분류번호 :	2001070304_23v2
능력단위 명칭 :	인공지능 데이터 전처리
능력단위 정의 :	인공지능 데이터 전처리란 인공지능 서비스를 위하여 확보된 데이터를 정제, 변환, 통합, 축소, 라벨링, 비식별화를 통해 데이터를 가공하는 능력이다.

능 력 단 위 요 소	수 행 준 거
2001070304_23v2.1 인공지능 데이터 정제하기	1.1 확보한 인공지능 데이터의 품질을 분석하여 정상 데이터와 오류 데이터를 정량적으로 측정할 수 있다. 1.2 발견된 오류 데이터를 분석하고 오류의 원인을 파악하여 인공지능 데이터의 정제 필요 여부를 결정할 수 있다. 1.3 정제가 필요한 인공지능 데이터에 대해서 정제작업을 수행할 수 있다. 1.4 정제된 인공지능 데이터를 확인하고 정제 수행 결과를 검증하여 정제보고서를 작성할 수 있다.
	【지식】 • 유형별 데이터 처리 기법 • 메타데이터 활용 방법론 • 기술 통계론 • 인공지능 윤리 가이드북 • 인공지능 FATE 지식
	【기술】 • 전처리 실행도구 사용기술 • 데이터 품질 분석 능력 • 데이터 처리 기술 • 데이터 유형 변환기술
	【태도】 • 인공지능 윤리기준을 준수하려는 노력 • 데이터 정제 방안을 논리적이고 구체적으로 수립하려는 자세
2001070304_23v2.2 인공지능 데이터 통합하기	2.1 다중 소스로부터 정제된 정상 데이터를 확보할 수 있다. 2.2 확보한 인공지능 데이터의 특성을 분석할 수 있다. 2.3 분석된 특성에 따라 목적에 맞는 데이터로 분류할 수 있다. 2.4 분류된 인공지능 데이터를 인공지능 데이터 확보 계획 수립 단계에서 제시된 기준에 따라 통합할 수 있다. 2.5 통합된 인공지능 데이터의 결과를 검증하고 통합 결과보고서를 작성할 수 있다.
	【지식】 • 인공지능 모델 학습 과대적합 및 과소적합 회피 방법론 • 기술 통계론 • 표본 추출 방법 • 인공지능 학습용 데이터 품질관리 가이드
	【기술】 • 전처리 실행도구 사용기술 • 데이터 품질 분석 능력 • 데이터 처리 기술 • 데이터 유형 변환기술
	【태도】 • 비정상 데이터를 최대한 파악하려는 의지 • 데이터 정제 방안을 논리적이고 구체적으로 수립하려는 자세

2001070304_23v2.3 인공지능 데이터 변환하기	3.1 설계된 인공지능 학습모델의 구성을 파악할 수 있다. 3.2 파악된 구성에 따라 인공지능 학습모델에 필요한 입력 데이터를 확인할 수 있다. 3.3 통합된 데이터를 인공지능 학습모델에 적합한 입력 데이터의 형태로 변환할 수 있다. 3.4 변환한 인공지능 데이터에 대해서 검증을 수행하여 결과보고서를 작성할 수 있다.
	【지식】 • 데이터 정규화 이론 • 정적, 동적 데이터 표현 방법 • 메타데이터 활용 방법론 • 인공지능 학습용 데이터 품질관리 가이드
	【기술】 • 데이터 유형별 전환도구 사용 기술 • 데이터 품질 분석 능력 • 데이터 처리 기술 • 데이터 유형 변환기술 • 데이터 분산기술
	【태도】 • 데이터 변환을 개선하려는 의지 • 데이터 변환 방안을 논리적이고 구체적으로 수립하려는 노력
2001070304_23v2.4 인공지능 데이터 축소하기	4.1 인공지능 학습의 최적화를 위해 변환된 인공지능 데이터를 확보할 수 있다. 4.2 확보한 인공지능 데이터의 크기가 인공지능 학습모델에 부합하는지 여부를 판단할 수 있다. 4.3 인공지능 학습결과가 부합하도록 인공지능 데이터의 크기를 축소할 수 있다.
	【지식】 • 인공지능 학습모델 • 기술 통계론 • 데이터 축소 방법
	【기술】 • 데이터 전환도구 사용 기술 • 데이터 품질 분석 능력 • 데이터 처리 기술 • 데이터 유형 변환기술
	【태도】 • 학습 결과에 부합하게 데이터 크기를 축소하려는 의지 • 데이터 축소 방안을 구체적으로 수립하려는 노력
2001070304_23v2.5 인공지능 데이터 라벨링하기	5.1 인공지능 데이터 라벨링 가이드라인의 기준과 목적을 파악할 수 있다. 5.2 인공지능 데이터 라벨링 유형과 저작도구의 종류, 기능을 파악할 수 있다. 5.3 데이터 유형에 적합한 데이터 라벨링 저작도구를 활용할 수 있다. 5.4 인공지능 데이터 라벨링 가이드라인을 기반으로 데이터 라벨링을 수행할 수 있다. 5.5 인공지능 데이터 라벨링 가이드라인을 기반으로 데이터 라벨링 수행 결과를 검사할 수 있다.
	【지식】 • 인공지능 학습데이터 라벨링 기초 이론 • 학습데이터 라벨링 저작도구 사용 방법

2001070304_23v2.5 인공지능 데이터 라벨링하기	【기술】 • 인공지능 학습데이터 구축 프로세스 파악 능력 • 학습데이터 저작도구 활용 능력 • 라벨링 가이드라인 파악 능력 • 인공지능 학습데이터 라벨링 프로젝트 배경, 목적 파악 능력
	【태도】 • 작업 가이드라인에 대한 수용적인 자세 • 작업 가이드라인을 적극적으로 이해하려는 노력
2001070304_23v2.6 인공지능 데이터 비식별화하기	6.1 인공지능 데이터의 비식별화 대상을 선별할 수 있다. 6.2 선별된 대상의 비식별화 방안을 결정할 수 있다. 6.3 인공지능 데이터의 비식별화 프로세스를 정의할 수 있다. 6.4 인공지능 데이터의 비식별화를 수행할 수 있다.
	【지식】 • 데이터 3법 • 비식별화 기법 • 개인정보활용동의서 작성 방법 • 비식별화 방법론
	【기술】 • 인공지능 학습데이터 비식별화 기술 • 개인정보활용동의서 작성 능력 • 텍스트(Text)데이터 비식별화 기술 • 멀티모달 데이터 비식별화 기술 • 비식별화 프로세스 작성 기술
	【태도】 • 비식별화 대상을 정확하게 식별, 파악하려는 노력 • 개인정보보호 문제에 대해 심도 있게 분석하려는 자세

□ 적용범위 및 작업상황

고려사항

-이 능력단위는 인공지능 서비스를 위하여 데이터의 정제, 통합, 변환, 축소를 통해 학습대상인 데이터를 가공하는 업무에 적용한다.

-인공지능 FATE란 인공지능의 공정성(Fairness), 책임성(Accountability), 투명성(Transparency), 윤리의식(Ethics)을 의미한다.

-인공지능 윤리기준이란 3대 기본원칙에 따라 인간성(Humanity)을 구현하기 위해 인공지능의 개발 및 활용 과정에서 인간의 존엄성 원칙, 사회의 공공선 원칙, 기술의 합목적성 원칙을 지키는 것을 의미한다.

-인공지능 데이터는 사운드, 영상, 문서, 자연어, 채팅 로그, 데이터베이스 레코드 등 다양한 유형이 될 수 있으므로 유형별 데이터에 대한 전문적 지식과 처리기술이 요구된다.

-인공지능 데이터 정제는 데이터의 모순점(데이터 입력에서 사람의 실수로 발생, 데이터 표현의 모순, 일치하지 않는 코드의 사용, 원래의 의도와 다른 목적으로 사용)을 포착하여 수정하는 것을 의미한다.

-인공지능 데이터 통합 작업은 데이터 통합, 스키마 통합, 개체 식별 문제, 데이터 값 충돌 감지 및 해결, 데이터 통합에서 중복 처리 등을 포함한다.

-인공지능 데이터 변환 작업은 데이터에서 노이즈 제거, 새로운 속성 추가, 데이터에 요약 작업 또는 집계작업, 데이터 정규화 등을 포함한다.

-인공지능 데이터 축소하기는 데이터 전처리 상황에 따라 선택적으로 수행할 수 있다.

-데이터를 축소를 하는 이유는 계산 및 메모리 효율성, 잡음 제거, 일반화 능력 향상, 데이터 분석 용이성을 위해서이다.

-인공지능 데이터 라벨링 작업은 학습데이터 구축과 관련된 직무 중 학습데이터 라벨링 분야에 대하여 라벨링을 직접 수행하는 업무에 적용한다.

-데이터 라벨링이란 이미지, 영상, 음성, 텍스트, 센서 등 다양한 종류의 데이터에 인공지능이 기계학습을 위하여 기능이나 목적에 부합하는 정보를 원천데이터에 부여하는 활동을 말한다.

-원천데이터란 원시데이터를 라벨링 공정에 투입하기 위해 필요한 전처리 등 정제 작업을 수행한 데이터로 라벨링 데이터가 부여되지 않은 상태의 데이터를 말한다.

-원시데이터란 기계학습을 목적으로 획득 단계에서 수집 혹은 생성한 음성, 이미지, 영상, 텍스트 등의 데이터를 말한다.

-데이터 라벨링 가이드라인은 프로젝트의 목적과 의도를 고려하여 필요한 저작도구와 저작도구 사용방법, 학습데이터 라벨링 방법, 기준 작업시간, 품질 기준 등을 포함하는 문서를 말한다.

-저작도구란 데이터 라벨링 작업을 위한 소프트웨어를 말한다.

-데이터 라벨링 기법은 비전, 음성, 텍스트, 센서 등에 따라 상이하나, 비전분야에서는 바운딩 박스, 키포인트, 폴리곤(폴리라인), 시맨틱 세그맨테이션 등의 기법이 있다.

-데이터 비식별화란 데이터에 포함된 개인정보를 삭제하거나 다른 정보로 대체하여 데이터 내에서 특정 개인을 식별하지 못하도록 행하는 데이터처리를 말한다. 여기에는 가명처리, 총계처리, 데이터값 제거, 범주화, 마스킹 등이 있다.

-비식별화 방안에는 개인정보활용동의서 작성을 포함한다.

-데이터 3법이란, 데이터 이용을 활성화하는「개인정보 보호법」,「정보통신망 이용촉진 및 정보보호 등에 관한 법률(약칭 : 정보통신망법)」,「신용정보의 이용 및 보호에 관한 법률(약칭 : 신용정보법)」등 3가지 법률을 통칭한다.

자료 및 관련 서류

- 인공지능 데이터 정의서

- 인공지능 학습 모델 설계서
- 인공지능 학습 모델 데이터 구조 정의서

장비 및 도구

- 컴퓨터, 문서작성 프로그램
- SQL 활용 프로그램
- 데이터 정제 프로그램
- 인공지능 데이터 전처리 프로그램

재료

- 해당사항 없음

□ 평가지침

권장평가방법

• 평가자는 능력단위 인공지능 데이터 전처리의 수행준거에 제시되어 있는 내용을 평가하기 위해 이론과 실기를 나누어 평가하거나 종합적인 결과물의 평가 등 다양한 평가 방법을 사용할 수 있다. • 평가자는 다음 사항을 평가해야 한다.

권 장 평 가 방 법	평 가 유 형	
	과 정 평 가	결 과 평 가
A.포트폴리오		
B.문제해결 시나리오	V	
C.서술형시험	V	V
D.논술형시험		V
E.사례연구	V	
F.평가자 질문		V
G.평가자 체크리스트	V	V
H.피평가자 체크리스트		
I.일지/저널		
J.역할연기		
K.구두발표	V	
L.작업장평가		
M.기타		

평가시 고려사항

• 수행준거에 제시되어 있는 내용을 성공적으로 수행할 수 있는지를 평가해야 한다.
• 평가자는 다음 사항을 평가해야 한다.
- 인공지능 학습 모델에 대한 개념
- 인공지능 데이터의 개념
- 인공지능 데이터의 전처리 활용 능력
- 인공지능 데이터의 전처리 결과 검증 능력
- 인공지능 데이터의 비식별화 결과 검증 능력

□ 관련기초능력

순번	관 련 기 초 능 력	
	주 요 영 역	하 위 영 역
1	의사소통능력	경청 능력, 기초외국어 능력, 문서이해 능력, 문서작성 능력, 의사표현 능력
2	수리능력	기초연산 능력, 기초통계 능력, 도표분석 능력, 도표작성 능력
3	문제해결능력	문제처리 능력, 사고력
4	정보능력	정보처리 능력, 컴퓨터활용 능력
5	기술능력	기술선택 능력, 기술이해 능력, 기술적용 능력

□ 개발·개선 이력

구 분		내 용
직무명칭(능력단위명)		인공지능모델링(인공지능 데이터 전처리)
분류번호	기존	2001070304_19v1
	현재	2001070304_23v2
개발·개선연도	현재	2023
	최초(1차)	2019
버전번호		v2
개발·개선기관	현재	
	최초(1차)	정보기술·사업관리 인적자원개발위원회(한국IT비즈니스진흥협회)
향후 보완 연도(예정)		-