숭실대학교 Soongsil University

인공지능과 데이터 윤리

숭실대학교 베어드교양대학
서유화 교수
yhsuh@ssu.ac.kr

목차

- 머신러닝에서 범할 수 있는 오류
- 데이터의 불완전성과 결함에 따른 예측 오류와 차별
  - 데이터 안의 차별
  - 데이터 부족으로 인한 차별
  - 민감한 정보누락에 따른 차별
  - 다이나믹 학습을 통한 차별

학습 목표

- 머신러닝에서 범할 수 있는 오류에 대해 설명할 수 있다.
- 데이터의 불완전성과 결함에 따른 예측 오류와 차별에 대해 설명할 수 있다.
- 편향적 데이터에 따른 인공지능 예측의 위험성을 인식하고 인공지능의 한계와 윤리적 관점을 갖는다.
- 인공지능의 개발자 및 사용자로써 윤리적인 데이터의 생성과 사용에 대한 책임과 신중한 자세를 갖는다.

지난시간 배운 내용

주	주제	온라인	오프라인
1	인공지능의 과거 현재와 미래	1. 강의 및 교과목 소개(공통, 핵심만) 2. 인공지능의 과거와 현재 3. 인공지능의 미래와 다양한 시선 4. 인공지능 개발환경 구축과 사용법(Anaconda/Colab)	1. 강의 및 교과목 소개(분반별, 자세히) 2. 다양한 인공지능 기술 경험하기 (자연어처리, 시각, 음성,) 3. 인공지능 챗봇만들기(IBM 왓슨 어시스턴트)
2	공공데이터를 이용한 사회문제 발견과 해결책 모색	1. 빅데이터의 정의와 가치 2. 공공데이터 수집하기 3. 공공데이터로부터 새로운 인사이트 발견하기 - 행정구역별 인구 데이터와 공공의료기관 현황 데이터 분석	1. 서울시 CCTV설치 현황 분석하기 2. 서울시 범죄발생 현황 분석하기
3	인공지능의 개요 및 머신러닝을 이용한 예측	1. 인공지능의 정의와 분류 2. 인공지능 학습방법 이해하기 3. 인공지능 알고리즘 소개	1. 머신러닝을 이용한 이미지 식별(구글 티쳐블 머신) 2. 머신러닝을 이용한 보스톤 집값 예측
4	인공지능과 데이터 윤리	1. 데이터의 불완전성과 결함에 따른 예측 오류와 차별 2. 데이터 왜곡에 따른 분석과 예측 결과 비교 - 지도학습(SVM)을 이용한 타이타닉호 생존자 예측	1. 데이터 편향성이 예측에 미치는 영향 2. 데이터 왜곡에 따른 예측 결과 비교 - 타이타닉호 생존자 예측
5	인공지능과 알고리즘 윤리	1. 알고리즘 기반 의사결정 시스템의 한계 2. 윤리가 적용된 인공지능 알고리즘	1. 알고리즘에 따른 예측 결과 비교 - 보스톤 집값 예측 - 폐암환자 생존 여부 예측
6	인공지능에 대한 다양한 이슈와 우리의 자세 고찰	1. 인공지능의 윤리적/법적 쟁점 (자율주행자동차, AI로봇, 트랜스 휴먼 등) 2. 인공지능시대 사회, 경제적 불평등 문제 3. 인공지능과 프라이버시 4. 인공지능의 윤리적 대응과 규제	1. 자율주행 자동차의 행동학습 시나리오 경험하기 2. 비윤리적 데이터 생성과 수집(웹 크롤링을 이용한 데이터 수집)
7		기말고사	

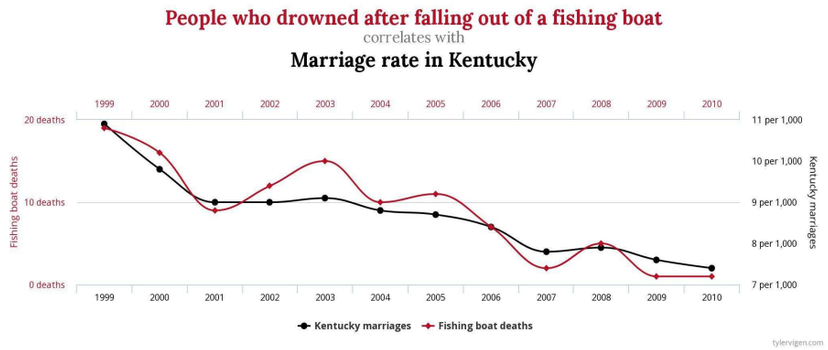
숭실대학교 Soongsil University

머신러닝에서의 범할 수 있는 오류

켄터키주의 결혼율과 고깃배에서 떨어져 익사한 사람 수

- 상관계수 : 0.95

  일반적으로 상관계수가 0.3이상이면 약한상관 관계, 0.7이상 1에 가까울 수록 높은 상관관계

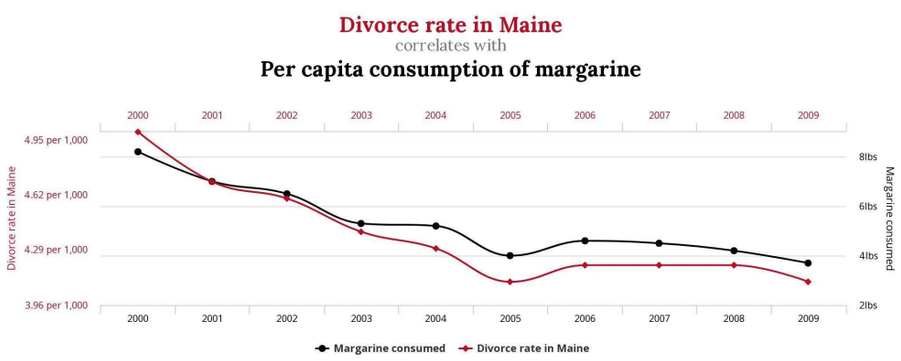


출처 : https://tylervigen.com/

Soongsil University

메인주의 이혼율과 1인당 마가린 소비량

- 상관계수 : 0.99

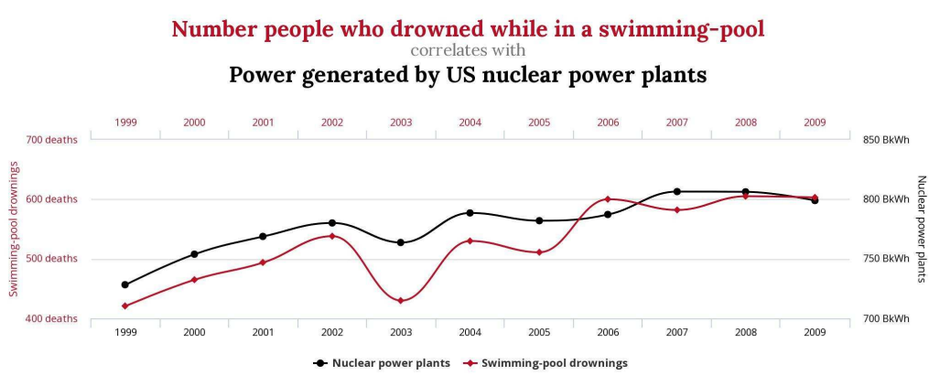


출처 : https://tylervigen.com/

Soongsil University

수영장에서 익사한 사람과 미국 핵발전소 생산량

- 상관계수 : 0.90



출처 : https://tylervigen.com/

Soongsil University

상관관계 vs 인과관계

- 상관관계
  - 데이터에서 어떠한 변인(변수)과 다른 여러 변인(변수)들 간의 관련성
  - 한 변수가 변화함에 따라 다른 변수가 어떻게 변화하는가?와 같은 변화의 강도와 방향
  - 상관계수 : 두 변수 사이의 상관 관계의 정도를 수치적으로 나타낸 계수
    - -1~1 까지의 값, 음수이면 음의상관, 양수이면 양의 상관, 0이면 관계없음
  - 활용 예)
    - 자동차 보험가입 시 운전자 정보제공(나이, 성별, 결혼)
    - 대학입시에서 내신성적과, 수학능력 고려
- 인과관계
  - 원인과 결과의 관계로 하나의 원인이 다른 결과를 일으키는 관계
  - "상관은 인과를 함축하지 않는다. (Correlation does not imply causation)"
    - 상관이 있는 것만으로 인과를 단정하지 못하고 인과의 전제에 지나지 않음

상관관계 vs 인과관계

- 인공지능의 예측
  - 머신러닝은 관찰된 행동에 대한 데이터를 근거로 상관관계가 충분하면, 새로운 데이터에 대해서도 결정을 내릴 수 있다는 가정에 기반
- 레모네이드 판매량 예측의 예

학습데이터 셋 (training data set)

날짜	온도	방문고객수	판매량
2021.1.3	20	20	40
2021.1.4	21	54	42
2021.1.5	22	21	44
2021.1.6	23	30	46
2021.1.7	24	10	48

판매량과 상관관계가 높은 학습피쳐 선별

온도	판매량
20	40
21	42
22	44
23	46
24	48

학습을 통한 모델 생성

새로운 데이터에 대한 예측

날짜	온도	판매량
2021.1.8	26	?
2021.1.9	27	?

상관관계 vs 인과관계

- "검증되지 않은 순수한 가설은 팩트로 여겨지지 않는다."
  - 팩드를 찾는 학문적 방법
    - 여러 번의 검증을 거쳐, 실험에서 반박할 수 없는 결과가 나온가설들만 이론이 됨
    - 이 이론의 예측이 통제된 반복된 실험에서 여러 번 옳은 것으로 입증되어야 팩트로 받아들여짐
  - 학문적 방법을 무시하고, 머신러닝 알고리즘으로부터 나온 상관관계만을 신뢰하여 실제 예측에 활용하는데는 위험성있으며 예측의 활용에 신중함이 필요함

숭실대학교 Soongsil University

데이터의 불완전성과 결함에 따른 예측 오류와 차별

데이터 안의 차별

- 아마존 AI 채용시스템
  - 2014년 아마존 자동평가 시스템을 구축하여 활용하고자 했으나 폐기
  - 인풋으로 이전 10년간 지원서류가 활용
  - 지원자들의 이력서에 1개에서 5개 사이의 별점을 부여하도록 고안
  - 100개의 이력서를 주고 프로그램이 상위 5개를 추천하면 채용하는 방식

→ 남성 지원자를 선호하는 패턴 발견

  - 10년간 지원서류(이 시기 성공적인 지원자들은 거의 남성)
  - 성별을 인풋으로 넣지 않아도 성별과 상관관계가 있는 특성을 찾아냄
  - 지원서류에 성별관련 특성이 나타나면 나쁜점수 부여
    - 이력서에 '여자 체스서클 ' 에서 활동했다고 되어있으면 부정적 평가
    - 여대 졸업증명서는 더 부정적으로 평가

데이터 안의 차별

- 학습데이터의 편향성이 미치는 영향
  - 이전 학습데이터에서 민감한 특성과 관련해 편향된 결과가 있었다면 알고리즘이 배후에 놓인 민감한 특성을 알지못한다 해도 그 특성과 다른 특성의 상관관계를 통해 편향을 찾아낼 수 있음
  - 머신러닝은 지금까지의 선호 경향을 더 강화시켜 더더욱 단일한 문화로 나가아게끔 할 수 있어, 편향을 더 강화시킬 수도 있음

데이터 부족으로 인한 차별

- 특정그룹의 데이터 부족으로 인한 차별

조이 부올람위니
(Joy Buolamwini) 테트토그(2017)

"How I'm fighting bias in algorithm"

출처 : https://www.youtube.com/watch?v=UG_X_7g63rY

데이터 부족으로 인한 차별

- 안면인식 시스템 편향성 사례

구분	세부내용
국립표준기술연구소(NIST)	- 안면인식 알고리즘이 여성, 성소수자, 흑인을 포함한 유색인종을 잘 인식해내지 못함 (2019년, 12월) - 백인 대비 흑인과 아시아인의 정확도가 10~100배 떨어짐
콜로라도 대볼더 캠퍼스	- 기업들이 제공하는 안면인식 서비스가 생물학적 성과 정체성이 일치하지 않는 '비시스젠더(non-cisgender)'의 인식에 많은 오류를 보임(2019년 10월) - 시스젠더를 인식하는 오류는 5%, 트랜스 남성을 여성으로 잘못인식하는 경우가 38%
BBC 보도	- 영국의 여권사진 검사시스템은 흑인 여성 사진을 백인 남성사진보다 여권 규정에 부적합하다고 판정 내릴 가능성이 두 배 이상 높음 (2020년 10월)
미국시민자유연맹(ACLU)	- 아마존의 안면인식기술 '레코그니션'을 이용해 미국 상하원 의원을 식별한 결과, 28명을 '범죄자'로 판별, 이중 유색인종이 11명이었고 이는 유색인종에 대한 불공정 가능성 존재 (2018년 8월)

출처 : 안면인식 도입의 사회적 논란과 시사점: 미국사례 중심으로, ERTI Insight 기술정책 브리프 2020.12

데이터 부족으로 인한 차별

- MIT Gender Shades 프로젝트
  - 상용화된 안면인식기술의 성별 분류 정확도를 모니터링하는 프로젝트
  - 조사결과 IBM, 마이크로소프트 등과 같은 회사의 안면인식 시스템은 백인일수록, 남성일수록 더 좋은 식별 성능을 나타냄
  - 가장 정확도가 낮은 흑인 여성과의 가장 높은 백인남성과의 정확도 차이는 약 35%p 수준

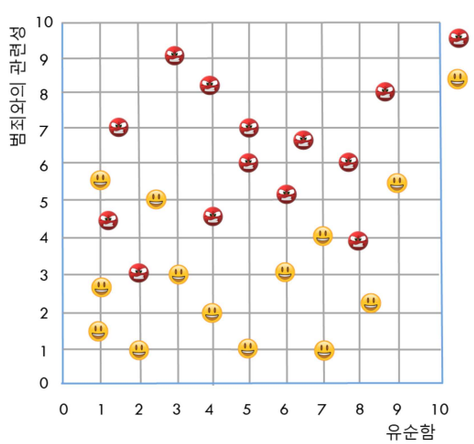
분류	흑인남성	흑인여성	백인남성	백인여성	최대격차
마이크로소프트	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

데이터 부족으로 인한 차별

- 이미지(영상) 인식 시스템
  - 의료영역의 다양한 인종의 데이터 부족으로 인한 차별
  - 선진국에서 이루어진 연구들을 토대로 주로 백인 대상이 많음
  - 인공지능을 활용한 의료분야에서도 데이터부족으로 특정 그룹은 의료 혜택을 받지 못하는 차별이 발생할 수 있음
- 음성인식 시스템
  - 음성인터페이스는 자판이나 마우스보다 더 상용화 될 것으로 예상
    - 엘리베이터, 비자신청 등
  - 악센트가 강한 사람들, 사투리를 쓰는 사람, 언어장애가 있는 사람들이 음성인식 서비스를 받는데 차별이 발생할 수 있음
- 인터넷과 디지털기기 접근이 낮은 국가나 여성들의 데이터 부족이로 인한 차별 가능성이 존재

민감한 정보를 누락시킴으로써 빚어지는 차별

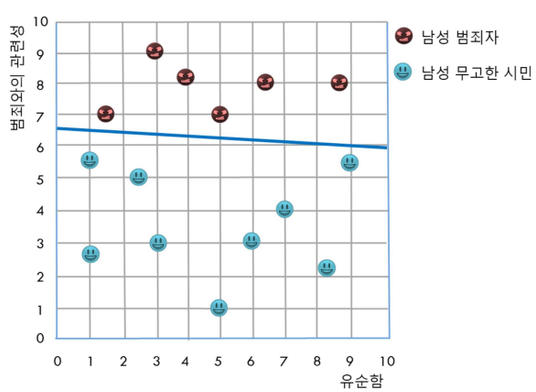
- 범죄자와 무고한 시민을 구별하는 예

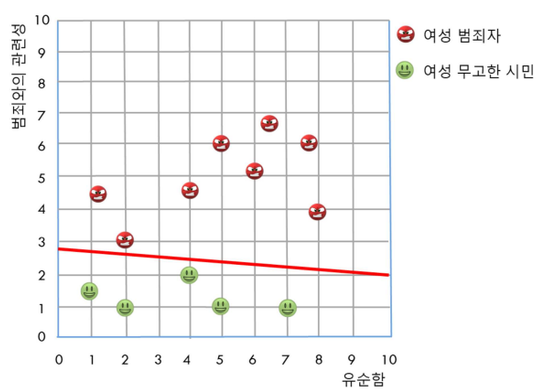


← 명백한 분할선을 긋기 어려움

민감한 정보를 누락시킴으로써 빚어지는 차별

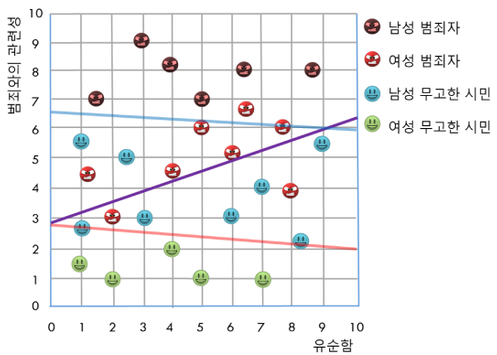
- 범죄자와 무고한 시민을 구별하는 예





민감한 정보를 누락시킴으로써 빚어지는 차별

- 범죄자와 무고한 시민을 구별하는 예
  - 머신이 민감한 특성을 대변하지 못하는 상태에서 두 집단의 행동이 차이가 날때 차별적인 결정이 빚어질 수 있음



← 남성 차별의 분할선
: 두 명의 무고한 남성을 재범자 측에 분류
재범을 저지른 두 여성을 무고한 시민으로 분류

※ 민감한 특성(성별)을 알고리즘에 알려주지 않았지만 불이익이 빚어질 수 있음

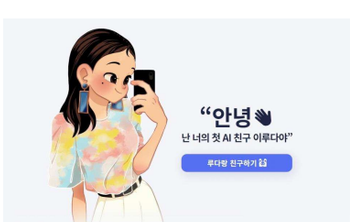
현실의 마이너리티 리포트

- 예측 치안(Predictive Policing)
  - 기계학습을 통한 특정 시기, 특정 장소에서 피해자가 될 가능성이 높은 사람과 가해자가 될 가능성이 높은 사람을 구별
  - 판결이나 가석방시 재범 가능성 계산

구분	세부내용
컴파스 (COMPAS)	- 미국 노스포인트사에서 개발한 인공지능 - 유사한 다른 범죄자들의 기록과 특정 범죄자의 정보를 빅데이터 분석을 통해범죄자의 재범 가능성을 예측 - 위스콘신 주에서 활용, 유타, 버지니아, 인디애나 주 등 유사한 소프트웨어 활용
프레트폴 (PrePol)	- 미국 캘리포니아주립대(UCLA)에서 개발한 범죄 정보를 분석해 10~12시간 뒤 범죄가 일어날 시간과 장소를 도출하는 프로그램 - 캘리포니아, 워싱턴, 사우스 캐롤라이나, 아리조나, 테네시, 일리노이 등 미국 일부지역과 영국, 네덜란드 등에 적용
갱단 범죄 예방 프로그램 (Complete Analytics Pilot Program to Fight Gang Grime)	- 액센추어사에서 개발 영국 경찰에서 운용하는 강력 범죄 예측 프로그램 - 5년 내 범죄 기록을 수집, 갱 조직원이 저지른 개인 범죄 기록의 날짜, 장소, 행동, SNS 게시물, 조직 내 다른 멤버를 욕하는 듯한 발언 등 세세히 수집범죄

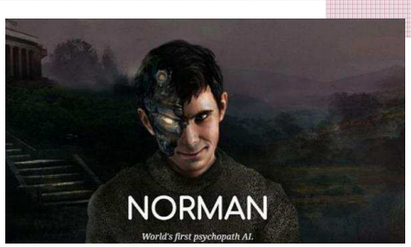
다이나믹 학습을 통한 차별

- 2016년 마이크로소프트 챗봇 '테이tay'
  - 컴퓨터가 인간 언어를 이해할 수 있도록하기 위한 MS실험 프로젝트
  - 미국 18~24세 연령층 사용자를 겨냥
  - 메시지 킥, 그룹미, 트위터를 통해 사람과 대화
  - AI 신경망 기술 기반으로 인간들의 대화의 단어사용법, 질문에 대답하는 방식, 특정사안에 관한 정보, 의견 등을 학습하여 반응에 반영
  - 차별 발언으로 공개 16시간 만에 운영 중단
- 2020년 국내 스캐너랩 '이루다'
  - 혐오, 혐별 발언 및 개인정보유출 등 유사한 문제로 약 3주 만에 종료
  - 인공지능 윤리에 대한 논의를 본격화 시킴
- 혐오적인 표현을 블랙리스트로 관리하며 자동필터링 하는 것은 한계가 있음
  - 우회적인 표현 등은 필터링이 어려움



편향적 데이터를 학습한 인공지능

- 2018년 MIT에서 개발한 세계최초 사이코패스 로봇 '노먼'
  - http://norman-ai.mit.edu/
  - 편향된 데이터가 기계학습 알고리즘에 사용될 때 어떤 결과를 야기하는지를 보여주기 위한 연구목적으로 개발
  - 미국 대표 소셜뉴스사이트 '레딧(Reddit)'에서 주로 죽음 등을 다루는 어둡고 부정적인 게시물로 학습



로드샤흐(Rorschach) 테스트 결과

- 표준 AI로봇 : 나무 가지 위에 앉아 있는 새들
- 노먼 : 한 남자가 감전되어 죽음에 이른다.

기술의 윤리적 사용을 위한 움직임

- 기술의 윤리적 사용을 위한 자성과 규제의 움직임
  - 완전하지 않은 이미지/음성 인식 기술의 성급한 사용에 대한 정부의 책임 강조
  - 현재 미국의 연방의회 및 일부 지방정부에서는 정부 기관의 안면, 음성, 바이오, 보행 인식 사용의 법안 통과 또는 금지
  - 영국은 세계 최초로 경찰 안면인식 기술 사용의 위법성을 인정
  - 산업계(마이크로소프트, 아마존, IBM, 트위터, 페이스북 등)에서는 서비스 나품 및 사용을 유예하거나 철수하고 문제 해결을 위한 전담팀 발족하여 개선

다음시간에 배울 내용

주	주제	온라인	오프라인
1	인공지능의 과거 현재와 미래	1. 강의 및 교과목 소개(공통, 핵심만) 2. 인공지능의 과거와 현재 3. 인공지능의 미래와 다양한 시선 4. 인공지능 개발환경 구축과 사용법(Anaconda/Colab)	1. 강의 및 교과목 소개(분반별, 자세히) 2. 다양한 인공지능 기술 경험하기 (자연어처리, 시각, 음성) 3. 인공지능 챗봇만들기(IBM 왓슨 어시스턴트)
2	공공데이터를 이용한 사회문제 발견과 해결책 모색	1. 빅데이터의 정의와 가치 2. 공공데이터 수집하기 3. 공공데이터로부터 새로운 인사이트 발견하기 - 행정구역별 인구 데이터와 공공의료기관 현황 데이터 분석	1. 서울시 CCTV설치 현황 분석하기 2. 서울시 범죄발생 현황 분석하기
3	인공지능의 개요 및 머신러닝을 이용한 예측	1. 인공지능의 정의와 분류 2. 인공지능 학습방법 이해하기 3. 인공지능 알고리즘 소개	1. 머신러닝을 이용한 이미지 식별(구글 티쳐블 머신) 2. 머신러닝을 이용한 보스톤 집값 예측
4	인공지능과 데이터 윤리	1. 데이터의 불완전성과 결함에 따른 예측 오류와 차별 2. 데이터 편향성이 예측에 미치는 영향 (구글티쳐블머신) 3. 지도학습(SVM)을 이용한 타이타닉호 생존자 예측	1. 타이타닉호 생존자 예측 - 데이터 편향성이 예측에 미치는 영향 - 데이터 왜곡에 따른 예측 결과 비교
5	인공지능과 알고리즘 윤리	1. 알고리즘 기반 의사결정 시스템의 한계 2. 윤리가 적용된 인공지능 알고리즘	1. 알고리즘에 따른 예측 결과 비교 - 보스톤 집값 예측 - 폐암환자 생존 여부 예측
6	인공지능에 대한 다양한 이슈와 우리의 자세 고찰	1. 인공지능의 윤리적/법적 쟁점 (자율주행자동차, AI로봇, 트랜스 휴먼 등) 2. 인공지능시대 사회, 경제적 불평등 문제 3. 인공지능과 프라이버시 4. 인공지능의 윤리적 대응과 규제	1. 자율주행 자동차의 행동학습 시나리오 경험하기 2. 비윤리적 데이터 생성과 수집(웹 크롤링을 이용한 데이터 수집)
7		기말고사	