



국민대학교
소프트웨어융합대학
소프트웨어학부


캡스톤 디자인 I

종합설계 프로젝트

| | |
|--------|------------------------------|
| 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 |
| 팀 명 | T-San(티끌모아 태산) |
| 문서 제목 | 계획서-집단지성을 이용한 데이터 라벨링 보상 플랫폼 |

| | |
|---------|-------------|
| Version | 1.9 |
| Date | 2020-MAR-22 |

| | |
|----|----------|
| 팀원 | 이정하 (조장) |
| | 박상일 |
| | 박지희 |
| | 윤여환 |
| | 이다은 |
| | 장태진 |

| | | | |
|--|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |


CONFIDENTIALITY/SECURITY WARNING

이 문서에 포함되어 있는 정보는 국민대학교 소프트웨어융합대학 소프트웨어학부 개설 교과목 캡스톤 디자인 수강 학생 중 프로젝트 “집단지성을 이용한 데이터 라벨링 보상 플랫폼”을 수행하는 팀 “T-San(티끌모아 태산)”의 팀원들의 자산입니다. 국민대학교 소프트웨어융합대학 소프트웨어학부 및 팀 “T-San(티끌모아 태산)”의 팀원들의 서면 허락없이 사용되거나, 재가공 될 수 없습니다.

문서 정보 / 수정 내역


| | |
|-----------------|------------------------------------|
| Filename | 계획서-집단지성을 이용한 데이터마ining 보상 플랫폼.doc |
| 원안작성자 | 이정하, 윤여환, 이다은, 장태진 |
| 수정작업자 | 이정하, 박지희, 박상일, 윤여환, 이다은, 장태진 |

| 수정날짜 | 대표수정자 | Revision | 추가/수정 항목 | 내 용 |
|------------|-------|----------|----------|----------------------------|
| 2020-03-05 | 이정하 | 1.0 | 최초 작성 | |
| 2020-03-08 | 장태진 | 1.1 | 내용 수정 | 일정 및 역할분담 재조정 |
| 2020-03-10 | 박지희 | 1.2 | 내용 작성 | 개요 및 목표 작성 |
| 2020-03-11 | 박상일 | 1.3 | 내용 작성 | 연구/ 개발 및 개발 결과 작성 |
| 2020-03-13 | 윤여환 | 1.4 | 내용 작성 | UX/UI 설계, 기대 효과 및 활용 방안 작성 |
| 2020-03-14 | 이다은 | 1.5 | 내용 작성 | 배경기술 작성 |
| 2020-03-17 | 장태진 | 1.6 | 내용 수정 | 배경기술 수정 |
| 2020-03-19 | 이정하 | 1.7 | 내용 수정 | 배경기술 수정 |
| 2020-03-20 | 전원 | 1.8 | 내용 작성 | 팀 구성 및 전반적인 프로젝트 비용 작성 |
| 2020-03-23 | 전원 | 1.9 | 1차 점검 | 전반적인 내용 및 양식 점검 |

| | | | |
|--|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

목 차

| | | |
|----------|--------------------------|-----------|
| 1 | 개요 | 4 |
| 1.1 | 프로젝트 개요 | 4 |
| 1.2 | 추진 배경 및 필요성 | 5 |
| 2 | 개발 목표 및 내용 | 7 |
| 2.1 | 목표 | 7 |
| 2.2 | 연구/개발 내용 | 8 |
| 2.3 | 개발 결과 | 10 |
| 2.3.1 | 시스템 기능 요구사항 | 10 |
| 2.3.2 | 시스템 비기능(품질) 요구사항 | 14 |
| 2.3.3 | 시스템 구조 | 16 |
| 2.3.4 | 결과물 목록 및 상세 사양 | 17 |
| 2.4 | UX/UI 설계 | 18 |
| 2.5 | 기대효과 및 활용방안 | 23 |
| 3 | 배경 기술 | 24 |
| 3.1 | 기술적 요구사항 | 24 |
| 3.2 | 현실적 제한 요소 및 그 해결 방안 | 28 |
| 3.2.1 | 하드웨어 | 28 |
| 3.2.2 | 소프트웨어 | 28 |
| 3.2.3 | 기타 | 29 |
| 4 | 프로젝트 팀 구성 및 역할 분담 | 29 |
| 5 | 프로젝트 비용 | 30 |
| 6 | 개발 일정 및 자원 관리 | 31 |
| 6.1 | 개발 일정 | 31 |
| 6.2 | 일정별 주요 산출물 | 32 |
| 6.3 | 인력자원 투입계획 | 34 |
| 6.4 | 비 인적자원 투입계획 | 35 |
| 7 | 참고 문헌 | 35 |

| | | | |
|--|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

1 개요


1.1 프로젝트 개요

글로벌 AI 소프트웨어 시장은 2018년 약 95억달러 규모에서 연 평균 43.4%씩 성장하여 2025년에는 1,186억 달러규모에 이를 것으로 보인다. AI 음성인식 기술이 스마트폰, AI스피커 등을 통해 확산되고 있는데, 2019년 이후에는 자율주행차, 웨어러블기기, 주요 가전 등 다양한 제품으로 확대될 전망이다.

(출처: [지식정보] 인공지능(AI) 시장 글로벌 동향)

AI 시장이 확대되면서 AI 기술 개발에 사용할 수 있는 데이터셋의 필요성이 증가하고 있다. 그로 인해 데이터셋 생성의 가장 기본적인 단계인 데이터 라벨링 또한 중요해지고 있다. 데이터 라벨링이란 데이터에 대한 결과값을 붙여서 AI의 학습 단계에 사용할 수 있도록 하는 작업이다. 오늘날 데이터 라벨링 작업은 대부분 사람들에 의해 수작업으로 진행된다. 그래서 데이터셋을 생성하는 작업은 많은 비용을 필요로 한다. 따라서 많은 기업과 연구 기관들은 AI 기술 개발에 사용할 데이터셋을 생성하는데 많은 비용을 지불하고 있으며 비용 문제로 인해 AI 기술 개발을 진행하는 것에 어려움을 겪기도 한다. 그러므로 이와 같이 데이터셋을 생성하는데 많은 비용이 드는 문제를 해결하기 위해 비용을 크게 절감 시킬 수 있는 효율적으로 데이터셋을 생성할 수 있도록 해주는 데이터 마이닝 플랫폼이 필요하다.

T-SAN 프로젝트는 클라우드소싱이 포함된 데이터 마이닝 플랫폼을 개발하는 것을 목표로 한다. 플랫폼 사용자는 의뢰자와 라벨링 참여자로 나뉜다. 의뢰자는 데이터셋과 라벨링 방법, 두 가지 정보를 제공해야 한다. 라벨링 참여자는 라벨링 작업을 할 때 보상으로 포인트를 받을 수 있다. 또한 라벨링 된 데이터를 검수하는 과정에서 머신러닝 알고리즘을 이용하여 검수 작업에 필요한 인건비를 줄이면서 신뢰도 있는 데이터셋을 수집할 수 있다. 결과적으로 T-SAN은 다수의 집단 지성을 이용한 라벨링 작업과 머신 러닝 알고리즘을 이용한 검수 과정을 통해 기존의 방법 대비 적은 비용으로 데이터셋을 생성해주는 플랫폼 역할을 수행할 수 있다.

| | | | |
|--|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

1.2 추진 배경 및 필요성

1.2.1 기술의 시장 현황

데이터라벨링의 필요성이 커지면서 월마트, 구글, 마이크로소프트, 글라스도어 등 경쟁한 글로벌 IT(정보기술) 기업 또한 라벨링 회사를 이용하여 데이터를 모으고 있다. 인도의 라벨링 회사 아이메리트(iMerit)는 2500명의 직원이 사진 및 동영상을 분류한다. 또 다른 라벨링 기업인 사마소스(Samasource)는 케냐, 우간다 등에서 저임금 노동자를 이용하여 데이터 라벨링 작업을 수행한다.


이러한 방식의 라벨링은 정확도가 높다는 장점이 있지만 국내에서 사용할 수 없고, 노동자의 교육과 고용 등의 과정에서 비용이 발생할 수 있다는 한계를 가지고 있다. 앞선 기존 기업들에서 나타나는 한계들로 인해 최근 데이터 라벨링 시장에서 클라우드소싱을 활용한 기업들이 생겨나고 있다. 대표적인 예로는 우리나라의 클라우드웍스가 있다. 클라우드웍스는 국내 라벨링 기업 중에 하나로, 기존 기업들의 한계를 극복하기 위해 클라우드 소싱 개념을 활용하였다. 이 회사는 사용자들이 라벨링을 하면 이에 대한 보상을 현금으로 지급한다. 하지만 진행되는 프로젝트수가 매우 적고 라벨링을 진행하는 사용자들 사이의 수의 균형이 맞지 않아 라벨링을 진행하고 싶은 사람들이 쉽게 라벨링을 진행 할 수 없는 경우가 존재한다. 또한 교육을 받은 노동자가 아닌 사용자들이 라벨링을 하기 때문에 라벨링 결과에 대한 검수가 필수적이다. 이로 인해 검수를 위한 사용자들이 존재하며, 이러한 구조로 인해 비용이 더 발생한다.

아마존에서 운영하는 아마존 메커니컬 터크(Amazon Mechanical Turk)라는 플랫폼은 클라우드소싱을 이용하여 다양한 문제를 해결하는 플랫폼으로, 여러 카테고리 중에 하나도 라벨링 작업을 포함하고 있으며, 세계적으로 이러한 플랫폼은 현재 계속해서 만들어지고 있는 추세이다.

본 프로젝트에서 클라우드소싱 개념을 도입하여 데이터 라벨링을 의뢰하는 의뢰자와 데이터를 라벨링을 수행하는 라벨러 사이의 플랫폼을 개발하는 것을 목표로 한다.


1.2.2 필요성

첫 번째, 클라우드 소싱 개념을 이용하면 참여를 원하는 모든 사람들이 라벨링 작업에 참여할 수 있으며, 또한 고용과 교육 등에서 나오는 비용을 절감할 수 있다. 하지만 교육된 노동자가 아닌 모든 유저들이 참여하는 만큼 신뢰도가 불명확하다는 단점이 있다. 이를 위해 기존 클라우드소싱 기업들은 유저들이 처리한 라벨링의 결과를 검수하기 위해 검수자 역할의 유저들을 고용하여 보상함으로써 이 문제를 해결하고 있다. 이는 사람이 직접 확인하기 때문에 정확도가 높다는 장점이 있지만 많은 비용이 든다. 본 프로젝트에서는 클러스터링 기능을 이용하여 라벨링된 결과의 신뢰도를 유지하고, 검수를 위한 비용을 획기적으로 줄일 수 있다.

| | | | |
|--|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

두번째, 기존 기업에서 현금을 이용하여 라벨링의 의뢰와 보상을 하는 것과 달리 본 프로젝트에서는 포인트와 블록체인을 이용한다. 라벨러는 작업에 대한 보상으로 포인트를 지급받을 수 있으며 지급받은 포인트는 플랫폼과 계약되어 있는 기업과의 거래에서 사용이 가능하다. 의뢰자가 작업을 의뢰하거나, 기업에서 포인트를 현금화하는 과정에서 블록체인이 사용되는데 사용되는 암호화폐는 Ethereum의 ERC20 토큰 표준을 사용한다. 대부분의 이더리움 기반 토큰들이 ERC20을 따르고 있고 많은 거래소와 지갑에서 이 방식을 지원하고 있으므로 상장이 진행된 이후에는 토큰 판매 및 구매에 대한 문제가 적을 것이다. 더 나아가 토큰의 상용화를 이끌어 낸다는 점에서 큰 장점을 갖고 있다.

세 번째, 공급되는 데이터와 라벨러의 수요에서 불균형이 일어나는 것을 방지하기 위해 유저들의 신뢰도, 활동수준에 따라 등급을 나누어 수행할 수 있는 작업에 제한을 두어 데이터의 공급과 수요의 불균형을 해소할 수 있다.

| | | | |
|---|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

2 개발 목표 및 내용

2.1 목표

본 프로젝트는 의뢰자에게 받은 데이터를 사용자(라벨러)들이 라벨링하여 유의미한 정보를 생성하는 플랫폼을 개발하는 것을 목표로 한다.

반응형 웹을 통해 모바일환경을 같이 지원하는 것을 목표로 한다. React를 통해 프론트엔드를 구현하도록 한다. 보다 빠른 개발 속도를 위해 state 관리 모듈인 mobx와 디자인 프레임 워크인 react-semantic-ui를 적극 사용함으로써 프론트 엔드 개발의 효율을 높일 수 있다.


라벨러가 라벨링을 통해 토큰을 획득 할 수 있으며, 보상으로 받은 토큰을 거래소와 개인 거래 등을 통해 판매 한다. 라벨링을 요청하는 의뢰자는 의뢰를 위한 보상 토큰을 라벨러에게 구매함으로써 공급과 수요의 발생으로 토큰의 가치가 형성할 수 있도록 한다. 대부분의 이더리움 기반 토큰들이 ERC20을 따르고 있기 때문에 보상을 위해 지급되는 암호화폐는 Ethereum의 ERC20 토큰 표준을 사용하도록 한다.

전반적인 대부분의 기술들은 Back-end에서 각각의 시스템으로 작동한다. 회원 정보 관리 및 게시물 업로드 등의 일반적인 기능들은 기존 웹사이트 개발 관례에 맞게 개발하며 특수 기능들은 각각의 시스템이 데이터베이스에 접근하여 처리한다. 웹페이지의 백엔드는 기본 기능 외에 토큰 입금 / 토큰 출금 등의 보상 관련 API들을 처리한다. Django를 통해 서버를 개발하며 GraphQL을 통해 프론트엔드와 통신 하는 것을 목표로 한다.

회원DB, 결산 내역 등 데이터가 많지 않거나 중요할 것으로 예상되는 데이터들은 Maria DB에, 크롤링과 라벨링 등으로 방대한 데이터가 저장 될 것으로 예상이 되는 부분은 MongoDB를 통해 데이터를 저장하여 효율적인 DB관리를 하도록 한다.

클러스터링 방법을 이용하여 유저 및 데이터라벨링의 신뢰도를 파악하도록 한다. 정확한 라벨링을 많이 진행한 라벨러일 수록 “신뢰도”라는 가중치를 높게 부여한다. 정확한 라벨링 여부를 활용 하는데 사용하여 라벨링 결과의 신뢰도를 높이는 것을 목표로한다.

또한, T-SNE로 라벨링 된 데이터셋을 2차원 공간에 투영하여 라벨링 결과와 데이터에 대한 클러스터링 결과를 사용자에게 나타낸다.

| | | | |
|---|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |


2.2 연구/개발 내용

2.2.1 웹

- 웹 프론트엔드
 - React.js로 반응형 Web Front End를 개발한다.
 - 사용자가 짧은 시간 내에 많은 데이터를 라벨링할 수 있도록 편리한 UX/UI 를 제공한다.
 - 웹 페이지는 데이터에 대한 라벨링을 작업하는 기능과 데이터 라벨링을 의뢰하는 기능으로 구분되어 있다.
 - 라벨링된 결과를 분석하여 그래프를 통해 시각화하여 보여준다.
 - GraphQL을 통해 Back-end에 데이터를 요청하며, 전달받은 데이터를 보여준다.
- 웹 백엔드 (API서버, 커맨드 센터)
 - Django로 서버를 개발한다.
 - GraphQL을 통해 Front-end로 부터 입력 데이터와 반환해야하는 데이터의 질의를 받아 반환한다.
 - MongoDB를 Driver 단위에서 직접 제어함으로써 방대한 양의 데이터의 입출력을 처리하며 검색이 빈번히 발생하는 필드는 인덱싱 처리한다.
 - MySQL은 SQLAlchemy를 통해 질의문을 직접 작성하지 않고 함수 호출 및 연산기호를 통해 간접적으로 질의문을 생성하여 사용한다. GraphQL과 함께 사용함으로써 개발의 효율과 두 질의문 간의 연동성을 높인다.

2.2.2 데이터 크롤링 및 전처리

- 크롤러 데몬
 - 뉴스 기사, 댓글 등의 문서 데이터와 여러 이미지 데이터를 주기적으로 크롤링하여 공공 데이터셋을 플랫폼 자체적으로 수집하며, 수집한 데이터셋은 필요에 따라 의뢰자에게 제공한다.
 - 크롤러는 Python으로 개발한다.
 - 수집한 데이터들은 1차적으로 가공 한 뒤 MongoDB에 저장한다.
- 문서 벡터 데몬
 - 크롤러를 통해 수집된 문서 데이터는 원본 데이터와 Python의 KoNLPy 라이브러리로 문서 데이터에 대한 형태소 분석을 진행하여 단어들을 추출한다.

| | | | |
|--|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |


- 추출한 단어들에 대해 기존 수집된 데이터를 이용하여 문서벡터를 생성한다.
- 각 문서 별 벡터와 단어 빈도수들은 문서 레코드를 외래키로 하여 MongoDB에 저장한다.

2.2.3 블록체인을 이용한 포인트 관리 시스템

- **블록체인(Smart Contract)**
 - 라벨러에게 보상으로 지불하는 포인트의 발행 및 사용을 투명하게 볼 수 있도록 한다.
 - 사용자의 익명성을 지키는 동시에 투명성을 높이기 위한 방안을 연구한다.
 - 이더리움 네트워크 상에서 솔리디티를 이용하여 개발한다.
 - 실제 이더리움을 소비하여 메인넷에서 작업을 하는 것은 비용 문제가 발생 할 수 있으므로 테스트넷에서 개발을 진행한 후, 실제 배포시에 메인넷에 업로드한다.
 - 포인트의 트랜잭션 기록이 모두 체인 상에 기록되므로 포인트의 발행 및 순환 과정을 투명하게 관리 및 제어한다.
- **블록체인 모듈**
 - 파이썬에서 블록체인을 사용 할 수 있는 라이브러리로 제작한다.
 - 스마트 컨트랙트를 통해 구현된 함수들을 사용 할 수 있도록 Web3를 사용한다.

2.2.4. 데이터 클러스터링 및 시각화

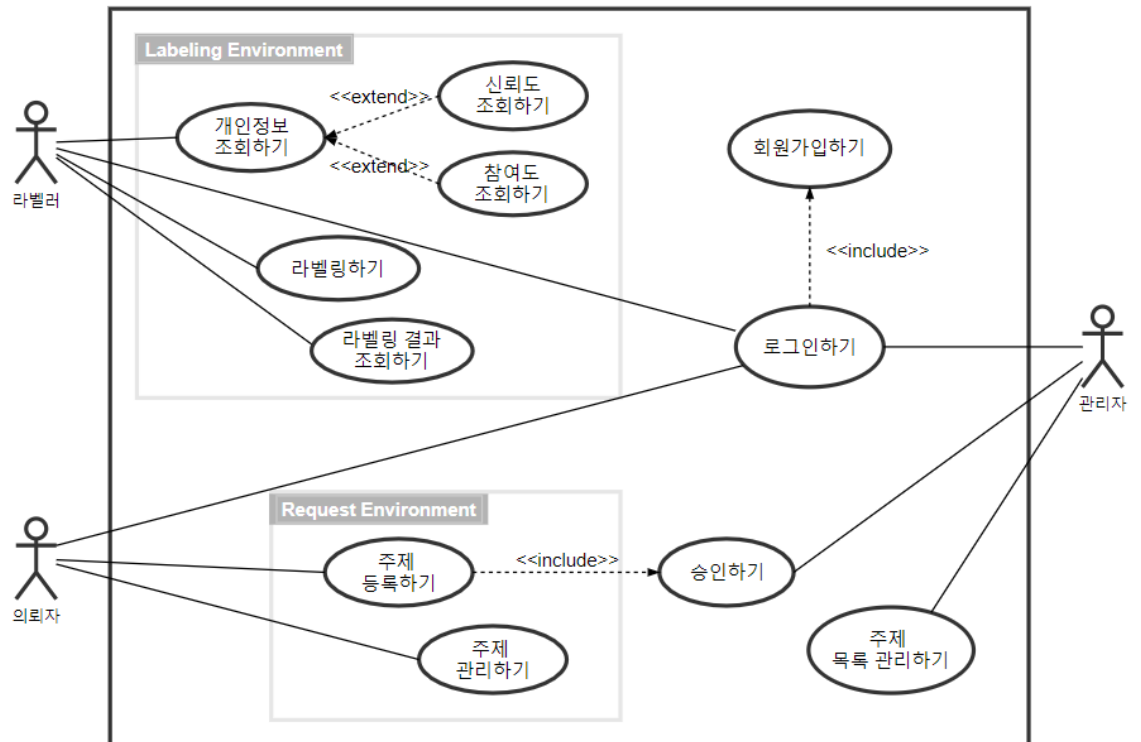
- **클러스터링 데몬**
 - 라벨링한 데이터에 대한 클러스터링 작업을 진행한다.
 - 우선, 데이터에 대한 라벨 없이 Python의 Scikit-learn 라이브러리를 이용하여 클러스터의 수가 정해져 있다면 K-means, 그렇지 않다면 AffinityPropagation 클러스터링을 수행하여 데이터를 클러스터링 한 뒤, 데이터의 클러스터링 결과와 라벨 결과를 비교하여 라벨링이 잘 되었는지 확인한다.
- **데이터 라벨링 T-SNE데몬**
 - Scikit-learn 라이브러리의 T-SNE로 라벨링 된 데이터셋을 2차원 공간에 투영하여 라벨링 결과와 데이터에 대한 클러스터링 결과를 나타낸다.

| | | | |
|---|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |


2.3 개발 결과

2.3.1 시스템 기능 요구사항

UseCase Diagram




[그림1] Use case 다이어그램

| | | | |
|--|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |


| | | |
|---------------------|---|--|
| Usecase name(ID) | 라벨링하기(UC1) | |
| Triggering event | 라벨러 계정의 사용자가 라벨링 시작하기 버튼을 클릭한다. | |
| Brief description | 라벨러 계정으로 로그인한 사용자는 웹상에서 보여지는 간단한 질문들에 대해 버튼 혹은 텍스트 형식으로 답변함으로써 라벨링을 진행할 수 있다. | |
| Actors | 라벨러 | |
| Preconditions | 사용자가 라벨러 계정으로 웹에 로그인 된 상태여야한다. | |
| Post conditions | 현 질문사항을 저장하고 다음 라벨링 질문사항 페이지로 이동한다. | |
| Flow of activities | Actor | System |
| | 1. 사용자는 라벨링 작업을 요청한다. 2. 사용자는 질문에 대한 답변을 한다. 3. 사용자는 “다음” 버튼을 클릭한다. | 1.1. 시스템은 해당 라벨링 주제에 대한 작업을 이행할 수 있는지 확인한다. 1.2. 시스템은 해당 라벨링 주제를 사용자 계정 리스트에 추가한다. 1.3. 시스템은 라벨링 질문 페이지를 화면에 표시한다. 3.1. 시스템은 사용자가 응답한 답변을 저장한다. |
| Exception condition | 1.a. 사용자가 선택한 주제가 참여인원 초과 혹은 만기일 완료의 이유로 이행할 수 없는 경우 1.a.1. “이 주제는 참여할 수 없습니다”라는 메시지를 화면에 표시한다. 3.a. 사용자가 “그만하기” 버튼을 클릭한 경우 3.a.1. 시스템은 현재 화면을 종료한 후 라벨링 리스트 화면을 표시한다. | |

| | |
|-------------------|---|
| Usecase name(ID) | 라벨링 결과조회하기(UC2) |
| Triggering event | 라벨러 계정의 사용자가 결과 조회하기 버튼을 클릭한다. |
| Brief description | 라벨러 계정으로 로그인한 사용자는 웹 페이지를 통해 자신이 라벨링한 데이터에 대한 결과 등의 정보를 확인할 수 있다. |
| Actors | 라벨러 |

| | | | |
|--|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |


| | | |
|---------------------|--|---|
| Preconditions | 사용자가 1개 이상의 데이터에 대해 라벨링을 한 상태여야 한다. | |
| Post conditions | 사용자가 선택한 주제의 데이터 라벨링 결과 상세 정보가 화면에 표시된다. | |
| Flow of activities | Actor | System |
| | 1. 사용자는 라벨링 결과 조회하기를 클릭한다. 2. 사용자는 목차에서 특정 주제를 클릭한다. | 1.1. 시스템은 해당 사용자가 라벨링한 데이터에 대한 정보를 가져온다. 1.2. 시스템은 해당 계정으로 참여 이력이 있는 주제목차를 화면에 표시한다. 2.1. 시스템은 사용자가 선택한 주제에 대한 상세 정보를 가져온다. 2.2. 시스템은 사용자가 선택한 주제의 상세정보를 화면에 출력한다. |
| Exception condition | 1.a. 해당 사용자가 라벨링한 데이터가 없다면 1.a.1. "데이터 결과가 없습니다"를 화면에 출력한다. | |

| | | |
|--------------------|--|---|
| Usecase name(ID) | 주제 등록하기(UC3) | |
| Triggering event | 의뢰자 계정의 사용자가 주제 등록하기 버튼을 클릭한다. | |
| Brief description | 의뢰자 계정으로 로그인한 사용자는 웹 페이지를 통해 원하는 주제와 이와 연관된 데이터, 키워드를 등록할 수 있다. | |
| Actors | 의뢰자 | |
| Preconditions | 사용자가 의뢰자 계정으로 웹에 로그인 된 상태여야한다. | |
| Post conditions | 주제를 등록한 후, 등록된 주제 관리하기 페이지로 이동한다. | |
| Flow of activities | Actor | System |
| | 1. 사용자는 "주제 등록하기" 버튼을 누른다. 2. 사용자는 해당 필드에 주제, 데이터, 키워드를 입력한다. 3. "저장" 버튼을 누른다. | 1.1. 주제 등록 페이지를 화면에 표시한다. 2.1. 주제, 데이터, 키워드를 작성하는 란이 모두 채워져있는지 확인한다. 3.1. 사용자가 작성한 데이터를 저장한다. 3.2. 등록된 주제에 대한 상세 |

| | | | |
|--|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

| | |
|---------------------|--|
| | 관리페이지를 화면에 표시한다. |
| Exception condition | <p>2.a. 주제, 데이터, 키워드를 모두 입력하지 않은 경우</p> <p>2.a.1. "입력란을 모두 기재해 주세요"라는 메시지를 화면에 표시한다.</p> <p>2.a.2. 해당 필드를 빨간색으로 표시한다.</p> <p>2.a. 특정 필드를 형식에 맞지 않게 입력한 경우</p> <p>2.a.1. "형식에 맞게 입력해주세요."라는 메시지를 화면에 표시한다.</p> <p>2.a.2. 해당 필드를 빨간색으로 표시한다.</p> |

| | | |
|--------------------|--|---|
| Usecase name(ID) | 등록한 주제 관리하기(UC4) | |
| Triggering event | 의뢰자 계정의 사용자가 등록한 주제 관리하기를 클릭한다. | |
| Brief description | 의뢰자 계정으로 로그인한 사용자는 본인(해당기업)이 등록한 주제에 대한 진행정보를 일괄적으로 볼 수 있고 삭제할 수 있다. 그 정보는 참여현황(진행도), 수치와 표를 통해 시각적으로 보여지는 라벨링 결과를 포함한다. | |
| Actors | 의뢰자 | |
| Preconditions | 사용자가 의뢰자 계정으로 웹에 로그인 된 상태여야한다. 이전에 등록된 주제가 있어야한다. | |
| Post conditions | 참여현황(진행도), 도식화된 라벨링 결과를 확인할 수 있다. | |
| Flow of activities | Actor | System |
| | 1. 사용자는 “주제 관리하기” 버튼을 클릭한다. | <p>1.1. 시스템은 사용자가 등록한 주제 정보를 가져온다.</p> <p>1.2. 시스템은 해당 계정으로 등록된 주제에 대한 목차를 화면에 표시한다.</p> |
| | 2. 사용자는 목차에서 특정 주제를 클릭한다. | <p>2.1. 시스템은 사용자가 선택한 주제에 대한 상세 정보를 가져온다.</p> <p>2.2. 시스템은 사용자가 선택한 주제의 상세정보와 삭제 버튼을 디스플레이한다.</p> |
| | 3. 사용자는 “삭제하기” 버튼을 클릭한다. | 3.1. 시스템은 “정말 삭제하시겠습니까?” 문구의 메시지를 화면에 표시한다. |
| | 4. 사용자는 “예” 버튼을 클릭한다. | <p>4.1. 시스템은 해당 주제에 대한 정보를 삭제한다.</p> <p>4.2. 시스템은 삭제된 항목이 빠진 목차 페이지를 화면에 표시한다.</p> |

| | | | |
|--|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

| | |
|---------------------|--|
| Exception condition | 1.a. 등록된 주제가 없을 경우 1.a.1. 빈 목차를 화면에 표시한다. 3.a. 해당 주제에 대한 라벨링 작업이 이미 진행중인 경우 3.a.1. 경고 및 추후 발생할 수 있는 상황을 안내하는 메시지를 화면에 표시한다. |
|---------------------|--|

2.3.2 시스템 비기능(품질) 요구사항

* 우선순위로 정렬

1) Interface(인터페이스)


- 사용자 인터페이스는 웹을 통한 GUI로 구성한다.
- 사이트는 특정 브라우저에 종속되지 않고 환경 지원이 가능하다.
- 특정 단말기의 화면 또는 비율에 제약되지 않는 반응형 웹을 제공한다.
- 사용자의 입력/수정/삭제 동작에 대한 확인 메시지를 제공한다.

2) Quality(품질)

- (신뢰성) 시스템은 정상 상태에서 서버 시행시간 동안 사용자에게 무중단 서비스를 제공한다.
- (사용성) Web, WAS, DB 서버 등은 기존 인프라 및 소프트웨어에 대해 호환을 제공한다.
- (복구성) 시스템은 장애 발생 시 3시간 이내에 정상 상태로 복구한다
- (사용성) 사용자 기능에 대한 도움말 및 이용설명을 제공한다.
- (학습성) 사용자가 원하는 기능을 쉽게 찾아서 이용할 수 있도록 제공한다.
- (이식성) 다양한 사용자 운영체제(Windows, Linux, MacOS 등)에 영향을 받지 않는다.

3) Performance(성능)


- 시스템은 웹페이지를 사용자가 요청한 시간으로부터 1초 이내에 디스플레이 한다.
- 시스템은 검색 요청에 대해 사용자가 요청한 시간으로부터 3초 이내에 결과 페이지를 화면에 출력한다.
- 시스템은 사용자가 요청한 라벨링 작업에 대한 응답을 2초 이내에 디스플레이 한다.

| | | | |
|--|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

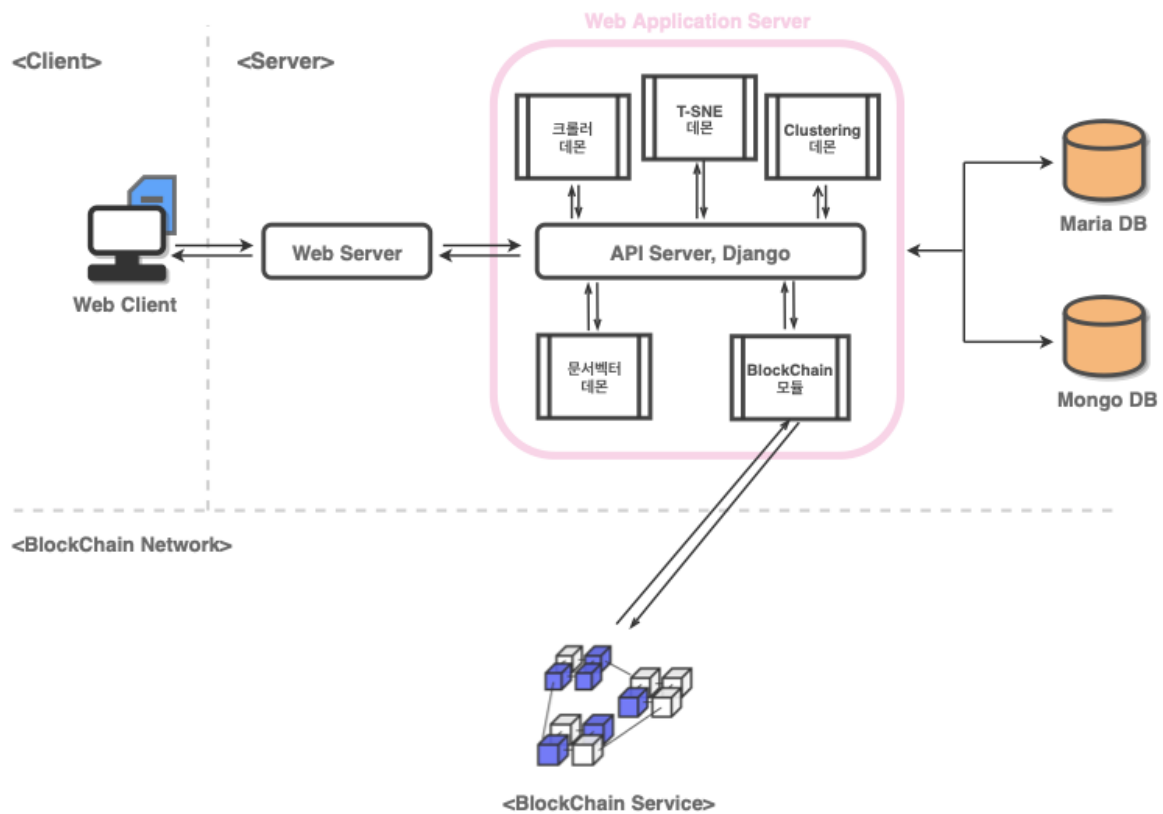
- 시스템은 사용자가 입력한 데이터 형식의 오류에 대해 1초 이내에 오류 메시지를 디스플레이 한다.
- 시스템은 사용자의 로그인 요청에 대해 1초 이내에 처리하여 로그인된 창을 디스플레이한다.
- 시스템은 200명 이상의 동시 접속 클라이언트에 안정적인 QoS를 제공한다.
- 시스템은 초당 50건의 입력 기능을 처리할 수 있어야한다.
- 시스템 자원 평균 사용률은 최대 90%를 초과하여 사용하지 않는다.

4) Security(보안성)


- 사용자의 역할에 따라 정보 접근 권한을 제한한다.
- 사용자 접근 및 정보 수정에 대한 로그를 기록한다.
- 다수의 사용자가 하나의 계정으로 동시 접속을 하지 못하도록 차단한다.
- 모듈 및 서버 간 자료교환 시 무결성을 제공한다.
- 시스템은 주기적인 백업을 실시하고, 데이터가 손상되었을 경우 백업한 정보를 사용한다.

| | | | |
|---|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

2.3.3 시스템 구조




[그림 2] 시스템 구성도

| | | | |
|--|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

2.3.4 결과물 목록 및 상세 사양

| 대분류 | 소분류 | 기능 | 형식 | 비고 |
|-----------|-------------|--|----------|-----------|
| Front-end | UI | 사용자가 신속하게 많은 양의 데이터 라벨링을 할 수 있는 UI를 제공 | React.js | |
| | Data API | 데이터 API를 연동 | GraphQL | API 문서 제작 |
| Back-end | 문서 벡터 데몬 | 요청된 문서에 대한 벡터를 생성 | Python | Server |
| | Crawling 데몬 | 라벨링 프로젝트를 위한 데이터 셋 크롤링. | Python | |
| | 블록체인 모듈 | 블록체인 관련 연동을 위한 라이브러리 | Python | |
| | 스마트 컨트랙트 | ERC20 표준 토큰 | Solidity | Ethereum |
| | MongoDB | 회원 정보와 지급 내역 등을 저장 및 관리 | Database | |
| | MariaDB | 라벨링된 데이터, 데이터셋 등을 저장 및 관리 | Database | |
| | 클러스터링 데몬 | Text에 대한 클러스터링 진행 | Python | |
| | T-SNE 데몬 | 데이터셋의 차원을 축소시켜서 데이터셋 시각화 진행 | Python | |

| | | | |
|---|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |


2.4 UX / UI 설계

1. 메인 페이지

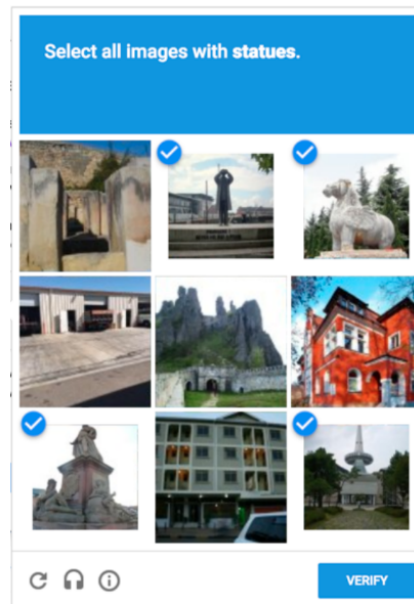


[그림 3] T-SAN 메인 페이지

T-San의 메인 페이지로 사용자들이 진행중인 라벨링 프로젝트에 참여할 수 있도록 화면에 출력한다. 상단 설명 이미지에서는 이 플랫폼을 소개하는 내용과 간단한 사용법을 제시함으로써 사용자가 해야 하는 행위를 요구 할 수 있다. 하단에는 현재 진행중인 라벨링 프로젝트를 카드형식으로 표현하며 의뢰자가 등록한 주제를 카드형태로 보여준다.


| | | | |
|---|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

2. 이미지 데이터 라벨링 데모 페이지

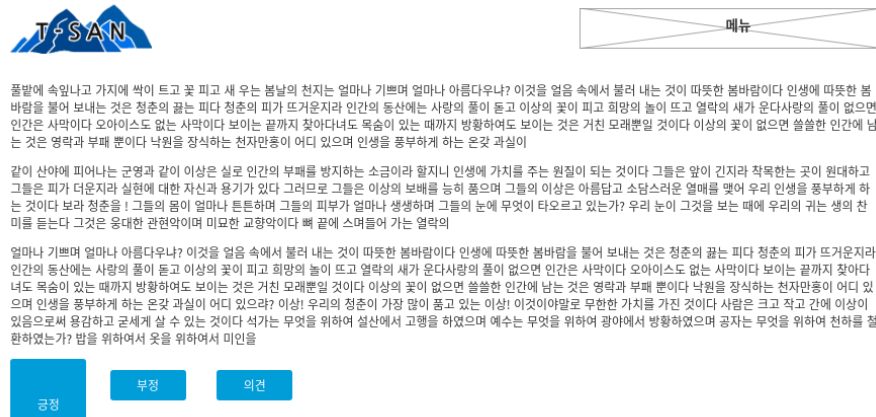


[그림 4] T-SAN 이미지 데이터 라벨링 데모

이미지 라벨링 데모 페이지로 의뢰자가 요구하는 형식에 맞게 라벨링 참여자가 이미지 데이터에 대한 라벨링을 할 수 있도록 UI를 제공한다. [그림 4] 과 같이 여러 개의 이미지에서 의뢰 요구에 맞는 이미지를 선택하는 형식을 제공하며, 의뢰의 유형에 따라 하나의 이미지에서 특정 부분을 드래그하여 선택하는 형식(예: 다음 사진에서 강아지 코 부분을 선택하시오.) 등 여러가지의 라벨링 기술들을 추가 한다.

|  <div> 국민대학교 소프트웨어학부 캡스톤 디자인 I </div> | 계획서 | | |
|--|-------------------------|--------------------------|-------------|
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

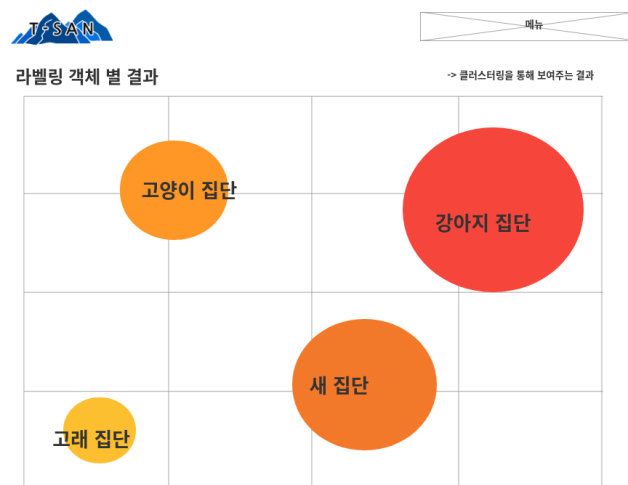
3. 텍스트 데이터 라벨링 데모 페이지



[그림 5] T-SAN 텍스트 데이터 라벨링 데모


텍스트 라벨링 데모 페이지로, 의뢰자가 요구하는 형식에 맞게 라벨링 참여자가 텍스트 데이터에 대한 라벨링을 할 수 있도록 UI를 제공한다. 의뢰 유형에 따라 텍스트 데이터의 내용에 대해 라벨을 선택 할 수 있다. (예: 다음 글이 긍정, 부정 인지 선택하시오)

4. 군집 별 데이터 라벨링 결과 페이지(군집화)



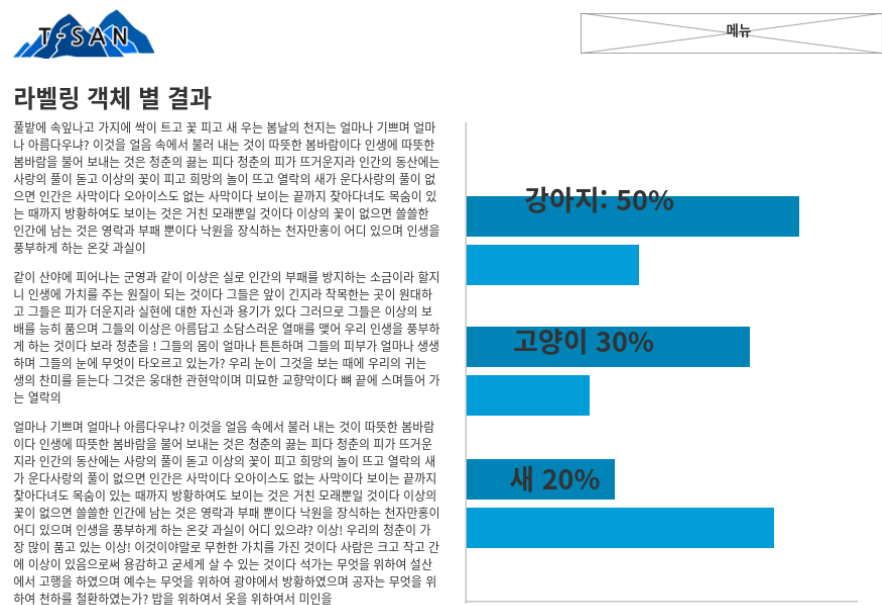
-> 비지도학습(클러스터링), TSNE 사용

[그림 6] 이미지 데이터 라벨링 결과(군집화)

| | | | |
|--|-------------------------|--------------------------|-------------|
|  <div> 국민대학교 소프트웨어학부 캡스톤 디자인 I </div> | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |


[그림 6] 는 이미지 데이터 라벨링 결과 페이지로 라벨링된 이미지 데이터를 T-SNE를 이용하여 이미지 데이터셋을 분석하여 군집화 한 결과를 시각화하여 그래프로 보여준 결과이다. 이와 같이 텍스트 데이터 또한 라벨링 결과와 비지도 학습을 결과를 같이 시각화를 통해 보여줌으로써 의뢰자가 쉽게 군집별 데이터를 이해 할 수 있고, 사용할 수 있도록 한다.

5. 라벨링 대상 객체의 데이터 라벨링 결과 페이지

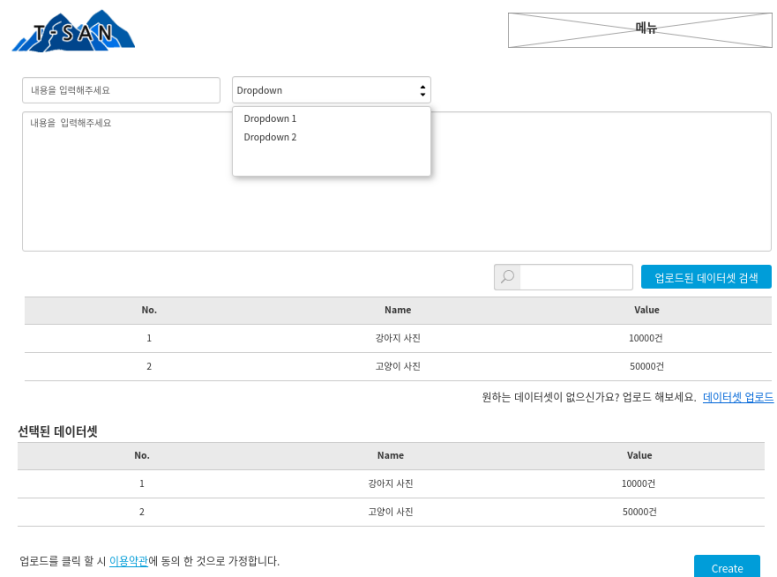


[그림 7] 텍스트 데이터 라벨링 결과 페이지

라벨링 대상 객체 1개에 대한 라벨링 현황 및 결과가 어떠한 분포가 되는지 자세한 정보를 보여준다. (예: 질문이 ‘이 글은 어떤 동물에 관련된 문서입니까?’ 이고 왼쪽 글이 데이터셋의 일부 일 때, 오른 쪽 그래프는 해당 글에 대해 라벨링 결과를 그래프로 보여준다.)

| | | | |
|---|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

6. 라벨링 프로젝트 등록 페이지




The screenshot shows the T-SAN web interface for project registration. At the top, there's a header with the T-SAN logo and a menu button. Below the header, there are two input fields for content, each with a dropdown menu. The first dropdown is open, showing 'Dropdown 1' and 'Dropdown 2'. Below these fields is a search bar and a button labeled '업로드된 데이터셋 검색'. A table displays uploaded datasets with columns 'No.', 'Name', and 'Value'. The table has two rows: '강아지 사진' (10000건) and '고양이 사진' (50000건). Below this table is a link '원하는 데이터셋이 없으신가요? 업로드 해보세요. 데이터셋 업로드'. Another section titled '선택된 데이터셋' shows a similar table with the same two rows. At the bottom, there's a note about terms of service and a 'Create' button.

| No. | Name | Value |
|-----|--------|--------|
| 1 | 강아지 사진 | 10000건 |
| 2 | 고양이 사진 | 50000건 |

| No. | Name | Value |
|-----|--------|--------|
| 1 | 강아지 사진 | 10000건 |
| 2 | 고양이 사진 | 50000건 |

[그림 8] 라벨링 프로젝트 등록 페이지

라벨링 프로젝트 등록 페이지는 의뢰자가 라벨링 프로젝트를 신청할 수 있도록 한다. 의뢰자는 프로젝트에 대한 간단한 설명과 함께 라벨링하고자 하는 데이터 종류(이미지 또는 텍스트)와 데이터셋, 라벨링 목록들을 작성하여 라벨링 페이지를 생성 할 수 있다.

| | | | |
|---|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

2.5 기대효과 및 활용방안

오늘날 인공지능은 더 많은 양의 데이터와 보다 빠른 처리 능력, 그리고 더 강력한 알고리즘이 결합되어 더욱 널리 보급되고 있다. 인공지능의 영역이 실제 세계로 확장, 실생활에서 실시간 데이터를 확보하고, 이를 기반으로 서비스 하는 시대가 온 것이다. 그러나 뛰어난 환경과 리소스를 갖춘 소수의 기업과 조직을 제외하면, 현실적인 비즈니스 현장에는 인공지능은 **데이터의 부족, 기술적 이슈, 인력 부족 등 많은 이슈**를 동시에 갖고 있다.


(출처: 인공지능 신문)



[그림 9] 데이터 라벨링 분야의 회사들

최근 들어 '데이터 라벨링' 분야에서 창업과 투자도 늘고 있으며 슈퍼브에이아이, 셀렉트스타, 크라우드웍스 등 데이터 라벨링 기업이 사업을 확대하고 있다. 일례로 크라우드웍스의 경우, 진행 가능한 작업 리스트 중 하나를 골라 데이터에 라벨링을 하는 것이다. 그리고 이에 대한 보상을 현금으로 지급하는 것으로 알려져 있다. 그리고 사람들이 라벨링을 한 데이터에 대해 사람이 일일이 검수하는 작업으로 진행된다. 검수자의 경우, 라벨링하는 사람보다 2배로 돈을 지급한다. 많은 사람들이 먼저 라벨링 작업을 완료하면 남은 작업이 많지 않아 처음 도전하는 사람들은 라벨링 작업을 체험해볼 수 없다. 이를 통해 특정 사람들이 독점하여 여러 사람들이 참여하기에 진입장벽이 높다는 것을 알 수 있다.

우리는 이를 해결하기 위해 많은 사람들에게서 라벨링 된 데이터를 제공받으면 검수자를 따로 고용하지 않고 인공지능을 이용하여 이를 검수한다. 검수 후, 라벨링한 참여자들에게 포인트를 이용하여 보상한다. 포인트 적립 및 사용에 대한 검증은 블록체인에 기록된다. 그러므로 발행 및 순환되는 포인트가 모두 투명하게 관리 및 제어된다 또한 포인트의 총량으로 "보상을 많은 받은 사람 = 신뢰가 높은 사람", "보상을 적게 받은 사람 = 신뢰도가 낮은 사람"으로 계산할 수 있다.

| | | | |
|--|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

라벨링 참여자들에게 주어진 포인트는 블록체인 네트워크를 형성하는 플랫폼에서도 사용할 수 있도록 한다. 이를 가능케 하기 위해 서버에서는 T-SAN 플랫폼이 다른 플랫폼들과 블록체인으로 거래를 하며 B2B모형을 형성한다. 이를 통해 효과적인 블록체인 기술의 활용을 기대할 수 있다.

이 프로젝트는 신뢰도에 따라 하루의 최대 참여 횟수를 제한하여, 라벨링의 신뢰도를 높이고 참여 가능한 프로젝트들을 고르게 분배 할 수 있도록 할 것 이며 공공프로젝트를 주기적으로 생성함으로써 사용자들의 이탈을 막을 수 있도록 함으로써 이 문제를 해결 할 것이다.

3 배경 기술

3.1 기술적 요구사항

3.1.1 개발 환경

1) 컴퓨팅 리소스

a) AWS EC2

AWS EC2는 AWS에서 제공하는 클라우드 서버 컴퓨터이다. EC2는 처리 데이터의 크기에 따라 적절한 자원을 할당하여 작업을 진행할 수 있으며, GPU 컴퓨팅을 지원하기 때문에 머신러닝 알고리즘을 안정적으로 수행하기에 적합한 환경이다.

b) Jupyter Notebook


Jupyter Notebook는 Python 코드를 작성하고 바로 실행시켜 결과를 확인할 수 있도록 해주는 웹 어플리케이션이다. 데이터 분석과 시각화 알고리즘을 구현하고 수행 결과를 Markdown으로 문서화 하기 편리하므로 공동 작업에서 활용 가능하다.

2) 데이터 저장 및 관리

데이터를 저장 및 관리하기 위해 데이터 유형에 따라 MongoDB와 MariaDB 중 적합한 데이터베이스에 저장한다.

a) MongoDB

크롤링과 라벨링 등으로 방대한 데이터가 저장 될 것으로 예상되는 부분은 MongoDB를 통해 데이터를 저장한다. NoSQL의 한 종류인 MongoDB의 특성상 매우 빠른 검색이 가능하므로 방대한 데이터를 검색하기에 매우 적합하다. Mongo DB를 쓰는 경우 Node.js의 Mongoose 등과 같은 ORM Framework를 쓰는 경우 속도가 현저히 저하되기 때문에 Mongo DB Driver 자체를 사용하여 개발 하는 것을 목표로 한다. 스키마가 정확히 정해지지 않거나 빈번히 스키마가

| | | | |
|--|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

변경 될 수 있는 데이터들을 저장하기에도 매우 적합하다.

b) MariaDB

전반적인 회원 DB 및 정산 관리는 MariaDB를 통해 관리한다. MongoDB와의 연동은 MariaDB의 Primary Key를 기반으로 한다. MariaDB는 스키마가 정해져 있으므로 데이터가 많지 않거나 중요한 데이터(회원 DB, 결산 내역)등의 데이터를 저장한다.

3) 운영체제

- a) Ubuntu: 개발 환경, 배포 환경
- b) macOS Catalina: 개발 환경
- c) Windows 10: 개발 환경

4) 개발 언어

a) JavaScript

i) React.js

React.js는 Facebook에서 개발한 node.js 라이브러리로, Virtual DOM 개념을 도입하여 가상의 DOM 객체와 변경된 데이터로 생성된 Virtual DOM을 비교하여 변경된 요소에 대해서만 기존의 DOM 객체에 적용하는 방식을 사용한다.

ii) Mobx

MobX는 React.js state 관리 라이브러리로 최소한의 공수로 상태관리 시스템을 설계할 수 있게 해준다. Observable State, Computed Value, Reactions, Actions의 주요 개념을 이용하여 효율적으로 React.js의 state 관리를 할 수 있다.

iii) React Semantic-UI


React Semantic-UI는 기존 JQuery와 함께 사용한 Semantic-UI의 React 프레임워크 버전이다. 이 프레임워크는 디자인 프레임워크로, 미리 디자인된 컴포넌트들을 사용함으로써 웹 사이트 디자인 때문에 개발 시간이 길어지는 문제를 해결할 수 있고, 모바일 환경에 적합한 반응형 웹 사이트를 쉽게 개발할 수 있다.

iv) GraphQL

GraphQL은 SQL과 비슷한 역할을 하는 API를 위한 쿼리(질의) 언어이다. SQL은 데이터베이스 시스템에 저장된 데이터를 효율적으로 가져오는 것을 목적으로 하지만 우리가 사용 할 GraphQL은 웹 클라이언트에서 데이터를 서버로부터 질의문을 통해 원하는 필드의 데이터를 효율적이고 직관적으로 가져오는 것을 목적으로 한다. 주로 클라이언트 시스템에서 작성하고 호출한다.

b) Python

i) Keras

| | | | |
|--|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

Keras는 Python으로 작성된 딥러닝 라이브러리이다. 내부적으로 TensorFlow, CNTK 혹은 Theano의 딥러닝 전용 엔진으로 구동된다. 시퀀스 모델 구조로 레이어를 순차적으로 쌓아 딥러닝 모델을 만들 수 있도록 하는 등 딥러닝 모델의 생성과 학습, 사용에 있어서 매우 직관적이고 간결한 API를 제공한다.

ii) Scikit-learn

Scikit-learn는 Python으로 작성된 머신러닝 라이브러리로, 분류, 회귀, 군집화, 의사결정 트리 등의 다양한 머신러닝 알고리즘과 데이터 처리 알고리즘을 제공한다.

iii) Django

Python의 Web Framework로 빠르고 보안이 철저하며 확장성 또한 높다. Web App에서 제공되어야하는 사용자 인증, 사용자 관리 등의 기술이 기본적으로 구현되어 있다.

iv) SQLAlchemy

SQLAlchemy는 Python 데이터베이스 툴킷으로, MariaDB를 질의문으로 직접 사용하지 않고 Python Class와 함수 호출을 통해 Database에 접근 및 제어가 가능하도록한다. 또한 프레임워크 단위에서 SQL Injection을 원천적으로 차단하여 서비스 보안이 한층 더 쉬워진다.

v) Graphene

Python GraphQL 라이브러리로, Python에서 SQLAlchemy와 함께 데이터베이스 접근에 사용 할 수 있다. 이 프로젝트에서는 GraphQL 질의를 Web Client에서 직접 작성 할 수 있게 함으로써, 사용자에게 직접 SQL을 제어 할 수 있는 기능을 이를 통해 지원함과 동시에 데이터 접근에 대한 제어를 할 수 있어 효율적이며 더욱 안전한 API 개발이 가능하다.

vi) PyMongo


Python에서 MongoDB에 관련된 작업을 하기 위한 라이브러리이다. 이 프로젝트에서는 빠른 속도를 위해 MongoDB를 사용하므로 속도를 저하시키는 원인이 되는 ORM Framework를 사용하지 않는다. Mongo Client Driver를 직접 사용 할 수 있는 PyMongo를 사용한다.

vii) PyMysql

Python에서 Mysql에 관련된 작업을 하기 위한 라이브러리로, MariaDB 또한 지원 하므로 이를 사용한다. 이 프로젝트에서는 SQLAlchemy에서 Mariadb에 접속하기 위한 수단으로 PyMysql을 사용한다.

viii) KoNLPy

KoNLPy파이썬에서 한국어 자연어 처리를 위한 파이썬 라이브러리이다. 텍스트 요약, 자동 질의응답 시스템, 챗봇, 기계 번역 등을 위해 사용되며, 이 프로젝트에서는 단어 추출 및 문서 벡터 제작을 위한 용도로 사용한다.

| | | | |
|--|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

c) Solidity

i) SafeMath

ERC20 토큰 개발에 앞서, 수 계산에서 발생 할 수 있는 Overflow, Underflow를 통한 토큰 부풀리기를 방지하기위해 SafeMath 라이브러리를 활용함으로써 이에 대한 시큐어코딩을 진행한다.

ii) ERC20

어느 거래소에서나 거래를 진행 할 수 있도록 ERC20 표준을 사용하기로한다.

5) 배포 환경

a) Ubuntu

i) Docker

Dockerfile과 docker-compose.yml을 통해 의존성 패키지 설치 및 서비스 실행을 진행한다. 데이터베이스 및 데이터셋 등의 중요한 파일들은 volume을 통해 host의 지정된 폴더에 저장하며, mariadb와 mongodb등 이미 환경구성이 되어있는 이미지들을 적극 활용하도록 한다.

ii) Gunicorn


Python 웹 서버 게이트웨이 인터페이스 HTTP 서버이다. 우리는 성능과 안전성 이슈를 해결하기 위해 장고의 runserver를 직접 사용하지 않고 gunicorn을 통해 장고 서버를 관리한다.

iii) Nginx

React로 작성된 웹 프론트 엔드는 빌드 이후 생성된 html, css, js 등의 static 파일들을 nginx를 통해 서비스하며, api 등의 작업은 특정 폴더의 하위 디렉토리로 접속한 경우 nginx의 리버스 프록시를 통해 Gunicorn으로 작동중인 api 서버로 연결합니다.

3.1.2 프로젝트 결과물 확인 환경

- 1) 본 서비스는 정상적으로 인터넷에 연결된 환경이 필수적이다.
- 2) 본 서비스는 AOS 8.0/IOS 11.0.1 버전 이상이어야 한다.
- 3) 본 서비스는 웹 브라우저를 통해 접속할 수 있는데, Chrome, Safari, Firefox, Microsoft Edge 이 네 가지 브라우저를 권장한다.

| | | | |
|--|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

3.2 현실적 제한 요소 및 그 해결 방안

3.2.1 하드웨어 제한 요소


- 확보된 데이터를 인공지능을 통해 학습 및 분석을 하려면 어느 정도 성능이 보장된 GPU 서버가 필요하다. 이를 위해 우선 국민대학교 소프트웨어학부에서 제공해주는 DLPC(딥러닝 프라이빗 클라우드, dlpc.cs.kookmin.ac.kr) 서버를 이용하고, 필요에 따라 고성능 그래픽 카드가 부착된 PC를 확보하여 학습 서버로 사용하도록 한다. 또한 필요하다면 AWS EC2 인스턴스를 이용하여 학습을 시킬 계획이다.

3.2.2 소프트웨어 제한 요소

- React 16 버전은 컬렉션 자료형인 Map과 Set을 사용한다. 이 기능을 자체적으로 지원하지 않거나 지원은 하지만 잘 호환되지 않는 오래된 브라우저(ex. IE 11 이하)나 기기가 있을 수 있다. 그런 버전의 브라우저를 사용하는 유저까지 수용하려면 'core-js'나 'babel-polyfill' 같은 전역 폴리필을 포함해서 구현한다. 이렇게 개발해도 놓치는 부분이 있을 수 있으니 최대한 많은 브라우저(IE, Chrome, Firefox, Microsoft Edge 등)에서 테스트 후에 배포하도록 한다.

3.2.3 기타 제한 요소

- 데이터를 인공지능을 활용하여 분석하는 단계에서 수많은 데이터가 필요한데, 데이터를 확보하기 위해서 우리가 스스로 라벨링 작업을 반복하여 분석 및 학습할 데이터를 확보해야 한다.
- 이미지 데이터에 대한 클러스터링은 ImageNet 등의 데이터를 바탕으로 학습된 pre-trained CNN 모델을 가지고 이미지의 특성을 추출하여 클러스터링에 활용하는 방법을 생각하였으나, 학습되지 않은 새로운 이미지 데이터 셋에 대한 클러스터링의 정확도에 대한 우려가 생겼다. 이를 해결하기 위해 다수에 의해 라벨링 되어 라벨링에 대한 신뢰도가 높은 데이터들로 모델을 학습시킨 뒤, 전체 데이터셋에 대한 클러스터링을 진행하는 방법을 연구해보도록 한다. 이 때, 적은 데이터를 이용한 학습 단계에서 모델의 성능을 극대화하도록 Selfie(Self-supervised Pretraining for Image Embedding)[4] 방법을 통해 진행하도록 한다.
- 상업 목적의 프로젝트가 아니고 학교 프로젝트이기 때문에 자본이 없기 때문에 라벨링에 대한 보상으로 현금을 지급하는 것이 아니라 블록체인으로 지급한다.
- 서비스의 특성상 개인의 독점을 막기 위해 1인당 1개의 계정을 원칙으로 한다. 그러므로 회원 가입시 NICE ID 를 이용한 실명 확인을 통한 본인 인증 절차가 필요하다. 하지만 NICE ID 를

| | | | |
|---|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |


이용한 실명 확인 서비스를 이용하기 위해서는 비용이 발생하는데, 그 부분이 현실적으로 감당하기 어렵기 때문에 1인당 1개의 계정을 가입한다고 가정을 하고 프로젝트를 진행한다.

4 프로젝트 팀 구성 및 역할 분담

* 굵은 글씨: 주 역할

* 일반 글씨: 보조(부) 역할

| 이름 | 역할 |
|-----|--|
| 이정하 | <ul style="list-style-type: none"> - Team Leader - Front-end 개발 - API Doc 작성 및 구현 |
| 박상일 | <ul style="list-style-type: none"> - ML(클러스터링, T-SNE)를 이용한 데이터 분석 및 데이터셋 시각화 - Dockerize |
| 박지희 | <ul style="list-style-type: none"> - ERC20 토큰 제작 - 블록체인 기반 신뢰도 알고리즘 고안 및 관련 모듈 개발 - Front-end 개발 |
| 윤여환 | <ul style="list-style-type: none"> - Front-end 개발 - ML(클러스터링, T-SNE)를 이용한 데이터 분석 및 데이터 셋 시각화 |
| 이다은 | <ul style="list-style-type: none"> - Back-end 개발 - API Doc 작성 및 구현 - DB 데이터 처리 및 관리 |
| 장태진 | <ul style="list-style-type: none"> - Software Project Leader - ERC20 토큰 제작 - 신뢰도 알고리즘 개발 - DB 설계 및 Back-end 모듈 개발 - 데이터에 대한 라벨링 판단 알고리즘(신뢰도 측정 알고리즘) 개발 - 문서 벡터 제작 및 분석 알고리즘 개발 |

| | | | |
|---|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |


5 프로젝트 비용

| 항목 | 예상치 (MD) |
|---------------|----------|
| 이슈 분석 | 5 |
| 아이디어 구상 | 5 |
| 프로젝트 분석 및 설계 | 10 |
| 전체 시스템 설계 | 10 |
| API 설계 및 구현 | 20 |
| UX/UI 설계 및 구현 | 15 |
| DB 설계 및 구현 | 15 |
| 검수시스템 설계 및 구현 | 15 |
| 블록체인 구현 | 10 |
| 서버 개발 | 20 |
| 시스템 관리 및 유지보수 | 5 |
| 회의 및 교수님 상담 | 5 |
| 합 | 135 |


6 개발 일정 및 자원 관리

6.1 개발 일정

| 항목 | 세부내용 | 1월 | 2월 | 3월 | 4월 | 5월 | 6월 | 비고 |
|--------|-------|----|----|----|----|----|----|----|
| 요구사항분석 | 이슈 분석 | | | | | | | |


| | | | | | | | |
|--|-------------------------|--------------------------|--|--|--|-------------|--|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | | | | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | | | | | |
| | 팀 명 | T-San(티끌모아 태산) | | | | | |
| | Confidential Restricted | Version 1.9 | | | | 2020-MAR-23 | |

| | | | | | | | | |
|--------|----------------|--|--|--|--|--|--|--|
| | 아이디어 구상 | | | | | | | |
| | 프로젝트 분석 및 설계 | | | | | | | |
| 관련분야연구 | 블록체인 연구 | | | | | | | |
| | Front-end 연구 | | | | | | | |
| | Back-end 연구 | | | | | | | |
| | 인공지능 연구 | | | | | | | |
| | API 설계 연구 | | | | | | | |
| 설계 | 전체 시스템 설계 | | | | | | | |
| | API 설계 | | | | | | | |
| | UX/UI 설계 | | | | | | | |
| | DB 설계 | | | | | | | |
| | 문서 벡터 데몬 설계 | | | | | | | |
| | Crawling 데몬 설계 | | | | | | | |
| | 블록체인 모듈 설계 | | | | | | | |
| | 스마트 컨트랙트 설계 | | | | | | | |
| | 클러스터링 데몬 설계 | | | | | | | |
| | T-SNE 데몬 설계 | | | | | | | |
| 구현 | 웹 개발 | | | | | | | |
| | 서버 개발 | | | | | | | |
| | 문서 벡터 데몬 개발 | | | | | | | |
| | Crawling 데몬 개발 | | | | | | | |
| | 블록체인 모듈 개발 | | | | | | | |
| | 스마트 컨트랙트 개발 | | | | | | | |
| | 클러스터링 데몬 개발 | | | | | | | |
| | DB 개발 | | | | | | | |
| | UX/UI 개발 | | | | | | | |
| 테스트 | 시스템 테스트 | | | | | | | |
| 최종 발표 | 발표 준비 및 마무리 | | | | | | | |


| | | | |
|---|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

6.2 일정별 주요 산출물

| 마일스톤 | 개요 | 시작일 | 종료일 |
|----------------|--|------------|------------|
| 계획서 발표 | 개발 환경 완성 (AWS 설치, 기본 응용 작성 및 테스트 완료) 산출물 : 1. 프로젝트 수행 계획서 2. 프로젝트 기능 일람표 | ~ | 2020-03-27 |
| 중간 자문 평가 1차 | Front-end <ul style="list-style-type: none"> - UX/UI MVP 구현 완료 Back-end <ul style="list-style-type: none"> - 웹 서버 설계 및 구현 - 각 데몬 및 모듈 설계 - API 설계 및 작성 - 문서 백터화 서버 설계 산출물 : 1. 프로젝트 1차 중간 보고서 2. 프로젝트 진도 점검표 3. 1차 기능별 소스 코드 | 2020-03-28 | 2020-04-24 |
| 중간 자문 평가 2차 | Front-end <ul style="list-style-type: none"> - UX/UI 세부 디자인 구현 - API 연동 Back-end <ul style="list-style-type: none"> - 각 데몬 및 모듈 구현 - API 구현 - 문서 백터화 서버 구현 산출물 : 1. 프로젝트 2차 중간 보고서 2. 프로젝트 진도 점검표 3. 2차 기능별 소스코드 | 2020-04-25 | 2020-05-29 |


| | | | |
|--|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

| | | | |
|------------------|---|------------|------------|
| 구현 완료 | 프로젝트 구현 완료 산출물: 1. 각 기능 소스 코드 | 2020-05-30 | 2020-06-09 |
| 전시용 자료 제출 | 전시용 자료 준비 산출물: 1. 포스터 및 소개 책자 | 2020-05-30 | 2020-06-10 |
| 온라인 평가용 자료 제출 | 온라인 평가용 자료 준비 산출물: 1. 온라인 평가용 자료 | 2020-05-30 | 2020-06-09 |
| 최종 결과 보고서 | 최종 결과 보고서 준비 산출물: 1. 최종 보고서 2. 프로젝트 최종 소스코드 | 2020-05-30 | 2020-06-19 |

| | | | |
|---|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

6.3 인력자원 투입계획

| 이름 | 개발항목 | 시작일 | 종료일 | 총개발일(MD) |
|-------------------|---|------------|------------|----------|
| 전원 | 이슈 분석 | 2020-01-20 | 2020-02-01 | 5 |
| 전원 | 아이디어 구상 | 2020-02-01 | 2020-03-01 | 10 |
| 전원 | 프로젝트 분석 및 설계 | 2020-02-15 | 2020-03-15 | 20 |
| 윤여환 이정하 박지희 | React.js를 이용한 Web Front-End 구현 | 2020-03-20 | 2020-06-09 | 20 |
| 이다은 장태진 | Django를 이용한 Web Back-End 구현 | 2020-03-20 | 2020-06-09 | 20 |
| 박상일 | Keras를 이용한 데이터 분석 시스템 구현 | 2020-03-20 | 2020-06-09 | 20 |
| 장태진 박지희 | Solidity를 이용한 블록체인 구현 | 2020-03-20 | 2020-06-09 | 15 |
| 이다은 이정하 | 서버 API 구현 및 명시 | 2020-04-01 | 2020-06-09 | 30 |
| 이다은 장태진 | 회원 데이터 및 결산 데이터 관리를 위한 MariaDB 설계 및 구현 | 2020-03-20 | 2020-06-09 | 10 |
| 장태진 윤여환 | 라벨링된 데이터 저장 및 관리를 위한 MongoDB 설계 및 구현 | 2020-03-20 | 2020-06-09 | 10 |
| 전원 | 서비스 최종 테스트 | 2020-05-31 | 2020-06-09 | 10 |

| | | | |
|---|-------------------------|--------------------------|-------------|
|  국민대학교 소프트웨어학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 집단지성을 이용한 데이터 라벨링 보상 플랫폼 | |
| | 팀 명 | T-San(티끌모아 태산) | |
| | Confidential Restricted | Version 1.9 | 2020-MAR-23 |

6.4 비 인적자원 투입계획

| 항목 | Provider | 시작일 | 종료일 | Required Options |
|------------|----------------------------|------------|------------|------------------|
| AWS | Amazon | 2020-03-20 | 2020-06-12 | |
| 개발용 PC 6대 | Lenovo, Samsung, Apple, LG | 2020-03-20 | 2020-06-12 | |
| 인공지능 학습 서버 | DLPC | 2020-03-20 | 2020-06-12 | |

7 참고 문헌

| 번호 | 종류 | 제목 | 출처 | 발행년도 | 저자 | 기타 |
|----|------|---|---|------|--|----------|
| 1 | 서적 | 리액트를 다루는 기술 (개정판) | 길벗 | 2019 | 김민준 | |
| 2 | 서적 | Mastering Blockchain | Packt Publishing | 2018 | Imran Bashir | |
| 3 | 서적 | 이더리움 베이직 | BookStar | 2017 | 고려대 | 블록체인 연구회 |
| 4 | 논문 | Selfie: Self-supervised Pretraining for Image Embedding | https://arxiv.org/abs/1906.02940 | 2019 | Trieu H. Trinh, Minh-Thang Luong, Quoc V. Le | |
| 5 | 기사 | AI 레벨업 시키는 '데이터 라벨링' 기업 뜀다 | 전자신문 | 2020 | 김시소 | |
| 6 | 기사 | [칼럼] 2020년 인공지능 주요 이슈와 트렌드...? | 인공지능 신문 | 2019 | 최창현 | |
| 7 | 웹페이지 | 인공지능(AI) 시장 글로벌 동향 | 델코지식정보 | 2019 | 델코 | |