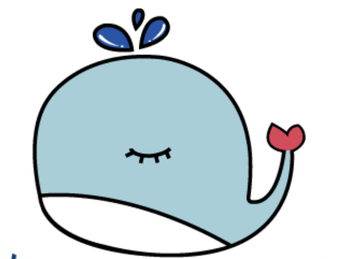


미니언즈 팀

CONTENTS

1 프로젝트 소개 및 필요성

- 4차 산업혁명의 시대를 대표할 수 있는 키워드는 단연 '데이터'일 것
- 이의 활용이 높아지자 다양한 분야에서 데이터를 활용한 텍스트 마이닝 사례가 증가하고 있음
  (챗봇을 이용한 고객관리 서비스, 악플 감지 시스템 등)
- 텍스트 분석의 성능은 올바른 모델링과 방대한 양의 학습 데이터 구축에 의해 좌우됨

➡ **따라서, 본 팀은 프로젝트를 통해 높은 정확도를 띄는 효과적인 학습 데이터 구축 방법을 제시하고자 함**

1 프로젝트 소개 및 필요성

스포츠	피겨 스케이팅, 축구, 야구, 농구 …
패 션	아메카지룩, 가죽자켓, 카고바지 …
반려동물	강아지, 고양이, 펫, 사료, 캣타워 …
여 행	전주여행, 해외여행, 부산, 강원도 …

- 학습데이터 구축을 위해서는 텍스트에 대한 라벨링이 필요
- 일반적으로 라벨링은 사람이 일일이 해야 하는 작업이기에 지루한 절차

1 프로젝트 소개 및 필요성

오락적인 요소를 라벨링에 결합

클라이밍

+

- 기존의 따분한 라벨링 과정을 오락적인 요소를 집어 넣어 즐겁게 작업할 수 있는 환경 구축
- 이를 기반으로 학습데이터의 양이 빠르게 증가할 수 있을 것으로 예상
- 이와 같은 데이터가 축적되어 학습이 이루어지면 분석 성능 역시 향상될 수 있을 것으로 기대

➡ **이에 대한 방안으로 라벨링 어플리케이션을 제안**

2 선행연구

- 현재 SNS 텍스트를 데이터로 활용고자 하는 연구들이 활발히 진행되고 있는 실정임

1. 인스타그램 기반 이미지와 텍스트를 활용한 사용자 감정정보 측정 (남민지 · 김정인 · 신주현, 2014)

- 인스타그램의 이미지와 텍스트(영어)를 이용하여 이용자의 감정상태를 분석

- 이미지 대표 색상을 추출해 색상에 맞는 감정 형용사를 댓글과 비교한 후 포스트를 대표하는 감정 형용사를 추출

2. Twitter user profiling based on text and community mining for market analysis (Kazushi Ikeda et al., 2014)

- 트위터 텍스트를 데이터화 하여 이용자의 프로필 예측

- 즉 연령대, 성별, 거주지, 직업, 기혼여부를 텍스트 데이터를 통해 유의미하게 구별함

2 선행연구

➡ 앞선 연구들은 SNS 텍스트 데이터의 유효성을 증명하고 있음

BUT

- 두 논문 모두 외국어를 기반으로 연구되어 한국어 텍스트에 대한 성능을 보장할 수 없음

3 학습데이터 구축

3.1 라벨링을 위한 데이터 수집

이용자에 대한 편향이 없도록 사전에 카테고리를 정의

➡ 분류별로 약 7천명의 이용자 포스트를 수집

- 카테고리 : 성별, 연령대, 직업, 관심사, 지역
- 성별, 연령대, 관심사 : 각 카테고리별로 대표할 수 있는 키워드를 선별한 후, 해시태그(#) 검색을 통해 이용자 추출
- 지역 : 수도권, 충청, 경상, 전라, 강원의 5개 지역별 확실한 중심 유저를 선별한 후, 해당 이용자의 커뮤니티 분석

3 학습데이터 구축

3.2 카테고리별 해시태그(#) 검색 키워드 예시

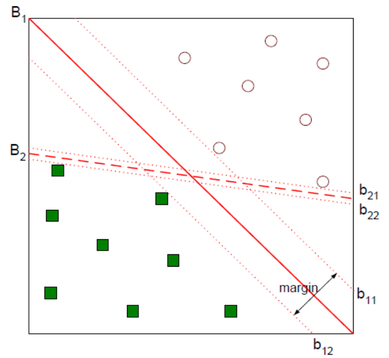
카테고리	카테고리 분류	해시태그 검색 키워드
성별	남성	남중, 남고, 군대, 입대, 신병휴가, 전역 등
	여성	여중, 여고, 여대, 이화여대, 숙명여대, 성신여대 등
연령대	미성년자	초1, 초4, 초딩스타그램 등
	20대	20대, 개강, 대학생, 대학생공스타그램, 학사모 등
	30대	30대, 계란한판, 서른, 스물열살 등
	40대 이상	40대, 마흔, 마흔틴, 50대, 50대몸짱, 50대아줌마, 꽃중년 등
관심사	게임	롤, 메이플스토리, 모바일게임, 게임스타그램, 게임추천 등
	맛집	맛스타그램, 맛있다, 먹방, 먹스타그램, 맛집그램 등
	미디어감상	유튜브, 유튜버, 구독, 넷플릭스, 미드, 영화스타그램 등
	반려동물	반려견, 반려묘, 강아지, 댕댕이, 펫스타그램 등
	스포츠	운동, 등산, 라이딩, 자전거, 축구, 야구, 헬스, 서핑 등
	여행	여행, 여행스타그램, 여행에미치다 유디니 등
	카페	카페스타그램, 카페투어, 감성카페, 카페추천, 커피맛집 등
	패션	오늘뭐입지, ootd, 오오디티, 남자코디, 여자코디 등

3 학습데이터 구축

3.3 모델링

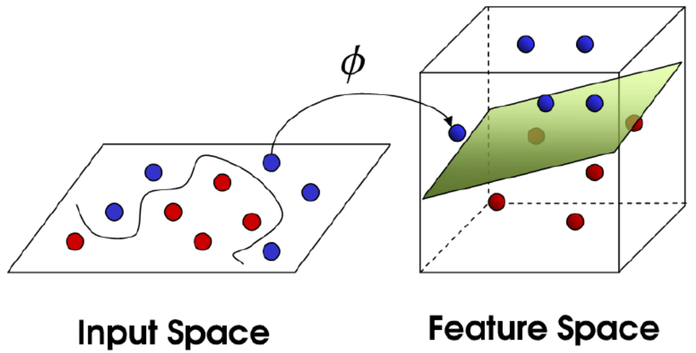
서포트 벡터 머신(support vector machine, SVM)

- 기계 학습의 분야 중 하나
- 패턴 인식, 자료 분석을 위한 지도 학습 모델
- 주로 분류와 회귀 분석을 위해 사용

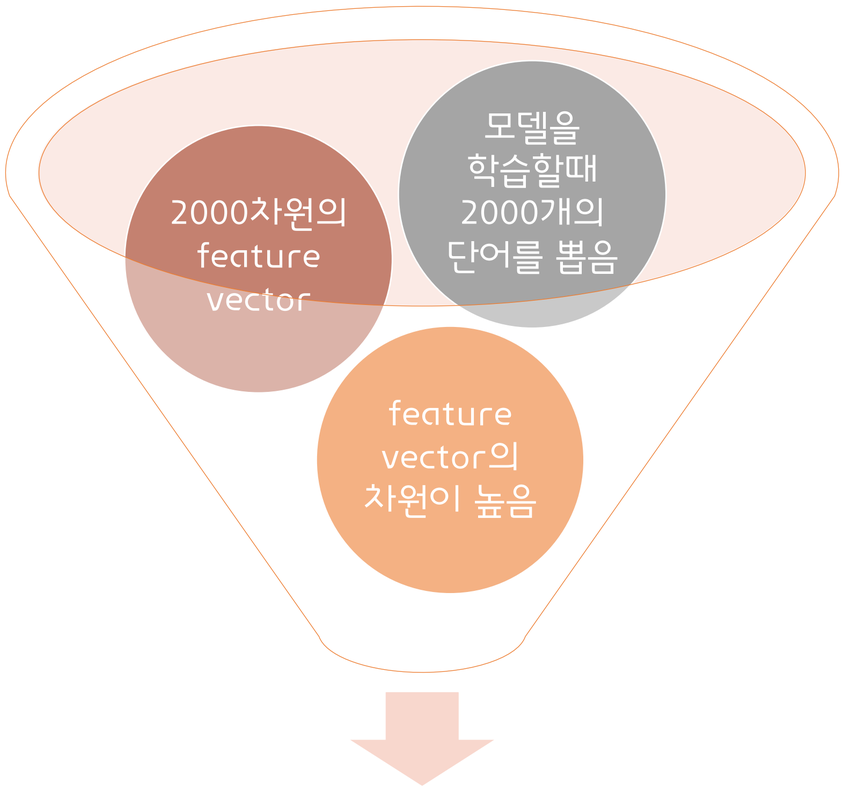


커널 서포트 벡터 머신(Kernel - support vector machine, Kernel-SVM)

- 원공간(Input Space)의 데이터를 선형분류가 가능한 고차원 공간(Feature Space)으로 매핑한 뒤, 두 범주를 분류하는 초평면을 찾음



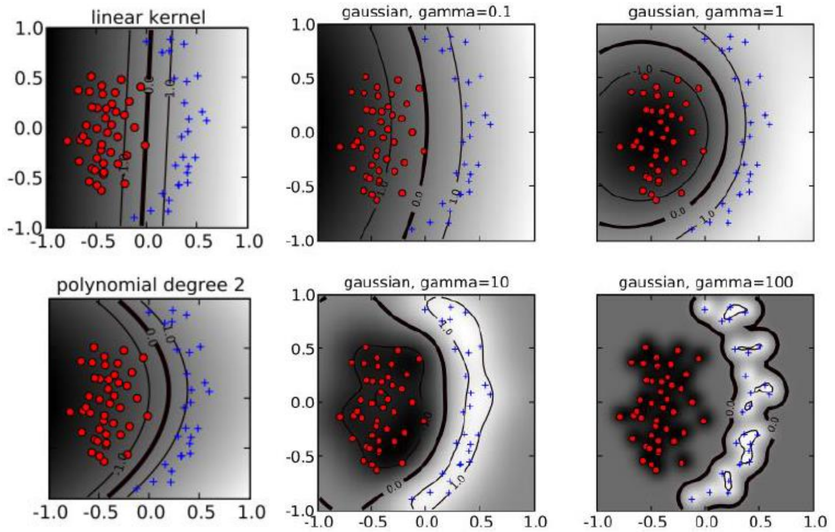
3 학습데이터 구축



gaussian svm 적절, 실제로 Linear, gaussian, polynomial svm 중 정확도가 가장 높게 나와서 사용

Gaussian Kernel

가우시안 커널은 Input Space가 몇 차원이 됐든 무한대 차원의 Feature Space로 매핑



3 학습데이터 구축

3.4 라벨링 어플리케이션

‘Whale’ + ‘Labeling’ = ‘Whaleling’

- 방대한 양의 국어 말뭉치를 엄청난 크기의 바다에 비유하였고 하나의 라벨링 과정을 바다에서 해산물을 얻는 낚시에 대입
- 바다에서 얻을 수 있는 최고의 해산물인 ‘고래’가 어플리케이션의 주요 테마

3 학습데이터 구축

기존 라벨링에 ‘Gamification’ 을 접목

- Gamification : ‘game(게임)’ + ‘-fication’ (무엇이 되기 만든다는 뜻의 접미사). 게임적인 요소를 활용해 게임이 아닌 분야에서 문제해결, 관심유도, 지식 전달 및 교육, 행동변화 등에 활용하는 것을 의미

- 라벨링을 ‘일’이 아닌 ‘놀이’로 만들어 관심을 유도하고 몰입을 이끌어냄

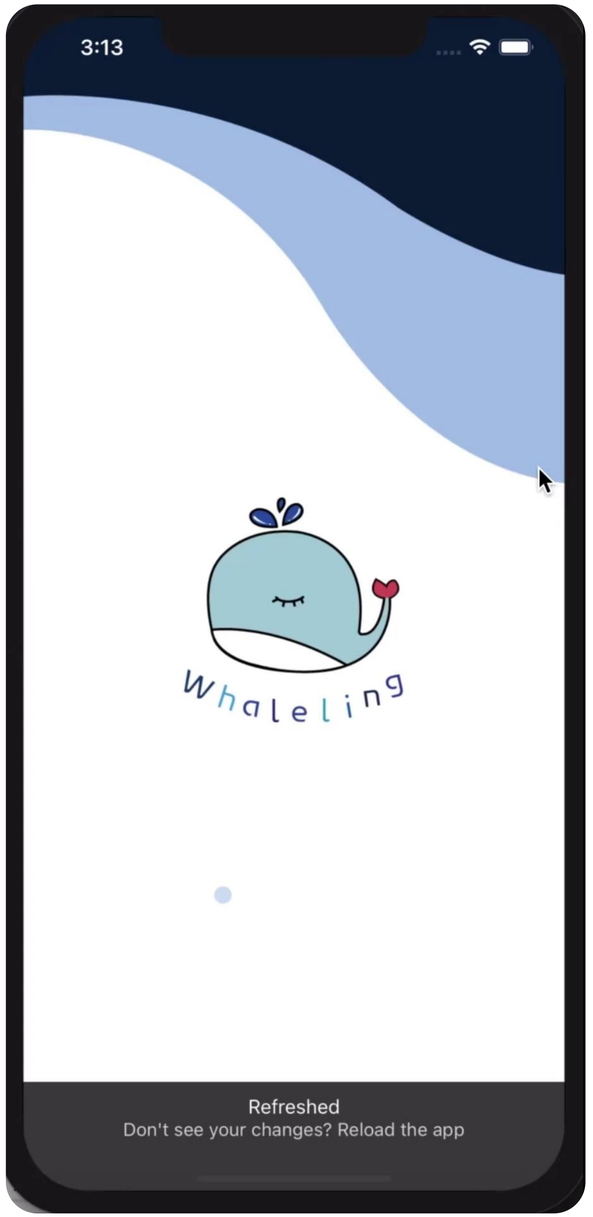
3 학습데이터 구축

3.4.1 Gamification 활용 방안 예시

원하는 카테고리 선택 → 라벨링을 한 만큼 포인트를 부여 → 누적된 포인트를 분석과정에서 활용 가능 + 랭킹(순위)을 이용해 경쟁심 유도

3 학습데이터 구축

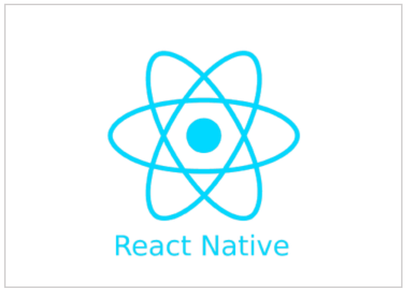
3.4.2 어플 시연



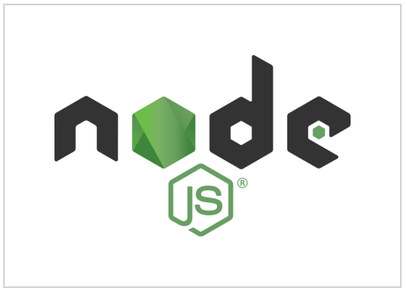
3 학습데이터 구축

3.4.3 used environment & techniques

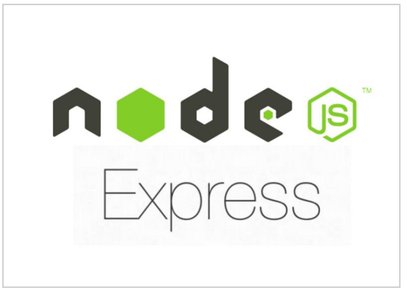
python 3.6



React Native



node js



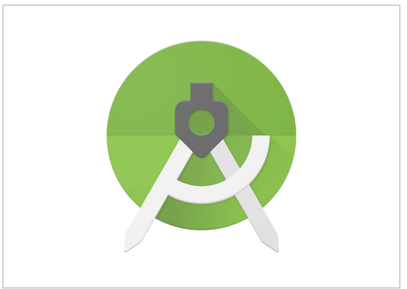
node js express

MySQL

Scrapy

PM2

VSCode

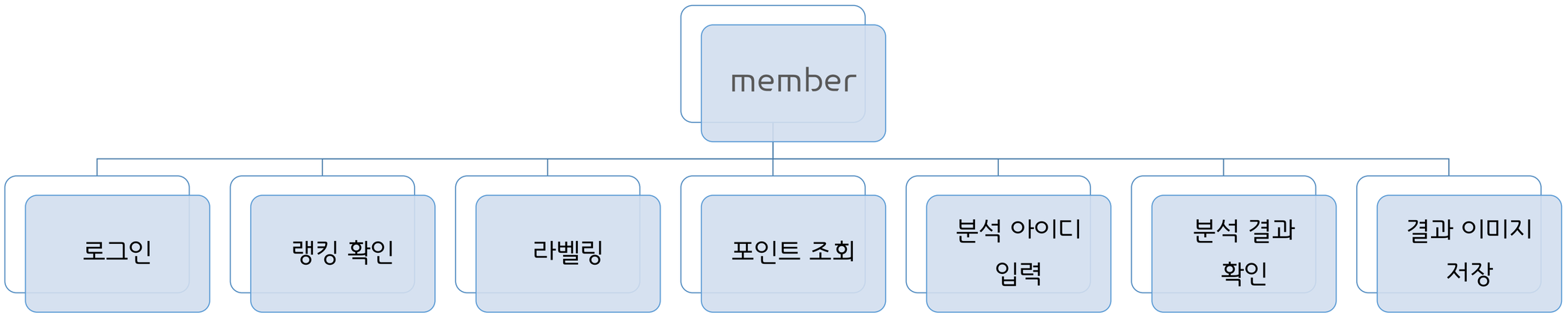


Android Studio

POSTMAN

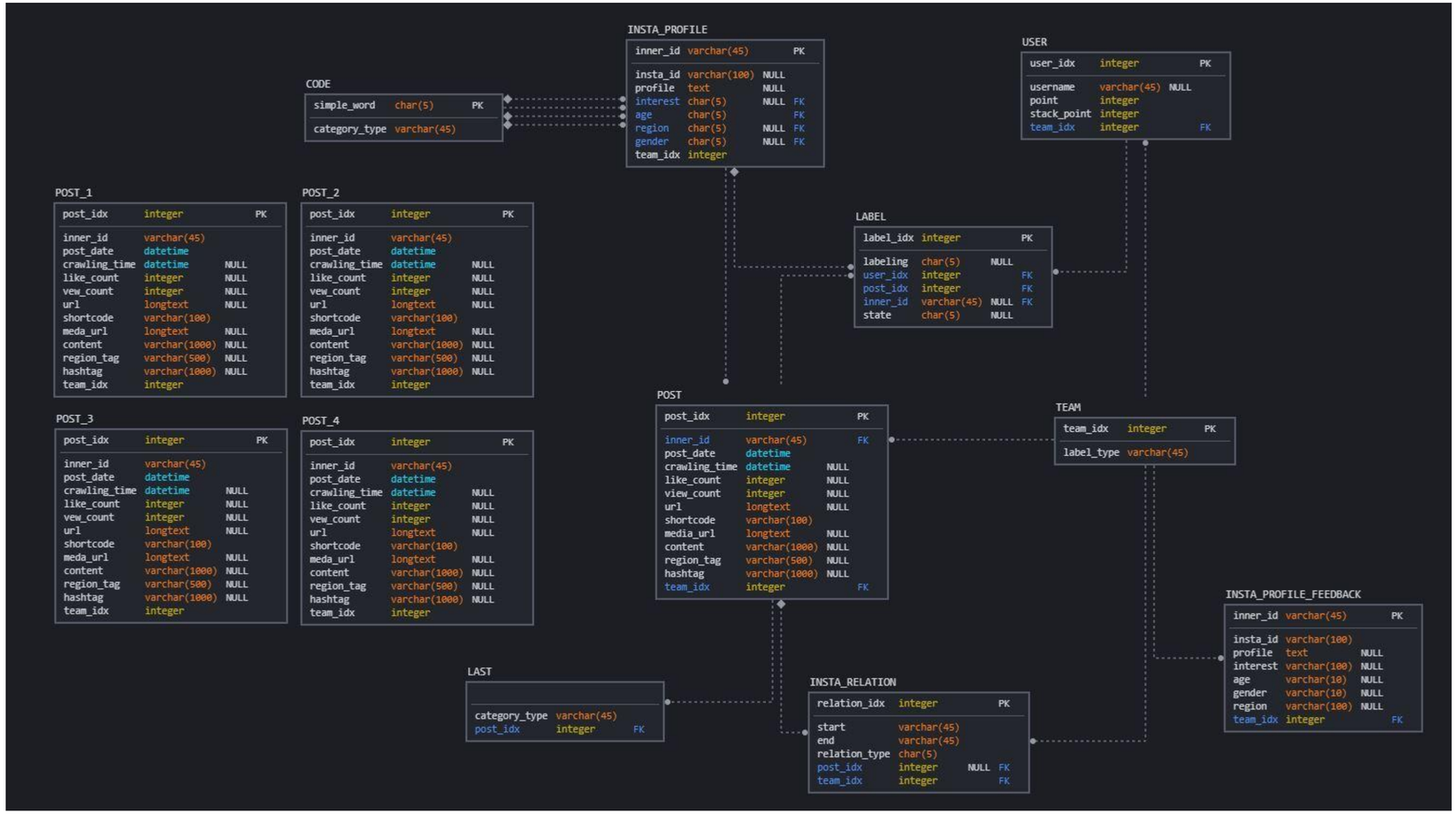
3 학습데이터 구축

3.4.4 어플 Usecase



3 학습데이터 구축

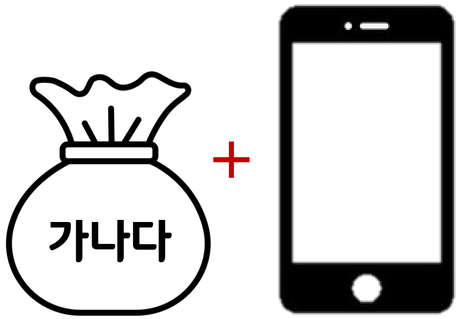
3.4.5 Database Design(ERD)



4 결론

4.1 활용방안

- 이용자의 간단한 프로필 분석 뿐만 아니라 애플리케이션의 주요 기능인 프로필 예측을 통해 타겟 마케팅이 가능할 것



- 국어 말뭉치와 결합함으로써 성능향상 가능

4 결론

4.2 프로젝트 성과와 한계

➡ 성과

- 어플리케이션 내에 적절한 보상을 통한 경쟁적인 라벨링 생태계 구축으로 인해 보다 효과적이고 신속한 학습데이터 구축 가능
- 이는 한국어 말뭉치 구축에 속도를 부여해 학습데이터를 이용한 후속 연구에 유의미한 결과를 도출할 것으로 예상

➡ 한계

- 해시태그 기반의 검색에 대한 편향성
- 이용자 수와 누적 라벨링이 적은 초기에는 활용가치가 떨어질 수도 있음

감사합니다