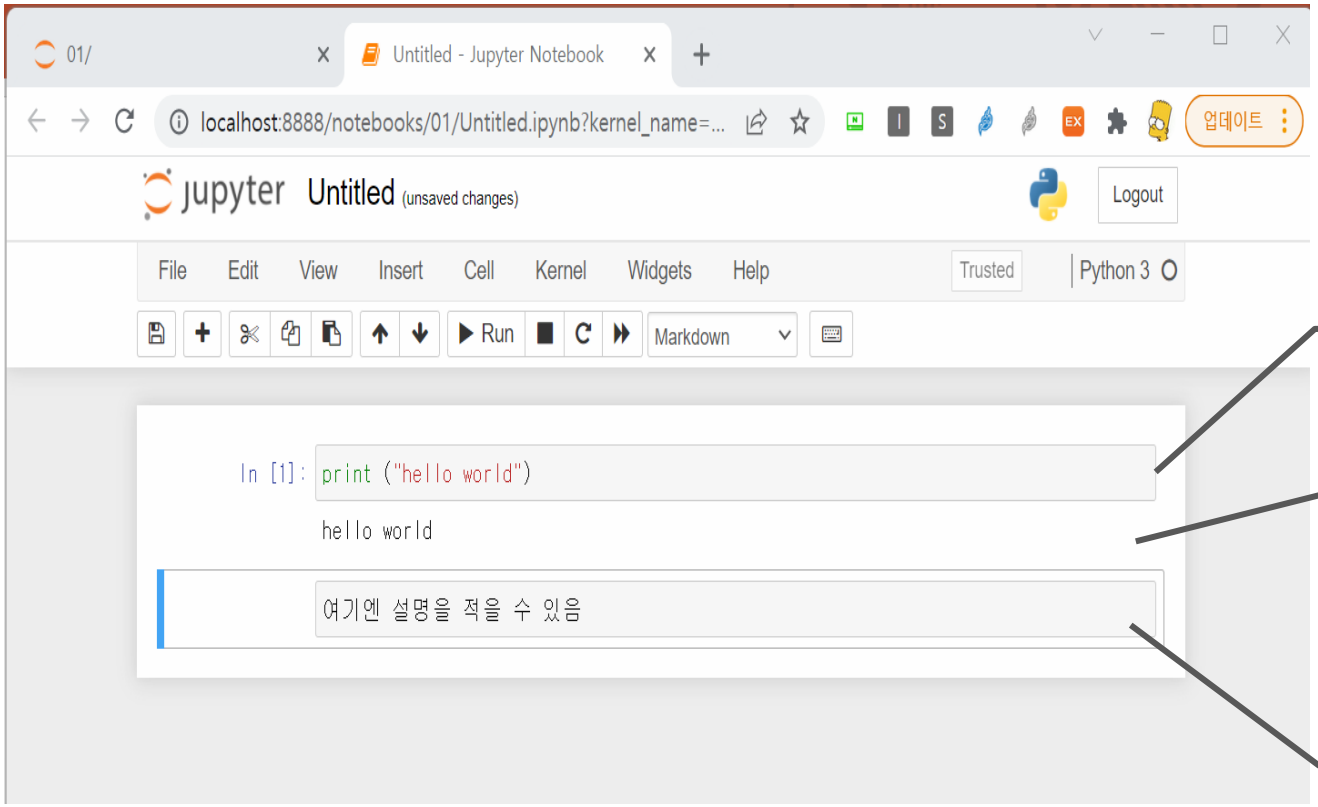


JUPYTER NOTEBOOK 사용

- 폴더 만들기 : 강의 주차 # 01
- 파이썬 파일 만들기



JUPYTER NOTEBOOK 구성



숭실대학교 Soongsil University

파이썬 데이터 처리 실습

- Test1.ipynb 파일 참조
- 파일 읽기
- 필드 분할
- 기술 통계
- Join
- Group by