KUBIG
DATA SCIENCE & AI

HANSOL DECO S.2 AI Competition

Team | NLP 3팀

18기 이수민, 장원준, 정해원

KUBIG
DATA SCIENCE & AI

CONTENTS

KUBIG
DATA SCIENCE & AI

01. About Contest

01. About Contest

한솔데코는 인공지능의 활용을 통해 건축 핵심 자재들의 품질관리와 하자 판단 과정을 더욱 정교하고 효율적으로 만들어, 자재들의 관리 및 운용의 질을 향상시키는 것을 목표로 하고 있다. 따라서 이번 한솔데코 시즌2 AI 경진대회를 통해 자연어 처리 기반의 QA 시스템을 통해 도배하자와 관련된 깊이 있는 질의응답 처리 능력을 갖춘 AI 모델 개발에 도움을 주고자 한다.

01. About Contest

Model

내진 설계의 종류 좀 알려줘

내진설계에는 내진구조, 제진구조, 면진구조가 있습니다. 내진구조는 건물 구조물이 지진의 힘에 따라 변형되는 것을 방지하기 위해 설계됩니다. 제진구조는 건물의 지반을 고정시키거나 지진의 힘을 줄이기 위해 설계되며, 면진구조는 건물을 지진에 노출되는 방향으로부터 보호합니다.

KUBIG
DATA SCIENCE & AI

02. DATA SET

02. DATA SET - 데이터 특징

Train Data

```
pd.set_option('display.max_rows', None)
pd.set_option('display.max_colwidth', None)
train
```

	id	질문_1	질문_2	category	답변_1	답변_2	답변_3	답변_4	답변_5
0	TRAIN_000	면진장치가 뭐야?	면진장치에 사용되는 주요 기술은 무엇인가요?	건축구조	면진장치란 지반에서 오는 진동 에너지를 흡수하여 건물에 주는 진동을 줄여주는 진동 격리장치입니다.	면진장치란 건물의 지반에서 발생하는 진동 에너지를 흡수하여 건물을 보호하고, 진동을 줄여주는 장치입니다. 주로 지진이나 기타 지반의 진동으로 인한 피해를 방지하기 위해 사용됩니다.	면진장치란 지반으로부터 발생하는 진동 에너지를 흡수하여 건물에 전달되는 진동을 줄여주는 장치를 말합니다. 이를 통해 건물의 안전성과 안정성을 향상시키고, 지진 등의 외부 충격으로부터 보호하는 역할을 합니다. 지진으로 인한 건물의 피해를 최소화하기 위해 주로 사용됩니다.	면진장치는 건물의 지반으로부터 오는 진동 에너지를 흡수하여 건물에 전달되는 진동을 최소화해 주는 진동 격리장치입니다. 이를 통해 건물 내부의 진동을 줄이고 안정성을 유지하는 데 도움을 줍니다.	면진장치는 건물에 오는 지반 진동의 영향을 최대한으로 흡수하여 건물에 전달되는 진동을 줄여주는 장치입니다. 지반으로부터 오는 진동 에너지의 영향을 완화시키기 위해 사용됩니다.

질문_1	질문_2
면진장치가 뭐야?	면진장치에 사용되는 주요 기술은 무엇인가요?

- 같은 내용에 대한 질문 2가지
  - 반말 / 존댓말 질문 존재

02. DATA SET - 데이터 특징

Train Data

```
pd.set_option('display.max_rows', None)
pd.set_option('display.max_colwidth', None)
train
```

	id	질문_1	질문_2	category	답변_1	답변_2	답변_3	답변_4	답변_5
0	TRAIN_000	면진장치가 뭐야?	면진장치에 사용되는 주요 기술은 무엇인가요?	건축구조	면진장치란 지반에서 오는 진동 에너지를 흡수하여 건물에 주는 진동을 줄여주는 진동 격리장치입니다.	면진장치란 건물의 지반에서 발생하는 진동 에너지를 흡수하여 건물을 보호하고, 진동을 줄여주는 장치입니다. 주로 지진이나 기타 지반의 진동으로 인한 피해를 방지하기 위해 사용됩니다.	면진장치란 지반으로부터 발생하는 진동 에너지를 흡수하여 건물에 전달되는 진동을 줄여주는 장치를 말합니다. 이를 통해 건물의 안전성과 안정성을 향상시키고, 지진 등의 외부 충격으로부터 보호하는 역할을 합니다. 지진으로 인한 건물의 피해를 최소화하기 위해 주로 사용됩니다.	면진장치는 건물의 지반으로부터 오는 진동 에너지를 흡수하여 건물에 전달되는 진동을 최소화해 주는 진동 격리장치입니다. 이를 통해 건물 내부의 진동을 줄이고 안정성을 유지하는 데 도움을 줍니다.	면진장치는 건물에 오는 지반 진동의 영향을 최대한으로 흡수하여 건물에 전달되는 진동을 줄여주는 장치입니다. 지반으로부터 오는 진동 에너지의 영향을 완화시키기 위해 사용됩니다.

- Category – 질문 내용의 분야
  - 마감재, 인테리어, 시공, 마감하자, 건축구조, 기타, 타 마감하자

02. DATA SET - 데이터 특징

Train Data

```
pd.set_option('display.max_rows', None)
pd.set_option('display.max_colwidth', None)
train
```

	id	질문_1	질문_2	category	답변_1	답변_2	답변_3	답변_4	답변_5
0	TRAIN_000	면진장치가 뭐야?	면진장치에 사용되는 주요 기술은 무엇인가요?	건축구조	면진장치란 지반에서 오는 진동 에너지를 흡수하여 건물에 주는 진동을 줄여주는 진동 격리장치입니다.	면진장치란 건물의 지반에서 발생하는 진동 에너지를 흡수하여 건물을 보호하고, 진동을 줄여주는 장치입니다. 주로 지진이나 기타 지반의 진동으로 인한 피해를 방지하기 위해 사용됩니다.	면진장치란 지반으로부터 발생하는 진동 에너지를 흡수하여 건물에 전달되는 진동을 줄여주는 장치를 말합니다. 이를 통해 건물의 안전성과 안정성을 향상시키고, 지진 등의 외부 충격으로부터 보호하는 역할을 합니다. 지진으로 인한 건물의 피해를 최소화하기 위해 주로 사용됩니다.	면진장치는 건물의 지반으로부터 오는 진동 에너지를 흡수하여 건물에 전달되는 진동을 최소화해 주는 진동 격리장치입니다. 이를 통해 건물 내부의 진동을 줄이고 안정성을 유지하는 데 도움을 줍니다.	면진장치는 건물에 오는 지반 진동의 영향을 최대한으로 흡수하여 건물에 전달되는 진동을 줄여주는 장치입니다. 지반으로부터 오는 진동 에너지의 영향을 완화시키기 위해 사용됩니다.

- 답변 – 각 질문에 대한 답변
  - 5가지 다른 답변
  - 길이/깊이가 다른 답변

답변_2	답변_3
면진장치란 건물의 지반에서 발생하는 진동 에너지를 흡수하여 건물을 보호하고, 진동을 줄여주는 장치입니다. 주로 지진이나 기타 지반의 진동으로 인한 피해를 방지하기 위해 사용됩니다.	면진장치란 지반으로부터 발생하는 진동 에너지를 흡수하여 건물에 전달되는 진동을 줄여주는 장치를 말합니다. 이를 통해 건물의 안전성과 안정성을 향상시키고, 지진 등의 외부 충격으로부터 보호하는 역할을 합니다. 지진으로 인한 건물의 피해를 최소화하기 위해 주로 사용됩니다.

02. DATA SET - Train/Test 전처리

Train_cleaned Data

```
def clean(text):
  cleaned = re.sub('[^a-zA-Zㄱ-힣()₩d.,?!₩s₩-]', ' ', text)
  return cleaned

train['질문_1']=train['질문_1'].apply(clean)
train['질문_2']=train['질문_2'].apply(clean)
train.head()
```

- 알파벳, 한글, 괄호, 문장부호, 공백 제외 글자 제거

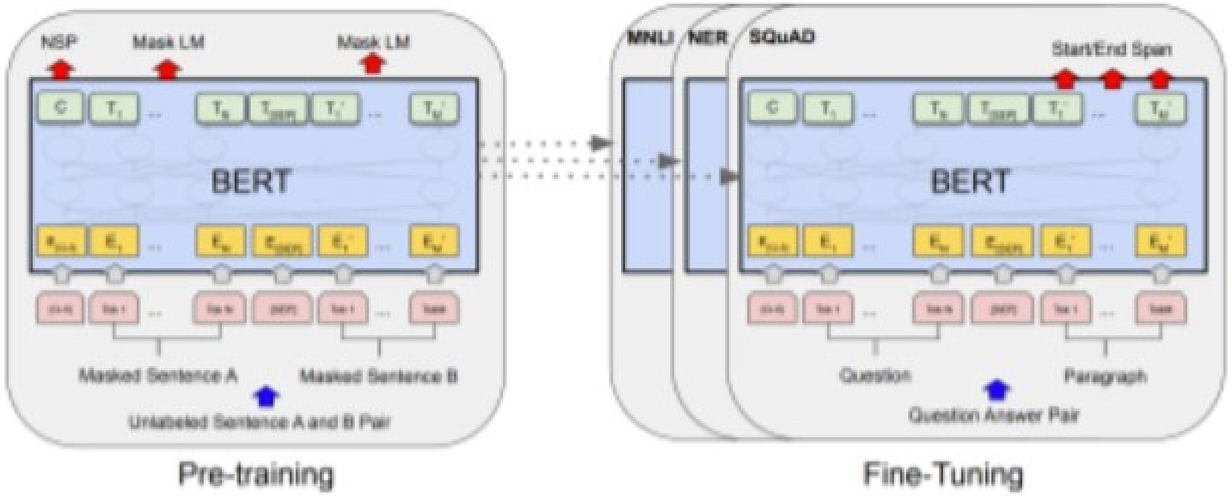
Test Data - 질문만 존재

test

	id	질문
0	TEST_000	방청 페인트의 종류에는 어떤 것들이 있는지 알고 계신가요? 또한, 원목사이딩을 사용...
1	TEST_001	도배지에 녹은 자국이 발생하는 주된 원인과 그 해결 방법은 무엇인가요?
2	TEST_002	큐블럭의 단점을 알려주세요. 또한, 압출법 단열판을 사용하는 것의 장점은 무엇인가요?
3	TEST_003	철골구조를 사용하는 고층 건물에서, 단열 효과를 높이기 위한 시공 방법은 무엇이 있...
4	TEST_004	도배지의 완전한 건조를 위해 몇 주 동안 기다려야 하나요?

- 또한/그리고 두 단어로 앞과 뒤 문장이 구별되는 경우 존재
- Test Data의 문장들을 분할하여 답변 생성 정확도 상승 예상

02. DATA SET - Category 예측



< 최종 모델 RoBERTa-large>

KLUE/Roberta-Large 모델 활용 (Val Accuracy 0.9845)

질문의 Category를 분류/예측 후 답변 생성

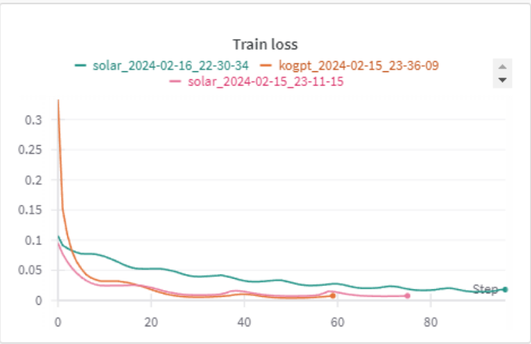
KUBIG
DATA SCIENCE & AI

03. Strategy & Model

03. Project Setting

W&B

tmux




```
Team 3/
│
├── train.py - main script to start training
├── inference.py - make submission with trained models
└── modules/ - functions and classes required to operate the model
    ├── dataloader.py
    ├── trainer.py
    └── utils.py
```

03. Trial & Error

Dataloader, Trainer 등 직접 구현

```python
def train_preprocessing(CFG):
    qa = QATemplate(CFG)
    train_tokenizer = CFG["TRAIN"]["TOKENIZER"]
    tokenizer = get_tokenizer(train_tokenizer)
    data
    colum
    colum
    forma
    atten
    for _
        f
```

```python
class BaiscTrainer:
    def __init__(self, CFG, model, train_loader, valid_loader):
        self.CFG = CFG
        self.model = model
        self.train_loader = train_loader
        self.valid_loader = valid_loader
        self.device = CFG["DEVICE"]
        self.epochs = CFG["TRAIN"]["EPOCHS"]
        self.es_patient = CFG["TRAIN"]["EARLY_STOPPING"]
        self.gradient_accumulation_steps = CFG["TRAIN"]["ACCUMUL_STEPS"]
        self.optimizer = self.get_optimizer(self.CFG)
        self.scheduler = self.get_scheduler()

    def get_optimizer(self):
        select_optimizer = self.CFG["TRAIN"]["OPTIMIZER"]
        learning_rate = self.CFG["TRAIN"]["LEARNING_RATE"]
        if select_optimizer.lower() == "adamw":
            optimizer = AdamW(self.model.parameters(), lr=learning_rate)
        return optimizer

    def get_scheduler(self):
        select_scheduler = self.CFG["TRAIN"]["SCHEDULER"]
        select_scheduler_cfg = select_scheduler["CFG"]
        if select_scheduler["NAME"].lower() == "cosineannealinglr":
            scheduler = CosineAnnealingLR(
                self.optimizer, T_max=select_scheduler_cfg["TMAX"]
            )
        return scheduler
```

Hugging Face

Transformers Library

- class transformers.**AutoModelForCausalLM**
- class transformers.**Trainer**
- class transformers.**BitsAndBytesConfig**
- class peft.**LoraConfig**

LLM 모델 학습에 필요한 기능 사용 가능

메모리 효율적인 학습 지원

```python
encoding = tokenizer(input_text, return_tensors="pt", padding="max_length",
                     max_length=CFG["TRAIN"]["MAX_SEQ_LEN"],truncation=True,add_special_tokens=False,)
```

Batch 단위로 같은 Data의 길이를 위해 Padding을 넣음

→ padding="max_length" 설정 시, max_length 값 설정 필요(중요)

03. Model Selection

A6000 48GB

7B : 7 x 10^9 x 2 byte = 14GB

학습 시엔 2~3배 여유 RAM 필요 (40GB↑)

Batch Size와 학습 시간 고려… 개선 방법이 필요

→

1. 모델의 크기를 줄이는 16bit → 8bit 모델 양자화
2. 업데이트 행렬의 크기를 줄여 RAM, 시간 줄이는 LoRA

Open Ko-LLM LeaderBoard

skt/kogpt2-base-v2

beomi/OPEN-SOLAR-KO-10.7B

LDCC/LDCC-SOLAR-10.7B

Instruction Tuning

Inference Test… → 효과적인 Instruction 형태는?

테스트 중…

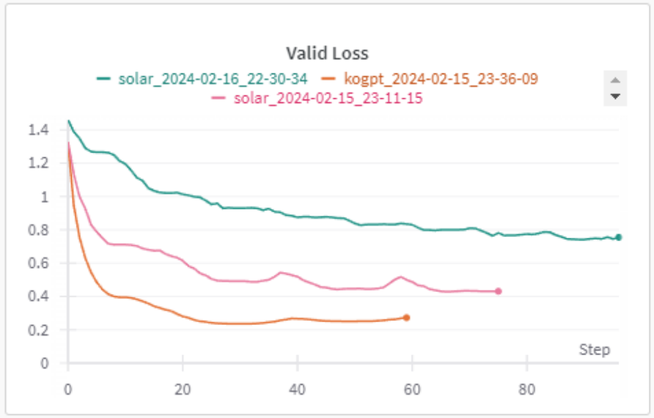
```
### 명령 : 아래 질문에 대해 답변을 단답식으로 하시오.
Q: <question>
A: <answer>
```

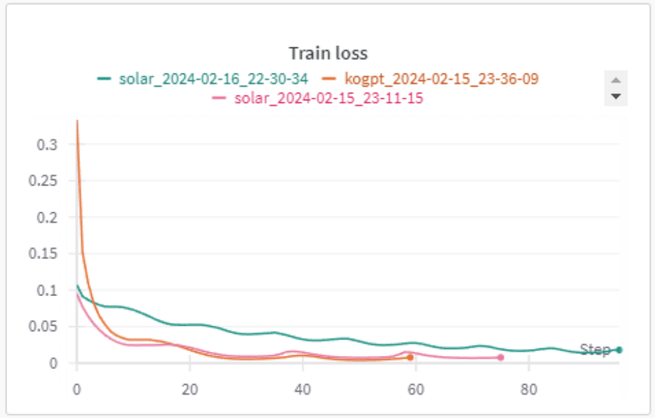
KUBIG
DATA SCIENCE & AI

04. Result

04. Result

Model	Public Score	비고
skt/kogpt2-base-v2	0.571	Baseline
skt/kogpt2-base-v2	0.598	Hyperparameter Tuning
Beomi/OPEN-SOLAR-KO-10.7B	0.479	
skt/kogpt2-base-v2	0.692	Test Set split





Valid loss와 align 되는 결과

Instruction Tuning의 중요성 파악 + 추가적인 시도 방향성 확립

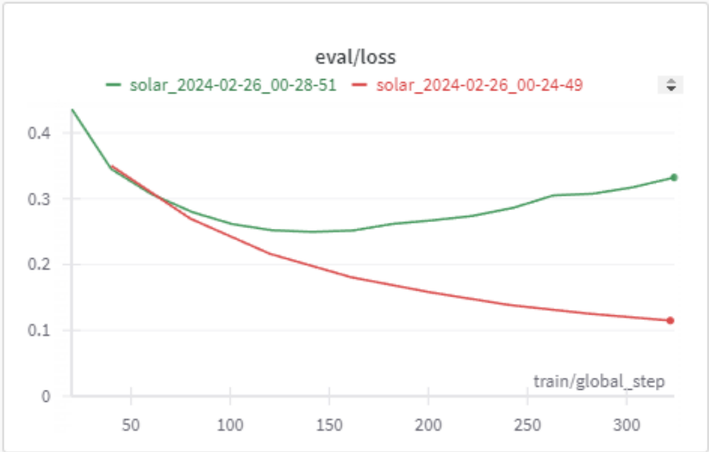
KUBIG
DATA SCIENCE & AI

05. Future Works

05. Back to the Data

1. 질문_1과 질문_2가 배치되는 경우

~의 장점은? ↔ ~의 단점은?



질문_2가 불확실하더라도

데이터의 양이 많은 것이 Good

2. Train Data의 답변에 특수문자가 포함된 경우

수정된 답변 ~

GPT가 대답할 때 내놓는 특별한 형식

Test Data의 정답은

완전한 문장 형태의 답변일까?

3. 데이터의 양이 부족

Papago API

2월 29일 종료, 하루에 10000개의 요청

Gemini

05. Whatever…

1.	2.	3.
RAG	Data Augmentation	Base Model Tuning
ChromaDB	Back Translation	Solar
Faiss		LLAMA
		MIXTRAL

KUBIG
DATA SCIENCE & AI

Thank You