




Boostcamp AI-Tech 2기

**KLUE RE 대회 토론게시판**

# 1. 자기소개



김채은 T2064

현) 기영이   
전) 요하리와 하이들   
전) 7Features 

🔥 핫게 🔥



[자유 토론] EDA로 발견한 데이터 중복 및 augmentation방향에 대한 인사이트

Posted by 김채은\_T2064 | 2021.09.28.20:10 | Comments (2)



Commented by 이요한\_T2166 | 2021.09.29.11:48

오우.. 좋은 인사이트 공유 감사합니다. 갓채은! 갓채은! 갓채은!

김채은\_T2064 | 2021.09.29.21:45 | 삭제

요하이님.... ㅎㅎㅎㅎㅎㅎㅎㅎㅎㅎㅎㅎㅎㅎ  
!!!~^!!!



Commented by 정전원\_T2206 | 2021.09.29.12:09

공유 감사합니다

중복 문장 중 label이 다른 데이터는 이정도인것 같네요!

6749	6749	대한항공은 5일 조양호 회장의 3자녀가 보유한 세이 버스카이 주식 9만9900주 전량...	['word': '대한항공', 'start_idx': 0, 'end_idx': 3,...]	['word': '조양호', 'start_idx': 9, 'end_idx': 11,...]	no_relation	wikipedia
12829	12829	대한항공은 5일 조양호 회장의 3자녀가 보유한 세이 버스카이 주식 9만9900주 전량...	['word': '대한항공', 'start_idx': 0, 'end_idx': 3,...]	['word': '조양호', 'start_idx': 9, 'end_idx': 11,...]	org.top_members/employees	wikipedia
8364	8364	배우 김병철 씨가 연기하는 정복동은 한리마마트를 당 하게 하기 위해 여러 계획을 세우...	['word': '정복동', 'start_idx': 15, 'end_idx': 17,...]	['word': '김병철', 'start_idx': 3, 'end_idx': 5, ...]	no_relation	wiktree
32299	32299	배우 김병철 씨가 연기하는 정복동은 한리마마트를 당 하게 하기 위해 여러 계획을 세우...	['word': '정복동', 'start_idx': 15, 'end_idx': 17,...]	['word': '김병철', 'start_idx': 3, 'end_idx': 5, ...]	per/alternate_names	wiktree
11511	11511	영화 '버즈 오브 프레이'는 베트남이 있는 고당사에서 할리퀸, 한트리스, 블랙 카나...	['word': '베트남', 'start_idx': 18, 'end_idx': 24, 'end_idx': 18,...]	['word': '고당사', 'start_idx': 24, 'end_idx': 26,...]	per/place_of_residence	wiktree
22258	22258	영화 '버즈 오브 프레이'는 베트남이 있는 고당사에서 할리퀸, 한트리스, 블랙 카나...	['word': '베트남', 'start_idx': 18, 'end_idx': 18,...]	['word': '고당사', 'start_idx': 24, 'end_idx': 26,...]	no_relation	wiktree
3296	3296	이날 프로그램 공개에서는 전복양산작법보존회와 강원 산·정상회의 사제동행 원소리, 광악...	['word': '강태환', 'start_idx': 62, 'end_idx': 64,...]	['word': '채소론', 'start_idx': 58, 'end_idx': 60,...]	per/title	wiktree
4212	4212	한편 전라남도도는 최근 확진자가 발생한 순천시와 여수 시에 마스크를 각각 2만장씩 총 ...	['word': '전라남도', 'start_idx': 3, 'end_idx': 9,...]	['word': '여수시', 'start_idx': 26, 'end_idx': 28,...]	org.members	wiktree
25094	25094	한편 전라남도도는 최근 확진자가 발생한 순천시와 여수 시에 마스크를 각각 2만장씩 총 ...	['word': '전라남도', 'start_idx': 3, 'end_idx': 9,...]	['word': '여수시', 'start_idx': 26, 'end_idx': 28,...]	org/place_of_headquarters	wiktree

김채은\_T2064 | 2021.09.29.21:46 | 삭제

공유 감사합니다 label이 다른 것은 직접 보고 판단하기 좋은 정도인 듯합니다!

Add Reply

👍 32

UP

## 2. EDA\_중복 문장 확인

int

약 3,660개 sentence 동일

	id	sentence	subject_entity	object_entity	label
count	32470.000000	32470	32470	32470	32470
unique	NaN	28803	12052	10195	30
top	NaN	도쿠가와 이에야스와 도쿠가와 히데타다가 20년에 걸쳐 안정시킨 막부를 이어받은 3대...	'민주당'	'대한민국'	no_relation
freq	NaN	3	144	276	9534
mean	16234.500000	NaN	NaN	NaN	NaN
std	9373.425957	NaN	NaN	NaN	NaN
min	0.000000	NaN	NaN	NaN	NaN

train / validation dataset split시  
leakage문제 발생!

### 3. EDA\_중복 문장 처리

#### CASE 1. sentence(=) , subj\_entity(≠) , obj\_entity(≠)

	guid	sentence	subject_entity	object_entity	label	source	subj_entity	subj_type	subj_start	obj_entity	obj_type	obj_start	dup
7989	7989	영원한 민주주의자 김근태 전 열린우리 당 의장의 부인이자 현직 민주당 국회의 원인 인재...	{'word': '인재근', 'start_idx': 44, 'end_idx': 46...	{'word': '김근태', 'start_idx': 10, 'end_idx': 12...	17	wikitree	인재근	PER	44	김근태	PER	10	False
12778	12778	영원한 민주주의자 김근태 전 열린우리 당 의장의 부인이자 현직 민주당 국회의 원인 인재...	{'word': '김근태', 'start_idx': 10, 'end_idx': 12...	{'word': '인재근', 'start_idx': 44, 'end_idx': 46...	17	wikitree	김근태	PER	10	인재근	PER	44	False



Augmentation흔적 > 문제 없음

### 3. EDA\_중복 문장 처리

#### CASE 2. sentence(=) , subj\_entity(=) , obj\_entity(=) , label(=)

278	277	이날 프로그램 공개에...	{'word': '강태환', 'start...	{'word': '색소폰', 'start...	no_relation	wikitree
10203	10202	이날 프로그램 공개에...	{'word': '강태환', 'start...	{'word': '색소폰', 'start...	no_relation	wikitree



첫번째 데이터만 보유하고 나머지 drop

```
dropped_dup = dataset_dup.drop_duplicates(['sentence', 'subj_entity', 'obj_entity', 'subj_start', 'obj_start', 'label'], keep='first')
```

### 3. EDA\_중복 문장 처리

### CASE 3. sentence(=) , subj\_entity(=) , obj(=) , label(≠)

- 개수 : `dropped_dup_first.duplicated(['sentence', 'subj_entity', 'obj_entity', 'subj_start', 'obj_start']).sum()`

6

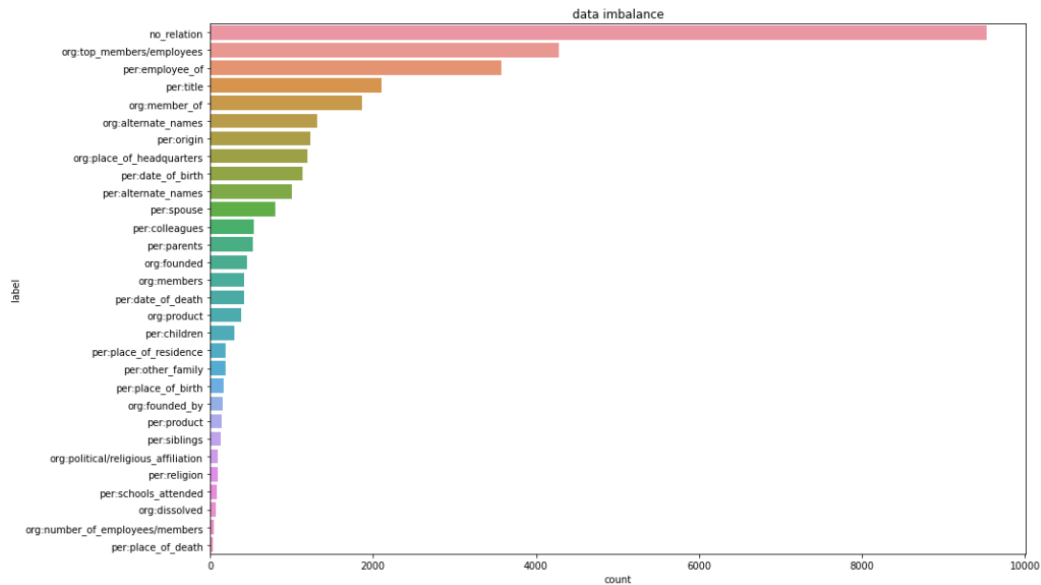
	guid	sentence	subject_entity	object_entity	label	source	subj_entity	subj_type	obj_entity	obj_type	subj_start	obj_start	dup	dup_label
4212	4212	한편 전라남도는 최근 확진자가 발생한 순천시와 여수시에 마스크를 각각 2만장씩 총 ...	{'word': '전라남도', 'start_idx': 3, 'end_idx': 6, ...}	{'word': '여수시', 'start_idx': 26, 'end_idx': 28, ...}	2	wikitree	전라남도	ORG	여수시	LOC	3	26	False	False
25094	25094	한편 전라남도에는 최근 확진자가 발생한 순천시와 여수시에 마스크를 각각 2만장씩 총 ...	{'word': '전라남도', 'start_idx': 3, 'end_idx': 6, ...}	{'word': '여수시', 'start_idx': 26, 'end_idx': 28, ...}	7	wikitree	전라남도	ORG	여수시	LOC	3	26	False	True

직접 Label 판별 후 drop

```
dup_diff_label_dataset_5 = dup_diff_label_dataset_4.drop(index=25094, axis=0)
```

## 4. Insight into Data Augmentation

### Data Imbalance문제



KLUE-RE 데이터를 Augmentation한 방식으로  
부족한 데이터를 늘리면 어떨까?

## 4. Insight into Data Augmentation

### ■ subject\_entity와 object\_entity 교체

guid	sentence	subject_entity	object_entity	label	source	subj_entity	subj_type	subj_start	obj_entity	obj_type	obj_start	dup
7989	영원한 민주주의자 김근태 전 열린우리 당 의장의 부인이자 현직 민주당 국회의 원인 인재...	{'word': '인재근', 'start_idx': 44, 'end_idx': 46...	{'word': '김근태', 'start_idx': 10, 'end_idx': 12...	17	wikintree	인재근	PER	44	김근태	PER	10	False
12778	영원한 민주주의자 김근태 전 열린우리 당 의장의 부인이자 현직 민주당 국회의 원인 인재...	{'word': '김근태', 'start_idx': 10, 'end_idx': 12...	{'word': '인재근', 'start_idx': 44, 'end_idx': 46...	17	wikintree	김근태	PER	10	인재근	PER	44	False

### ■ 다른 단어로 entity 변경

#### sentence

도쿠가와 이에야스와 도쿠가와 히데타다가 20년에 걸쳐 안정시키 막부를 이어받은 3대 쇼군 도쿠가와 이에미쓰는 중신들에게 유교 사상을 철저히 연구할 것을 지시했고, 한 편으로 도쿠가와 미쓰쿠니 등은 《대일본사》(大日本史)와 같은 역사서를 편찬하는 등 문치(文治)를 지향하였으며, 이를 바탕으로 5대 쇼군 도쿠가와 이에쓰나 대에 에도 막부는 겐로쿠 호황이라고 부르는 최대의 전성기를 맞이했다.

#### subject

{'word': '도쿠가와 이에미쓰', 'start\_idx': 50, 'end\_idx': 58, 'type': 'PER'}

{'word': '도쿠가와 히데타다', 'start\_idx': 11, 'end\_idx': 19, 'type': 'PER'}

{'word': '도쿠가와 이에야스', 'start\_idx': 0, 'end\_idx': 8, 'type': 'PER'}

#### object

{'word': '에도 막부', 'start\_idx': 181, 'end\_idx': 185, 'type': 'ORG'}

#### label

per:employee\_of

Augmentation 팀



- Entity변경을 위한 NER
- Labeling을 위한 예측 모델



## 공유 +1



“생각대로 되지 않는다는 건 정말 멋져요!  
생각지도 못했던 일이 일어나는걸요.”

