방문자들을 위한
박물관 챗봇 구현

2/26 최종발표

21.01.04 - 21.02.26

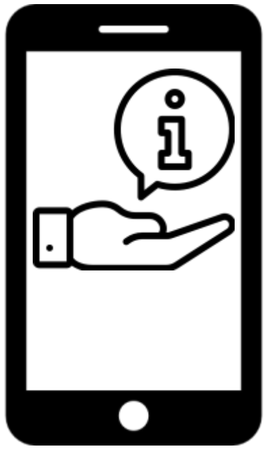
김주연

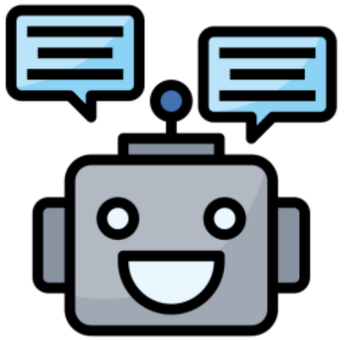
CONTENTS

/

01. 개요

기획 배경

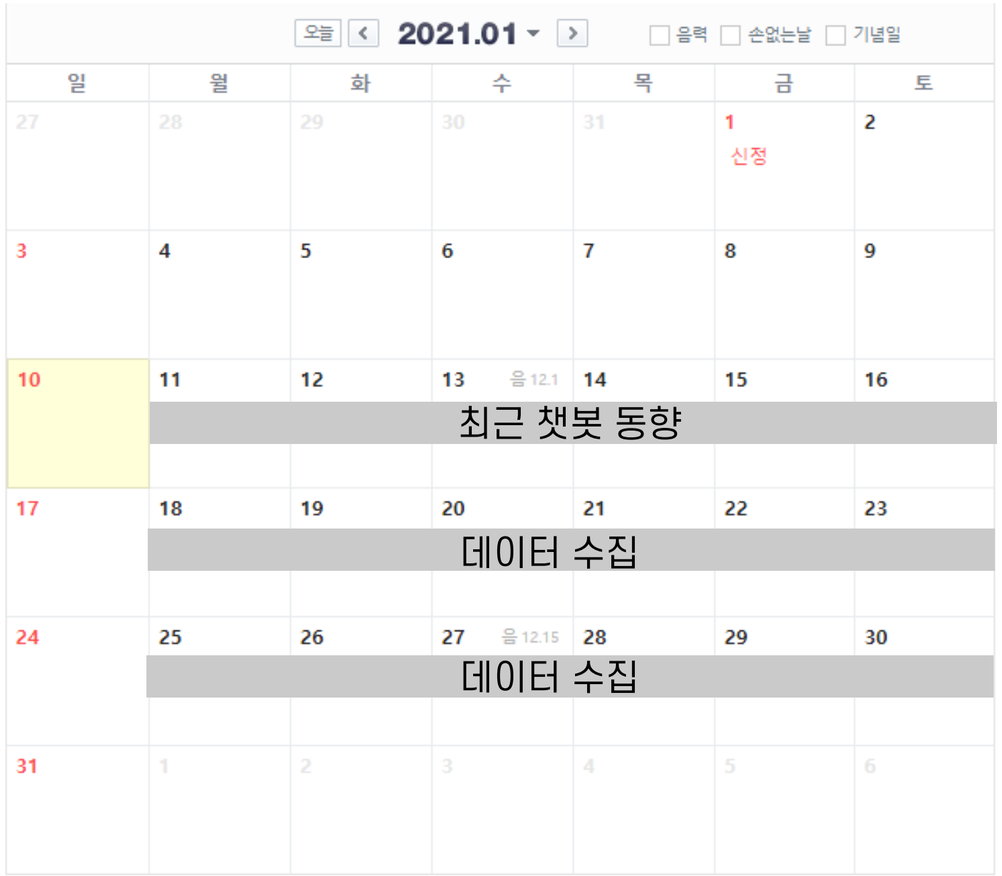


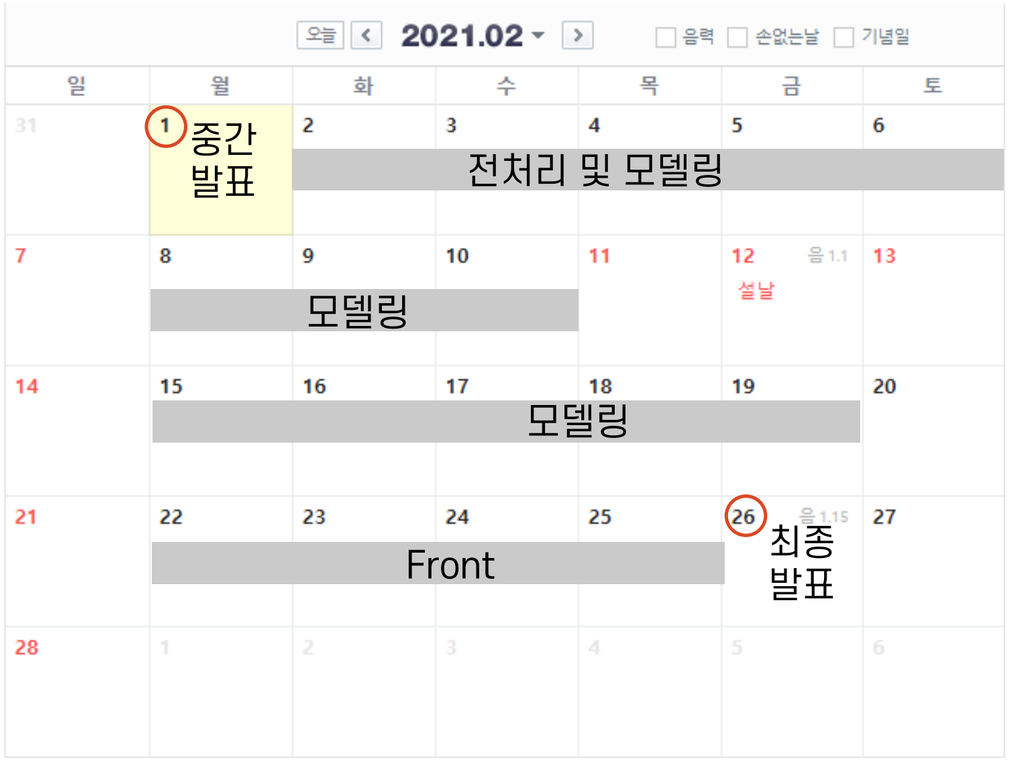


방문자들이 박물관 이용 시 조금 더 편리하게 이용할 수 있도록
정보 제공의 목적을 가진 챗봇 구현

01. 개요

진행 과정


오늘
2021.01
음력
손없는날
기념일
일 월 화 수 목 금 토
27 28 29 30 31 1 신정 2
3 4 5 6 7 8 9
10 11 12 13 음12.1 14 15 16
최근 챗봇 동향
17 18 19 20 21 22 23
데이터 수집
24 25 26 27 음12.15 28 29 30
데이터 수집
31 1 2 3 4 5 6


오늘
2021.02
음력
손없는날
기념일
일 월 화 수 목 금 토
31 1 중간 발표 2 3 4 5 6
전처리 및 모델링
7 8 9 10 11 12 음1.1 설날 13
모델링
14 15 16 17 18 19 20
모델링
21 22 23 24 25 26 음1.15 최종 발표 27
Front
28 1 2 3 4 5 6

02. 데이터

도자기 모델	
출처	직접 제작
참고	민족의 문화유산과 업적을 정리·집대성한 한국민족문화대백과사전 / 나무위키 / e뮤지엄 / 지식iN
형식	Q & A 형식의 CSV
개수	질문 개수 : 약 6,000개 (label 900개)

<내용>

1. 도자기의 기본과 개념
   - 도자기는 뭘로 만드나요?
   - 도자기를 만드는 방법을 알려줄래?

2. 국립 광주박물관 도자실의 도자기
   - 강진에서 고려청자가 많이 발견되는 이유는?
   - 순백자를 대표하는 도자기를 알려주세요.

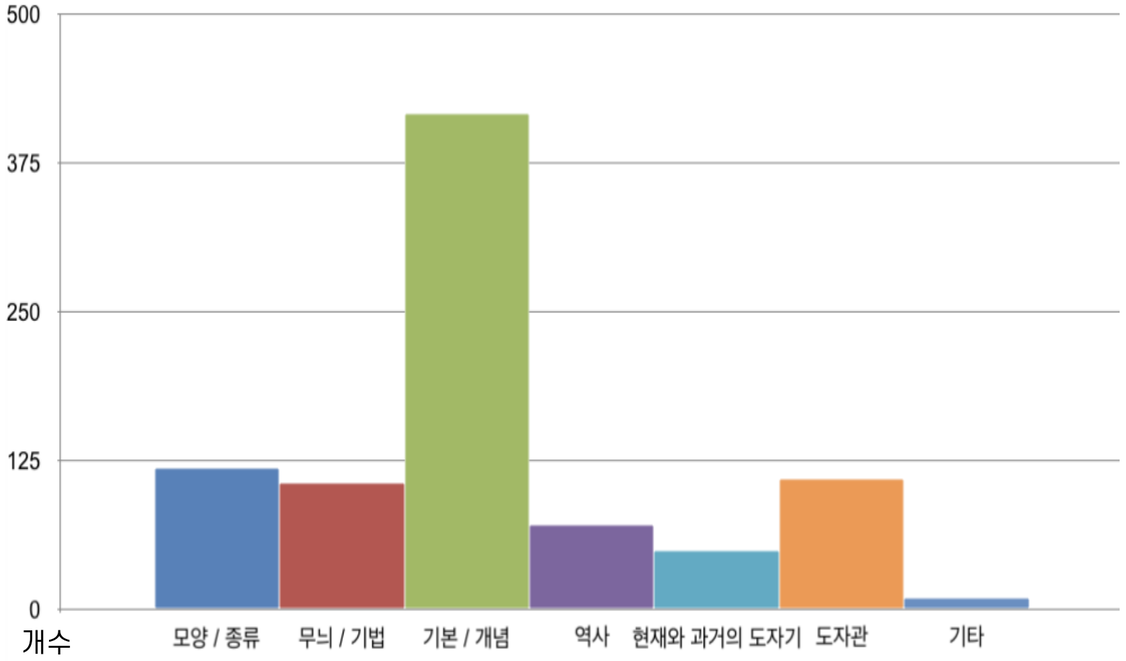
3. 도자기의 무늬와 종류
   - 귀갑무늬가 뭐야?
   - 질그릇은 어떻게 생겼어?

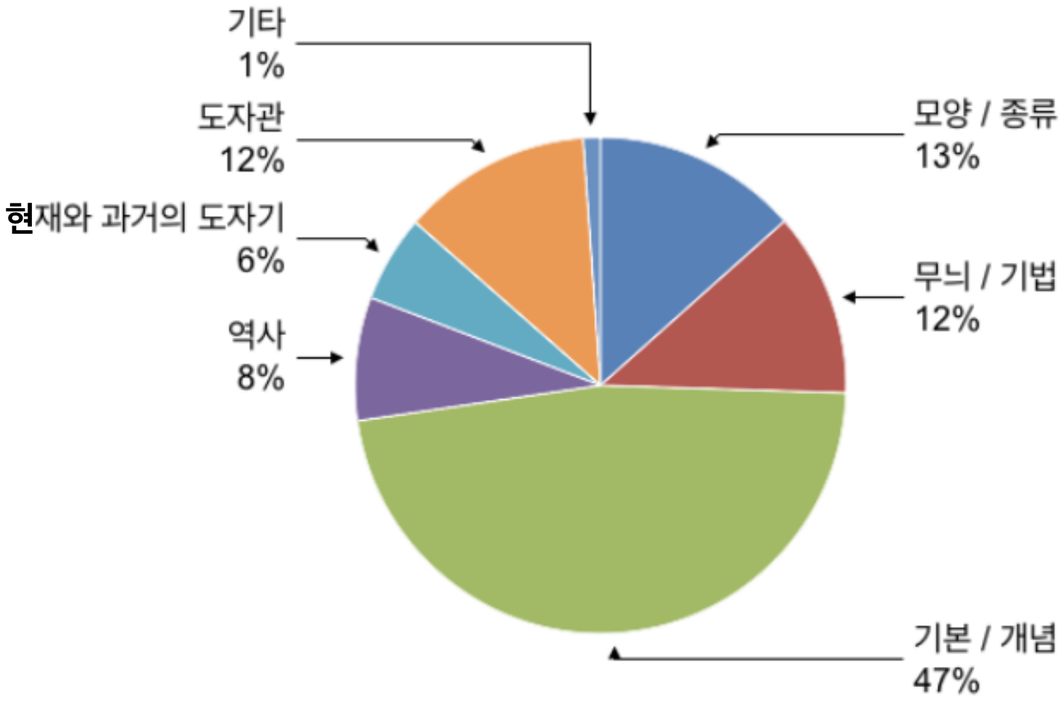
4. 도자기의 역사
   - 고려시대 초기에는 어떤 도자기가 유행했어?
   - 조선시대에 일본군이 우리나라 도자기 장인들을 잡아갔다는 것이 사실인가요?

5. 현재와 과거의 도자기
   - 우리나라에서 가장 큰 도자기는?
   - 미래의 도자기 산업은 어떻게 발전될까요?

02. 데이터

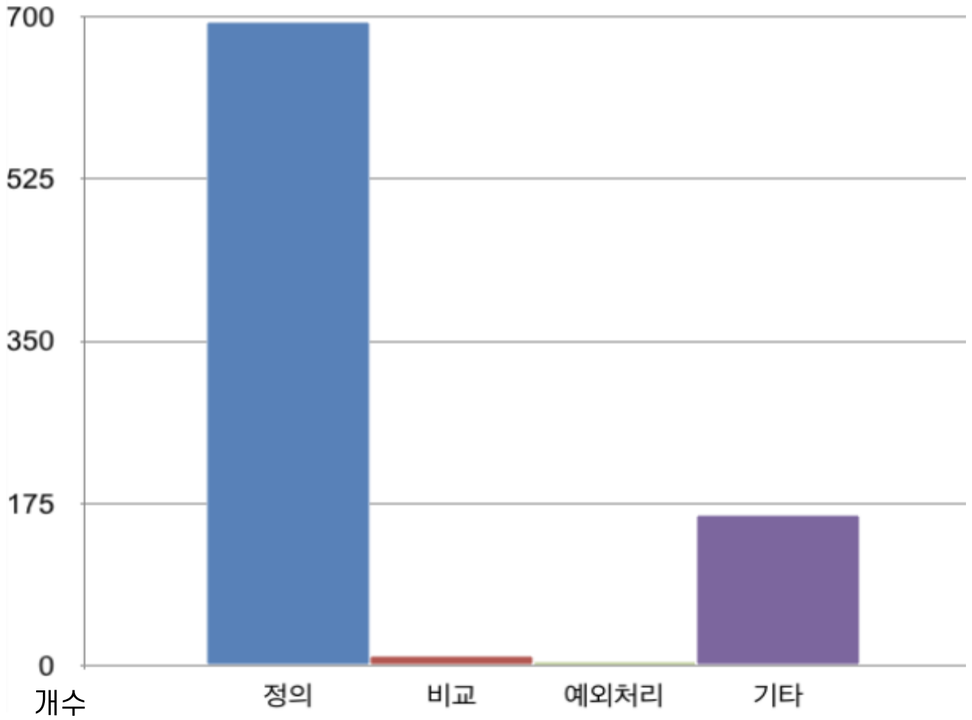
데이터 분석 : 질문의 내용에 따른 분류

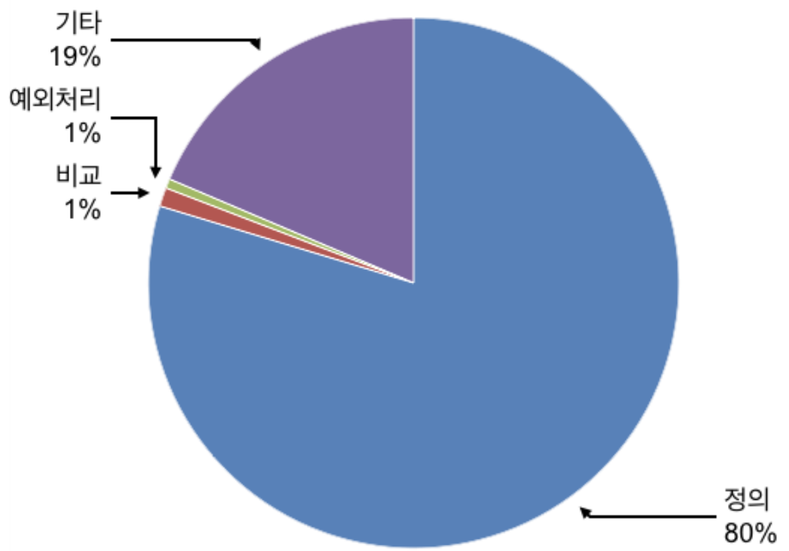




02. 데이터

데이터 분석 : 질문의 종류에 따른 분류





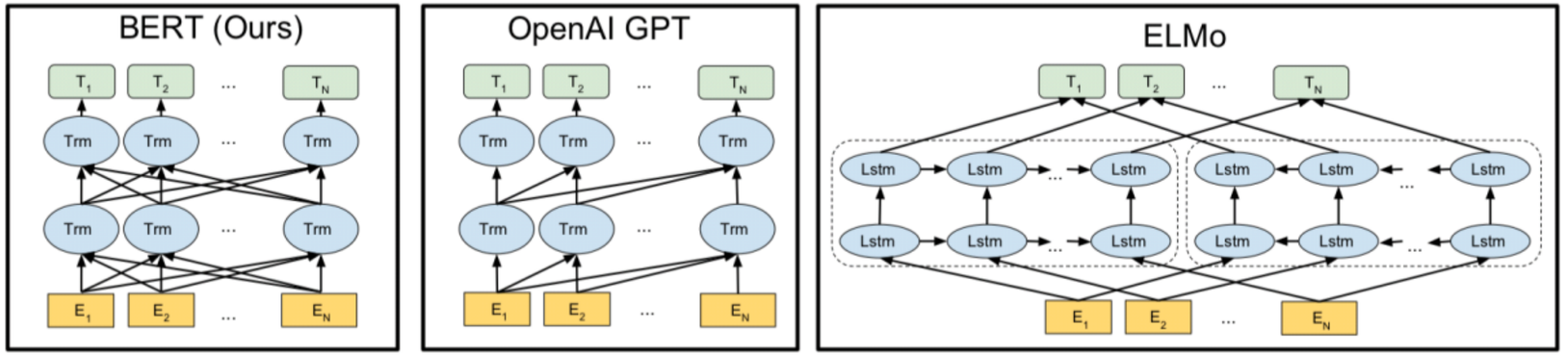
02. 데이터

Wordcloud

03. 모델

BERT

- transformer 구조를 활용
- 기본적으로 대용량 unlabeled data로 모델을 미리 학습시킨 후, 특정 task를 가지고 있는 labeled data로 transfer learning을 하는 모델
- BERT 이전의 pre-trained 모델은 대용량 unlabeled data를 통해 model을 학습하고 이를 토대로 뒤쪽에 특정 task를 처리하는 network를 붙이는 방식으로 진행(ELMo, openAI GPT..) -> 하지만 이런 방식들은 shallow bidirectional 또는 unidirectional



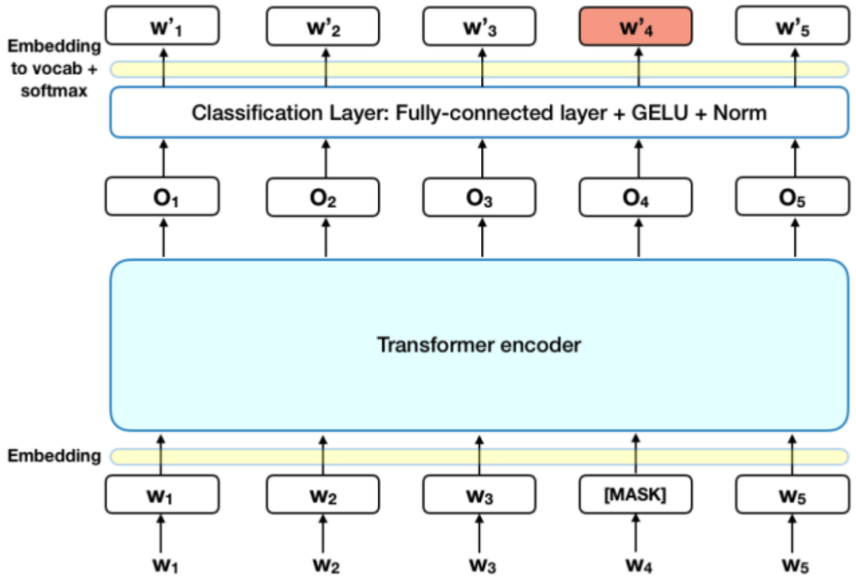
- 특정 task를 처리하기 위해 새로운 network를 붙일 필요 없이, bert model 자체의 fine-tuning을 통해 해당 task의 SOTA 달성

03. 모델

BERT

1. Masked LM

- MLM 단어중의 일부를 [mask] token으로 바꾸어 준다. 바꾸어주는 비율은 15%
- 이를 통하여 LM의 left-to-right (혹은 r2l)을 통하여 문장 전체를 predict하는 방법론과는 달리, [MASK] token 만을 predict하는 pre-training task를 수행

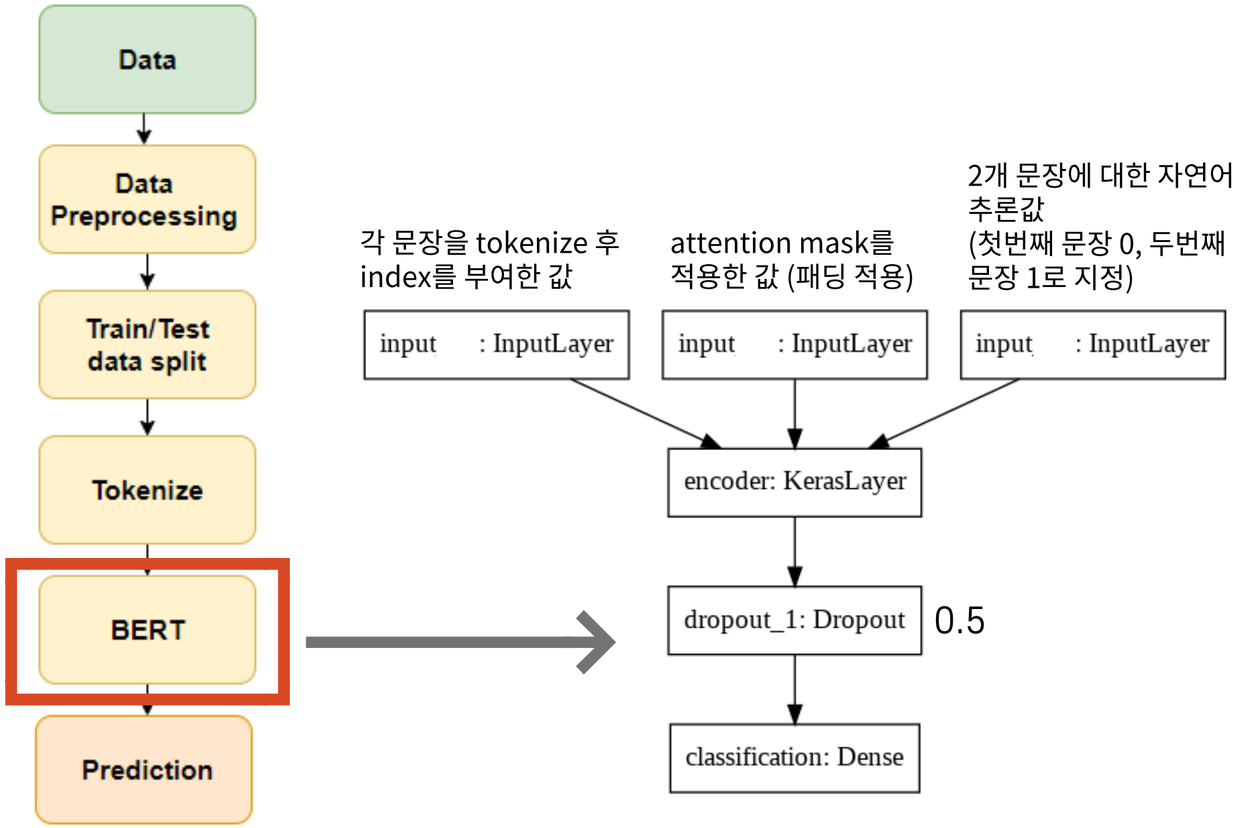


2. NSP(Next Sentence prediction)

- 이 pre-training task 수행하는 이유는, 여러 중요한 NLP task중에 QA나 Natural Language Inference(NLI)와 같이 두 문장 사이의 관계를 이해하는 것이 중요한 것이기 때문
- 그래서 BERT에서는 corpus에서 두 문장을 이어 붙여 이것이 원래의 corpus에서 바로 이어 붙여져 있던 문장인지를 맞추는 binarized next sentence prediction task를 수행
- 50% : sentence A, B가 실제 next sentence,
  50% : sentence A, B가 corpus에서 random으로 뽑힌(관계가 없는) 문장으로 구성
- pre-training이 완료되면, 이 task는 97~98%의 accuracy를 달성

03. 모델

Flowchart

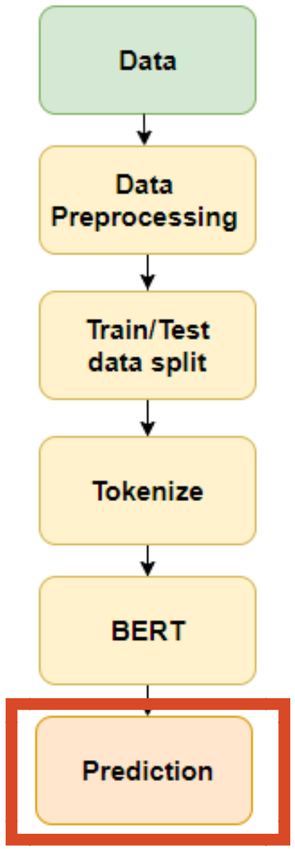


1. Data Preprocessing
 - data 증가

2. Tokenize
- bert의 word piece tokenizer를 사용
 huggingface 라이브러리의 encode_plus 기능 사용
 (특정문장을 버트에 필요한 입력형태로 변환 / 문장을 최대 길이에 맞게 패딩 / 결괏값은 딕셔너리로 출력)

3.Bert
- Pretrained model을 사용하여 학습

03. 모델



Binary classification

- 두 문장이 같다면 1
  다르다면 0

- activation function = sigmoid
- loss = MSE
- learning rate = 3e-5
- optimizer = Adam

1. pretrained model 사용 안한 경우

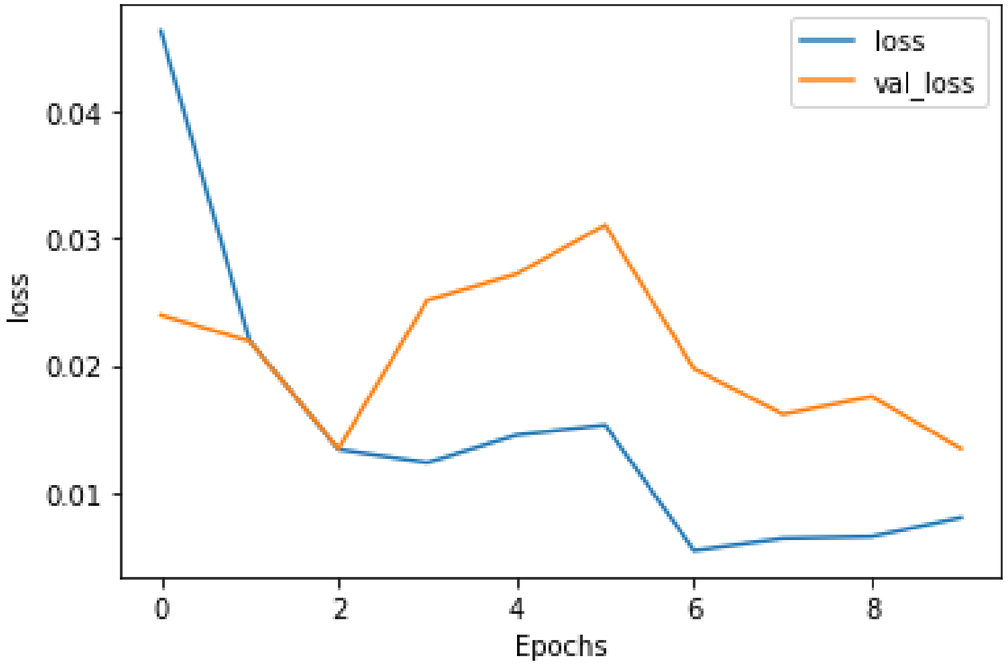
accuracy = 0.76
loss = 0.54

2. pretrained model 사용 한 경우

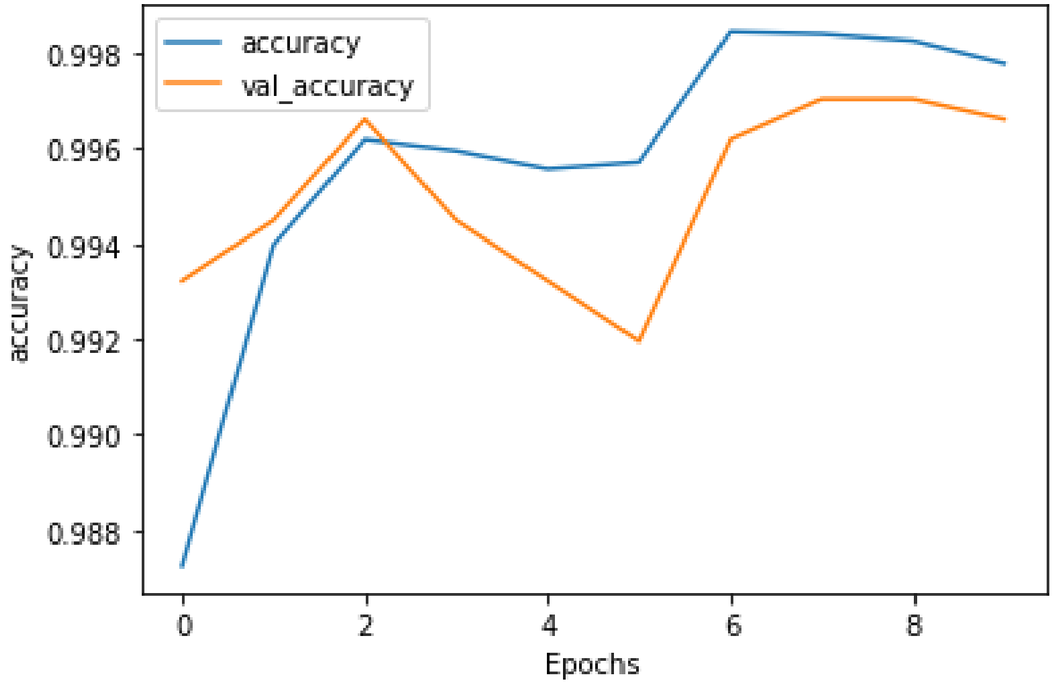
accuracy = 0.99
loss = 0.02

03. 모델

result : loss & acc



loss = 0.019

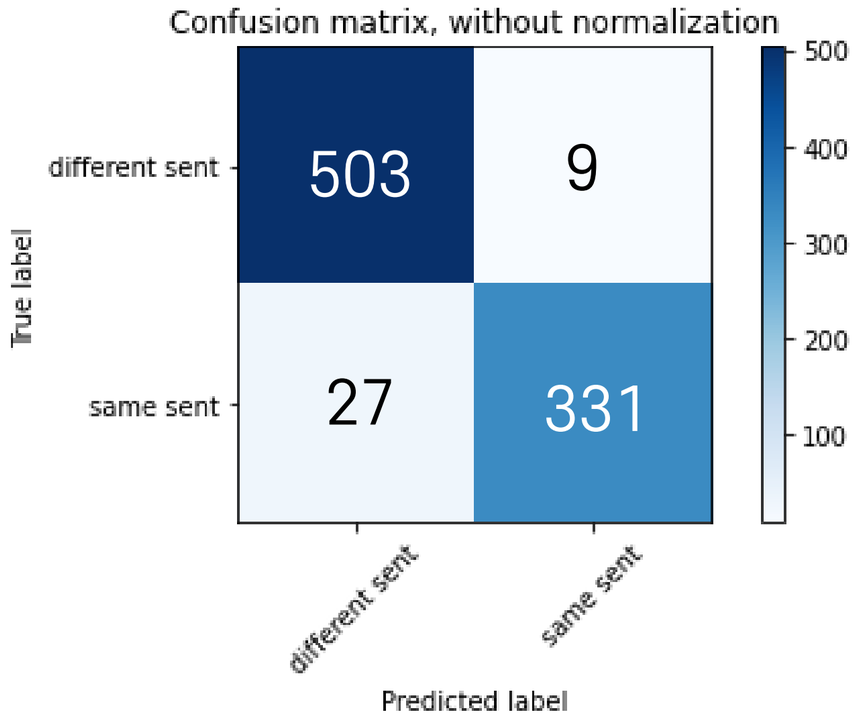


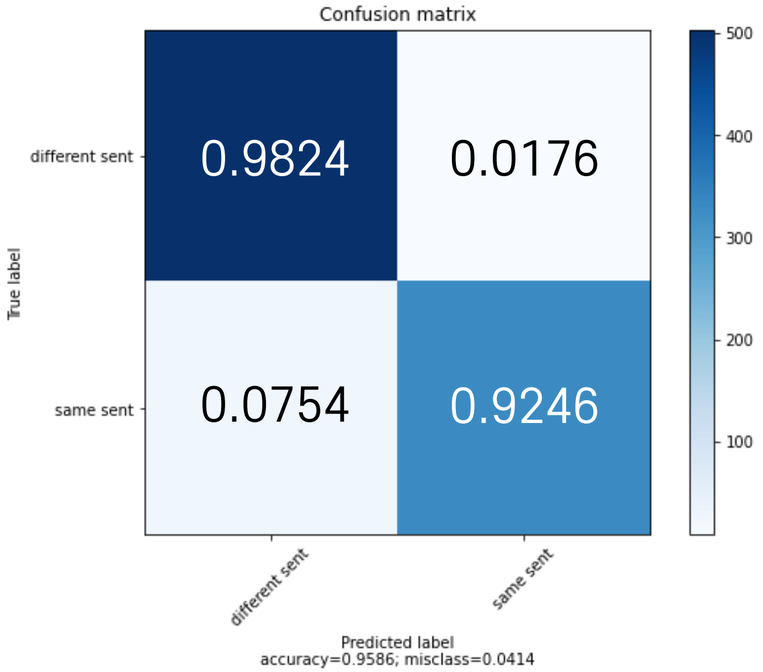
accuracy = 0.9949

03. 모델링

test

- 870쌍으로 구성 (새로 추가된 질문 + train 시 사용했던 질문)
- loss: 0.2582 / accuracy: 0.9586





04. 시연


도자기봇
국립광주박물관의 도자기들을 소개합니다!
previous
next
<강진의 청자>
전라남도 강진군은 고려 시대 청자 생산의 중심지로서 188개소에 달하는 청자 가마터가 확인되어 사적 제 68호로 지정되었습니다. 강진의 용운리 가마터에 대한 발굴은 1980년부터 1982년까지 국립중앙박물관에 의해 진행되었습니다. 그 결과 10세기 후반에서 12세기에 음각 및 양각 청자는 물론 상감청자에 이르는 고려 최고 전성기의 청자를 제작한 가마였던 것으로 확인되었습니다. 굴뚝과 천장을 제외한 나머지 부분이 비교적 양호하게 남아 있던 용운리 청자 가마는 현재 국립광주박물관 옥외전시장에 이전, 복원 되어 있습니다. 2001년에서 2002년까지 국립광주박물관은 삼흥리 청자 가마터 5기와 토기 가마터 9기를 발굴하였습니다. 이때 보물 제 1023호 상약국이 새겨진 청자합과 동일한 글자가 세겨진 청자합 조각이 출토되었습니다.
1. 청자 음각 용무늬 매병
명칭 청자 음각 용무늬 매병
다른명칭 청자 음각 용문 매병(靑磁 陰刻 龍文 梅甁)
국적/시대 한국-고려
출토지 전라남도-강진군
분류 식 - 음식기 - 음식 - 병
도자봇 : 안녕! 도자기에 관해 궁금한 것이 있다면 무엇이든지 나에게 물어봐!
Send

05. 마무리

데이터 부족

inference 속도

Ko-bert model

감사합니다