



자연어 기반 기후기술분류 AI 경진대회

한예송, 홍재령

목차

#01 Introduction

#02 Baseline & AutoEDA

#03 Pororo 활용

#04 XLM-Roberta 활용



Introduction



#1 대회 소개

주제 : 국가 연구개발과제를 '기후기술분류체계'에 맞추어 라벨링하는 알고리즘 개발

데이터셋

- train.csv (174304, 13) - 기후기술분류 label 포함
- test.csv (43576, 12) - 기후기술분류 label 미포함
- sample_submission.csv (43576, 2)
- labels_mapping.csv - label과 기후기술분류체계를 mapping 한 meta data

평가산식 : Macro-F1

과제명	요약문_연구목표	요약문_연구내용	요약문_기대효과	요약문_한글키워드	요약문_영문키워드
유전정보를 활용한 새로운 ...	○ 새로운 해충분류군의 등...	(가) 외래 및 돌발해충의 발...	○ 새로운 돌발 및 외래해...	뉴클레오티드 염기서열, 분...	nucleotide sequence, mol...
대장암의 TRAIL 내성 표적 ...	최종목표: TRAIL 감수성 표...	1차년도 1) Microarray를 ...	1) TRAIL 내성 특이적 표적...	대장암,항암제 내성,세포사...	TRAIL,Colorectal cancer,TR...
비목질계 셀룰로오스 식물...	* 식물계자원 정련 및 최적...	* 식물계자원 정련 및 최적...	* 국내 독자적인 비목질계 ...	기능성 셀룰로오스 파이버,...	functional cellulose fiber,n...
소화기 암 진단용 분자영상...	# 암특이적 바이오마커 발...	# 소화기 암 진단용 분자영...	# 암 진단기술의 차별성: ...	분자 진단,형광 조영제,프...	Molecular diagnosis,Fluore...
위암환자의 항암제반응예...	수술이 불가능한 위암환자...	-In situ hybridization 검사...	-본 연구는 파라핀보관조...	BRCA,제자리부합법,조직미...	BRCA,Insituhybridization,ti...
국제 핵융합 재료조사시설(...	○ 기존 가속기 설계 및 운...	○ 1차년 (2017년): - IFMIF...	○ 현재 한국은 IFMIF 에 ...	국제 핵융합 재료 조사 시...	International Fusion Mater...
마이크로시스를 적용한 옥...	1. 2차년도 개발목표 2차년...	2. 2차년도 개발내용 2차년...	3. 기술적 및 경제적 기대...	마이크로시스,옥내케이블,...	Microsheath,Indoor cable,...
임상·오믹스 정보 통합 개...	본 연구의 최종 목표는 종...	1 단계 1) CDM 기반 종적 ...	- 암 정밀의료 관련 시스템...	개방형 플랫폼,통합 임상 ...	openplatform,Integratedcli...
IoT기반 수출배 선과장 물...	IoT기반으로 한 수출배 선...	수출배 원물보관 환경에 따...	IoT기반 수출배 입출고 관...	현장연구,생산단지,수출,현...	field,production area,expo...
지역 창조경제 생태계 활성...	○ 바이오산업 분야의 혁신...	<공동프로그램 추진 배경>...	(1) 바이오 분야 ○ 오송생...	의료기기,의약,헬스,화장품,...	Medical,Medicine,Health,C...

#1 기후기술 분류체계란?

국가과학기술자문회의 기후기술협력 중장기 추진계획을 바탕으로,
[감축], [적응], [융·복합]의 3개 분야의 45개 기술분류로 구분되어 활용

대분류	중분류				소분류 범위
감축	온실가스 저감	에너지 생산 & 공급	발전 & 전환	(1)비재생 에너지	1. 원자력 발전
					2. 핵융합 발전
					3. 청정화력 발전·효율화
				(2)재생 에너지	4. 수력
					5. 태양광
					6. 태양열
					7. 지열
					8. 풍력
					9. 해양에너지
					10. 바이오에너지
					11. 폐기물
				(3)신에너지	12. 수소제조
					13. 연료전지
		에너지 저장 & 운송	(4)에너지 저장	14. 전력저장	
				15. 수소저장	
			(5)송배전 & 전력 IT	16. 송배전 시스템	
				17. 전기지능화 기기	
		(6) 에너지 수요	18. 수송효율화		
			19. 산업효율화		
			20. 건축효율화		
	(7)온실가스 고정	21. CCUS			
		22. Non-Co2 저감			

Baseline & AutoEDA



#2-1 [Baseline] Random Forest

2. 데이터 EDA

okt Tokenizer + CounterVectorizer + Randomforest Classifier

```
test.head(2)
```

index	제출 년도	사업명	사업 _부 처명	계속 과제 여부	내역사업 명	과제명	요약문_연구목표	요약문_연구내용	요약문_기대효과	요약문_한글키워 드	요약문_영문키워드	
0	174304	2016	경제협력권 산업육성	산업 통상 자원 부	신규	자동차융합 부품	R-FSSW 기술 적용 경량 차체 부품 개 발 및 품질 평가를 위한 64채널 C-SC...	○ 차체 점용접부의 품질 검사를 위한 64 채널 무선 기반 C- Scan 탐촉자 개발W...	○ 1차년도WnWn . 개발 탐측 시스템의 성능 평 가 위한 표준 시편 제작 시...	○ 기술적 파급효과 WnWn - 본 연구에서 개 발된 R-FSSW 접합 기술 은 기존 ...	마찰교반점용접, 비 파괴 검사, 초음파 탐 상, 씨 스캔, 용접 품 질 평가	Friction Stir Spot Welding, Non- destructive ev...
1	174305	2018	개인기초연 구(과기정 통부)(R&D)	과학 기술 정보 통신 부	계속	신진연구 (총연구비5 천이상~1.5 억이하)	다입자계를 묘사하 는 편미분방정식에 대한 연구	자연계에는 입자의 개수가 아주 큰 다양 한 다입자계가 존재 한다. 이런 다입자계 의 효...	연구과제1. 무한입자계 의 동역학 / 작용소 (operator) 방정식에 대 한 연구Wn...	본 연구는 물리학에서 중요한 대상인 다입자 계를 묘사하는 모델방 정식의 정당성을 보장 하...	다체계 방정식,동역 학의 안정성,양자역 학,고전역학,평균장 극한,고전극한,비상 대론적 극한	many particle system,stability of dynamics,qua...

과제명 길이 최댓값: 229

과제명 길이 최솟값: 2

과제명 길이 평균값: 35.84252225995961

과제명 길이 중간값: 34.0

요약문_연구목표 길이 최댓값: 3951

요약문_연구목표 길이 최솟값: 1

요약문_연구목표 길이 평균값: 318.1008066366807

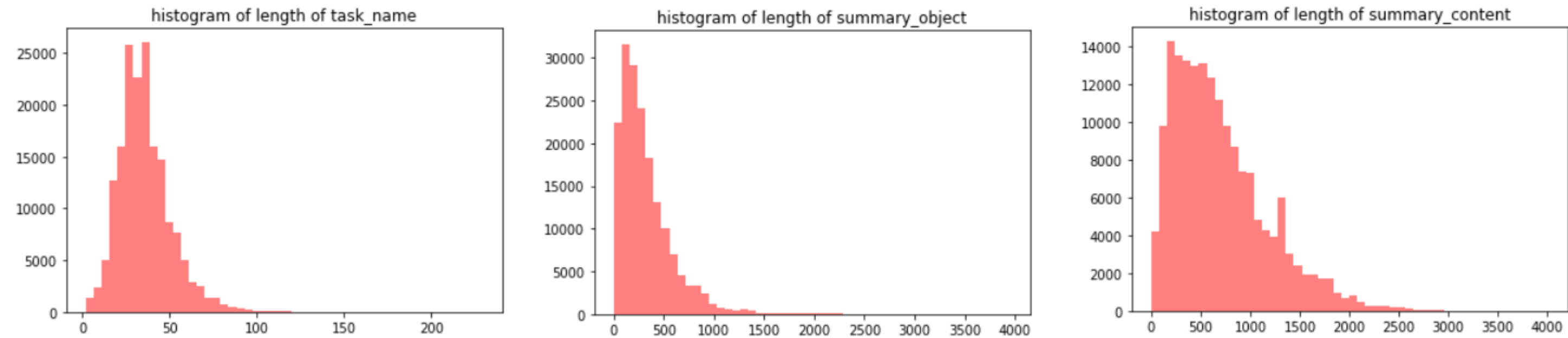
요약문_연구목표 길이 중간값: 249.0

요약문_연구내용 길이 최댓값: 3999

요약문_연구내용 길이 최솟값: 1

요약문_연구내용 길이 평균값: 699.2930282724435

요약문_연구내용 길이 중간값: 597.0



#2-1 [Baseline] Random Forest

3. 데이터 전처리

해당 baseline에서는 과제명 column만 활용

- 1) re.sub 한글 및 공백을 제외한 문자 제거
- 2) okt 객체를 활용해 형태소 단위로 나눔
- 3) remove_stopwords로 불용어 제거

```
train.head(2)
```

	과제명	label
0	유전정보를 활용한 새로운 해충 분류군 동정기술 개발	24
1	대장암의 TRAIL 내성 표적 인자 발굴 및 TRAIL 반응 예측 유전자 지도 구축...	0

```
def preprocessing(text, okt, remove_stopwords=False, stop_words=[]):
    text=re.sub("[^가-힣ㄱ-ㅎㅏ-ㅣ]", "", text)
    word_text=okt.morphs(text, stem=True)
    if remove_stopwords:
        word_review=[token for token in word_text if not token in stop_words]
    return word_review
```

```
stop_words=['은','는','이','가','하','아','것','들','의','있','되','수','보','주','등','한']
okt=Okt()
clean_train_text=[]
clean_test_text=[]
```

```
for text in tqdm.tqdm(train['과제명']):
    try:
        clean_train_text.append(preprocessing(text, okt, remove_stopwords=True, stop_words=stop_words))
    except:
        clean_train_text.append([])
```


#2-1 [Baseline] Random Forest

3. 데이터 전처리

4) tokenizer 인자에는 list를 받아서 그대로 내보내는 함수를 넣어줌
(소문자화를 하지 않도록 설정해야 에러가 나지 않음)

```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(tokenizer = lambda x: x, lowercase=False)
train_features=vectorizer.fit_transform(clean_train_text)
test_features=vectorizer.transform(clean_test_text)
```

4. 모델링

훈련 데이터 셋과 검증 데이터 셋으로 분리 후

```
TEST_SIZE=0.2
RANDOM_SEED=42

train_x, eval_x, train_y, eval_y=train_test_split(train_features, train['label'], test_size=TEST_SIZE, random_state=RANDOM_SEED)
```

랜덤포레스트로 모델링

```
from sklearn.ensemble import RandomForestClassifier

forest=RandomForestClassifier(n_estimators=100)

forest.fit(train_x, train_y)
```

#2-2 [Baseline] LSTM

okt Tokenizer + keras embedding + LSTM

이전과 동일한 1~2단계 진행

3. 데이터 전처리

해당 baseline에서는 과제명 column만 활용

- 1) re.sub 한글 및 공백을 제외한 문자 제거
- 2) okt 객체를 활용해 형태소 단위로 나눔
- 3) remove_stopwords로 불용어 제거

```
train.head(2)
```

	과제명	label
0	유전정보를 활용한 새로운 해충 분류군 동정기술 개발	24
1	대장암의 TRAIL 내성 표적 인자 발굴 및 TRAIL 반응 예측 유전자 지도 구축...	0

```
def preprocessing(text, okt, remove_stopwords=False, stop_words=[]):
    text=re.sub("[^가-힣ㄱ-ㅎㅏ-ㅣ]", "", text)
    word_text=okt.morphs(text, stem=True)
    if remove_stopwords:
        word_review=[token for token in word_text if not token in stop_words]
    return word_review
```

```
stop_words=['은','는','이','가','하','아','것','들','의','있','되','수','보','주','등','한']
okt=Okt()
clean_train_text=[]
clean_test_text=[]
```

```
for text in tqdm.tqdm(train['과제명']):
    try:
        clean_train_text.append(preprocessing(text, okt, remove_stopwords=True, stop_words=stop_words))
    except:
        clean_train_text.append([])
```

#2-2 [Baseline] LSTM

3. 데이터 전처리

4) 토큰나이징 객체를 만든 후 인덱스 벡터로 변환

```
tokenizer=Tokenizer()
tokenizer.fit_on_texts(clean_train_text)

train_sequences=tokenizer.texts_to_sequences(clean_train_text)
test_sequences=tokenizer.texts_to_sequences(clean_test_text)
word_vocab=tokenizer.word_index
```

5) 패딩 처리

```
train_inputs=pad_sequences(train_sequences, maxlen=40, padding='post')
test_inputs=pad_sequences(test_sequences, maxlen=40, padding='post')
```

6) 추후 재사용 가능하도록 npy로 변환

```
DATA_IN_PATH='./data_in/'
TRAIN_INPUT_DATA = 'train_input.npy'
TEST_INPUT_DATA = 'test_input.npy'

import os
if not os.path.exists(DATA_IN_PATH):
    os.makedirs(DATA_IN_PATH)

np.save(open(DATA_IN_PATH+TRAIN_INPUT_DATA, 'wb'), train_inputs)
np.save(open(DATA_IN_PATH+TEST_INPUT_DATA, 'wb'), test_inputs)

data_configs={}
data_configs['vocab']=word_vocab
data_configs['vocab_size'] = len(word_vocab)+1
json.dump(data_configs, open(DATA_IN_PATH+'data_configs.json', 'w'), ensure_ascii=False)
```

#2-2 [Baseline] LSTM

4. 모델링

파라미터 설정

```
vocab_size = data_configs['vocab_size']  
embedding_dim = 32  
max_length = 40  
oov_tok = "<OOV>"
```

가벼운 NLP모델 생성 및 compile, fit

```
model = tf.keras.Sequential([  
    tf.keras.layers.Embedding(vocab_size, embedding_dim, input_length=max_length),  
    tf.keras.layers.GlobalAveragePooling1D(),  
    tf.keras.layers.Dense(128, activation='relu'),  
    tf.keras.layers.Dense(46, activation='softmax')  
])
```

```
model.compile(loss='sparse_categorical_crossentropy',  
              optimizer='adam',  
              metrics=['accuracy'])
```

```
num_epochs = 30  
history = model.fit(train_inputs, labels,  
                    epochs=num_epochs, verbose=2,  
                    validation_split=0.2)
```

#2-3 AutoEDA

Dataprep AutoEDA

AutoEDA를 통해 알 수 있는 것

1. 요약문 데이터들과 과제명을 제외한 데이터들 → Cardinality 작음 (=중복도 높음)
2. train 데이터 전체 안의 중복 데이터는 없지만, 요약문 데이터들과 과제명 데이터의 Cardinality가 train 데이터 174304와 같지 않음
→ 동일한 과제명, 동일한 요약문을 가진 데이터가 있음
→ train 데이터의 cardinality 비율 < test 데이터의 cardinality 비율
3. 기후기술 분류가 아닌 데이터(label==0)가 80%이상 존재
→ 불균형 데이터에 대한 적절한 처리 필요
4. 각 피처별 평균 길이 / 최소 길이/ Word Frequency

create_report(): 데이터셋에 대한 포괄적인 profile report 생성

```
create_report(train)
```

#2-3 AutoEDA – Train Dataset Cardinality

train 데이터의 cardinality 값들만 모아보기

```
# 전체 train 데이터 개수 대비 각 컬럼 당 unique 값 개수
print("*-----개수-----*")
print(train.drop("label", axis=1).apply(lambda x: x.nunique()))

print("\n", "*-----비율-----*", "\n")
# 전체 train 데이터 개수 대비 각 컬럼 당 unique 값 비율
display(train.drop("label", axis=1).apply(lambda x: x.nunique()).div(train.shape[0]).mul(100))
```

-----개수-----	*-----비율-----*
제출년도	4
사업명	1414
사업_부처명	28
계속과제여부	2
내역사업명	4324
과제명	106623
요약문_연구목표	133267
요약문_연구내용	146499
요약문_기대효과	136041
요약문_한글키워드	109125
요약문_영문키워드	115513
dtype: int64	
제출년도	0.002295
사업명	0.811226
사업_부처명	0.016064
계속과제여부	0.001147
내역사업명	2.480723
과제명	61.170713
요약문_연구목표	76.456650
요약문_연구내용	84.047985
요약문_기대효과	78.048123
요약문_한글키워드	62.606136
요약문_영문키워드	66.270998
dtype: float64	

→ 비율: 작을수록 Cardinality 작음

train, test unique 값 종류 확인

```
=====제출년도=====
제출년도의 train unique 개수:4
제출년도의 test unique 개수:4
test 제출년도에서 train 제출년도 unique 제거 후 개수 :0
test 데이터에만 있는 값의 비율 : 0.0%
test 제출년도에서 train 제출년도 unique 제거 후 나머지 :set()
```

```
=====사업명=====
사업명의 train unique 개수:1414
사업명의 test unique 개수:1158
test 사업명에서 train 사업명 unique 제거 후 개수 :26
test 데이터에만 있는 값의 비율 : 2.25%
test 사업명에서 train 사업명 unique 제거 후 나머지 :['정지계도복합위성개발
기술개발(R&D)', '한국천문연구원연구운영비지원(R&D)(운영경비)', '국립기상과
원자력통제기술원연구운영비지원(R&D)(운영경비)', '한국철도기술연구원연구운영
학원연구운영비지원(R&D)(운영경비)']
```

→ test 데이터에도 train 데이터의 값과 완전히 같은 데이터가 어느 정도 존재

```
=====과제명=====
과제명의 train unique 개수:106623
과제명의 test unique 개수:37857
test 과제명에서 train 과제명 unique 제거 후 개수 :13978
test 데이터에만 있는 값의 비율 : 36.92%
test 과제명에서 train 과제명 unique 제거 후 나머지 :['다제내성 감염병 치료를 위한 항생제의 집!
생화적 요인 구명', '섬유형 트랜지스터 삽입형 전자섬유 기반 인체신호 모니터링 기술 실용화를 위한
지 기능부 Water-proof형 친환경 절연 다분기시스템', '온라인 구매 여정 기반 E-Commerce AI 솔루션
정화 및 악취제거', '다차원 대용량 데이터들을 위한 다해상도 근사 기법 연구', '황칠나무 잎 추출물으
시스템과 IoT 시스템을 위한 혁신적 통신 기술', '고온 환경에서의 인광 기반 비접촉 온도/속도/압력,
```

```
=====요약문_연구내용=====
요약문_연구내용의 train unique 개수:146499
요약문_연구내용의 test unique 개수:40254
test 요약문_연구내용에서 train 요약문_연구내용 unique 제거 후 개수 :31785
test 데이터에만 있는 값의 비율 : 78.96%
```

```
=====요약문_기대효과=====
요약문_기대효과의 train unique 개수:136041
요약문_기대효과의 test unique 개수:39332
test 요약문_기대효과에서 train 요약문_기대효과 unique 제거 후 개수 :27229
test 데이터에만 있는 값의 비율 : 69.23%
```

#2-3 AutoEDA – Feature별 분석

1) 제출년도

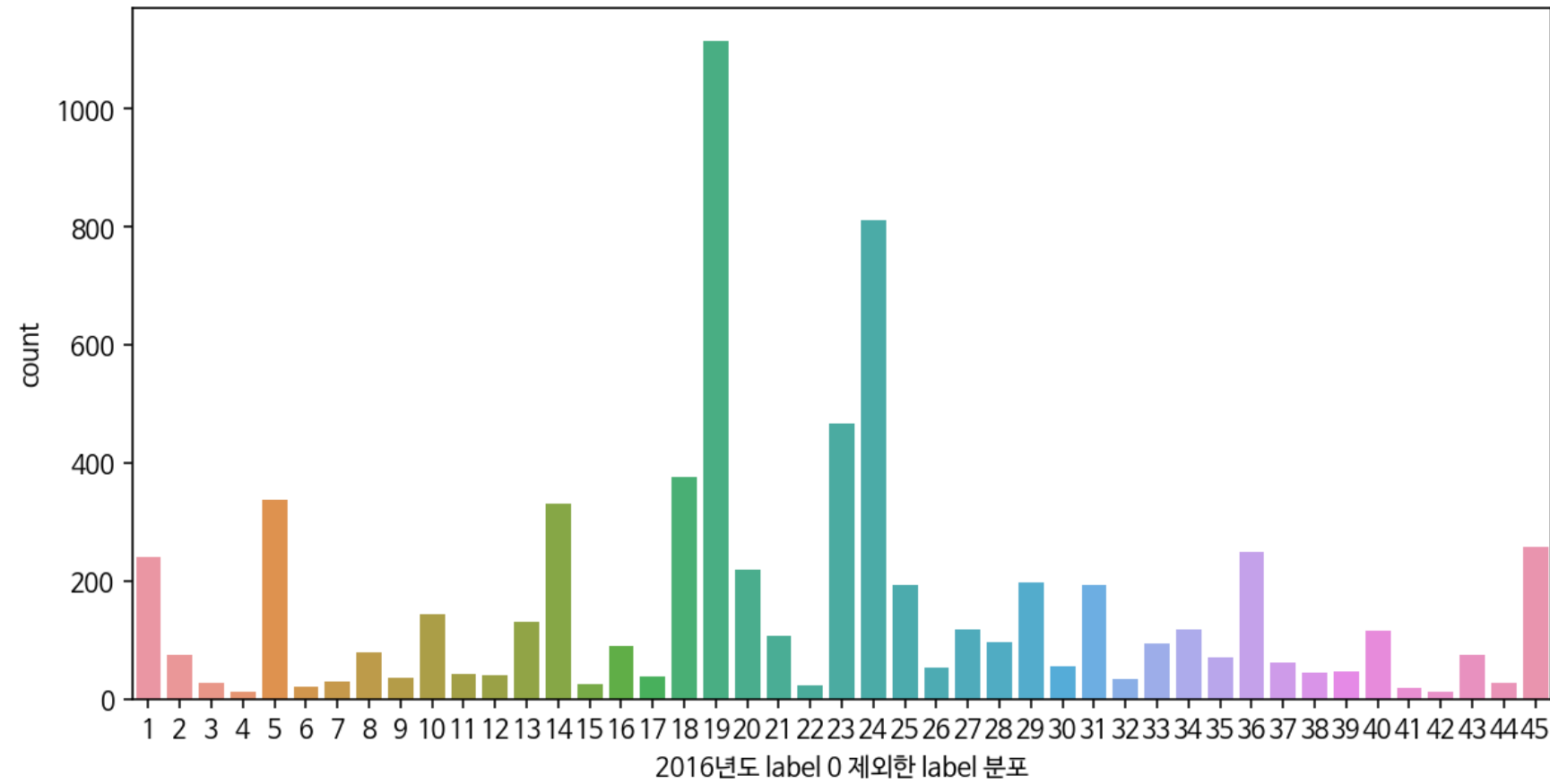
1. 제출년도 train / test 비율 분석

```
display(train.제출년도.value_counts(normalize=True).mul(100).round(2).to_frame())
print()
display(test.제출년도.value_counts(normalize=True).mul(100).round(2).to_frame())
```

제출년도		제출년도	
2019	28.02	2019	28.18
2018	25.83	2018	25.63
2017	24.54	2017	24.35
2016	21.62	2016	21.84

2. 연도별 label 분포 분석

```
year = '2016'
plt.figure(figsize=(10,5))
sns.countplot(data=train.query(f'label != "0" and 제출년도 == @year'), x='label').set_xlabel(f'{year}년도 label 0 제외한 label 분포')
```



#2-3 AutoEDA – Feature별 분석

2) 사업명

1. 사업명 train / test 비율 top 10 분석

```
display(train.사업명.value_counts(normalize=True).mul(100).round(2).to_frame().head(10))
print()
display(test.사업명.value_counts(normalize=True).mul(100).round(2).to_frame().head(10))
```

사업명	
개인기초연구(과기정통부)(R&D)	9.34
개인기초연구(교육부)(R&D)	6.72
개인기초연구(미래부)	3.63
개인기초연구(교육부)	3.24
개인연구지원	2.54

2. 사업명 label 하나인 데이터 중 개수가 많은 순서대로 20개 확인

```
train.groupby("사업명").agg({'label':['nunique','count','get_mode']}).droplevel(0, axis=1).sort_values(["nunique", "count"], ascending=[True, False]).head(20)
```

사업명	nunique	count	get_mode
질환극복기술개발(R&D)	1	539	0
암연구소및국가암관리사업본부연구운영비지원	1	419	0
지방대학육성사업(0.5)	1	415	0
첨단의료기술개발	1	367	0
한국고등과학원연구운영비지원	1	331	0
첨단의료기술개발(R&D)	1	317	0
뇌과학위성기술개발(R&D)	1	302	0

➔ 가장 첫 번째 데이터는 같은 사업명인 데이터 539개가 모두 label 0으로만 분류됨

#2-3 AutoEDA – Feature별 분석

3) 사업_부처명

1. 사업_부처명 train / test 비율 분석

```
display(train.사업_부처명.value_counts(normalize=True).mul(100).round(2).to_frame())
print()
display(test.사업_부처명.value_counts(normalize=True).mul(100).round(2).to_frame())
```

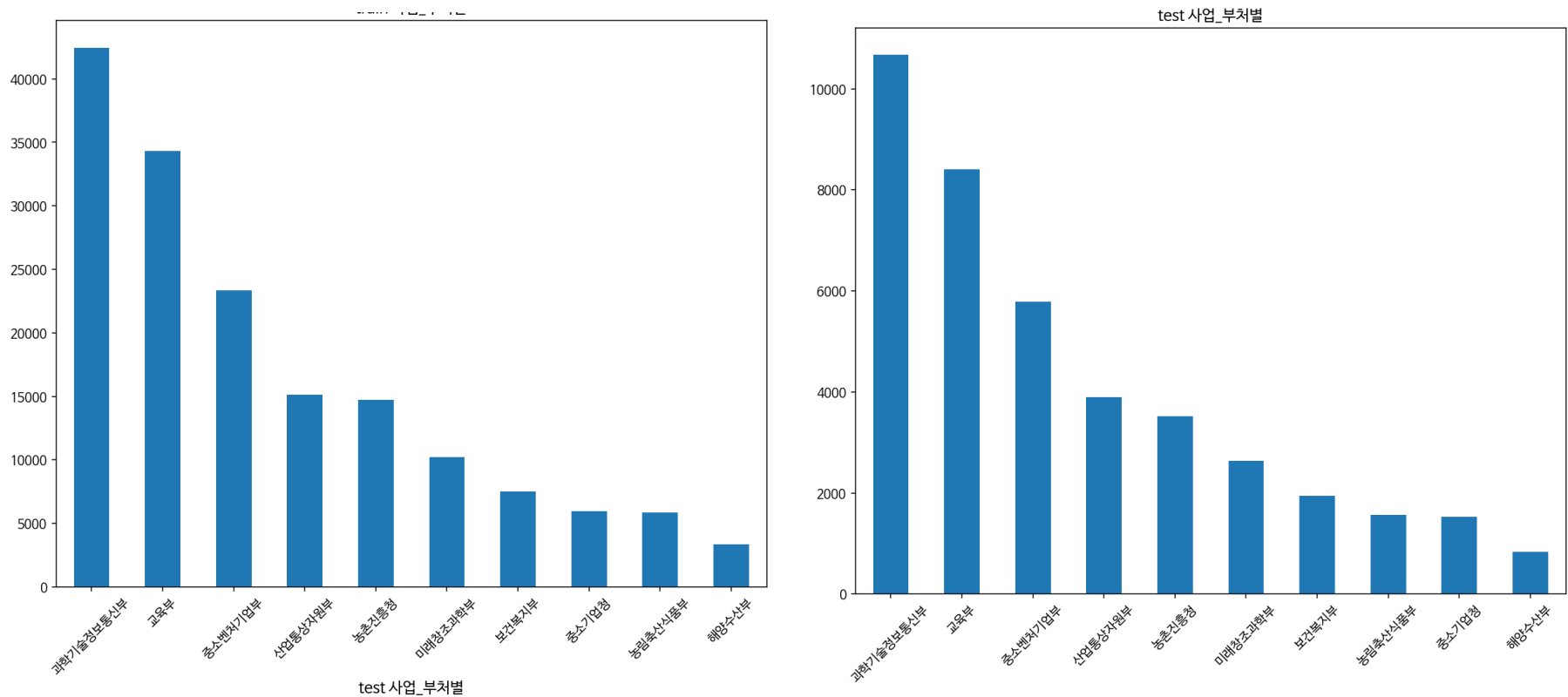
사업_부처명	
과학기술정보통신부	24.36
교육부	19.70
중소벤처기업부	13.40
산업통상자원부	8.68
농촌진흥청	8.44
미래창조과학부	5.97

사업_부처명	
과학기술정보통신부	24.50
교육부	19.31
중소벤처기업부	13.28
산업통상자원부	8.93
농촌진흥청	8.08
미래창조과학부	6.06

2. 사업_부처별 분포 분석

```
fig, axs = plt.subplots(2,1,figsize=(10,18))
train.사업_부처명.value_counts()[10].plot.bar(rot=45, ax = axs[0], title = "train 사업_부처별")

test.사업_부처명.value_counts()[10].plot.bar(rot=45, ax = axs[1], title = "test 사업_부처별")
```



#2-3 AutoEDA – Feature별 분석

4) 계속과제여부

1. 계속과제여부 train / test 비율 분석

```
display(train.계속과제여부.value_counts(normalize=True).mul(100).round(2).to_frame())
print()
display(test.계속과제여부.value_counts(normalize=True).mul(100).round(2).to_frame())
```

계속과제여부		계속과제여부	
계속	58.05	계속	58.32
신규	41.95	신규	41.68

5) 내역사업명

1. 제출년도 train / test 비율 분석

```
display(train.내역사업명.value_counts(normalize=True).mul(100).round(2).to_frame().head(10))
print()
display(test.내역사업명.value_counts(normalize=True).mul(100).round(2).to_frame().head(10))
```

내역사업명	
자유공모	3.40
기본연구(1년~3년)	2.93
기본연구지원사업	2.75
기본연구지원	2.20
기본연구(1년~5년)	2.10
글로벌박사펠로우십사업	1.63

내역사업명	
자유공모	3.39
기본연구(1년~3년)	2.83
기본연구지원사업	2.75
기본연구지원	2.20
기본연구(1년~5년)	1.95
글로벌박사펠로우십사업	1.72

#2-3 AutoEDA – Feature별 분석

6) 과제명, 요약문

```
# 컬럼별 중복 데이터 개수 확인 및 라벨 중복 확인 함수
def show_col_val_counts(col, threshold):

    dic_study_name = {}
    col_val_cnt = train[col].value_counts()

    # threshold 이상의 데이터를 가지고 있는 value만 선택
    for study_name in col_val_cnt.loc[col_val_cnt.ge(threshold)].index:

        # 연구내용 등 긴 문장은 50개로 줄이기
        print(f"*-----{study_name[:50]}-----*")
        study_name_val_cnt = train.loc[train[col]==study_name].label.value_counts()

        # 해당 col의 value 값이 고유 라벨 하나만을 가지고 있는 경우 저장
        if study_name_val_cnt.shape[0]==1:
            dic_study_name[study_name] = study_name_val_cnt.index[0]
            display(study_name_val_cnt)
            print()

    return dic_study_name
```

```
_ = show_col_val_counts("과제명", 10)
```

-----여성참여활성화과제-----

```
0      10
Name: label, dtype: int64
```

-----기초과학연구소-----

```
0         6
27        2
33        2
Name: label, dtype: int64
```

-----양식어류 건강성 신속진단을 위한 기술개발-----

```
34      10
Name: label, dtype: int64
```

#2-3 AutoEDA – Feature별 분석

6) 과제명, 요약문

```
_ = show_col_val_counts("요약문_연구목표", 10)
```

```
_ = show_col_val_counts("요약문_기대효과", 30)
```

- 요약문_한글키워드, 요약문_영문키워드
으로도 반복
- 각 요약문/ 과제명마다
정형화된 형식이 있는 것 확인 가능

-----대학 및 연구기관이 보유한 연구장비 공동활용율 증가 및 고가의 연구장비를 활용함에 있어 6-----

0	1944
19	43
45	11
18	9
5	8
16	7
29	6
28	3
27	3
25	1
24	1
23	1
20	1
14	1
8	1
6	1

Name: label, dtype: int64

-----보안과제 정보-----

0	711
10	2
40	1
13	1
8	1

Name: label, dtype: int64

#2-3 AutoEDA – Feature별 분석

6) 과제명, 요약문

```
_ = show_col_val_counts("요약문_연구목표", 10)
```

```
_ = show_col_val_counts("요약문_기대효과", 30)
```

- 요약문_한글키워드, 요약문_영문키워드
으로도 반복
- 각 요약문/ 과제명마다
정형화된 형식이 있는 것 확인 가능

-----대학 및 연구기관이 보유한 연구장비 공동활용율 증가 및 고가의 연구장비를 활용함에 있어 6-----

0	1944
19	43
45	11
18	9
5	8
16	7
29	6
28	3
27	3
25	1
24	1
23	1
20	1
14	1
8	1
6	1

Name: label, dtype: int64

-----보안과제 정보-----

0	711
10	2
40	1
13	1
8	1

Name: label, dtype: int64

#2-3 AutoEDA – Feature별 분석

‘보안과제정보’ ?

```
# 보안과제정보라는 표현이 한 샘플에서 8개 이상은 나오지 않는 것으로 판단
display(train.loc[train.isin(['보안과제정보']).sum(axis=1).ge(8)])
```

	제출년 도	사업 명	사업_부처 명	계속과제여 부	내역사업 명	과제 명	요약문_연구목 표	요약문_연구내 용	요약문_기대효 과	요약문_한글키워 드	요약문_영문키워 드	label
index												

```
# 보안과제정보가 7개 써있는 샘플
secu_7_dup = train.loc[train.isin(['보안과제정보']).sum(axis=1).eq(7)]
# 보안과제정보가 0개 써있는 샘플
secu_0_dup = train.loc[train.isin(['보안과제정보']).sum(axis=1).eq(0)]

secu_7_dup
```

➔ 7개 컬럼 모두 써 있거나 하나도 안 써있거나로 나뉘져 있음
보안과제정보 ➔ 직관적으로 보안상 정보를 나타낼 수 없다는 의미?
➔ 일단 이상값으로 판단

	제출년 도	사업명	사업_부처명	계속과제 여부	내역사업 명	과제명	요약문_연구 목표	요약문_연구 내용	요약문_기대 효과	요약문_한글키 워드	요약문_영문키 워드	label
index												
132	2018	이공학학술연구기반구축 (R&D)	교육부	계속	보안과제 정보	보안과제 정보	보안과제정보	보안과제정보	보안과제정보	보안과제정보	보안과제정보	0
274	2017	바이오.의료기술개발	과학기술정보통신부	신규	보안과제 정보	보안과제 정보	보안과제정보	보안과제정보	보안과제정보	보안과제정보	보안과제정보	0
298	2017	BK21플러스사업(0.5)	교육부	신규	보안과제 정보	보안과제 정보	보안과제정보	보안과제정보	보안과제정보	보안과제정보	보안과제정보	0

#2-3 AutoEDA – 결측치

plot_missing(train)

Stats

Bar Chart

Spectrum

Heat Map

Dendrogram

Missing Statistics

Missing Cells	15169
Missing Cells (%)	0.7%
Missing Columns	5
Missing Rows	3166
Avg Missing Cells per Column	1264.08
Avg Missing Cells per Row	0.09

5개의 컬럼에 대략 2%의 비율로 결측값 존재

```
# 다섯 개 모두 결측값인 데이터
nan_5 = train.loc[train.isna().sum(axis=1).eq(5)]

#nan_5
print("nan_5 개수 : %d" %nan_5.shape[0])

# 결측값이 없는 데이터
nan_0 = train.loc[train.isna().sum(axis=1).eq(0)]
print("nan_0 개수 : %d" %nan_0.shape[0])

# 1~4개의 결측값을 가지고 있는 데이터 개수
print("nan_1,2,3,4 개수 : %d"%(train.shape[0]-nan_5.shape[0]-nan_0.shape[0]))
```

```
nan_5 개수 : 2972
nan_0 개수 : 171138
nan_1,2,3,4 개수 : 194
```


#2-3 AutoEDA – 결측치

다양한 방식으로 결측치 확인

1) 결측치가 1~5개인 데이터 확인

데이터 확인

```
for i in range(1,5):  
    display(train.loc[train.isna().sum(axis=1).eq(i)])  
    print()  
    print('='*100)  
    print()
```

2305	2017	농업기술 경영연구	농촌 진흥청	계속	농업기술경 영연구	농업 R&D 사전 경 제성 분석 연구	농업 R&D 사전경제성 분석WnWn 농업 R&D 사 전경제성 분석 개선 방안 도출	농업 R&D 사전경제성분 석 평가체계 구축WnWn . 타기관 평가시스템 조사 및 진...	NaN	사전경제성, 사 후경제성	ex-ante economic analysis, post economic econo...	0
4814	2017	농업기술 경영연구	농촌 진흥청	신규	농업기술경 영연구	온라인 직거래 유형 별 특성 분석 및 거 래매뉴얼 개발	온라인 직거래 유형별 특 성 분석 및 거래매뉴얼 개발WnWn 농식품의 국 내외 온라인 ...	농․식품의 국내 ․외 온라인 직거래 변화 트렌드 분석WnWn 농...	NaN	전자상거래, 온 라인 쇼핑, 온라 인 구매, 온라인 거래, 농산물 직 거래	Electronic commerce, Online shopping, Online p...	0

15773	2018	지역연구 개발혁신 지원 (R&D)	과학 기술 정보 통신부	계속	연구개발지 원단 육성지 원	전북연구개발 지원단 지원사 업	o 2018년 전북연구개발 지원단은 그간 추진해온 기능별 중점과업의 파급 력을 극대화하...	o 과학기술 정책 기획WnWn - 전북 과학기술위원회 중심의 정책 개발 심의 및 기획 ...	o 지역 R&D 기획 ·조정체제 정립 및 효율적 R&D 투자 를 구현하고 지역 의 혁신역...	NaN	NaN	0
24533	2016	산업전문 인력역량 강화	산업통 상 자 원부	계속	인적자원생 태계조성	2016 산업별 인적자원개발 협업체 활성화 지원사업	산업계가 주도적으로 산 업계의 수요를 체계적으 로 발굴하여 정부 및 인력 양성기관 등에 ...	산업 변화에 신속 대응WnWn * 미지 정 주요 산업분야 에 SC 지정WnWn * ...	산업계 인력 수요 를 반영한 인력양 성 . 공급을 위해 업종별 인력수급 조사 및 교육훈 련...	NaN	NaN	0

3681	2017	개인기초연 구(미래부)	과학 기술 정보 통신부	신규	전략공 모	금융 위기 극복을 위한 인공지능 및 오피니언 마이닝을 이용한금융 의사결정지원시스템 개발	NaN	NaN	NaN	위기 예방,시그널 감 지,인공지능,오피니언 마이닝,준지도 기계학 습,정보 필터링,워드 ...	Crisis prevention,Signal detection,Artificial ...	0
11921	2017	개인기초연 구(미래부)	과학 기술 정보 통신부	신규	전략공 모	형광형 나노자성비드 와 결합된 항체를 이용 한 경락연결망 가시화 기술 개발	NaN	NaN	NaN	단클론 항체,경락 순 환시스템,프리모 순환 시스템,프리모 연결 망,형광 나노자성비 드,	Mono clone antibody,Kyungrak circulatory syste...	0
17696	2017	개인기초연 구(미래부)	과학 기술 정보 통신부	신규	전략공 모	미토콘드리아 안에서 의 펩타이드 자기조립 에 의한 선택적 노화세 포 제거를 통한 생체조 직...	NaN	NaN	NaN	재생의약,항노화신약, 자기조립,노화세포,펩 타이드,	regenerative medicine,ati-aging drug,self-asse...	0

#2-3 AutoEDA – 결측치

다양한 방식으로 결측치 확인

2) 결측값 label 확인

```
nan_5.label.value_counts().to_frame()
```

	label
0	2912
45	9
23	8
14	5
25	4
19	4
16	4
5	3
34	3

3) 결측 값이 세 개인 데이터 중 다섯 개 추출

```
dup_list=train.loc[train.isna().sum(axis=1).eq(3)].index[:5]

#
for idx in dup_list:
    check_data = train.loc[idx][['사업명','내역사업명','과제명']].values
    display(train.query("사업명==@check_data[0] and 내역사업명 ==@check_data[1] and 과제명==@check_data[2]"))
    print("="*100, "\n")
```

	제출 년도	사업명	사업_부 처명	계속 과제 여부	내역 사업 명	과제명	요약문_ 연구목 표	요약문_ 연구내 용	요약문_ 기대효 과	요약문_한글키워드	요약문_영문키워드	label
index												
3681	2017	개인기초 연구(미래 부)	과학기술 정보통신 부	신규	전략공 모	금융 위기 극복을 위한 인공지능 및 오피니언 마이닝을 이용한금융의사 결정지원시스템 개발	NaN	NaN	NaN	위기 예방,시그널 감지,인공지능, 오피니언 마이닝,준지도 기계학 습,정보 필터링,워드 ...	Crisis prevention,Signal detection,Artificial ...	0

	제출 년도	사업명	사업_부 처명	계속 과제 여부	내역 사업 명	과제명	요약문_ 연구목 표	요약문_ 연구내 용	요약문_ 기대효 과	요약문_한글키워드	요약문_영문키워드	label
index												
11921	2017	개인기초 연구(미래 부)	과학기술 정보통신 부	신규	전략공 모	형광형 나노자성비드와 결합된 항체를 이용한 경락연결망 가시 화 기술 개발	NaN	NaN	NaN	단클론 항체,경락 순환시스템,프리 모 순환시스템,프리모 연결망,형광 나노자성비드,	Mono clone antibody,Kyungrak circulatory syste...	0

#2-3 AutoEDA – 데이터 구조와 불용어

‘요약문-연구목표’의 구조

- 조선해양산업 인적자원 관련 정책수립을 위한 기초자료 생성\n\n
- 조선해양산업 선순환적 인력수급 체제 구축 및 안정화\n\n
- 지속가능한 주력산업으로서의 대외 인식 제고\n\n
- 조선해양산업 HRD기구로서의 대표성 확보 및 위상 강화\n\n
- 조선해양산업 인적자원개발 분야 정보 교류 네트워크 구축 및 강화’

-
- 조선해양산업 인적자원 관련 정책수립을 위한 기초자료 생성
 - 조선해양산업 선순환적 인력수급 체제 구축 및 안정화
 - 지속가능한 주력산업으로서의 대외 인식 제고
 - 조선해양산업 HRD기구로서의 대표성 확보 및 위상 강화
 - 조선해양산업 인적자원개발 분야 정보 교류 네트워크 구축 및 강화

① 특수문자(○, -, ?,② ..etc)

○와 같은 자음도 특수문자 형식(?)처럼 들어가 있음

② 영문이 괄호 안에 들어가 있기도 아니기도

③ 줄 바꿈(\n) 표현이 다수 들어가 있는 데이터도 있고 아닌 데이터도 있음

\n이 없는 데이터는 한 문장으로 표현된 데이터? → \n이 없어도 마침표(.)로 문장이 나뉘어져 있음

‘요약문-연구내용’의 구조도 비슷

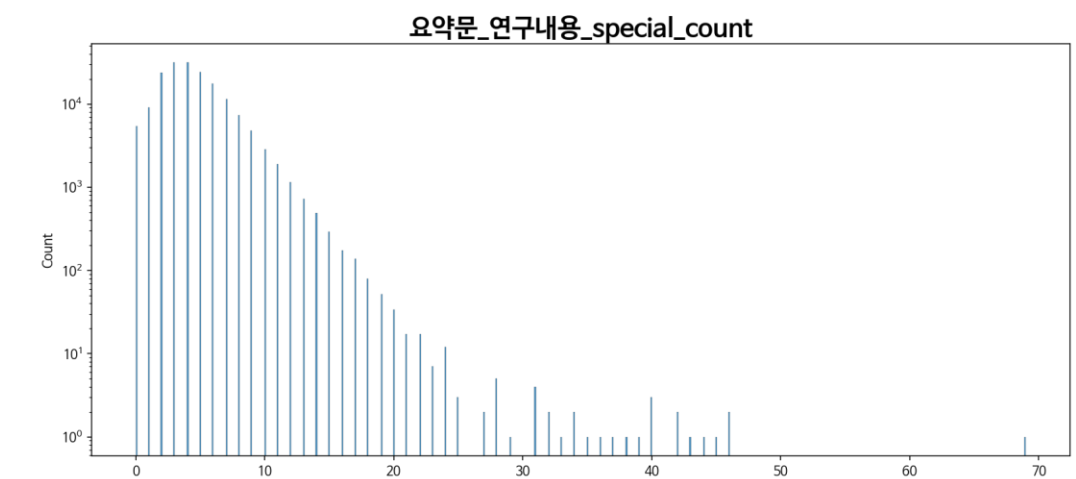
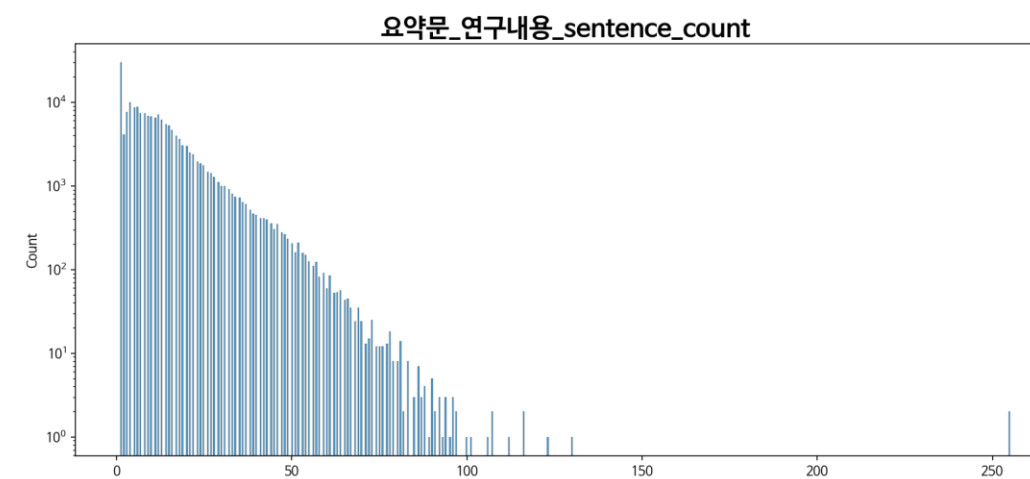
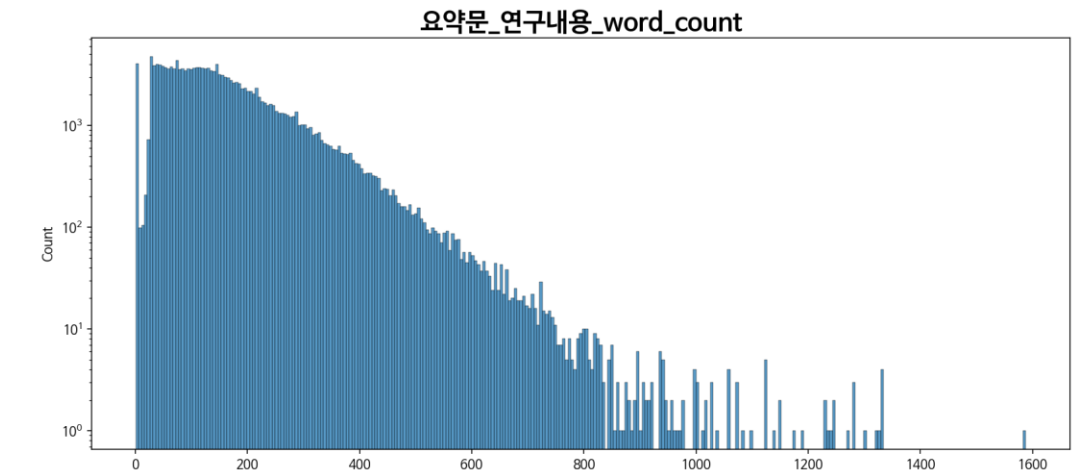
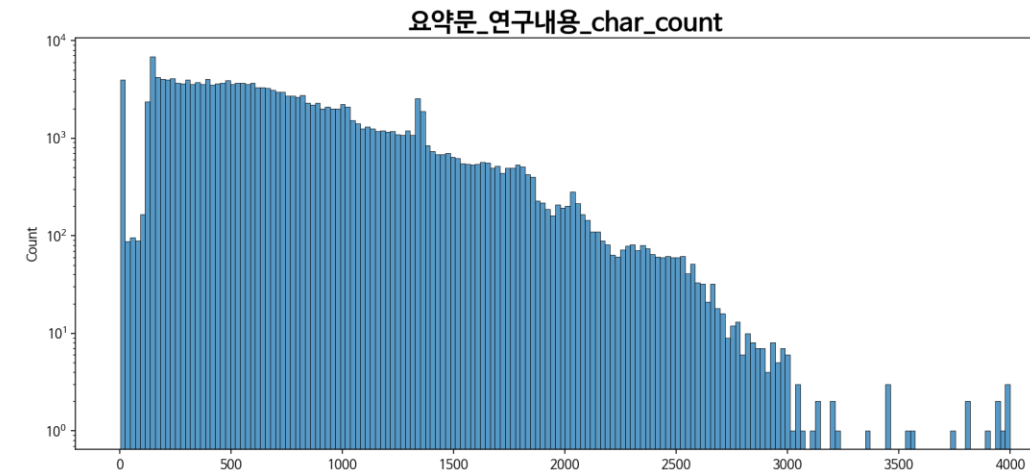
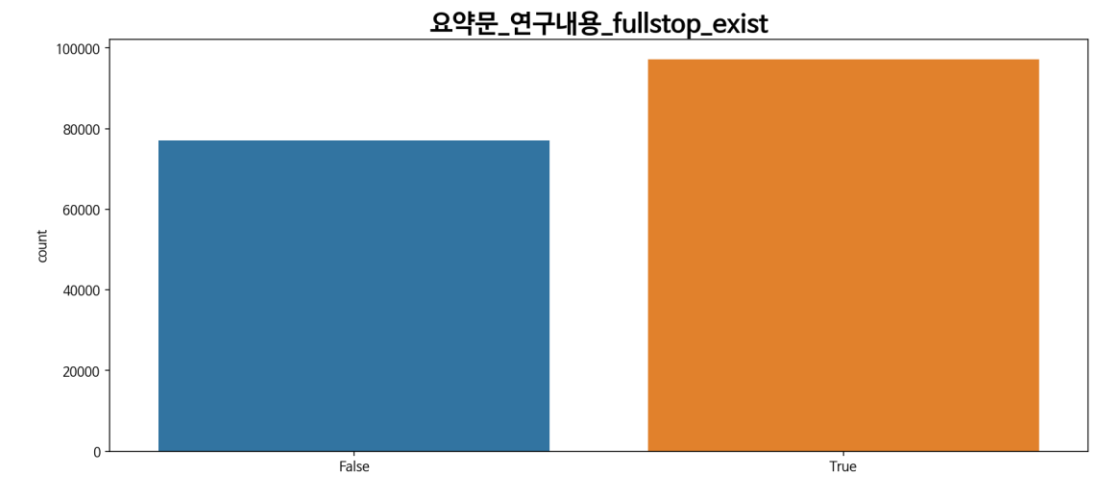
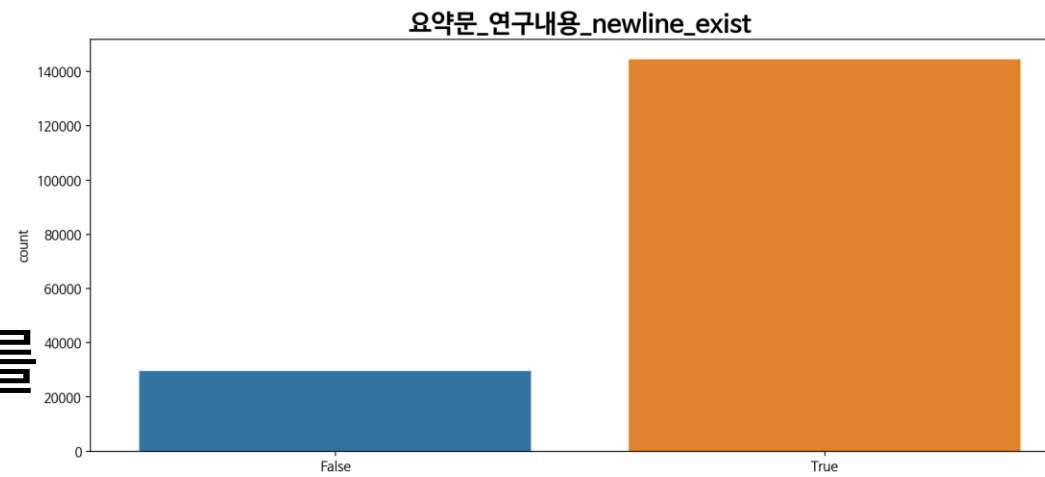
- 제조공정 연구
 - 생물학적 활성과 수율 등을 고려한 최적의 개뿔 추출물 제조공정 확립
 - 액상 및 분말원료 대량생산
- 품질관리 연구
 - 개뿔 추출물의 영양성분 분석 및 유해물질 기준 규격 설정
 - 개뿔 추출물의 지표성분 함량시험 및 밸리데이션
- 기능성 연구
 - 간 손상 동물모델 실험을 통한 개뿔 추출물의 간 손상 억제 효과 확인
 - 지방간 동물모델 실험을 통한 개뿔 추출물의 지방간 억제 효과 확인
 - 외부전문기관 시험을 통한 경쟁원료 대비 객관적 효능 평가

#2-3 AutoEDA – 데이터 구조와 불용어

요약문_연구내용 추가 분석

특수문자 개수 구하는 함수와
\n 기준으로 문장or 문단 수 확인하는 함수를
활용하여 분석 함수 생성 및 시각화

- newline_exist : \n 존재 여부
- fullstop_exist : 마침표(.) 존재 여부
- char_count : 문자 수
- word_count : 단어 수 (띄어쓰기 기준)
- sentence_count : 문장 수
- special_count : 특수문자 개수 확인




Pororo 활용



#03-1 Pororo

#1 카카오 브레인에서 다양한 한글 자연어 처리 작업을 위해 개발한 파이썬 라이브러리

#2 BERT, Transformer등 파이토치로 구현된 최신 NLP 모델을 사용해 30여 가지의 자연어 처리 작업을 수행 가능

 PORORO: Platform Of neuRal mOdelS for natuRal language prOcessing

NOTES

Installation and Usage

Configuration

TEXT CLASSIFICATION

Automated Essay Scoring

Age Suitability Prediction

Natural Language Inference

Paraphrase Identification

Review Scoring

Semantic Textual Similarity

Sentence Embedding

Sentiment Analysis

Zero-shot Topic Classification

SEQUENCE TAGGING

Contextualized Embedding

Dependency Parsing

Fill-in-the-blank

Machine Reading Comprehension

Named Entity Recognition


Part-of-Speech Tagging

Semantic Role Labeling

SEQ2SEQ

Constituency Parsing

Grammatical Error Correction

 » Welcome to PORORO's documentation!

[View page source](#)

Welcome to PORORO's documentation!

Notes

- Installation and Usage
- Configuration

Text Classification

- Automated Essay Scoring
- Age Suitability Prediction
- Natural Language Inference
- Paraphrase Identification
- Review Scoring
- Semantic Textual Similarity
- Sentence Embedding
- Sentiment Analysis
- Zero-shot Topic Classification

Sequence Tagging

- Contextualized Embedding
- Dependency Parsing
- Fill-in-the-blank
- Machine Reading Comprehension
- Named Entity Recognition
- Part-of-Speech Tagging
- Semantic Role Labeling

Seq2Seq

- Constituency Parsing
- Grammatical Error Correction

SEQUENCE TAGGING

Contextualized Embedding

Dependency Parsing

Fill-in-the-blank

Machine Reading Comprehension

Named Entity Recognition

Part-of-Speech Tagging

Semantic Role Labeling

SEQ2SEQ

Constituency Parsing

Grammatical Error Correction

MISC.

Automatic Speech Recognition

Image Captioning

Collocation

Lemmatization

Morphological Inflection

Optical Character Recognition

Speech Synthesis

Tokenization

Word Translation

Word Embedding

Seq2Seq


- Constituency Parsing
- Grammatical Error Correction
- Grapheme-to-Phoneme
- Phoneme-to-Grapheme
- Machine Translation
- Paraphrase Generation
- Question Generation
- Text Summarization
- Word Sense Disambiguation

Misc.

- Automatic Speech Recognition
- Image Captioning
- Collocation
- Lemmatization
- Morphological Inflection
- Optical Character Recognition
- Speech Synthesis
- Tokenization
- Word Translation
- Word Embedding

Indices and tables

- Index
- Module Index
- Search Page

Next 

© Copyright 2021, Kakao Brain Corp.
Built with Sphinx using a theme provided by Read the Docs.

EMWHA
EURON

#03-2 Pororo를 활용한 문장 유사도를 활용한 분류 시도

#1 Pororo 라이브러리의 Sentence Embedding를 사용하여 문장 유사도를 활용

```
sembed = Pororo(task="sentence_embedding", lang="ko")
```

```
def predict_corpus(self, corpus):
    """Text Data들을 Embedding"""
    corpus_embeddings = self._model.encode(corpus, convert_to_tensor=True)
    return corpus_embeddings

def embeddings_to_embeddings(self, embedding, embeddings, cand):
    """Embedding 한 corpus를 비교하여 유사도 추출하는 함수 """
    total_result_list = []
    for embed in embedding :
        cos_scores = util.pytorch_cos_sim(embed, embeddings)[0]
        cos_scores = cos_scores.cpu()
        k = min(len(cos_scores), 10)
        top_results = np.argsort(-cos_scores, range(k))[0:k]
        top_results = top_results.tolist()
        result = list()
        for idx in top_results:
            result.append(
                (idx, cand[idx].strip(), round(cos_scores[idx].item(), 2)))
        total_result_list.append(result)
    return total_result_list
```


#03-2 Pororo를 활용한 문장 유사도를 활용한 분류 시도

#1 test data와 train data의 과제명, 요약문_연구목표, .. 내용에 대한 sentence의 유사도를 비교하여 출력

```
sbjt_embeddings = predict_courpus(sembed, list(train_data['과제명'].astype(str)))
objt_embeddings = predict_courpus(sembed, list(train_data['요약문_연구목표'].astype(str)))
cont_embeddings = predict_courpus(sembed, list(train_data['요약문_연구내용'].astype(str)))
efft_embeddings = predict_courpus(sembed, list(train_data['요약문_기대효과'].astype(str)))
hankey_embeddings = predict_courpus(sembed, list(train_data['변형_한글키워드'].astype(str)))
```

	sbjt_res_list	objt_res_list	cont_res_list	efft_res_list	hankey_res_list	enkey_res_list	mean_res_list
0	[(82984, 소면적작물 농약 직권등록 작물잔류성 시 험 (한경대), 0.86), (...	[(66582, 소면적 및 수출 유망작물의 작물잔류성 시험 (2016년 경북대), 1...	[(15101, 절화작약 축성재 배 병해충 방제기술 개발, 0.84), (99625,...	[(135809, 농약직권등록시 험(작물잔류성) 및 안전사 용․잔류허용기준 설...	[(28729, 소면적작물 농약직 권등록 작물잔류성 시험 (호 서대학교), 1.0), ...	[(28729, 소면적작물 농약직 권등록 작물잔류성 시험 (호 서대학교), 1.0), ...	[(136259, 소면적작물 농약 직권등록 작물잔류성 시험 (안전성평가연구소), 0...
1	[(133104, 분자소재 사업 단, 1.0), (164, 분자과학기 술사업단, 0.9...	[(114358, 친환경 복합기 능 해안항만 구조시스템 창의인재양성 사업팀, 0.79...	[(1870, 멀티 오믹스 융합 기술 기반 혁신 신약 연구 전문인력 양성 사업팀, ...	[(133104, 분자소재 사업 단, 1.0), (20543, 분자과학 기반창의인재양성...	[(133104, 분자소재 사업단, 1.0), (57589, 분자측매 설계 및 응용...	[(133104, 분자소재 사업단, 0.98), (104984, 창의소재 인 력양성 ...	[(133104, 분자소재 사업단, 1.0), (20543, 분자과학기 반 창의인재양성...
2	[(36207, 식물유래 폴리페 놀/탄닌 유도체 기반 다기 능 접착소재 발골 및 소재 활...	[(8333, 새로운 이온저장 및 kinetic이 발현된 수퍼 하이브리드 제조를 위 한...	[(69247, 고접착성 속경화 형 방식 코팅제 기술 개발, 0.82), (14460...	[(125740, Bio based polyurethane적용 내구성 이 향상된 나노...	[(37420, 지속가능한 기능성 코팅 원료로써 페 PET병의 재활용 기술 및 적용...	[(10359, 친환경 난연 연속 패널(PIR) System 기술 개 발, 0.88),...	[(27295, 건축용 패널 및 스 프레이용 친환경 수발포 폴 리우레탄폼 제조기술 개발...
3	[(19517, 포도 비가림시설 의 현황 분석 및 내재해형 표준설계서 개발, 0.97...	[(18812, 풍력에너지 전력 망 적용 기술 연구센터, 0.85), (119488,...	[(18812, 풍력에너지 전력 망 적용 기술 연구센터, 0.84), (29092, ...	[(42011, 모종의 수분스트 레스 모니터링 시스템 고 도화, 0.86), (6813...	[(19517, 포도 비가림시설의 현황 분석 및 내재해형 표준 설계서 개발, 1.0)...	[(19517, 포도 비가림시설의 현황 분석 및 내재해형 표 준설계서 개발, 1.0)...	[(19517, 포도 비가림시설의 현황 분석 및 내재해형 표 준설계서 개발, 0.97...
4	[(89910, 환경친화 용액법 기반 RGO 복합소재의 고 기능 응용 연구, 1.0)...	[(5554, 환경친화 용액법 기반 RGO 복합소재의 고 기능 응용 연구, 1.0),...	[(5554, 환경친화 용액법 기반 RGO 복합소재의 고 기능 응용 연구, 1.0)...	[(5554, 환경친화 용액법 기반 RGO 복합소재의 고 기능 응용 연구, 0.86)...	[(5554, 환경친화 용액법 기 반 RGO 복합소재의 고기능 응용 연구, 0.99)...	[(5554, 환경친화 용액법 기 반 RGO 복합소재의 고기능 응용 연구, 1.0)...	[(5554, 환경친화 용액법 기 반 RGO 복합소재의 고기능 응용 연구, 0.97)...
...
17426	[(65739, 섬유아세포를 이 용한 세포치료제 개발, 1.0), (139591, 섬...	[(329, 신물질 개발 및 적 용을 통한 기능성 화장품 의 개발, 1.0), (349...	[(329, 신물질 개발 및 적 용을 통한 기능성 화장품 의 개발, 1.0), (349...	[(329, 신물질 개발 및 적 용을 통한 기능성 화장품 의 개발, 1.0), (349...	[(139591, 섬유아세포를 이 용한 세포치료제 개발, 1.0), (127800, ...	[(139591, 섬유아세포를 이 용한 세포치료제 개발, 0.9), (130141, ...	[(139591, 섬유아세포를 이 용한 세포치료제 개발, 1.0), (65739, 섬...
17427	[(67622, 모터 권선의 인덕 턴스와 멀티레벨 회로구 조를 이용한 전기자동차 용 일체...	[(329, 신물질 개발 및 적 용을 통한 기능성 화장품 의 개발, 1.0), (349...	[(329, 신물질 개발 및 적 용을 통한 기능성 화장품 의 개발, 1.0), (349...	[(329, 신물질 개발 및 적 용을 통한 기능성 화장품 의 개발, 1.0), (349...	[(136942, 등급대비 30% 이 상 냉방에너지소비효율(EER) 이 향상된 변속도 ...	[(109274, IEC 61850 기반 GIS 현장조작반 반도체형 제어릴레이 Ki...	[(22155, 전기자동차용 인휠 모터 및 후륜샤시들렛폼 부 품 개발, 0.92), (...
17428	[(96114, 신체적 노쇠수준 이 치매위험에 미치는 영 향 연구, 0.84), (10...	[(20842, 농약 노출과 만 성퇴행성 질환의 연관성 규명(1), 0.79), (1...	[(17027, 신경병리 PET 뇌 영상기반 알츠하이머성 치매 조기진단 및 예측 기 술...	[(123989, 노년기 건강에 대한 사회경제적 요인의 영향-복합만성질환을 중 심으로...	[(47451, 인제기관의 조직별 텔로미어 길이와 FOCXO3, APOC3, LMN...	[(60642, 알츠하이머병 뇌에 서 해마 연결성의 지도화를 통한 광유전학적 치료, ...	[(153534, 노쇠의 정적수립 을 위한 과학적 근거 확보, 0.86), (9611...
17429	[(54848, 바이러스성출혈 성패혈증바이러스(VHSV) 의 넘치 어체내 침입 연구, ...	[(110356, 차세대 시퀀싱 을 이용한 넘치의 VHSV 감염과 관련된 coding...	[(148230, 해양의 포식기 생성 원생생물의 탐색 및 진단기법 개발 연구, 0.8...	[(111804, 제브라피쉬를 이용한 수산어류질병 감 염모델 구축, 0.82), (8...	[(139003, VHSV 감염시 생체 방어기전에 작용하는 nxr1 의 기능 연구,...	[(54848, 바이러스성출혈성 패혈증바이러스(VHSV)의 넘치 어체내 침입 연구, ...	[(54848, 바이러스성출혈성 패혈증바이러스(VHSV)의 넘치 어체내 침입 연구, ...
17430	[(38447, 산학협력력 기술 개발 사업, 0.85), (1495, 산학협력력 기술...	[(61761, 산-학 협력 연구 단(석유-가스 생산증진), 1.0), (39021,...	[(61761, 산-학 협력 연구 단(석유-가스 생산증진), 1.0), (13076,...	[(61761, 산-학 협력 연구 단(석유-가스 생산증진), 1.0), (39021,...	[(39021, 산-학 협력 연구단 (석유-가스 생산증진), 1.0), (61761,...	[(39021, 산-학 협력 연구단 (석유-가스 생산증진), 1.0), (61761,...	[(61761, 산-학 협력 연구단 (석유-가스 생산증진), 0.96), (39021...

XLM-Roberta 활용



#04-1 XLM-Roberta

#1 기존의 다국어 자연어 처리 모델: mBERT, XML

#2 어휘 희석 -> 사전학습 모델 성능 저하

어휘 희석 - 데이터가 적은 언어와 데이터가 많은 언어들이 섞여 학습되어 각각의 언어 모델에 비해 성능이 저하되는 현상

#04-1 XLM-Roberta

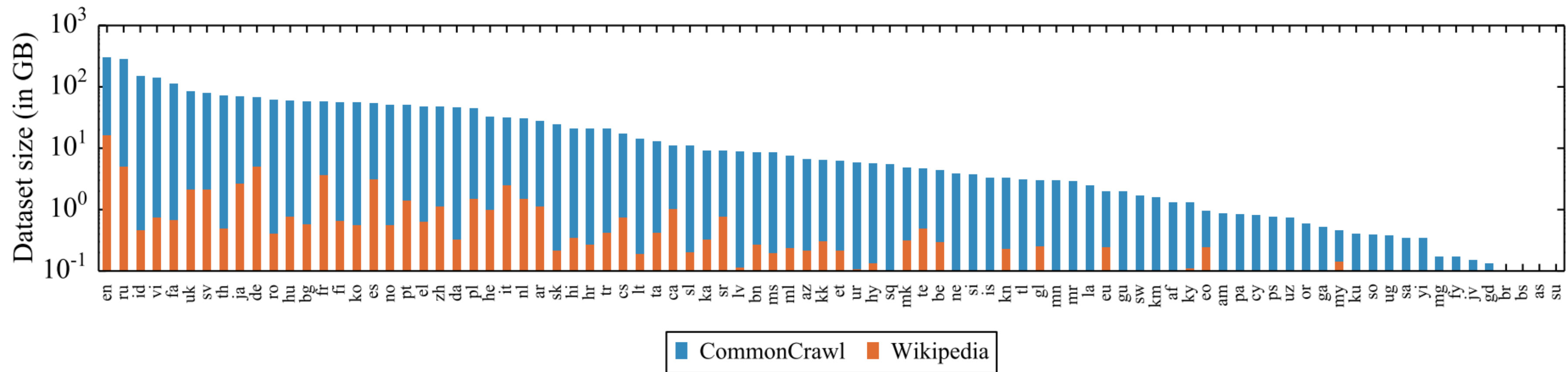
XLM-100

공통점

- language model 기반의 transformer
- Masked Language Model objective 사용
- 100개의 언어의 문자에 대해 처리 가능

향상된 점

- Wiki-100 → 2.5TB의 정제된 CommonCrawl 데이터로 학습
 - 양이 적었던 언어의 충분한 양의 데이터 획득
 - 양이 적은 언어에 대해서도 좋은 성능



#04-1 XLM-Roberta

#1 모델의 학습 루틴이 Roberta 모델과 동일

Model	D	#M	#lg	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
<i>Fine-tune multilingual model on English training set (Cross-lingual Transfer)</i>																			
Lample and Conneau (2019)	Wiki+MT	N	15	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
Huang et al. (2019)	Wiki+MT	N	15	85.1	79.0	79.4	77.8	77.2	77.2	76.3	72.8	73.5	76.4	73.6	76.2	69.4	69.7	66.7	75.4
Devlin et al. (2018)	Wiki	N	102	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
Lample and Conneau (2019)	Wiki	N	100	83.7	76.2	76.6	73.7	72.4	73.0	72.1	68.1	68.4	72.0	68.2	71.5	64.5	58.0	62.4	71.3
Lample and Conneau (2019)	Wiki	1	100	83.2	76.7	77.7	74.0	72.7	74.1	72.7	68.7	68.6	72.9	68.9	72.5	65.6	58.2	62.4	70.7
XLM-R _{Base}	CC	1	100	85.8	79.7	80.7	78.7	77.5	79.6	78.1	74.2	73.8	76.5	74.6	76.7	72.4	66.5	68.3	76.2
XLM-R	CC	1	100	89.1	84.1	85.1	83.9	82.9	84.0	81.2	79.6	79.8	80.8	78.1	80.2	76.9	73.9	73.8	80.9
<i>Translate everything to English and use English-only model (TRANSLATE-TEST)</i>																			
BERT-en	Wiki	1	1	88.8	81.4	82.3	80.1	80.3	80.9	76.2	76.0	75.4	72.0	71.9	75.6	70.0	65.8	65.8	76.2
RoBERTa	Wiki+CC	1	1	<u>91.3</u>	82.9	84.3	81.2	81.7	83.1	78.3	76.8	76.6	74.2	74.1	77.5	70.9	66.7	66.8	77.8
<i>Fine-tune multilingual model on each training set (TRANSLATE-TRAIN)</i>																			
Lample and Conneau (2019)	Wiki	N	100	82.9	77.6	77.9	77.9	77.1	75.7	75.5	72.6	71.2	75.8	73.1	76.2	70.4	66.5	62.4	74.2
<i>Fine-tune multilingual model on all training sets (TRANSLATE-TRAIN-ALL)</i>																			
Lample and Conneau (2019) [†]	Wiki+MT	1	15	85.0	80.8	81.3	80.3	79.1	80.9	78.3	75.6	77.6	78.5	76.0	79.5	72.9	72.8	68.5	77.8
Huang et al. (2019)	Wiki+MT	1	15	85.6	81.1	82.3	80.9	79.5	81.4	79.7	76.8	78.2	77.9	77.1	80.5	73.4	73.8	69.6	78.5
Lample and Conneau (2019)	Wiki	1	100	84.5	80.1	81.3	79.3	78.6	79.4	77.5	75.2	75.6	78.3	75.7	78.3	72.1	69.2	67.7	76.9
XLM-R _{Base}	CC	1	100	85.4	81.4	82.2	80.3	80.4	81.3	79.7	78.6	77.3	79.7	77.9	80.2	76.1	73.1	73.0	79.1
XLM-R	CC	1	100	89.1	85.1	86.6	85.7	85.3	85.9	83.5	83.2	83.1	83.7	81.5	83.7	81.6	78.0	78.1	83.6

#1 SOTA 모델

#04-2 7위 코드

자연어 기반 기후기술분류 AI 경진대회

녹색기술센터 | 자연어 | 환경 | Macro F1

₩ 상금 : 총 600만원

🕒 2021.06.21 ~ 2021.08.16 17:59

+ Google Calendar

👤 966명 📅 마감

대회안내

데이터

코드 공유

토크

리더보드

제출

[private 7위] BERT, XLM-RoBERTa, Logistic, LGBM



hotorch

3. 사용 모델

1, 2 과정을 적용 후 다음과 같은 모델을 만든 후 앙상블 했을 때 제일 좋은 점수를 얻었습니다.

1. xlm-roberta(5fold cv, 모델 수 3(input) * 5(fold) = 15)
2. bert-base-multilingual-cased(5fold cv, 모델 수 3(input) * 5(fold) = 15)
3. Logistic(10fold cv, 모델 수 1(input) * 10(fold) = 10)
4. LightGBM(단일 모델, 모델 수 1)

```
# bert
```

```
from transformers import BertForSequenceClassification, BertConfig, BertTokenizer
```

```
from transformers import XLMRobertaConfig, XLMRobertaTokenizer, XLMRobertaTokenizerFast, XLMRobertaModel, XLMRobertaForSequenceClassification
```

```
from transformers import AdamW, get_linear_schedule_with_warmup, get_cosine_schedule_with_warmup
```

THANK YOU

