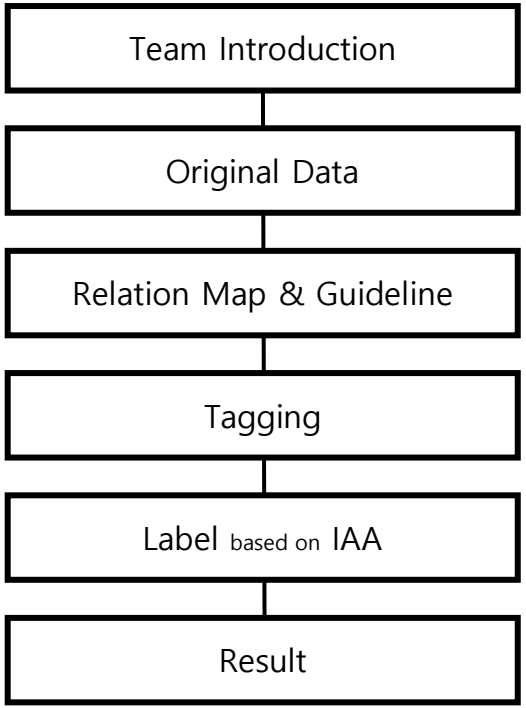
boostcamp ai tech

P-stage Level-3

NLP Data Production

NLP-team-12 : AI-it

목차


Team Introduction
Original Data
Relation Map & Guideline
Tagging
Label based on IAA
Result

Team Intro

T2163
이연걸

팀장
London Is Red

T2217
진혜원

말하는 감자
NORTH LONDON IS WHITE

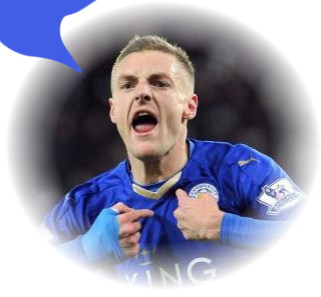
T2096
박진영

All_rounder
현재_솔로

T2130
양재욱

Round shoulder
현재_커플

T2211
조범준



1일 1노래 공유
노래_공유_받아요_DM_주세요

T2127
안성민

London Is Deep Blue
내_주식도_Deep_Blue

T2050
김재현

세계_최고의_듈_노션
🤟&🤙=♡

Original Data

Wrong Original Data

If 팀 이름 == 지역
축구팀이 아닌 지역에 대한 wiki 데이터 존재

Wrong Document Filtering

총 56개 중 18개

풀럼 FC, 포츠머스 FC, 크리스털 팰리스 FC, 첼시 FC
윔블던 FC, 웨스트브로미치 앨비언 FC, 왓퍼드 FC
에버턴 FC, 아스널 FC, 선덜랜드 AFC, 사우샘프턴 FC
블랙풀 FC, AFC 본머스 번리 FC, 반즐리 FC
미들즈브러 FC, 리버풀 FC, 레딩 FC

Document Crawling

```
import wikipediaapi
wiki = wikipediaapi.Wikipedia( language='ko',
                               extract_format=wikipediaapi.ExtractFormat.WIKI)
keyword = ['풀럼 FC','포츠머스 FC','크리스털 팰리스 FC',
           '첼시 FC','윔블던 FC','웨스트브로미치 앨비언 FC',
           '왓퍼드 FC','에버턴 FC','아스널 FC','선덜랜드 AFC',
           '사우샘프턴 FC','블랙풀 FC','AFC 본머스','번리 FC',
           '반즐리 FC','미들즈브러 FC','리버풀 FC','레딩 FC']
for w in keyword:
    page_py = wiki.page(w)
    print(w+" Page - Exists: %s" % page_py.exists())
    p_wiki = wiki.page(w)
    with open("/content/wikidata_crawing/"+w+".txt", "w") as f:
        f.write(p_wiki.text)
```

첼시(영어: Chelsea)에는 다음과 같은 뜻이 있다.

인명
첼시 클린턴(Chelsea Clinton, 1980 ~)은 빌 클린턴과 힐러리 클린턴의 외동딸이다.
로렌초 첼시: 1361년 7월 16일부터 그의 사망까지 제58대 베네치아의 도제로 역임했던 베네치아의 정치인
첼시 매닝: 위키리크스에서 최대 규모의 미국의 군사 기밀 사항이 포함된 내부 자료를 제공한 내부 고발자
첼시 핸들러: 미국의 희극인, 텔레비전 진행자, 작가, 배우
첼시 리: 미국의 프로농구 선수
첼시 리케츠: 미국의 배우
첼시 케인: 미국의 배우 및 가수
첼시 케인 (작가): 미국의 작가
첼시 노블: 미국의 배우
첼시 하프페니: 잉글랜드의 배우

음악
CHERRSEE는 일본의 걸그룹이다

지명
첼시(Chelsea)는 영국 런던의 지역이다.
첼시(Chelsea)는 미국 맨해튼의 지역이다.
첼시(Chelsea)는 미국 매사추세츠 주의 도시이다.
첼시(Chelsea)는 미국 앨라배마 주의 도시이다.
첼시(Chelsea)는 미국 버몬트 주의 도시이다.

스포츠
첼시 FC(Chelsea Football Club)는 영국 런던 연고의

기타
첼시 번(Chelsea bun)은 빵의 한 종류이다.
첼시 교(Chelsea Bridge)는 영국 런던의 다리이다.
호텔 첼시(Hotel Chelsea)는 뉴욕의 호텔이다.
위타드 오브 첼시(Wittard of Chelsea)는 홍차 브랜드

아스널(arsenal), 아르세날은 다음을 가리킨다.

무기고
아스널 FC(Arsenal F.C.)
아스널 WFC(Arsenal W.F.C.)
아스널 스타디움(Arsenal Stadium)
아르세날레 디 베네치아(Venetian Arsenal)
아르세날 데 사란디(Arsenal de Sarandí)
FC 아르세날 키예프(FC Arsenal Kyiv)
FC 아르세날 툴라(FC Arsenal Tula)
아스널 쉅

크리스털 팰리스(Crystal Palace)는 제1회 만국 박람회의 장소로 런던에 건설된 수정궁 및 그에 연관된 명칭이다.

수정궁
크리스털 팰리스 (런던) - 지명
크리스털 팰리스 FC

Relation Map Discussion Point

Entity Type 어떻게 선별하는 것이 좋을까

Relation Range 포함범위를 어떻게 설정할 것인지

Time Point / POS 룰 세팅을 어떻게 할 것인지

Entity Type NER Result + Checking Files (1 / N)

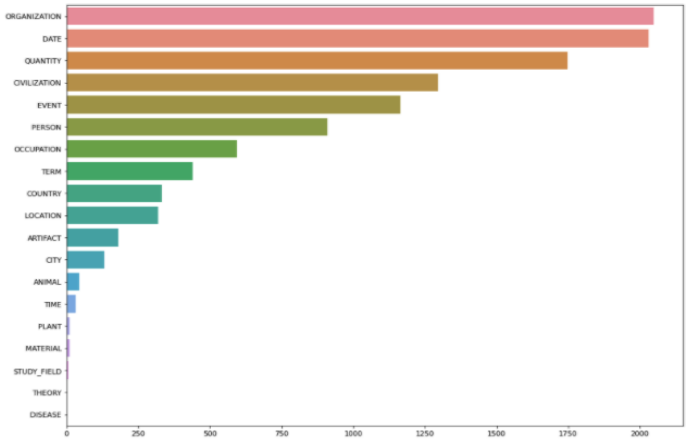
- 박진영 (AFC 본머스 - 리버풀)

org:hometown (단체:연고지) ——● **Relation 후보군 추출**

▼ ex

```
AFC 본머스(A.F.C. Bournemouth)는 잉글랜드 도싯주 본머스를 본거지로 하는 프로 축구 클럽이다.
리버풀 FC(영어: Liverpool Football Club)는 잉글랜드 머지사이드주 리버풀을 연고로 하는 프리미어리그 축구 클럽이다.
노리치 시티 풋볼 클럽(영어: Norwich City Football Club)은 잉글랜드의 축구 클럽으로, 노리치를 연고로 하고 있다.
노팅엄 포리스트 FC(Nottingham Forest FC)는 잉글랜드의 프로 축구 클럽팀으로 노팅엄을 연고로 삼고, 시티 그라운드를 홈 구장으로 이용하고 있다.
```

```
[('ORGANIZATION', 2049),
 ('DATE', 2030),
 ('QUANTITY', 1749),
 ('CIVILIZATION', 1295),
 ('EVENT', 1165),
 ('PERSON', 910),
 ('OCCUPATION', 594),
 ('TERM', 440),
 ('COUNTRY', 332),
 ('LOCATION', 320),
 ('ARTIFACT', 180),
 ('CITY', 131),
 ('ANIMAL', 45),
 ('TIME', 30),
 ('PLANT', 10),
 ('MATERIAL', 10),
 ('STUDY_FIELD', 5),
 ('THEORY', 1),
 ('DISEASE', 1)]
```



——● **대략적인 Entity 분포 확인**

· 분석 표지: 15개

- PERSON(PS), LOCATION(LC), ORGANIZATION(OG), ARTIFACTS(AF), DATE(DT), TIME(TI), CIVILIZATION(CV), ANIMAL(AM), PLANT(PT), QUANTITY(QT), STUDY_FIELD(FD), THEORY(TR), EVENT(EV), MATERIAL(MT), TERM(TM)

Relation Map Result

No	Relation(KR)	**Relation(ENG)**	**(Sub, Obj)**	Description
1	관계_없음	**no_relation**	-	관계를 유추할 수 없음 or 정의된 클래스 중 하나로 분류할 수 없음
2	단체:연고지	org:hometown	(ORG, LOC)	Object 는 Subject 의 연고지
3	단체:라이벌	org:rival	(ORG, ORG)	Object 는 Subject 의 라이벌 관계
4	단체:상대_단체	org:counterpart	(ORG, ORG)	Object 는 Subject 의 경기 상대
5	단체:상위_단체	org:member_of	(ORG, ORG)	Object 는 Subject 가 속하는 단체
6	단체:창단_일자	org:founded	(ORG, DAT)	Object 는 Subject 의 창립일 또는 설립일
7	단체:경기장	org:stadium	(ORG, LOC/ORG)	Object 는 Subject 가 사용하는 홈 경기장
8	단체:구성원	org:members	(ORG, PER)	Object 는 Subject 의 구성원(선수, 감독, 코치, 구단주, 단장 등)
9	이벤트:발생_시기	evt:happened	(EVT, DAT)	Object 는 Subject 의 발생 시기
10	사람:역할	per:role	(PER, ROL)	Object 는 Subject 의 역할

ORG:Hometown

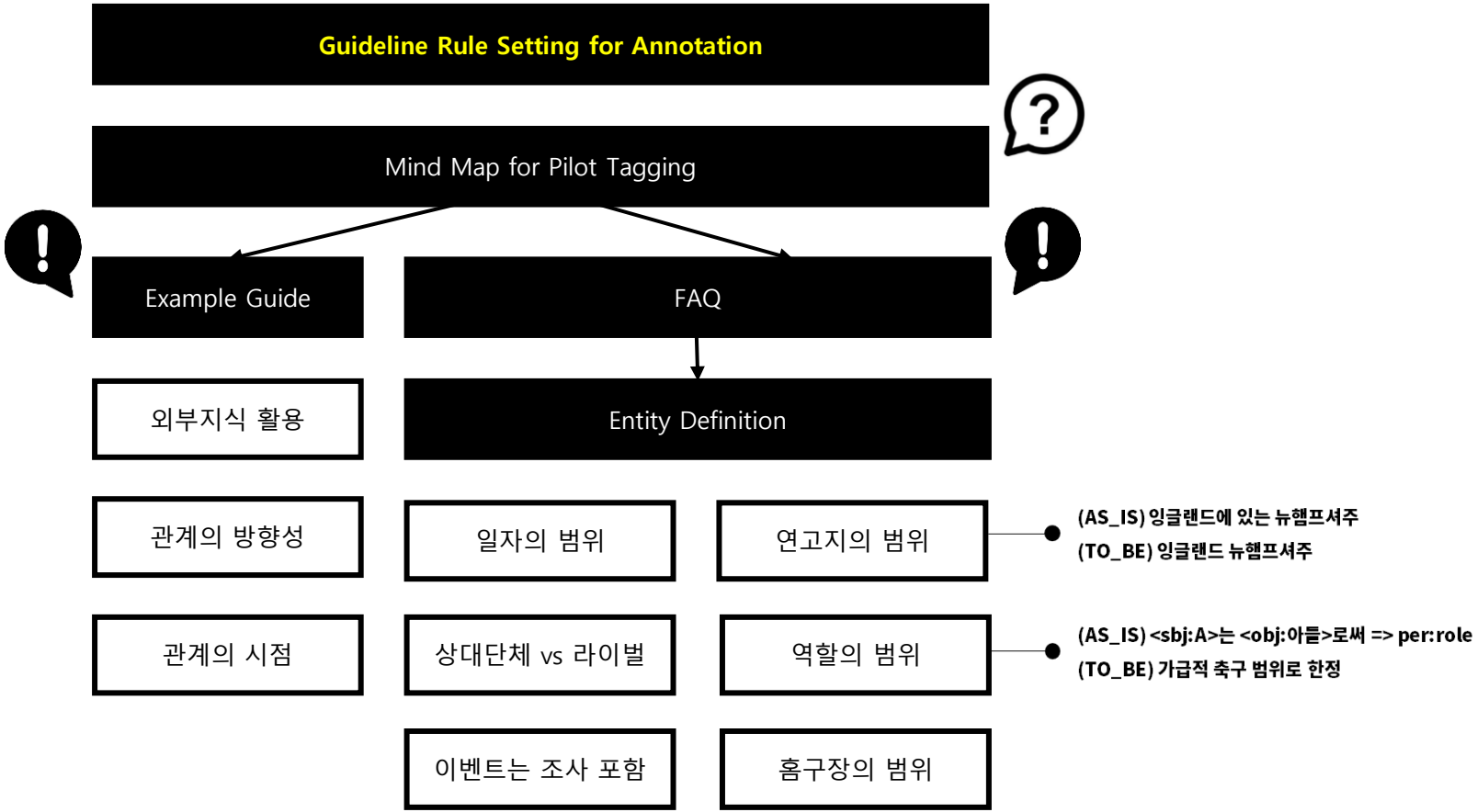
Subject Object

맨체스터 유나이티드 축구 클럽은 잉글랜드 **맨체스터**에 있는 잉글랜드 프로 축구 구단이다.

No_Relation

Subject Object

리버풀은 로저스의 경질 이후, 마인츠 05에서 뛰어난 지도력을 보여준 위르겐 클롭을 감독으로 선

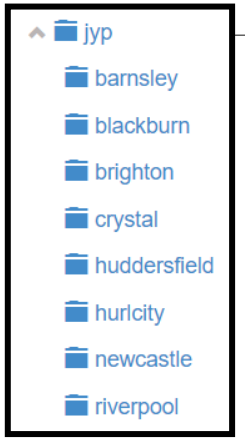

Guideline Rule Setting for Annotation
Mind Map for Pilot Tagging
Example Guide
FAQ
외부지식 활용
Entity Definition
관계의 방향성
일자의 범위
연고지의 범위
(AS_IS) 잉글랜드에 있는 뉴햄프셔주
(TO_BE) 잉글랜드 뉴햄프셔주
관계의 시점
상대단체 vs 라이벌
역할의 범위
(AS_IS) <sbj:A>는 <obj:아들>로써 => per:role
(TO_BE) 가급적 축구 범위로 한정
이벤트는 조사 포함
홈구장의 범위

Tagging procedure

조범준

맨체스터 시티.txt
레스터 시티.txt
수비수.txt
더비 카운티.txt
울버햄프턴 원더러스.txt
셰필드 웬즈데이.txt
골키퍼.txt
셰필드 유나이티드.txt

● 파일 분배 by 문장 개수



● 작업자별 디렉토리 구분

sentence_49.txt	조 페이건 감독 또한 생클리와 페이즐리 감독이 해낸 것을 물려받아 첫 시즌인 83 - 84시즌에 리그와 리그 컵, 유로피언 컵 우승을 석권하게 되었다.
sentence_168.txt	리버풀은 이를 위해 구장 인근의 주택 90채를 철거하고, 1억 5천만 파운드의 액수를 투자해 안필드를 확장하고 있다.

● 문장 단위 입력 by kss

조 페이건 감독 또한 생클리와 페이즐리 감독이 해낸 것을 물려받아 첫 시즌인 83 - 84시즌에 리그와 리그 컵, 유로피언 컵 우승을 석권하게 되었다.

● Tagging

IAA Pilot to Final

Pilot : 10%만 진행

	재현	진영	재욱	성민	연걸	혜원	범준
PA	0.753	0.861	0.891	0.882	0.861	0.893	0.875
PE	0.217	0.164	0.154	0.482	0.233	0.173	0.129
Fleiss' Kappa	0.684	0.842	0.872	0.772	0.819	0.871	0.856

Final

	재현	진영	재욱	성민	연걸	혜원	범준	total
PA	0.872	0.907	0.896	0.915	0.903	0.859	0.874	0.889
PE	0.183	0.164	0.118	0.626	0.218	0.151	0.130	0.194
Fleiss' Kappa	0.844	0.881	0.882	0.774	0.876	0.834	0.856	0.862

JBJ's Review

- **category (data) imbalance** 가 심할수록 값이 커짐
- fleiss' kappa 값에 **반비례**

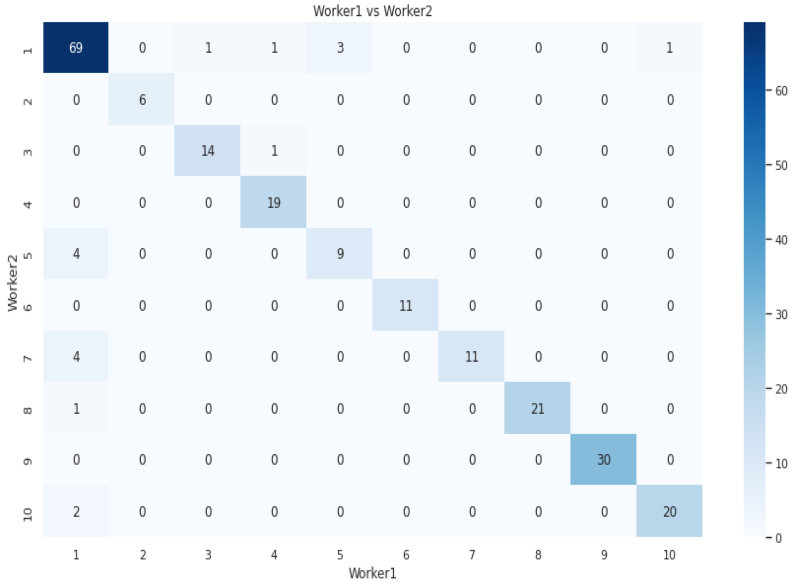
JBJ's Review

- **작업자간 tagging 일치도**가 높을 수록 값이 커짐
- fleiss' kappa 값에 **비례**

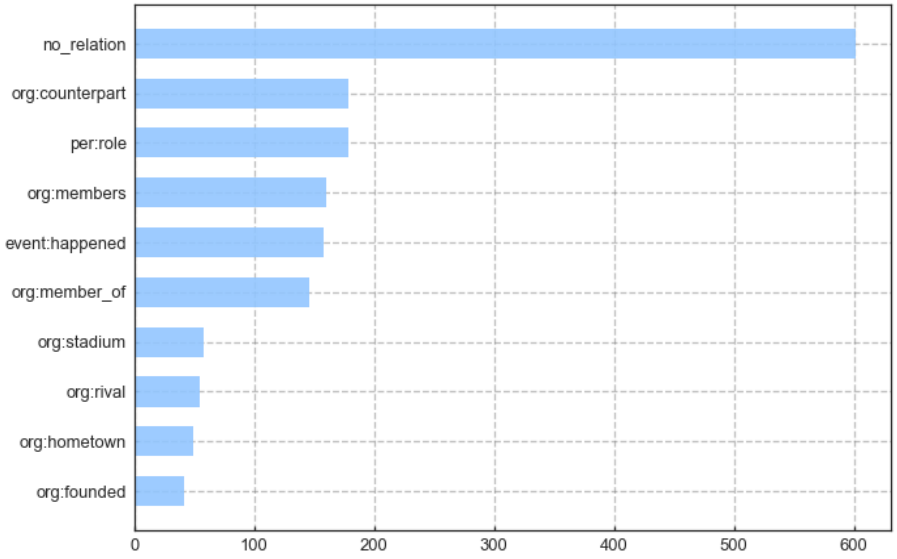
$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Label based on IAA

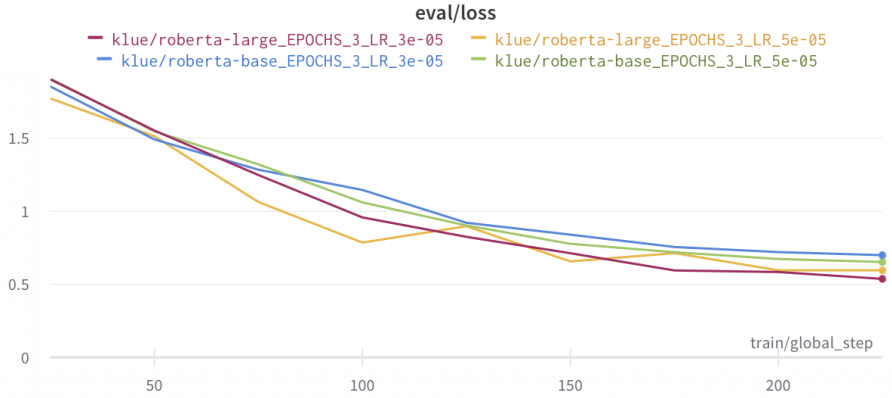
Confusion Matrix 작업자간 차이 확인

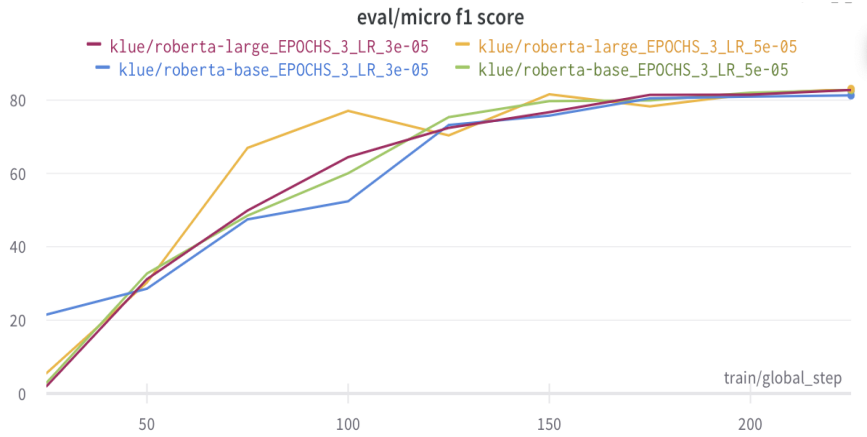


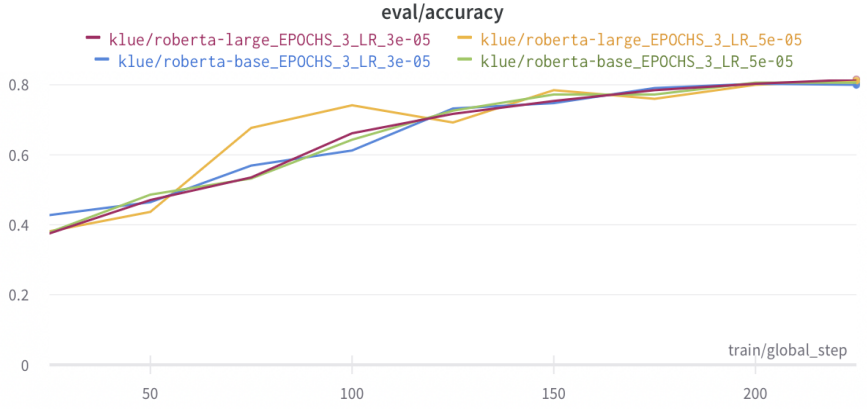
Final Label 최종 분포



Learning Result







Model	klue/roberta-base	klue/roberta-base	klue/roberta-large	klue/roberta-large
Epoch	3	3	3	3
Learning Rate	5e-5	3e-5	5e-5	3e-5
LR scheduler type	linear	linear	linear	linear
eval batch size	16	16	16	16
train batch size	16	16	16	16

감사합니다 ☺