경북대학교

빅데이터와 클라우드 인프라

목차 Contents

0 도입

1. 과정 소개
2. AI훈민정음 소개

1 과정 소개

경북대 K-Digital Training, 빅데이터 분석가 양성과정 2기

교 과 명		빅데이터와 클라우드 인프라	교육 수준 초급 중급 □ 고급
교육대상		K-Digital Training 빅데이터 분석가 교육과정 수강생 - 빅데이터 분석가로 취업하고자 하는 취업준비생 - 대학 졸업자 또는 예비졸업자(전공자/비전공자 포함)	
학습목표		-클라우드 컴퓨팅 주요 인프라 기술들을 알 수 있다. -빅데이터AI플랫폼 활용을 할 수 있다. -AI훈민정음 예제를 통해서 인공지능의 주요 모델과 알고리즘을 학습 할 수 있다. -빅데이터AI플랫폼을 활용하여 인공지능 사례를 개선 및 개발할 수 있다.	
사용교재		자체 재작	
활용도구		T3Q.ai, Python, Jupyter Notebook	
차시	시간	강의 내용	
1	4	**클라우드 컴퓨팅 소개 및 활용 (1)** - 퍼블릭 클라우드 이해 (AWS, NHN)	
2	4	**클라우드 컴퓨팅 소개 및 활용 (2)** - 퍼블릭 클라우드 이해 (AWS, NHN)	
3	4	**클라우드 컴퓨팅 소개 및 활용 (3)** - 프라이빗 클라우드 이해 (가상머신, 도커/쿠버네티스)	
4	4	**클라우드 컴퓨팅 소개 및 활용 (4)** - 프라이빗 클라우드 이해 (가상머신, 도커/쿠버네티스)	
5	4	**빅데이터 AI플랫폼 활용 (1)** - 빅데이터 기능 활용(T3Q.cep): 예제를 통한 데이터 저장/변환/실시간 추론 파이프라인 구성	
6	4	**빅데이터 AI플랫폼 활용 (2)** - 인공지능(T3Q.dl): 예제를 통한 데이터 전처리, 학습, 추론 구성	
7	4	**AI훈민정음을 통한 인공지능 학습 (1)** - (예제1) 이미지 분류: 손그림 이미지 분류	
8	4	**AI훈민정음을 통한 인공지능 학습 (2)** - (예제2) 바이너리 이상탐지: 악성코드(Malware) 이상탐지	
9	4	**AI훈민정음을 통한 인공지능 학습 (3)** - (예제3) 위성분류 : 허리케인 위성 이미지의 풍속 예측	
10	4	**AI훈민정음을 통한 인공지능 학습 (4)** - (예제4) 위성 회귀 : 다중레이블된 지표면 분류	
11	4	**미니 프로젝트 1 (1)** - 점자보도블록 이상탐지 개선하기(1~2인 과제)	
12	4	**미니 프로젝트 1 (2)** - 발표 및 토론	
13	4	**미니 프로젝트 2 (1)** - 텍스트 분류(AI작가) 기존 기능 개선하여 제출하기 (2~3인 조별 과제)	
14	4	**미니 프로젝트 2 (2)** - 발표 및 토론	

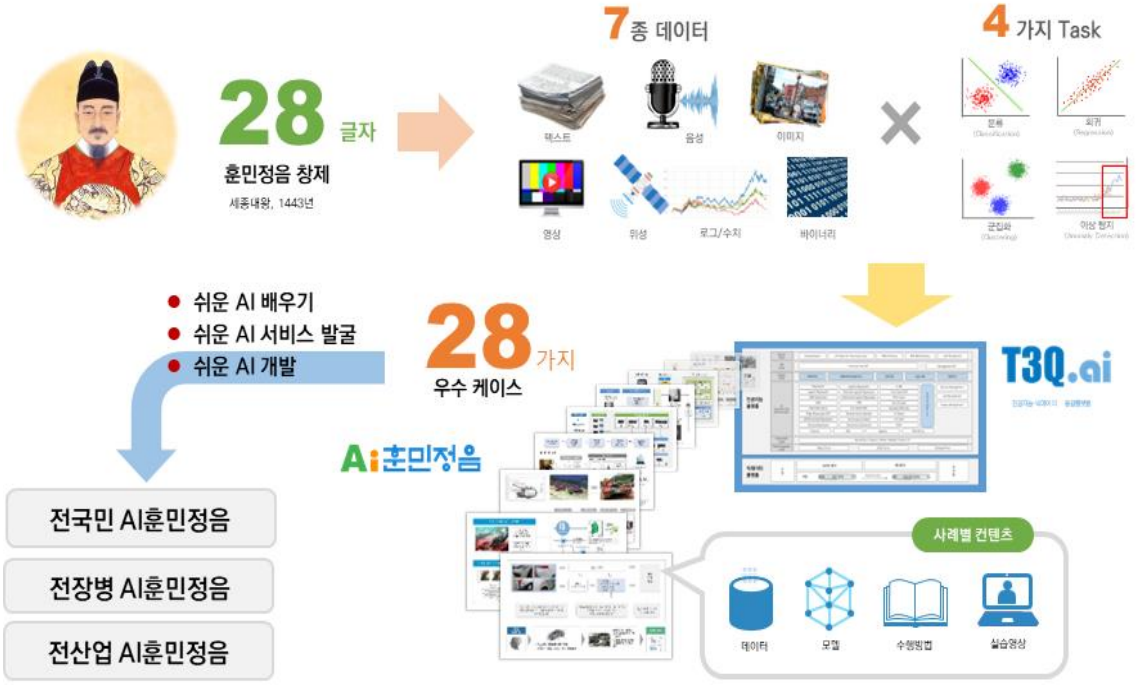
2. AI훈민정음 소개

티쓰리큐 AI훈민정음

◆ 세상에서 인공지능을 가장 잘 활용하는 대한민국!

1442년 경 세종대왕께서 글이 없어 자기 뜻을 제대로 표현하지 못하는 백성들을 위하여 28글자를 만들어 세상의 모든 소리와 뜻을 전할 수 있게 하였습니다.

이처럼 티쓰리큐(주)는 2021년 인공지능과 빅데이터를 누구나 쉽게 배우고 사용하기 위해 「AI훈민정음」을 만들었습니다. 이는 인공지능·빅데이터 통합플랫폼(T3Q.ai) 위에 인공지능에서 다루는 Data 7종과 인공지능이 하는 일 Task 4가지를 조합한 28가지 우수사례를 통합하여 탑재한 인공지능 디지털 사례집입니다.



「AI훈민정음」은 따라하기 방식으로 남녀노소 누구나 쉽게 AI를 배우고, 이들 28가지 우수사례를 기반으로 개인, 기업, 기관에서는 인공지능으로 무엇을 할 수 있는지 쉽게 AI 서비스를 발굴하며, 서비스가 발굴되면 해당 우수사례를 기반으로 쉽게 AI 서비스를 개발할 수 있습니다.

또한 T3Q.ai 플랫폼 사용자는 생산자이자 소비자로서 「AI훈민정음」을 활용하여 인공지능을 배우고, 만들고, 판매하는 인공지능 생태계 활성화에 기여하는 것입니다.

2. AI훈민정음 소개

티쓰리큐 AI훈민정음

AI훈민정음, Data 7종

인공지능에서 다루는 대표적인 Data는 텍스트, 음성, 이미지, 영상, 위성, 수치/로그, 바이너리 등 기본 7종입니다.(추가 가능).

텍스트	텍스트마이닝과 자연어처리를 필요로 하는 줄글
음성	재생 가능한 소리 형태의 자연어
이미지	사물(객체)을 식별할 수 있는 2차원 데이터, RGB 등의 채널 존재
영상	동영상 파일 혹은 비디오 스트리밍
위성	지역정보와 다수의 수치정보 레이어가 포함된 복합 위성데이터
수치/로그	주기적으로 수집되는 일련의 측정치 등 수치
바이너리	위 6가지 분류에 포함되지 않는 데이터이거나 비정형 자료

AI훈민정음, Task 4가지

인공지능이 하는 기본 Task는 분류(Classification), 회귀(Regression), 군집화(Clustering), 이상탐지(Anomaly Detection) 등 4가지입니다.(추가 가능)

분류	기존에 존재하는 데이터의 Category 관계를 파악하고, 새롭게 관측된 데이터의 Category를 스스로 판별하는 모델
회귀	연속된 값으로 표현되는 결과를 예측하는 모델
군집화/유사도	레이블이 지정 되어있지 않은 데이터를 그룹핑하는 분석 알고리즘
이상탐지	정상범위 외의 데이터를 식별하는 모델

Data 7종 x Task 4가지 = 우수사례 28가지

「AI훈민정음」은 Data 7종, Task 4가지를 조합한 28가지 우수사례에 대한 데이터, 모델, 수행방법, 가이드, 실습영상 등의 콘텐츠를 인공지능·빅데이터 통합 플랫폼(T3Q.ai)에 탑재한 프레임워크입니다. 우수사례는 계속적인 추가가 가능합니다.

2. AI훈민정음 소개

티쓰리큐 AI훈민정음

AI훈민정음 적용

「AI훈민정음」을 이용하면 인공지능 배우기, 서비스 발굴이 쉬워집니다. 또한 인공지능 서비스를 손쉽게 개발할 수 있습니다.

쉬운 AI 배우기

쉬운 AI 서비스 발굴

쉬운 AI 서비스 개발

AI훈민정음 활용 방안 (「全국민 AI」, 「全장병 AI」, 「全산업 AI」)

- 全국민의 AI 접근성 및 활용성 향상을 통한 지식 격차 해소
- 군 장병의 복무기간 중 맞춤형 인공지능·소프트웨어 교육과 개발
- 각급학교의 인공지능 교육 및 교원 AI 역량 강화와 활용
- 정부의 인공지능 100만 AI 인력 양성에 활용
- 중소벤처 제조 플랫폼에 적용하여 제조업 경쟁력 강화 (KAMP)
- 닥터앤서 의료 플랫폼으로 확장을 통해 AI 의료 서비스 제공

AI훈민정음 포털 활용

누구나 참여하여 아이디어, 데이터, 모델들을 발굴하고 만들어 유통할 수 있는 인공지능 생태계인 「AI훈민정음 포털」에서는 인공지능 플랫폼에 탑재된 AI 표준 사례를 중심으로, 따라하면서 쉽게 배우고(노코딩), 쉽게 서비스를 발굴하고 개발, 판매하여 이익을 창출할 수 있음

AI 비즈니스 사례 (티쓰리큐 AI훈민정음 활용 73사례 중 대표사례)

- 국가 AI 광주데이터센터 (광주광역시, '20~)
- AI 중소벤처 제조 플랫폼 (중소벤처기업부, '20~'22)
- 북한정보 인공지능/빅데이터 분석시스템 구축 (통일부, '21, '22)
- 지휘통제 지능정보 플랫폼 (국방부_국방기술품질원, '20~'23)
- 라이프로그 빅데이터 플랫폼 (원주연세의료원, 컨소시엄 '20)
- 범부처 인공지능 산업플랫폼 (과학기술정보통신부_NIPA, '20)

2. AI훈민정음 소개

티쓰리큐 AI훈민정음

AI훈민정음 28개 기본사례

구분	분류 존재하는 class(명목변수)에서 선택하는 모델	회귀 연속된 값으로 표현되는 결과를 만들어내는 모델	이상탐지 정상을 학습함으로써 정상범위 외의 데이터를 식별하는 모델	군집화 feature vector 추출(embedding 결과)의 유사성을 분석하는 모델
텍스트 텍스트 마이닝과 자연어 처리에 필요한 줄글	**텍스트 x 분류 :** **영화리뷰 텍스트 감정분석(긍정/부정)** ①IMDB 데이터셋 ②상업적 사용 불가 ③vectorize, standardization, Embedding Layer ④텍스트 정규화와 Embedding Layer를 활용한 영화리뷰 텍스트 이진분류 ⑤Tensorflow(Apache 2.0)	**텍스트 x 회귀 :** **RNN을 사용한 텍스트 생성(AI 작가)** ①셰익스피어의 저작 데이터셋 ②OPEN DATA ③RNN ④텍스트를 치환하여 RNN을 통한 입출력(생성) 시퀀스 구현 ⑤Tensorflow(Apache 2.0)	**텍스트 x 이상탐지 :** **SMS 스팸탐지(Spam, Ham)** ①Kaggle 스팸메일 데이터셋 ②UCI 인용문 삽입 ③LSTM ④class weight를 이용한 불균형 데이터 조정, LSTM을 이용한 모델 설계 ⑤Apache 2.0	**텍스트 x 군집화 :** **텍스트 문서 군집화** ①20 newsgroups 데이터셋 ②UCI 인용문 삽입 / 상업적 사용 명시되어있지 않음(Unknown) ③MiniBatchKMeans ④vectorizer, TruncatedSVD를 통한 전처리, MiniBatchKmeans로 군집화 ⑤BSD 3-Clause "New" or "Revised"(scikit-learn)
이미지 사물(객체)을 식별하는 2차원 데이터. RGB 등의 채널 존재	**이미지 x 분류 :** **손그림 이미지 분류** ①quickdraw 데이터셋 ②CC BY 4.0 ③CNN ④이미지데이터를 numpy 형태로 표현해 어떤 이미지인지 분류 ⑤자체제작	**이미지 x 회귀 :** **Style Transfer** ①mountain(김홍도의 금강사군첩 - 토왕폭).jpg, view.jpg ②김홍도 사진 상업적 이용 가능 ③CNN ④Content Image를 Style Image와 유사하게 바꿈 ⑤Tensorflow(Apache 2.0)	**이미지 x 이상탐지 :** **노후시설물 이미지를 이용한 이상탐지** ①AI Hub의 노후 시설물 이미지 ②AI hub ③SMOTE, VGG16 ④SMOTE를 이용한 불균형 데이터 처리, VGG16을 이용한 모델 설계 ⑤Apache 2.0	**이미지 x 군집화 :** **사람 얼굴 캐릭터 사진 군집화** ①Cartoon-set; Feature Pairs ②CC BY 4.0 ③VGG19, PCA, K-means ④이미지의 특성을 추출하여 군집화, VGG19 모델을 이용하여 이미지들의 특징 벡터추출, 차원 축소 후 군집화 ⑤자체제작
음성 재생 가능한 소리 형태의 자연어	**음성 x 분류 :** **WAV 형식의 단어 분류** ①Speech Commands 데이터셋 ②CC BY 4.0 ③CNN ④음성 데이터에서 스펙트로그램 생성, 생성한 스펙트로그램을 CNN으로 분류 ⑤Tensorflow(Apache 2.0)	**음성 x 회귀 :** **RNN을 사용한 음악 생성(AI 작곡가)** ①MAESTRO ②CC BY 4.0 ③RNN ④RNN을 이용한 모델로 음표 읽고 다음 음표 예측을 반복하여 음악을 생성 ⑤Tensorflow(Apache 2.0)	**음성 x 이상탐지 :** **산업 기계 소리 이상탐지** ①MIMII 데이터셋 ②CC BY-SA 4.0 ③AutoEncoder ④AutoEncoder를 이용한 산업기계(팬) 소리 이상탐지 ⑤MIT-0 License	**음성 x 군집화 :** **환경소리 군집화** ①ESC-50 데이터셋 ②CC BY 4.0 ③K-means ④K-means를 이용하여 환경소리 50가지 중 10가지 클래스를 군집화 ⑤Apache 2.0
영상 동영상(비디오) 파일 스트리밍	**비디오 x 분류 :** **전이학습 및 순환모델을 사용하여 비디오 분류** ①UCF101 데이터셋 ②인용문 삽입 ③CNN-RNN Architecture ④비디오를 이미지의 시계열 데이터로 보고 3D텐서에 넣어 GRU, CNN, RNN으로 행동 분류 ⑤Keras(Apache 2.0)	**비디오 x 회귀 :** **컨벌루션 LSTM을 사용한 다음 프레임 비디오 예측** ①Moving MNIST 데이터셋 ②인용문 삽입 ③Convolutional LSTMs ④일련의 과거 프레임이 주어지면 다음에 올 비디오 프레임을 예측 ⑤Keras(Apache 2.0)	**비디오 x 이상탐지 :** **보행자 통로 영상 이상탐지** ①UCSD 데이터셋 중 ped1 ②인용문삽입(kaggle) ③Convolusional LSTM Autoencoder ④영상 해상도 조정 및 데이터 증대, Convolusional LSTM Autoencoder로 이상탐지 ⑤MIT License	**비디오 x 군집화 :** **얼굴 키포인트가 있는 얼굴 데이터셋 군집화** ①얼굴 키포인트가 있는 YouTube 얼굴 데이터셋 ②CC0 ③K-means ④exploring-youtube-faces-with-keypoints-dataset ⑤Apache 2.0
로그(수치) 주기적으로 수집되는 일련의 측정치 등 수치	**로그수치 x 분류 :** **NBA 선수 포지션 분류** ①2015~2016-2020~2021 시즌 stats 정보 ②상업적 사용은 안되며, 사용시 NBA 페이지 링크 첨부 ③CNN ④NBA 선수들 포지션 예측 ⑤자체제작	**로그수치 x 회귀 :** **NBA 선수 연봉 예측** ①Player Info 데이터셋, Player Salary 데이터셋 ②상업적 사용은 안되며, 사용시 NBA 페이지 링크 첨부, CC0 ③RandomForestRegressor ④NBA 선수들 연봉 예측 ⑤자체제작	**로그수치 x 이상탐지 :** **신용카드 사기탐지** ①Credit Card Fraud Detection ②Open Data Commons ③Sequential을 사용하여 모델을 정의하고 학습 (클래스 가중치 설정 포함) ④데이터분할 및 StandardScaler를 이용한 불균형 데이터 조정, Sequential을 이용한 이상탐지 ⑤Tensorflow(Apache 2.0)	**로그수치 x 군집화 :** **Pokemon 스탯에 따른 군집화** ①pokemon 능력치 데이터셋 ②CC0 ③K-means ④K-means 알고리즘을 이용하여 pokemon을 능력치에 따라 군집화 ⑤자체제작
위성 지역정보와 다수 수치정보 레이어가 포함된 복합 위성데이터	**위성 x 분류 :** **다중레이블된 지표면 분류** ①UC Merced 데이터셋 ②인용문 삽입 ③EfficientNetB5 ④EfficientNetB5를 이용하여 다중레이블된 위성이미지 분류 ⑤Apache 2.0	**위성 x 회귀 :** **허리케인 위성 이미지의 풍속예측** ①폭풍에 대한 이미지 데이터 ②CC BY-4.0 ③ResNet50 ④ResNet50을 이용한 모델로, 허리케인 이미지에서 풍속을 예측 ⑤MIT License	**위성 x 이상탐지 :** **픽셀의 다중스펙트럼값을 이용한 이상탐지** ①Landsat Satellite 데이터셋 ②UCI 인용문 삽입 ③Sequential을 사용하여 모델을 정의하고 학습 ④위성 이미지 픽셀의 다중 스펙트럼 값으로 클래스를 나누어 적은 수의 클래스를 이상값으로 탐지 ⑤Apache 2.0	**위성 x 군집화 :** **위성이미지 군집화** ①Satellite Remote Sensing Image -RSI-CB256 ②CC0 ③VGG19, PCA, K-means ④위성 이미지 특성 추출 통한 산림과 사막화 군집화 ⑤자체제작
바이너리 위 6가지에 포함되지 않는 데이터 (비정형)자료	**바이너리 x 분류 :** **악성코드(Malware) 분류** ①Malware as Images : binary 파일(.bin)을 이미지 파일(.png)로 변환한 이미지 데이터 ②CC BY-SA 4.0 ③EfficientNetB5 ④EfficientNetB5을 이용한 악성코드 이미지 분류 ⑤Apache 2.0	**바이너리 x 회귀 :** **악성코드(Malware) 감염예측** ①Kaggle의 Malware Detection ②CC0 ③LightGBM ④LightGBM 모델을 이용하여 악성코드에 감염될 확률을 예측 ⑤자체제작	**바이너리 x 이상탐지 :** **악성코드(Malware) 이상탐지** ①Malware as Images : binary 파일(.bin)을 이미지 파일(.png)로 변환한 이미지 데이터 ②CC BY-SA 4.0 ③U-Net ④변환된 Malware 이미지를 U-Net을 이용하여 정상코드와 악성코드 탐지 ⑤자체제작	**바이너리x군집화 :** **악성코드(Malware) 군집화** ①Malware as Images : binary 파일(.bin)을 이미지 파일(.png)로 변환한 이미지 데이터 ②CC BY-SA 4.0 ③K-means ④변환된 Malware 이미지 데이터를 양성과 악성으로 군집화 ⑤자체제작

감사합니다.