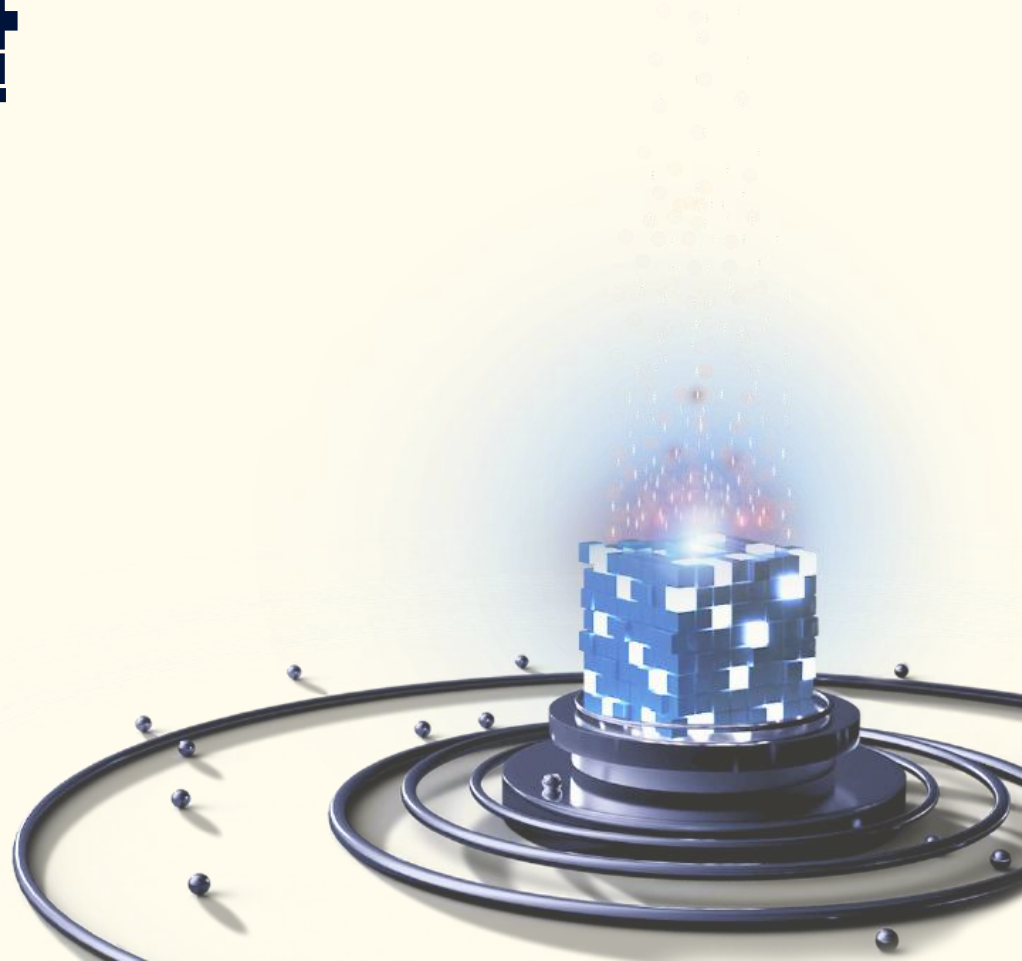
지속 가능한 성장을 위한 오픈소스 커뮤니티의 역할

개발자와 커뮤니티

송영숙

Sionic AI Inc.



Contents

01

자기 소개

01 송영숙

생계용 커리어

- 경희대학교 국어국문학과 국어학 전공(Ph. D.)
- (전) Naver - Campany AI
- (현)사이오닉 AI 리서치 / 정책 총괄
- 국립국어원 NIA 한국정보화진흥원 일반상식 데이터 구축 사업 총괄 PM 외 다수의 데이터 구축 사업 참여

커뮤니티 커리어

- 다수의 오픈 스터디 조직(2014~)
- 파이콘 2018 키노트
- 자연어 처리 컨퍼런스 LanCon Organizer (2019~)
- 저서 및 역서(모두의 한국어 텍스트 분석, 파이썬을 활용한 딥러닝 전이학습 등)

01 지속 가능한 성장을 위한 두 개의 커리어

생계용 업무

배워서 바로 사용

→ 현재에 투자

+

커뮤니티 참여

꼭 써먹지 않더라도 개념 정도는 알아두려고

→ 미래에 투자

02

진입하는 사람들을 위한 커뮤니티의 역할

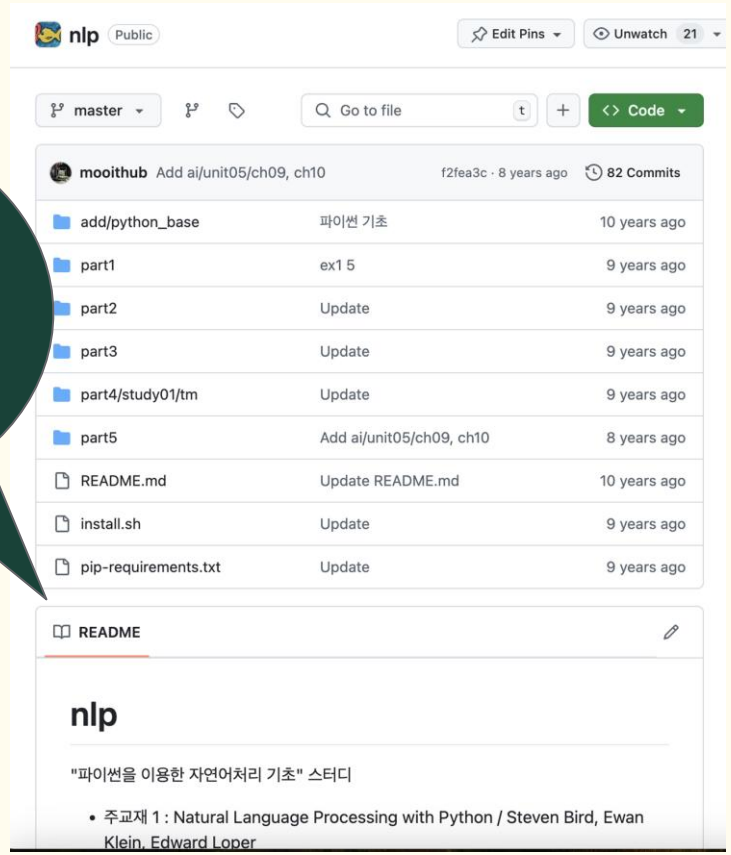
하나의 스터디에서 수많은 스터디 팀이 파생되기까지!

02 하나의 스터디에서 수많은 스터디 팀이 파생되기까지!

2014년 자연어처리 기초 스터디에서 시작

2017년 통계 분석, 자연어, 개발툴 등으로 확장

[가장 많은 한 말]
안녕하세요. 뼈문과 초보입니다.



[가장 많이 들은 말]
오늘 뵙고 싶은데 어느 스터디에 계시나요?

02 하나의 스터디에서 수많은 스터디 팀이 파생되기까지

팀 빌딩과 리딩

- 처음에는 하나의 팀이었으나 점점 자신의 관심에 따라 여러 팀으로 파생 된 후 연합 세미나 등도 개설
- 모르는 것(코딩) 알려 주시면 아는 것(음성학)을 공유하겠다는 취지로 모임 개설
- 3명 이상 모이면 시작한다고 했는데 약 20명 모임

02 하나의 스터디에서 수많은 스터디 팀이 파생되기까지

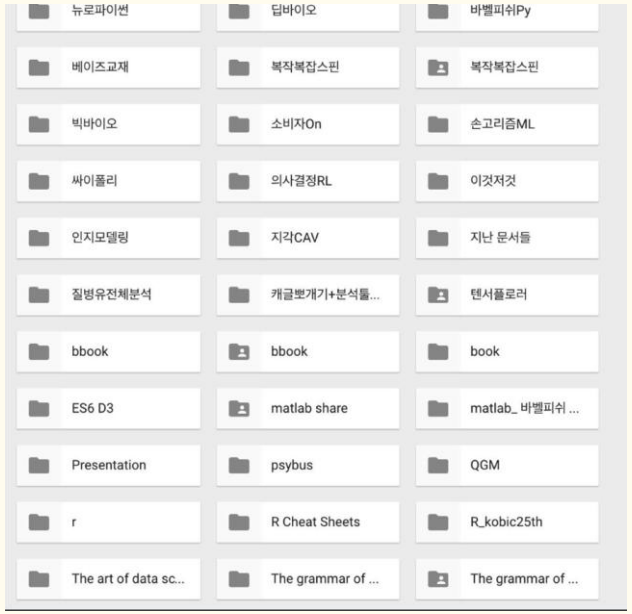
모임 운용

- 커뮤니티 모임은 선생님이 없음
- 모두가 모두에게 배우면서 가르치기 때문에 모르는 것을 모른다고 말하는 것이 중요!
- 덤으로 자기 전공이 아닌 사람에게 말하는 스킬이 늘어남
- 벽을 만들면 나의 지식이지만 벽을 허물면 모두의 지식으로 증강

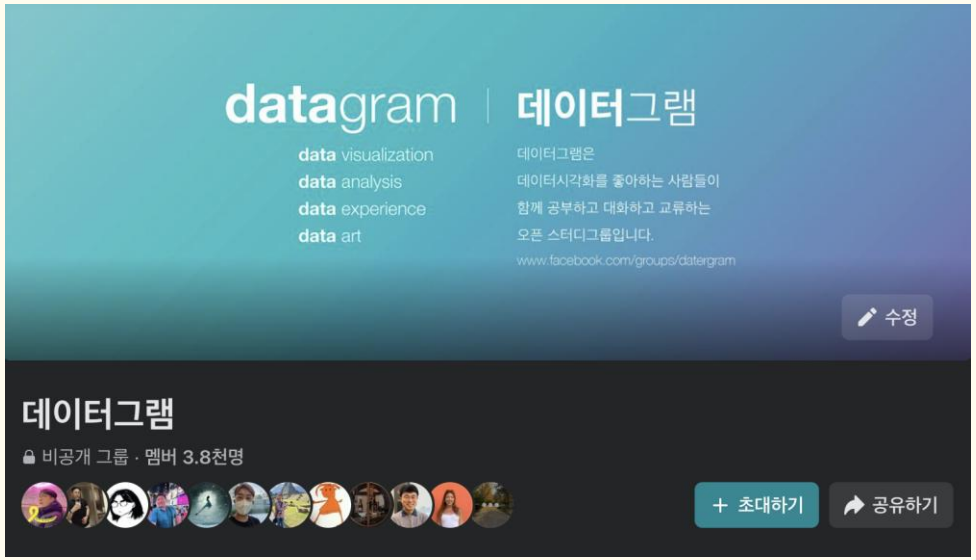
02 하나의 스터디에서 수많은 스터디 팀이 파생되기까지

기록은 기억을 지배한다는 마음으로 구글 드라이브, 깃허브에 기록을 남기고 페이스북 그룹을 통해 공유

기록과 공유



2016년 9월 14일 구글 드라이브 기록 폴더

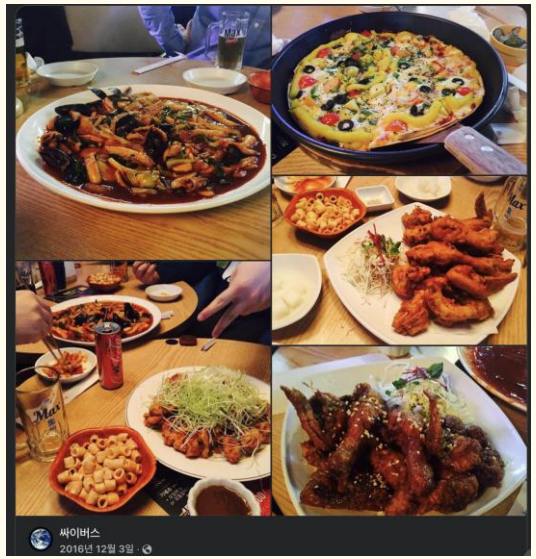


(현재) 페이스북 그룹을 통해 모임 시기 및 모임 내용 공유

02 하나의 스터디에서 수많은 스터디 팀이 파생되기까지

모임 팁 : 뒤풀이를 통해 나의 분노와 재미에 공감할 수 있는 사람 찾기

뒤풀이



2016년 모임 뒤풀이

2023년 모임 뒤풀이

뒤풀이 때 스터디 연합 모임 등을 기획하기도 하고 소모임이 생기기도 하고 무엇보다 지금은 10년 이상 만나는 친구가 됨

03

지속 가능한 성장을 위한 커뮤니티의 역할

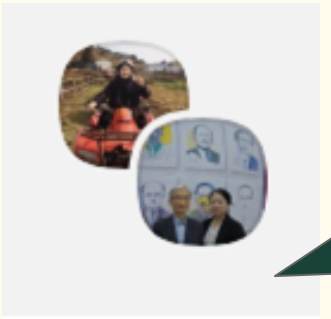
- 10년차 이상 커뮤니티 활동을 하고 난 후
- 현재 오픈소스 커뮤니티 발전을 위해 하고 있는 일

03 10년차 이상 커뮤니티 활동을 하고 난 후

- 각 개인의 지속 가능한 성장을 위해 커뮤니티에 역할
- AI 분야는 변화가 빠르기 때문에 서로의 상황을 공유할 공간과 그동안 쌓인 도메인 분야, AI 데이터 구축에서 기여할 수 있는 것들을 모색

03 10년차 이상 커뮤니티 활동을 하고 난 후

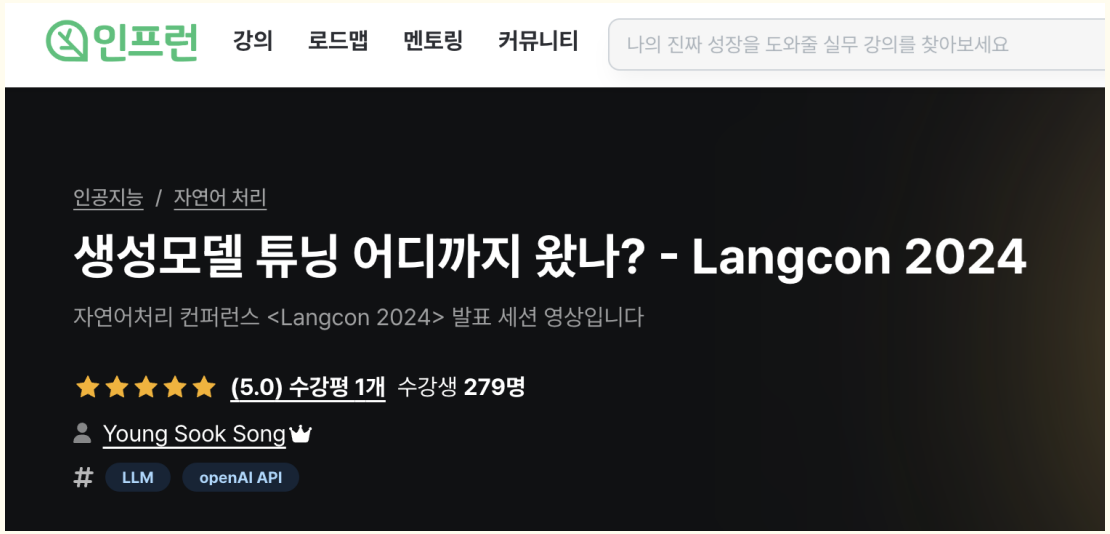
Langcon 컨퍼런스 : 현업에서 배운 지식을 공유하는 모임으로 성장
관심 있는 사람들간의 소모임

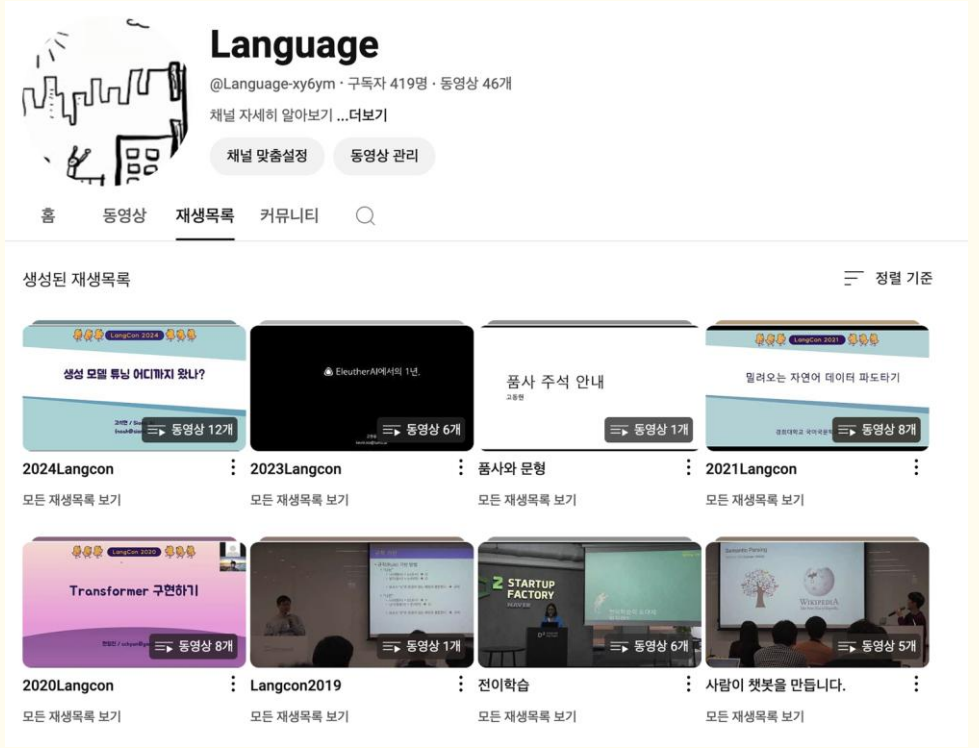


기술 문서 작성에 관심 있는 사람들의 모임 톡방

03 10년차 이상 커뮤니티 활동을 하고 난 후

- 초심자 및 전문가 초청 발표 등을 **발표** 자료 및 영상으로 기록하여 유튜브, 인프런(무료)에 공개
- 회사 대표님과 동료들도 발표!
- 일본에서의 IT 현황등 발표에 폭이 광범위해짐

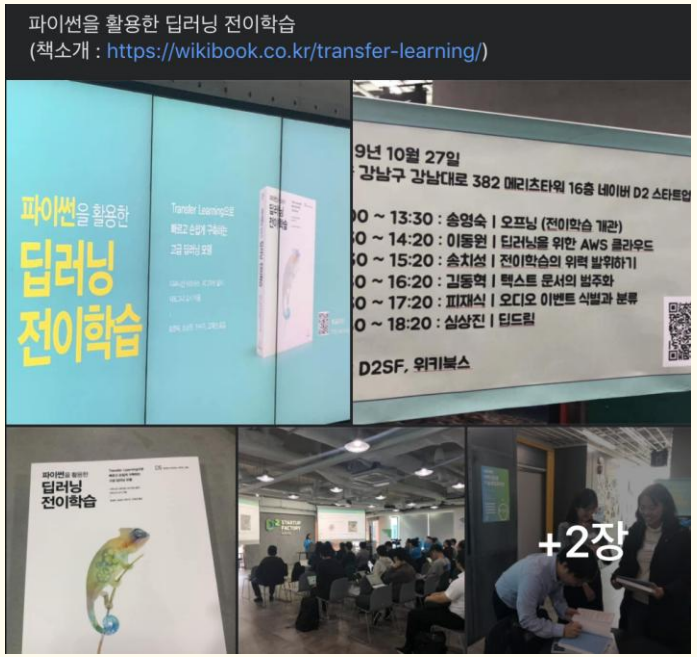




https://2024langcon.oopy.io/72eef1e5-da99-48ae-b18b-cb576a2d68ac
https://www.youtube.com/@Language-xy6ym/playlists

03 10년차 이상 커뮤니티 활동을 하고 난 후

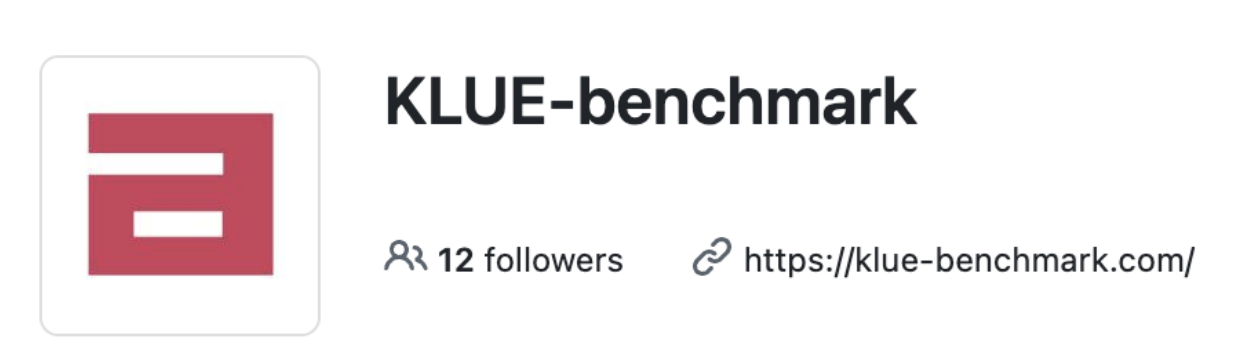
- 번역, 저술 등에 도전

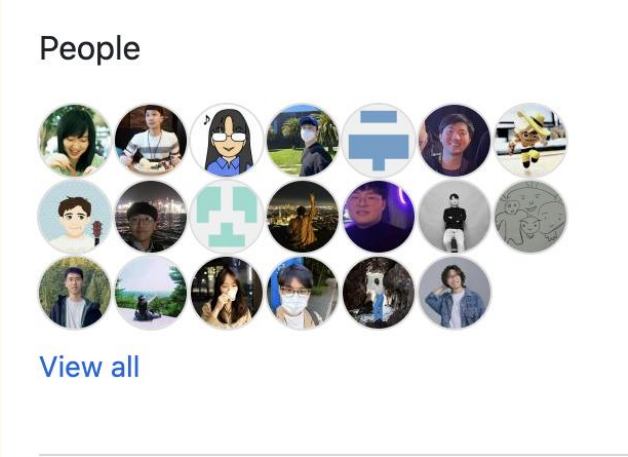


- 공개 스터디 또는 커뮤니티에서 만나서 손발이 잘 맞는 분들과 책 번역 및 저술
- 책의 한 챕터씩을 맡나 북토크 형식으로 발표

03 10년차 이상 커뮤니티 활동을 하고 난 후

- 논문 쓰기
- KLUE: Korean Language Understanding Evaluation약 30명의 서로 다른 회사에 소속되어 있는 사람들이 모여서 데이터를 만들고 학술대회 논문도 썼다는 점에서 커뮤니티적인 성격이 강했음

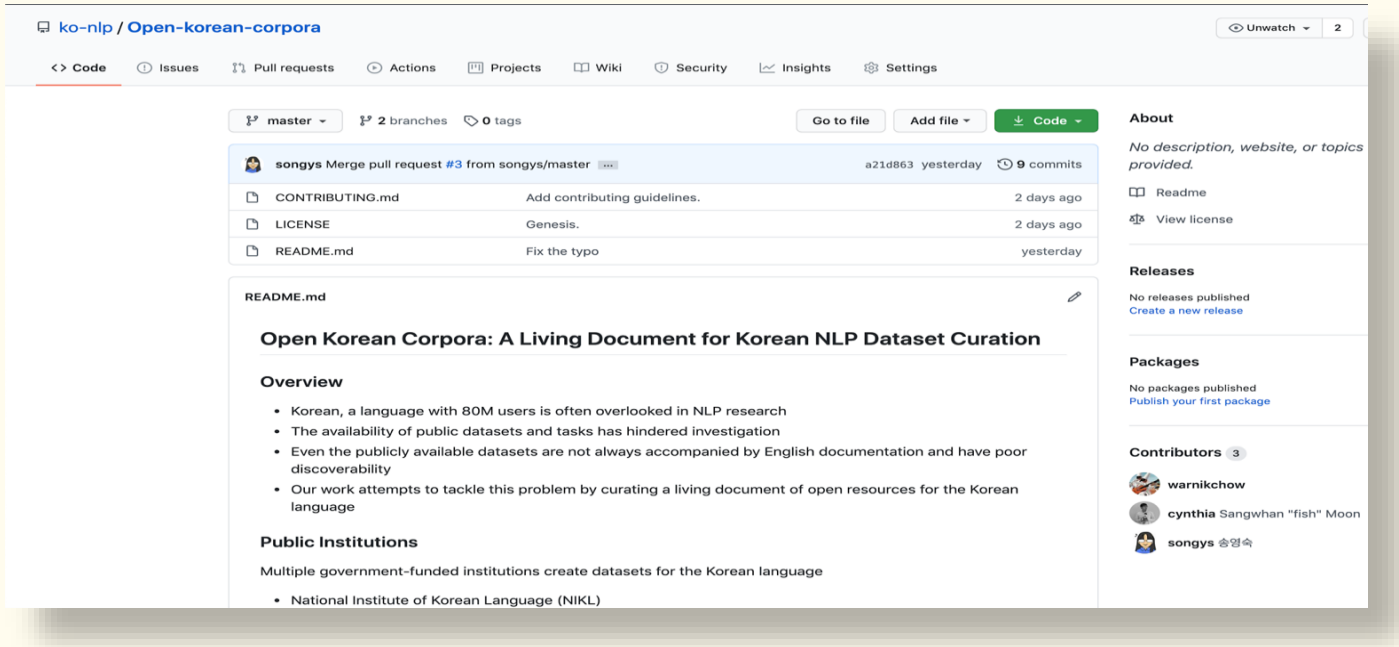




03 현재 오픈소스 커뮤니티 발전을 위해 하고 있는 일 : 데이터 큐레이션

데이터 큐레이션이란?

- 데이터 큐레이션의 정의: 데이터 구축과 생성뿐만 아니라 데이터의 활용 가치를 높이는 모든 활동을 포함
- 활동 배경: 각자 데이터를 만들고 Github에 공개하는 경험을 하면서 체계적 데이터 관리의 필요성을 실감



https://github.com/ko-nlp/Open-korean-corpora

03 현재 오픈소스 커뮤니티 발전을 위해 하고 있는 일 : 데이터 큐레이션

한국어 자연어처리가 어렵다?

- 데이터 세트를 구하는 것이 어렵다
- 누군가 데이터를 공개했지만 검색이 잘 되지 않는다
- 한국어의 특성을 파악하기가 어렵다

공개 데이터의 목록을 만들고 쉽게 접근할 수 있는 방법을 찾아보자는 목적에서 3명이서 주기적으로 업데이트

03 현재 오픈소스 커뮤니티 발전을 위해 하고 있는 일

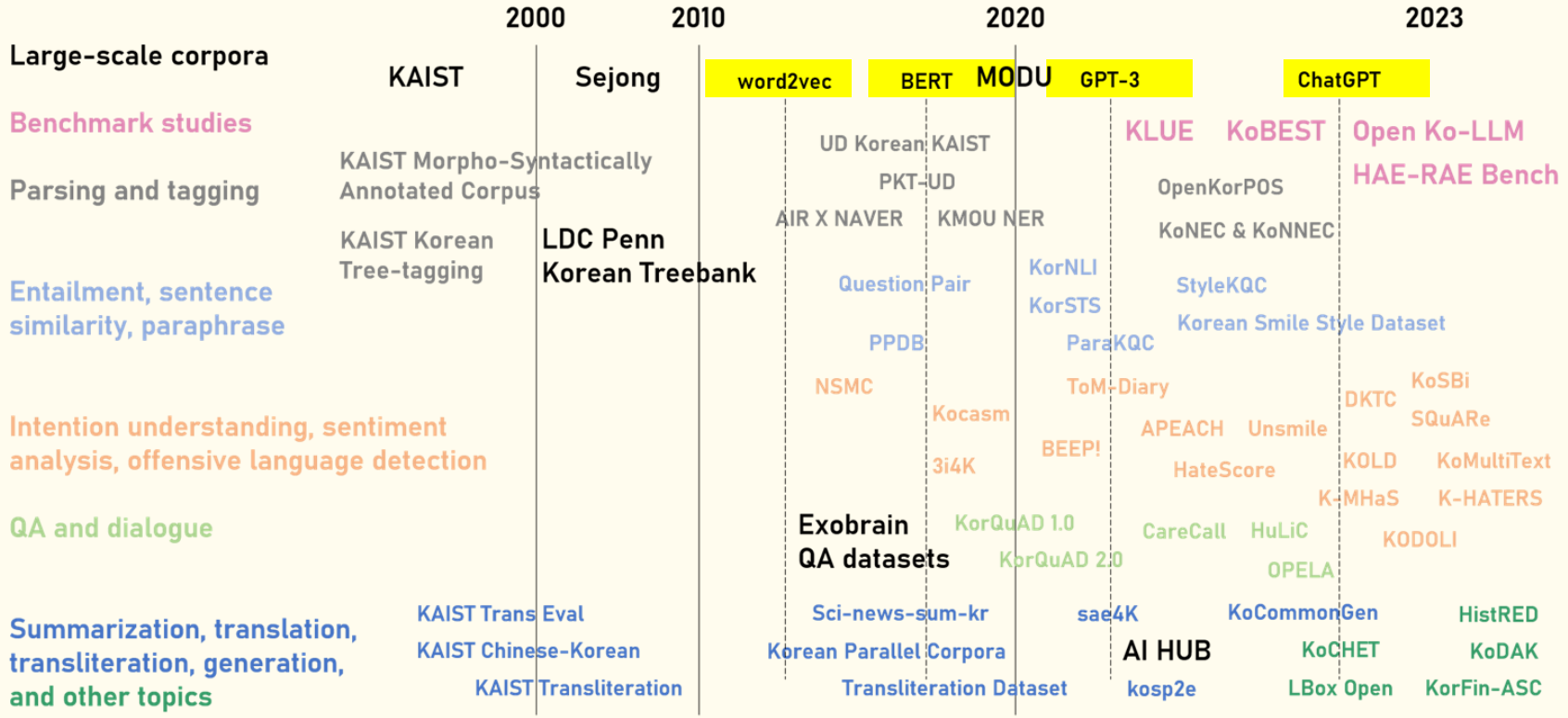
AI 학습 및 평가용 데이터 큐레이션

- 2019년 12월 15일 여기저기 저장되어 있던 공개 데이터 세트의 링크를 정리하고 깃허브에 공개
- 2020년 8월 21일 조원익님, 문상환님이 함께하면서 분석적으로 정리하고 NLP-OSS 2020@EMNLP에서 한국어 자연어처리 오픈 데이터 체계화의 contribution을 인정받아 게재 수락!
- 데이터 로더팀과 연결되어 지속 가능한 오픈소스의 체계화
- 2023 공개 데이터들을 재정비하여 정리
  2023 PACLIC 학회에서 관련 내용 발표
- 2024 Huggingface에 공개한 데이터 한국어 데이터 정리

https://github.com/ko-nlp/Open-korean-corpora,
https://github.com/songys/AwesomeKorean_Data
WI Cho, S Moon, Y Song(2020) Open Korean Corpora: A Practical Report, NLP-OSS 2020 @EMNLP, 85-93
Won Ik Cho, Sangwhan Moon and Youngsook Song(2023), Revisiting Korean Corpus Studies through Technological Advances, " in Proc. PACLIC 2023, Hong Kong

03 현재 오픈소스 커뮤니티 발전을 위해 하고 있는 일 : 데이터 큐레이션

시간 순서에 따른 한국어 코퍼스 구축



대규모 학습 데이터

단어 하나 하나 주석하는 정교한 형태의 학습/ 테스트 데이터 또는 법률과 헬스 케어 같은 특정 도메인의 데이터

03 현재 오픈소스 커뮤니티 발전을 위해 하고 있는 일 : 데이터 큐레이션

2024년 10월 Huggingface에 공개한 데이터 한국어 데이터 정리

No	id	author	downloads	likes	task_ids	language	multilinguality	library	format	arxiv	source_datasets	license	size_categories	task_categories
1	skt/kobest_v1	skt	30333	34	none	ko	monolingual	polars	json	2204.04541	original	cc-by-sa-4.0	10K<n<100K	none
2	maywell/korean_textbooks	maywell	2102	97	none	ko	none	polars	parquet	2306.11644	none	apache-2.0	1M<n<10M	none
3	beomi/KoAlpaca-v1.1a	beomi	1273	35	none	ko	none	polars	parquet	none	none	none	10K<n<100K	text-generation
4	sean0042/KorMedMCQA	sean0042	1262	21	none	ko	none	polars	parquet	2403.01469	none	cc-by-nc-2.0	1K<n<10K	question-answering
5	MarkrAI/KOpen-HQ-Hermes-2.5-60K	MarkrAI	1144	53	none	ko	none	polars	parquet	none	none	mit	10K<n<100K	text-generation
6	KorQuAD/squad_kor_v1	KorQuAD	943	18	extractive-qa	ko	monolingual	polars	parquet	1909.07005	original	cc-by-nd-4.0	10K<n<100K	question-answering
7	MarkrAI/KoCommercial-Dataset	MarkrAI	615	125	none	ko	none	polars	parquet	2107.06499	none	mit	100K<n<1M	none
8	CarrotAI/ko-instruction-dataset	CarrotAI	484	22	none	ko	none	polars	json	2304.12244	none	apache-2.0	1K<n<10K	text-generation
9	taeminlee/Ko-StrategyQA	taeminlee	431	12	document-retriev	ko	monolingual	polars	json	none	Ko-StrategyQA	none	10K<n<100K	text-retrieval
10	maywell/ko_wikidata_QA	maywell	410	36	none	none	none	polars	csv	none	none	none	100K<n<1M	none
11	LDCC/korag	LDCC	333	6	none	ko	none	polars	parquet	none	none	none	10K<n<100K	text-generation
12	KETI-AIR/korquad	KETI-AIR	310	1	none	none	none	none	none	none	none	none	none	none
13	smilegate-ai/kor_unsmile	smilegate-ai	293	3	none	none	none	polars	parquet	none	none	none	10K<n<100K	none
14	sionic/ko-dpo-mix-7k-trl-style	sionic	220	5	none	none	none	polars	parquet	none	none	none	1K<n<10K	none
15	kakaobrain/kor_nli	kakaobrain	214	16	multi-input-text-c	ko	monolingual	polars	parquet	none	extended|xnli	cc-by-sa-4.0	100K<n<1M	text-classification

전체 내용은 다음 링크 참조
https://github.com/songys/huggingface_KoreanDataset/blob/main/README.md

03

현재 참여할 수 있는 AI 관련 커뮤니티

04 현재 참여할 수 있는 AI 관련 커뮤니티

- 모두의 연구소
https://modulabs.co.kr/
  다양한 사람들이 모여서 랩 형식으로 운용 중.
  팀장 또는 팀의 성향에 따라 결과물이나 과정의 편차가 클 수 있음

- eleuther.ai/ : 영어를 중심으로 한국어를 포함한 다국어 작업 중
 https://www.eleuther.ai/

- 파이썬 한국 사용자 모임 : https://2024.pycon.kr/about/pyconkr2024
  컨퍼런스 2024년 10월 26-27일
  파이썬 사용자라면 한국, 일본이나 미국 등에서도 발표 가능

- SciPy Korea : https://scipy.kr/
  1년에 한 번 정도 발표자 모집
- Langcon2025 발표자 및 스탭 모집 중! 함께해요!

감사합니다

지속 가능한 성장을 위한 오픈소스 커뮤니티의 역할