**<(19/11/23) 프로젝트 맨토링>**

**(박인근 멘토님)**

* 현업 실무 분야 : 마케팅 및 정보분석, 비즈니스 기간계(legacy), big data 분석
* 머신러닝기반 빅데이터 분석 -> 기간계 분석에 기반이 됨.
* Java, python 등 언어능력이 기반 -> 머신러닝기반 빅데이터 분석이 가능
* 프로젝트를 통해 머신러닝 기반 빅데이터 분석의 개념을 알게 될 것
* Ex/ 청약서에 대한 데이터의 흐름

대리점 -> 본사의 흐름에서 청약서는 pdf의 형태로 청약검증센터로 가게 됨

해당 정보는 이름, 나이, 상품 가입정보, 가족관계증명서 등의 내용이 포함되어 있음

청약서 내 기제 정보는 청약검증센터의 정답DB와 비교되어 참인지 거짓인지 검증됨.

: 그런데 검증은? 사람이 직접 한다! -> 로봇이 대체할 수 있으나 머신러닝이 안되어있음

(토씨 하나 틀리거나, 필적이 알아보기 힘들다면 오청약 양식으로 인지하게 됨)

이런 사례가 기업에서 다수 일어나게 됨. -> 이런 현업에서의 ‘필요성’을 프로젝트로 찾는다면?

(물론 개인정보 등 공개된 정보는 제한적이겠지만)

**(홍지후 멘토님)**

* 현업 실무 분야 : Cloud(Cloudstack/Xen), 빅데이터플랫폼 기획, 아키텍트, 구성, 차세대 PJT 수행, 지능형 E2E 관제 플랫폼 구축
* **사례교육1 : 외국인 택시투어 프로젝트(2018.12.15)**

(KT 빅사이트 유동인구 & 공공데이터를 활용한 외국인택시투어 활성화 프로젝트)

택시투어란? 우리나라에는 없지만 관광산업들이 잘 발달한 나라에는 활성화 되어있는 산업

해당 사업내용을 reference 삼아 관광활성방안으로 삼으면 어떨까?

1. **현상 분석**: 문제점을 파악한다.(외국인 관광객 서울 집중 현상, 서울 관광자원 고갈현상)
2. **대략적 해결방안** : 지적되는 문제에 대한 직관적인 해결법 (경기도 택시투어를 게획하면?)
3. **Workflow 구성** : 대략적인 비즈니스 흐름 도시( ex/ 투어 서비스 신청, 정보 제안, 요금..)
4. **데이터 수집대상 및 방안 파악** : 정형데이터는 어디서, 비정형 데이터는 어디서… .
5. **프로젝트 방법론** : 빅데이터 분석 방법론 활용(분석계획 -> 데이터 준비 -> 데이터 분석 -> 시스템 구현, =>데이터 분석), 진행경과 가시적으로 명시

(1월 전까지는 가이드라인이 잡혀서 어던 프로젝트 주제를 진행할 것인지 설명 가능해야 할 것! )

* **인공지능/ 머신러닝/ 딥러닝** Overview

**인공지능이란?** : “뇌가 수행하는 기능, 특히 지능을 요하는 능력”을 컴퓨터가 수행하는 것 (Every aspect of learning of any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it)

**머신러닝이란?** : (인공지능의 큰 그림에서) 기계가 일일이 코드로 명시하지 않은 동작을 데이터로부터 데이터로부터 학습하여 실행할 수 있도록 하는 알고리즘을 개발하는 연구분야 (Machine Leaning is scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence

(또 다른 정의, TOM M. Mitchell)

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance as tasks in T, as measured by P, improves with experience E. (Experience -> Task -> performance ->Experience의 순환구조)

([https://playground.tensorflow.org/#activation=tanh&batchSize=10&dataset=circle&regDataset=reg-plane&learningRate=0.03&regularizationRate=0&noise=0&networkShape=4,2&seed=0.06193&showTestData=false&discretize=false&percTrainData=50&x=true&y=true&xTimesY=false&xSquared=false&ySquared=false&cosX=false&sinX=false&cosY=false&sinY=false&collectStats=false&problem=classification&initZero=false&hideText=false](https://playground.tensorflow.org/#activation=tanh&batchSize=10&dataset=circle&regDataset=reg-plane&learningRate=0.03&regularizationRate=0&noise=0&networkShape=4,2&seed=0.06193&showTestData=false&discretize=false&percTrainData=50&x=true&y=true&xTimesY=false&xSquared=false&ySquared=false&co)) playground Tensorflow: Neural Network(참조)

**딥 러닝이란?** : Deep learning is a class of machine learning algorithms that uses multiple layers to progressively extract higher level features from the raw input.

->기존에 머신러닝에서 사람이 개입하던 부분을 좀더 multiple layer과정을 통해 대체한다.

딥러닝 분야 : 딥러닝은 어떤 형태의 데이터를 다루는가?(AUDIO, IMAGES, TEXT, Etc)

가령, Audio data의 **ASR**(Automatic Speech Recognition) pipeline

“음성신호를 받아서 문장의 형태로 산출하는 역할을 하고 구현하기 위해서는 전처리, 패턴인식, 후처리 3단게가 필요하다.”

Images data 의 **Computer vision** (classification -> localization -> detection -> segmentation)

Text data – audio data 의 **NLP**(Natural language processing) (ex/ 시리나 빅스비)

음성데이터를 텍스트로, 텍스트를 음성데이터로

(front end : language areas<- feeling data -> backend-algorithm)

* 인공지능의 역사 Overview

컴퓨터 과학과 인지 과학의 역사

1936년 튜링의 튜링머신과 1939~1945폰 노이만의 컴퓨터 구조 (컴퓨터 구조 정립)

컴퓨터 구조 정립 -> **기호주의 인공지능**(규칙기반 인공지능 , A를 물었을 때 B를 답하라(정형적))

**연결주의 인공지능**(connectionism, 신경만 기반 인공지능 대두) : 컨셉은 좋았으나 기술이 달림

회귀하여 **순수통계적 인공지능**(통계기반 AI 대두) ->2010년 딥러닝의 대두로 **연결주의 인공지능** (Deep learning neural network) 다시 등장

Deep learning < Machine learning < Artificial Intelligence < Cognitive science (우로갈수록 포괄)

\*참고자료: 영상자료 FeiFei Li 교수의 Ted 강연 “어떻게 컴퓨터가 사진을 이해하게 되었는가”

(<https://www.ted.com/talks/fei_fei_li_how_we_re_teaching_computers_to_understand_pictures?language=ko#t-10937>) IMAGENET 을 통한 이미지 범주 분류, 합성공 신경망을 이용한 알고리즘 발전

\*참고자료: Jeopardy!: IBM Watson wins on Jeopardy(2011) (인공지능의 퀴즈프로그램 우승)

(<https://www.youtube.com/watch?v=lI-M7O_bRNg>)

\*참고자료: Intelligent Personal Assistant: apple launches siri(2011)

\*참고자료: Microsoft Research’s speech recognition breakthrough(2016)

(<https://www.youtube.com/watch?v=TnJzDb1er5M>) (demo speech translation : 연설의 동시통역)

\*참고자료: Machine Translation : RNN Encoder-Decoder Architecture(2014)

\*참고자료: Generative Adversarial Network(GANs) (2014) (Training set과 noise로 위작구분과 가상의 사람얼굴 생성 모델 활용)

\*참고자료: Facebook Deepface(2014) (딥러닝 얼굴 인식 시스템)

\*참고자료: Tensorflow - Google open sources its deep learning framework(2015)(Tensorflow의 시작)

\*참고자료: NVIDIA Launches deep learning supercomputer(2016) (NVIDIA DGX-1)

\*참고자료: Computer GO: DeepMind’s AlphaGo Defeats Sedol Lee(2016)(이세돌 9단과 알파고) (딥러닝의 관심을 끌어 모은 계기)

(알파고의 승리 알고리즘 참조: Mastering the game of Go with deep neural network and tree search)

\*참고자료 : Open source Deep Learning Framework War(2017)

( tensorflow, caffe, keras 등 딥러닝 프레임워크의 춘추전국시대)

\*참고자료: Google Assistant: Duplex(2018) (long-short term memory 활용) (문제없는 예약전화)

Autonomous Driving, Robotics, Computer vision, speech recognition, Q&A system 등, 사례에서 볼 수 있듯이 프로젝트로서 적용 가능한 분야는 무궁무진하다.

* **사례교육 2: 소프트웨어 영향 평가 사업(첨부파일 참조)**

각종 정부부처들은 국책사업을 실시하게 앞서 해당 사업으로 어떤 부작용 및 악영향((예)민간기업 수익 침해) 을 미칠지 사전 조사하고, 가이드라인을 배포한다. 이러한 자료 스크랩은 텍스트가 기반이 되며, 스크랩된 텍스트 중 키워드를 추출하여 미리 영향을 받을만한 사업들, 제품, SW 등을 사전에 식별한다. (정보시스템 구축, 운영사업, ISP, 소프트웨어 개발 사업 등등에도 적용)

수많은 지자체마다 각자 수천장에 이르는 자료를 지금까지는 사람이 일일이 식별했었음

* 머신으로 훨씬 효율적으로 대체할 수 있지 않을까?

(이런 프로젝트를 납품한다면? => 곧 국책사업, 창업이 될 수 있다. )

* **사례교육 3: 블록체인을 이용한 출판/ 유통구조 개선**

현재 출판사들이 독점하고 있는 유통구조 때문에 출판의 시작이 되는 원 저작자의 인세는 정작 낮음. 이에 블록체인 기술을 도입하면 중계자가 없다는 특성상 소비자와 원 저작자가 직접적으로 접촉할 수 있고, 수익구조 역시 바꿀 수 있다.

* **사례교육 4 : 경희대학교 더위 탈출 넘버 원**

1. 상황분석 – 더위의 심각성(더위가 심각해지는 상황에 대한 간략한 도표화된 searching)
2. 원인분석 – 비수혜원인 등 기존 지원제도에 대한 SWOT분석, W해결법에 대한 문제제기 및 보고
3. 솔루션 – 분류 모델과 무더위지수(사용지표(더위체감지수) 등 명시, 지역선정, 더위체감지수 산출, 최종 복지 우선 지역 선정 순의 문제해결 프로세스 명시)
4. 분석과정 – 사용한 데이터 (사용한 데이터 셋 분류 ; 모델 개발용, 그룹 인덱스 개발용, 무더위지수 계산용 데이터 셋 등 )

**<세미프로젝트 초안 토의하기 >**

**B조 :** 공공데이터를 활용한 고독사 예측

기초생활수급자, 홀몸노인, 1인 가구의 전기요금, 가스요금 미납 기간, 자살률 등 공공데이터를 이용한 고독사 에상

보완방향 지역설정 예측범위 설정

feedback/ 밀린 요금 미납기간이 특정 얼마나 되는지에 따라 고독사를 예측하고자 한다면 standard가 되는 기준을 생각해봐야함

**E조 :** 장애인을 위한 지하철은 없다.,

위험요소에 취약한 장애인들에 대하여 지하철 안전사고 발생량에 대한 데이터를 수집하여 해당 역에서 어떤 위험이 있는지 예고해주는 어플리케이션 개발

지하철 사고데이터, 역사정보, 사고원인데이터, 유동인구 데이터,

역사 정보 및 사고 사례에 대한 상관분석 및 회귀 분석을 통한 인과 관계 파악과 로컬 서비스를 통해 역 도착 시 주의할 정보를 알림.

Feedback/ 모집한 사고데이터가 일반인들에게서 발생한 것인지, 장애인들에 의하여 생긴 데이터인지 불분명함(사용자들에 대하여 별 의미가 없을 수 없다.)

**A조 :** optima(블로그 신뢰도 수치화 제공 프로그램)

텍스트 마이닝을 통해 블로그 포스팅 등 온라인 리뷰를 얼만큼 신뢰할 수 있는지 신뢰도를 수치화하여 제공하는 프로그램 개발

텍스트 마이닝을 바탕으로 글을 분석하고 신뢰성을 나타내는 지표를 만들어 신뢰성을 수치화함

**C조 :** 나랑 여행갈래?(외국인 관광객을 대상으로 한 한국여행 추천시스템)

관광지를 검색했을 때 주변의 유사한 관광지를 추천,취향에 따른 추천검색이 가능하도록 제공

한국으로 처음 여행을 온 외국인들에게 유용한 교통수단과 주변정보 등을 제공

공공 데이터 포털 활용, 관광지 정보 수집, 주소 시각화(구글맵), 파파고 API(외국어 지원)

Feedback: 지역적 식품군을 군집화하여 추천, 어느 것에 포커싱할것인지?(음식, 관광지?) 범위를 확실히(너무 많으면 어려울 것 같다)

**F조 :** 공공자전거 및 전동킥보드를 이용하여 사용자가 서울 및 근교의 핫 플레이스를 방문할 수 있도록 안내하는 서비스

공유경제에 대한 인식변화 및 공공자전거, 전동 킥보드의 이용 증가, 핫플레이스에 대한 소비 증가, 힐링 문화에 대한 개인적인 욕구 증가, 지역경제 활성화 기대

Feedback: 동선에 관련한 정보를 제공받을 수 있는지?