

# AI 기반 관광지 개선 솔루션





## Team Leader

- 프로젝트 기획 총괄
- [Lim Heejin](#)
- [dg961108@naver.com](mailto:dg961108@naver.com)
- <https://github.com/heejvely>



## Team Member

- Data 구축 총괄
- [Choi Yunjin](#)
- [cyunjin@gmail.com](mailto:cyunjin@gmail.com)
- <https://github.com/ete-llorona>



## Team Member

- BERT Fine-tuning 총괄
- [Lee Minchan](#)
- [leemc9955@naver.com](mailto:leemc9955@naver.com)
- <https://github.com/Leemc95>



## Team Member

- 웹 페이지 구축 총괄
- [Han A-Leum](#)
- [hal0576@naver.com](mailto:hal0576@naver.com)
- <https://github.com/zena-H>



## Team Member

- 데이터 시각화 총괄
- [Park Kibeom](#)
- [ssw4110@gmail.com](mailto:ssw4110@gmail.com)
- <https://github.com/KIBEOMP>

Project git hub:

<https://github.com/CAKD3-Intent-Classification>

## 분석 환경

- Python
- Google\_colab

## 모델링

- Tensorflow
- KorBERT(Etri)

## 대시보드

- Django
- Figma
- QGIS

# INDEX

---

1. 기획 의도
2. 모델 구축 및 평가
3. 대시보드



# 기획 의도

---



**고사 위기 제주 외국인 관광시장 '트레블 버블'로 도약 준비**

[위드코로나] 위드코로나로 여행 기대감 '쑥'..활기 띠는 관광업계

**경남도, '트레블 버블' 선제 대응...해외마케팅 시동 건다**

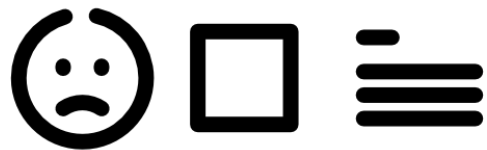
경기도, 트레블버블 대비 '외국인 관광객 유치' 간담회

고사 위기 제주 외국인 관광시장 '트레블 버블'로 도약 준비

[위드코로나] 위드코로나 여행 기대감 '쑥'..활기 띠는 관광업계  
**\*트레블버블(비격리안전권역):**

코로나19 상황에서 두 국가 이상의 방역 우수 지역이  
경남도, '트레블 버블' 선제 대응.. 해외마케팅 시동 건다  
**서로 자유로운 여행을 허용하는 것**

경기도, 트레블버블 대비 '외국인 관광객 유치' 간담회

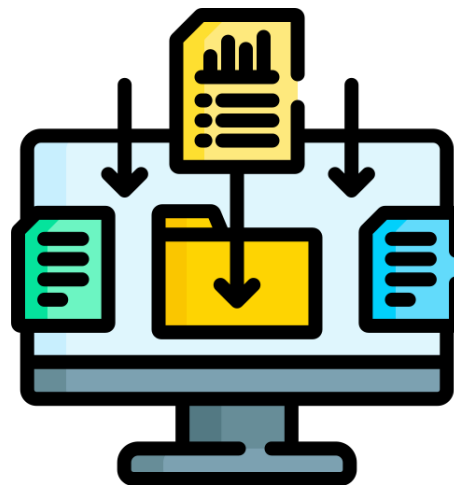


- 서비스 대상: 국내 관광 시설 관계자
- 서비스 기획 목적: 위드코로나 시대에 대비하여 기존 관광 시설의 재정비에 필요한 정보를 제공






관광객들의 실제 경험이 담긴  
온라인 리뷰를 수집하여 분석 및 솔루션 제공



이용자들의 경험과 의견이 구체적으로 담겨 있어  
관광지에 대한 많은 정보 파악 가능

## 기획 의도





김방한

지역 가이드 · 리뷰 31개 · 사진 70장

★★★★★ 3년 전

국립중앙박물관 에서 옛날 문화가 잘 보전 되었다.

 1





신혜숙

리뷰 2개

★★★★★ 2년 전

자세한 설명과 함께 역사를 배울수 있고 외국인들도 자세히 우리나라에 대해 역사를 잘들을수 있어 좋았습니다.  
경희루도 인터넷 예약해서 내부를 볼수있어서 너무 좋았습니다.

 2




HONGSHU JIN

리뷰 163개 · 사진 221장

★★★★★ 3년 전

경치가 정말 좋은 곳입니다

 1

- 리뷰와 별점 불일치로 인한 리뷰 신뢰성 하락
- 관광지 개선점 파악의 어려움

Ex) 긍정 리뷰지만 부정적 별점 부여

★★★★★

## 키워드 리뷰 정식 출시

이제 별점 대신 키워드 리뷰로 가게의 다양한 강점을 한눈에 확인해 보세요!

 뷰가 좋아요



리뷰와 별점 불일치로 인한  
N사의 평가 방식 변경  
[리뷰 중심의 평가]



자연어 딥러닝 모델을 통해

리뷰 문장의 긍/부정성을 가려내고

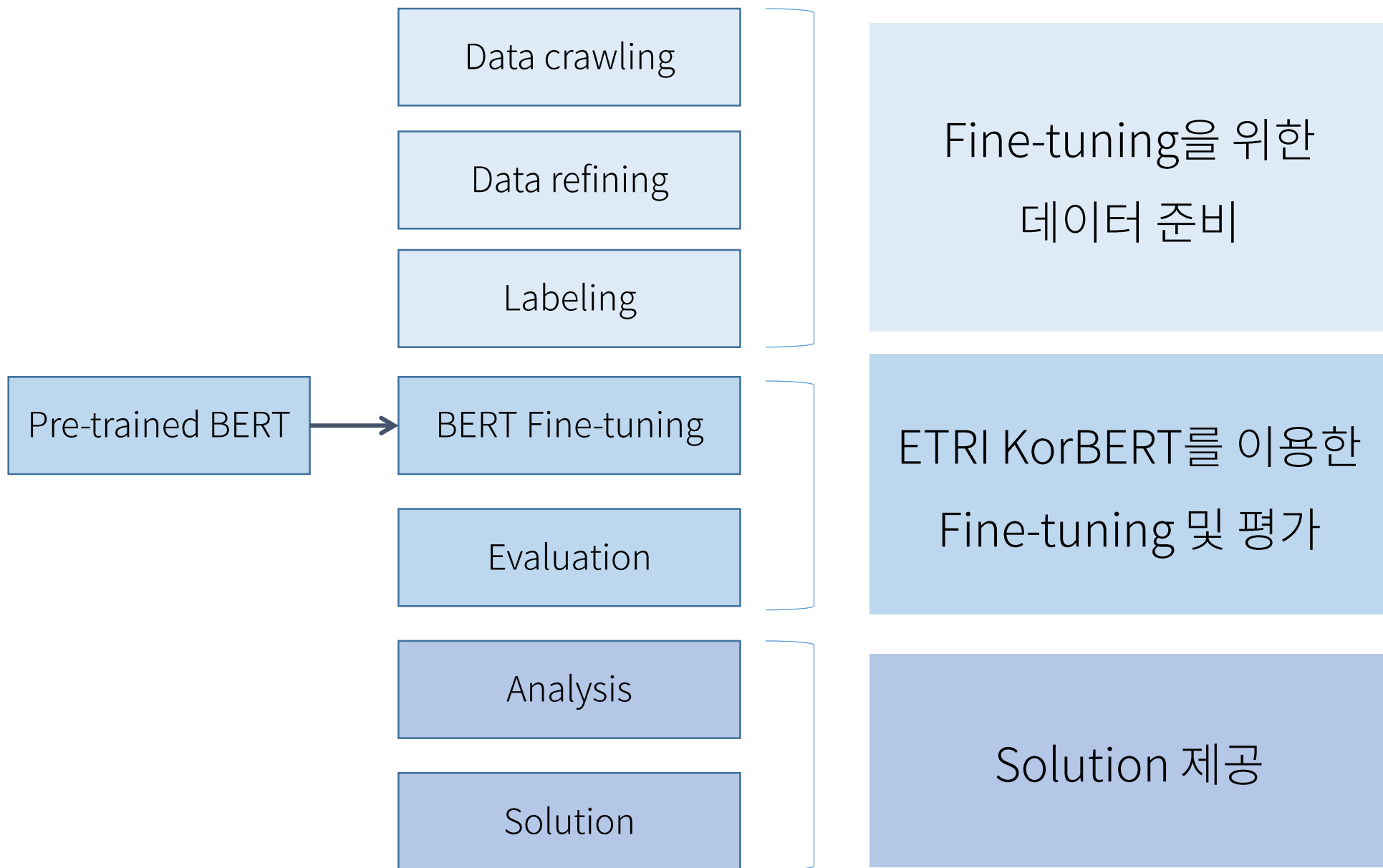
이를 기반으로 관광지 개선 솔루션을 제시

# 모델 구축 및 평가

---



## Modeling process



# BERT model

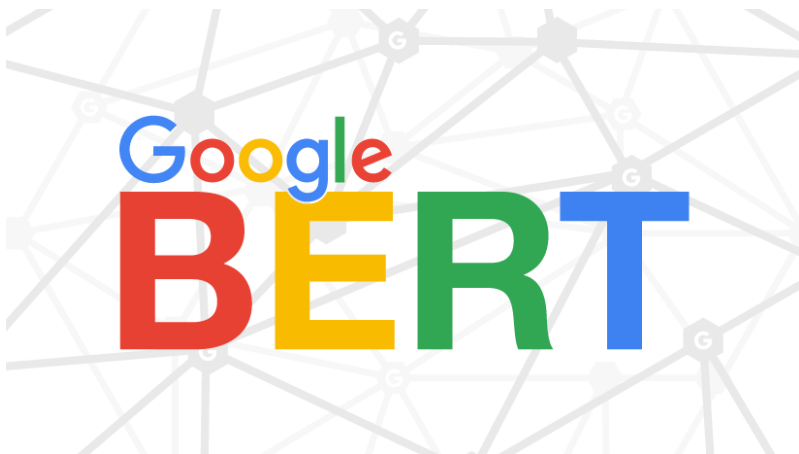
BERT: Bidirectional Encoder Representations from Transformers (2018, Google AI Language)



(25억단어)



Book Corpus  
(8억단어)



- 33억 개의 단어를 학습시킨 언어 모델
- transformer encoder 부분 사용
- Masked Language Modeling(MLM),  
Next sentence prediction을 통한 사전 학습 진행
- Feature-based approach:  
특정 task를 수행하는 network에 feature로 제공
- Fine-tuning approach:  
pre-trained model에 추가적인 task 학습

## Fine-Tuning tasks

**의도 분류**

**질의 응답**

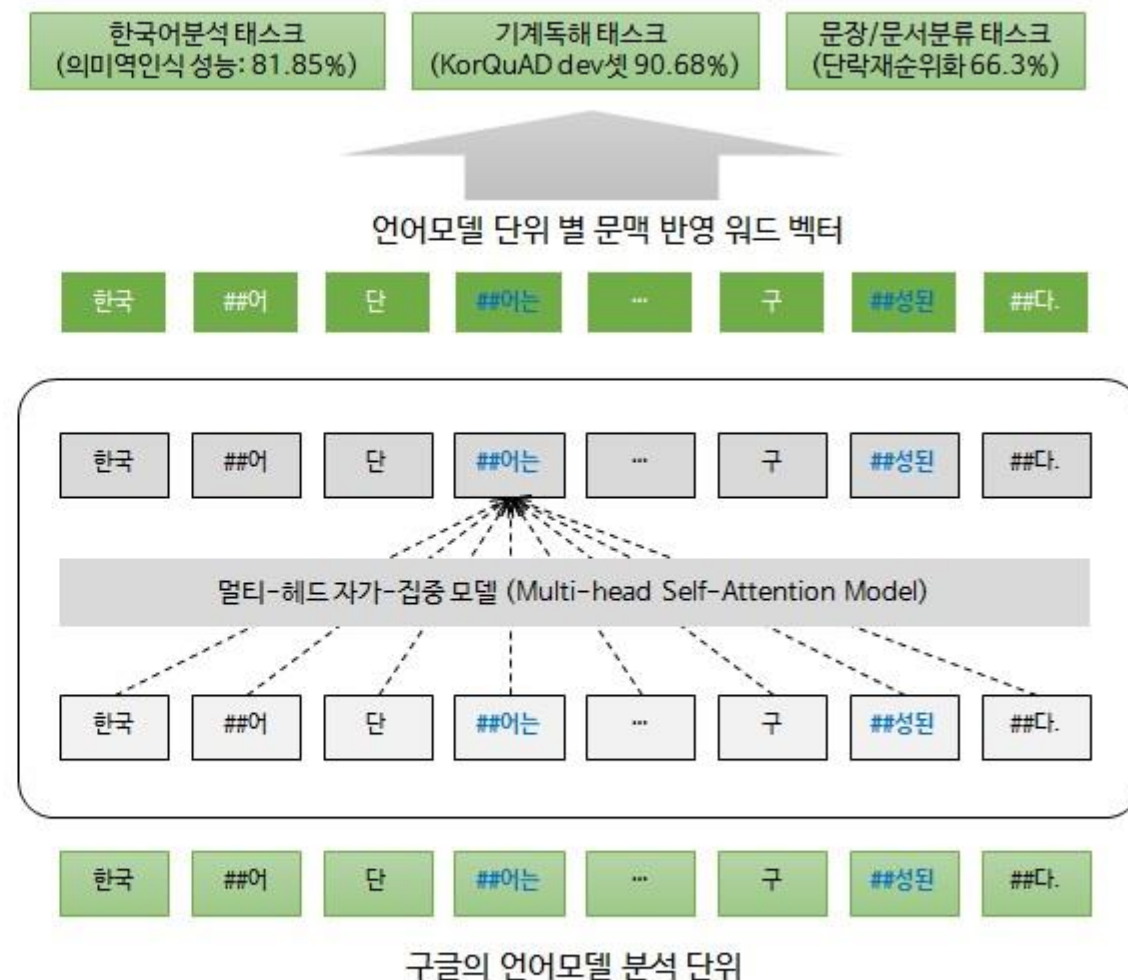
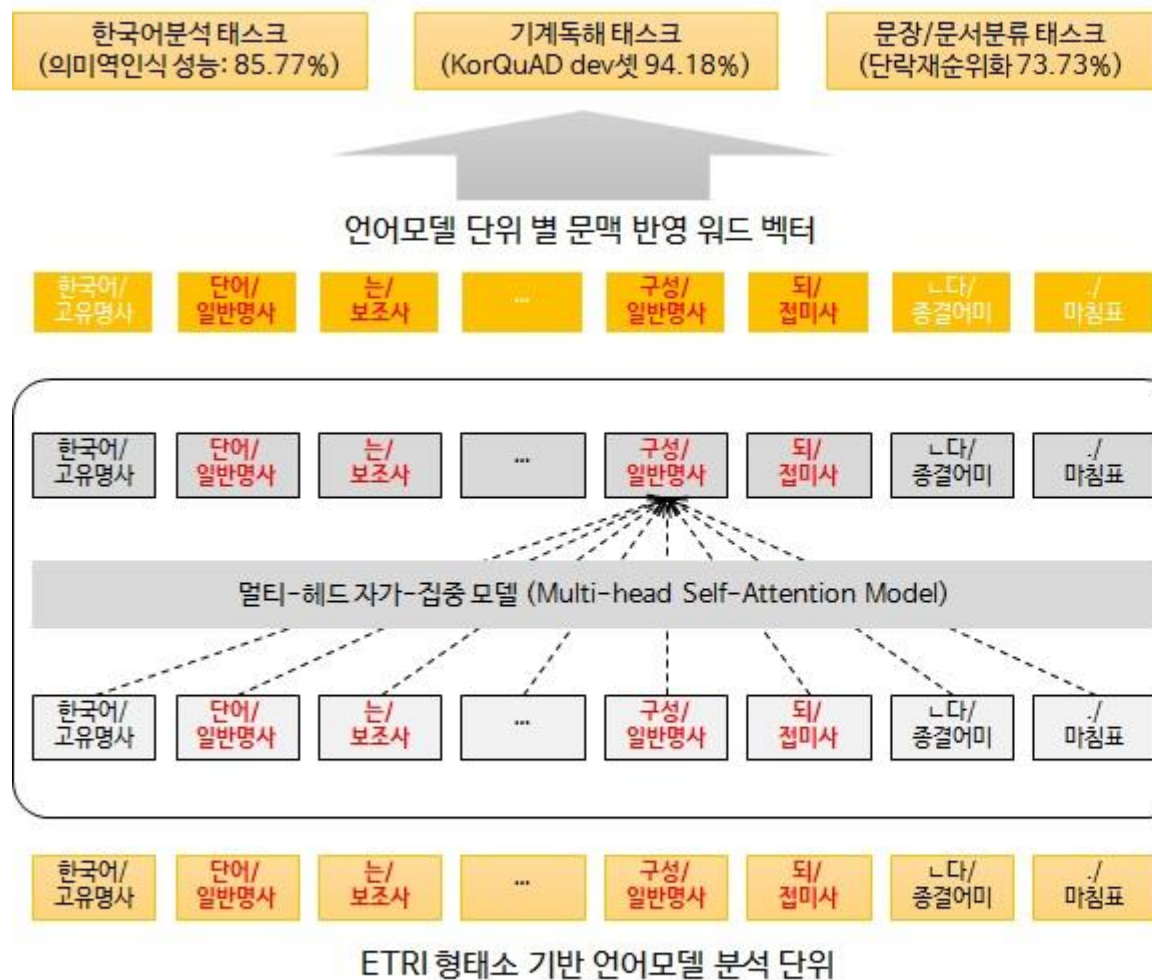
**슬롯 태깅**

**두 문장 관계 분류**



한국어 언어적 특성을 잘 반영하여  
훈련시킨 pre-trained model  
(23GB 원시 말뭉치 학습)

# BERT model



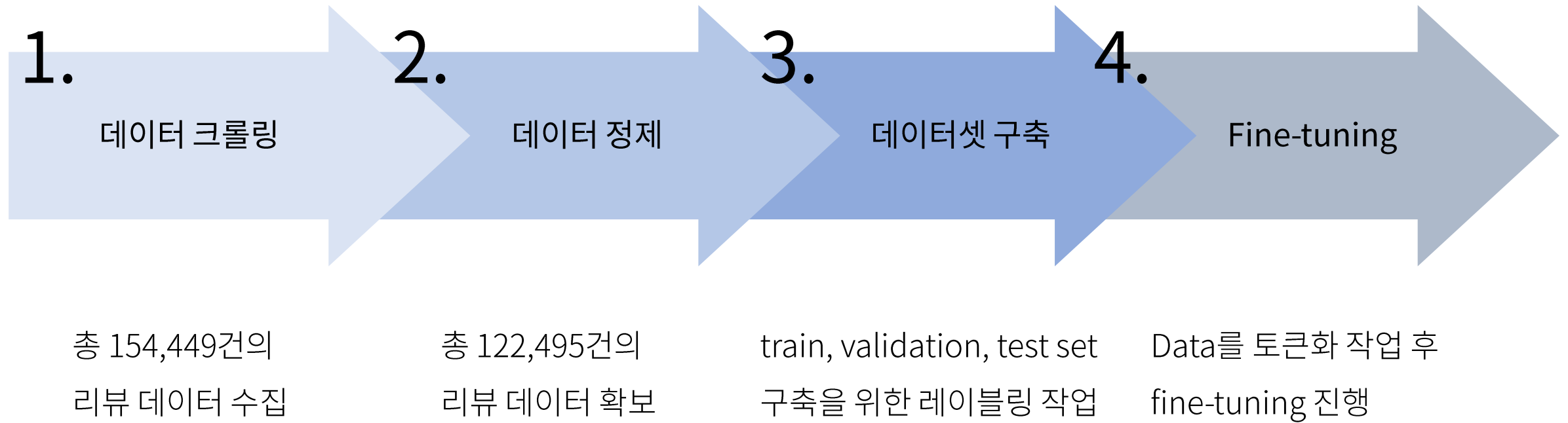
예문: 한국어 단어는 형태소로 구성된다.

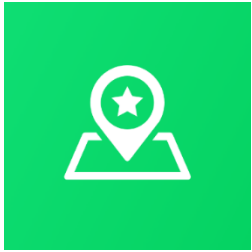
〈ETRI 형태소 기반 언어모델과 구글 언어모델 비교〉



## Data prepare process

---





Naver place

- 방문 인증된 이용자에 한해 리뷰 작성 가능
- 20년 1월 기준 순 이용자수 약 1,380만 명



Google maps

- 리뷰의 사실 관계 파악을 위한 검증 시스템 보유
- 지도 콘텐츠 참여에 따른 보상으로 지속적인 참여와 성의있는 리뷰 작성 유도

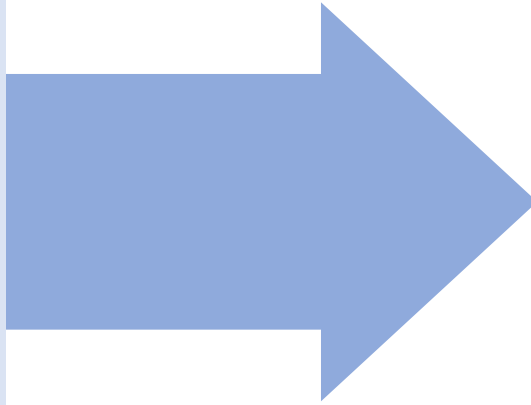


Trip advisor

- 엄격한 리뷰 게시 가이드라인을 통한 신뢰성
- 자체 리뷰 분석 시스템으로 부정 행위 색출 후 리뷰 게시를 통한 신뢰성 보유
- 세계 최대의 여행 플랫폼, 월간 이용자수 4억 6천만명



사전 조사 데이터 합계  
3,000건 이상 관광지 목록 선정



총 64곳 관광지  
154,449건의 리뷰 수집

### [ Refining list ]

1. 특수문자 제거
2. 이모티콘 제거
3. 개행(\n) 제거
4. Multi space(2번 이상 띄어쓰기) 변경
5. 10자 미만, 250자 초과 제거
6. 중복 리뷰 제거

총 122,495건 데이터 확보

## Labeling 기준

긍정	중립	부정
긍정 리뷰 100%	긍정, 부정 mix	부정 리뷰 100%
관광지 정보 + 긍정 리뷰 포함	관광지 정보만 있는 리뷰	관광지 정보 + 부정 리뷰 포함

긍정, 중립, 부정 각 4,000개씩 레이블링 진행 → 총 12,000개의 data 구축

학습에 도움되지 않는 리뷰(장소와 상관없는 리뷰, 정치적 견해 등)는 학습 데이터에서 제외

Train, Validation, Test = 7 : 1.5 : 1.5 비율로 데이터 구축



## Fine-tuning\_prepare data

- Fine-tuning을 위한 데이터 준비

Label	Review tokenization
0	'서울_', '다른_', '아', '쿠', '아', '리', '움', '에_', '비해_', '규모', '가_', '작', '음_', '금액_', '조정이_', '있다면_', '가', '볼', '만_', '함', '!_'
1	'어린', '아이들이_', '탈', '것이_', '많', '아요', '._', '시설', '점', '검', '으로_', '운', '행', '하지_', '않는_', '놀', '이_', '기', '구가_', '많', '네요_'
2	'규모', '가_', '작', '아요_', '아이', '들', '이', '좋아', '함_'

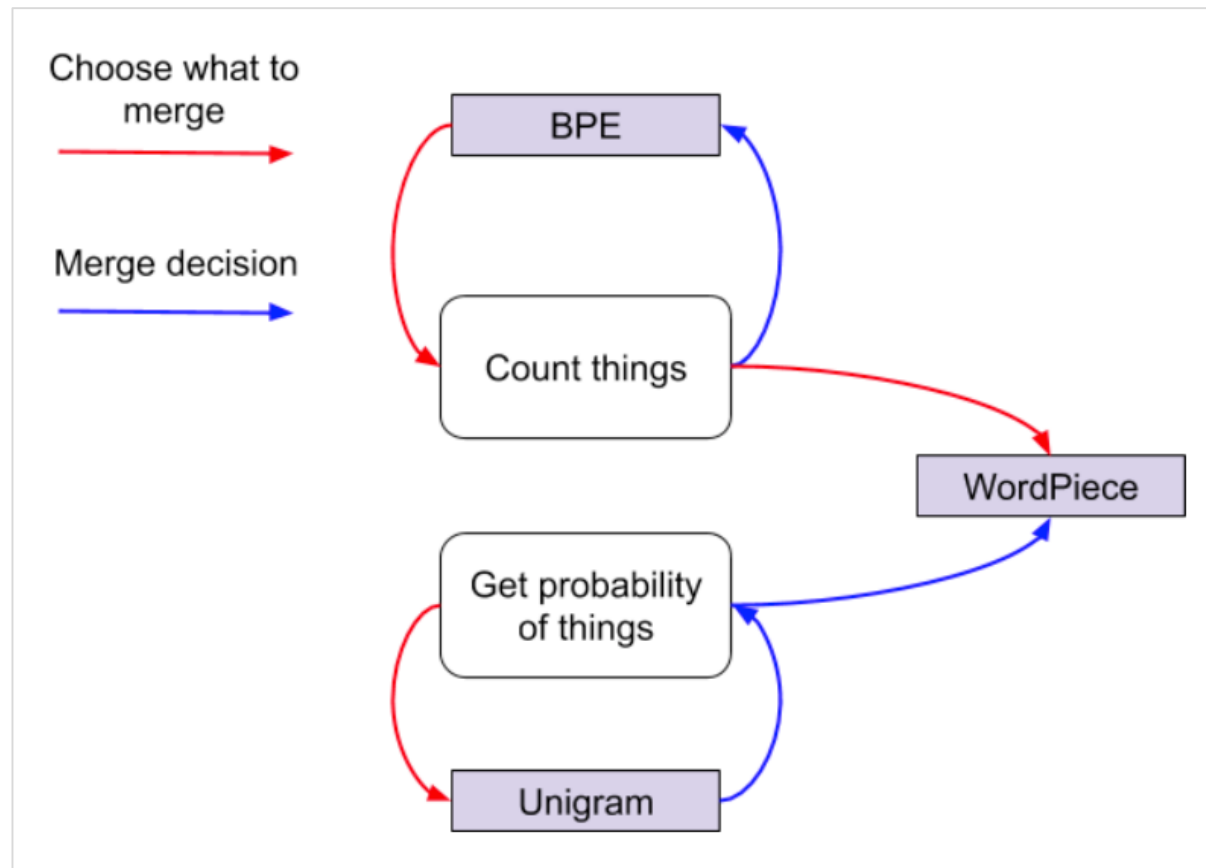
기존의 vocabulary에 없는 신조어나 오타가 일으키는 OOV(out of vocabulary) 문제를 해결하기 위해 word piece modeling 방식으로 tokenization 진행

### BPE(Byte pair embedding)?

OOV(out of vocabulary) 문제를 해결하기 위해 연속적으로 가장 많이 등장한 글자의 쌍을 찾아서 하나의 글자로 병합하는 방식 수행

### Word piece modeling?

BPE 변형 알고리즘으로, BPE와 달리 corpus의 우도(likelihood)를 가장 높이는 쌍을 병합하는 방식.  
BERT를 훈련하기 위해 사용된 모델



## Fine-tuning Hyper parameter 설정

### 필수

--train data

--model 저장 경로 지정

### 옵션

--validation data 적용

--epochs 설정

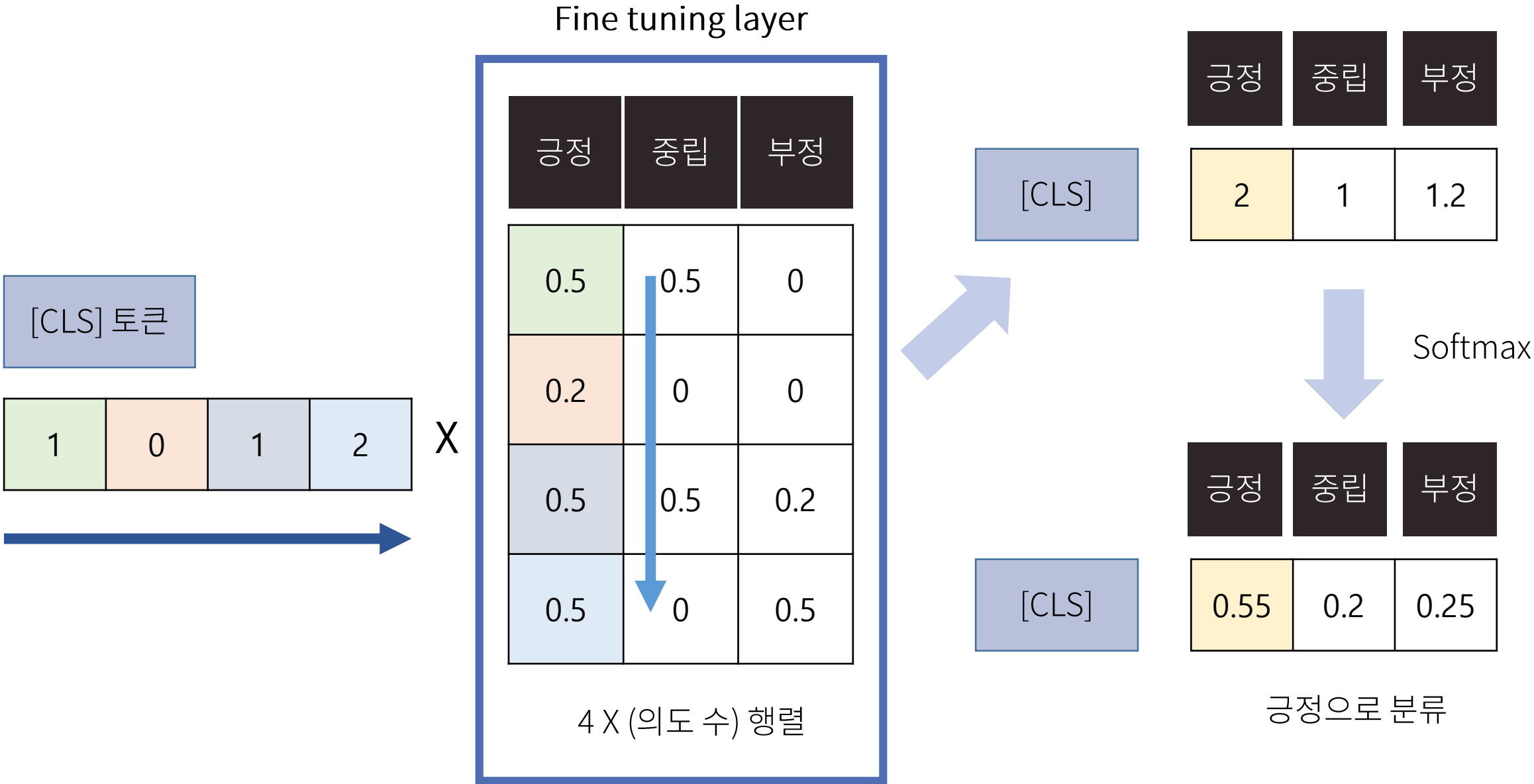
--batch size 설정

--type 선택 [bert or albert]

Label	Review vectorization
0	'[CLS]', '서울_', '다른_', '아', '쿠', '아', '리', '웁', '에_', '비해_', '규모', '가_', '작', '음_', '금액_', '조정이_', '있다면_', '가', '볼', '만_', '함', '._', '[SEP]'
1	'[CLS]', '어린', '아이들이_', '탈', '것이_', '많', '아요', '._', '시설', '점', '검', '으로_', '운', '행', '하지_', '않는_', '놀이_', '기', '구가_', '많', '네요_', '[SEP]'
2	'[CLS]', '규모', '가_', '작', '아요_', '아이', '들', '이', '좋아', '함_', '[SEP]'

- Vectorization 진행 → CLS, SEP 토큰 생성
- CLS(special classification token): BERT 내부의 transformer 층을 거친 후 토큰화된 문장의 의미 보유

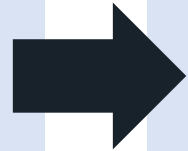
Fine-tuning layer 예시





**ETRI**

KorBERT (eojeol)



12,000개 data Fine-tuning

Train: Validation: Test = 7 : 1.5 : 1.5



Accuracy = **0.877**

## Model comparison

---

- Untrained model

DNN

Accuracy = 0.55

Epochs: 10
optimizer: Adam(1e-4)
tokenizer: Tensorflow SubwordTextEncoder (Wordpiece Model)

Text CNN(Conv1D)

Accuracy = 0.592

Epochs: 10
optimizer: Adam(1e-4)
tokenizer: Tensorflow SubwordTextEncoder (Wordpiece Model)

LSTM

Accuracy = 0.55

Epochs: 10
optimizer: Adam(1e-4)
tokenizer: Tensorflow SubwordTextEncoder (Wordpiece Model)

## Model comparison

ETRI KorBERT (eojeol)

Accuracy = 0.877

Pretrained data: 신문기사와 백과사전 등 23GB의 대용량 텍스트

Epochs: 4 (early stopping checkpoint)

Optimizer: RMSprop(learning rate = 1e-4)

Tokenizer: ETRI Wordpiece Model

attention\_probs\_dropout\_prob: 0.1 -> 0.3으로 변경

hidden\_dropout\_prob: 0.1 -> 0.3 으로 변경

SKT Brain KoBERT

Accuracy = 0.86

Pretrained data: 위키 문서의 문장 500만개(5400만단어)와 뉴스 문장 2000만개(2억7000만단어)

Epochs: 20

Optimizer: Adam(learning rate = 5.0e-5, decay = 0.0025)

Tokenizer: Sentencepiece Model

Google BERT

Word Piece 기반 다국어모델  
(bert-multilingual-cased)

Accuracy = 0.84

Pretrained data: 다국어 wikipedia 문서, 104 languages

Epochs: 3

Optimizer: Adam(learning rate = 5.0e-5)

Tokenizer: Wordpiece Model

## Model comparison

ETRI KorBERT (eojeol)

Accuracy = **0.877**

Pretrained data: 신문기사와 백과사전 등 23GB의 대용량 텍스트

Epochs: 4 (early stopping checkpoint)

Optimizer: RMSprop(learning rate = 1e-4)

Tokenizer: ETRI Wordpiece Model

attention\_probs\_dropout\_prob: 0.1 -> 0.3으로 변경  
hidden\_dropout\_prob: 0.1 -> 0.3 으로 변경

ETRI KorBERT  
(eojeol\_Albert)

Accuracy = **0.875**

Pretrained data: 신문기사와 백과사전 등 23GB의 대용량 텍스트

Epochs: 2

Optimizer: Adam(learning rate = 5.0e-5)

Tokenizer: ETRI Wordpiece Model

ETRI KorBERT (morp)

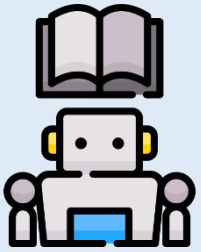
Accuracy = **0.819**

Pretrained data: 신문기사와 백과사전 등 23GB의 대용량 텍스트(47억개 형태소)

Epochs: 2

Optimizer: Adam(learning rate = 5.0e-5)

Tokenizer: ETRI Wordpiece Model



구축한 모델로  
110,495건 리뷰 예측



예측값으로 리뷰 분석



Dash board 구축

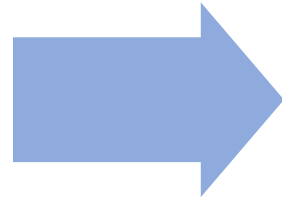
# 대시보드

---





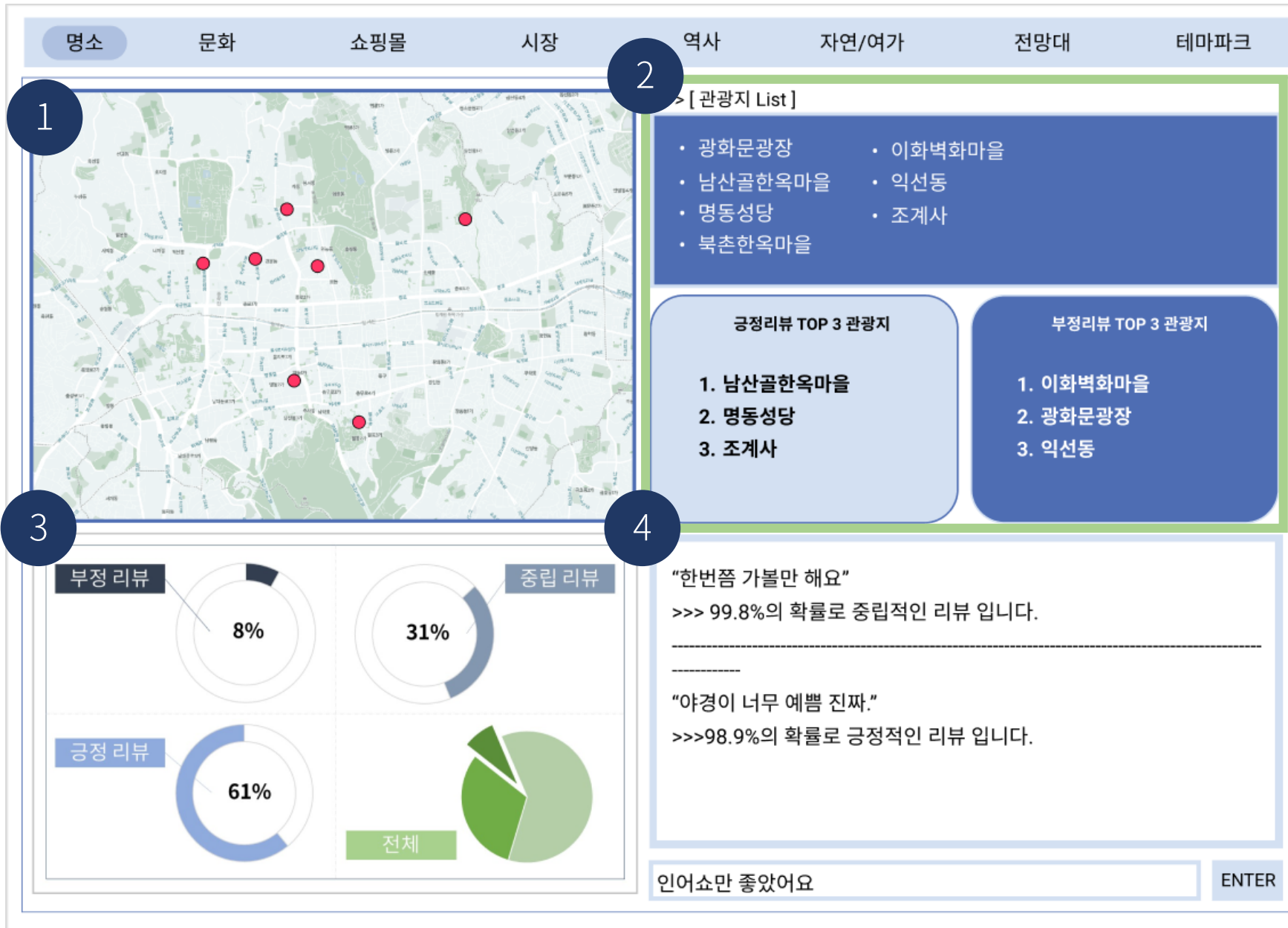
Figma로 UI 디자인



**django**

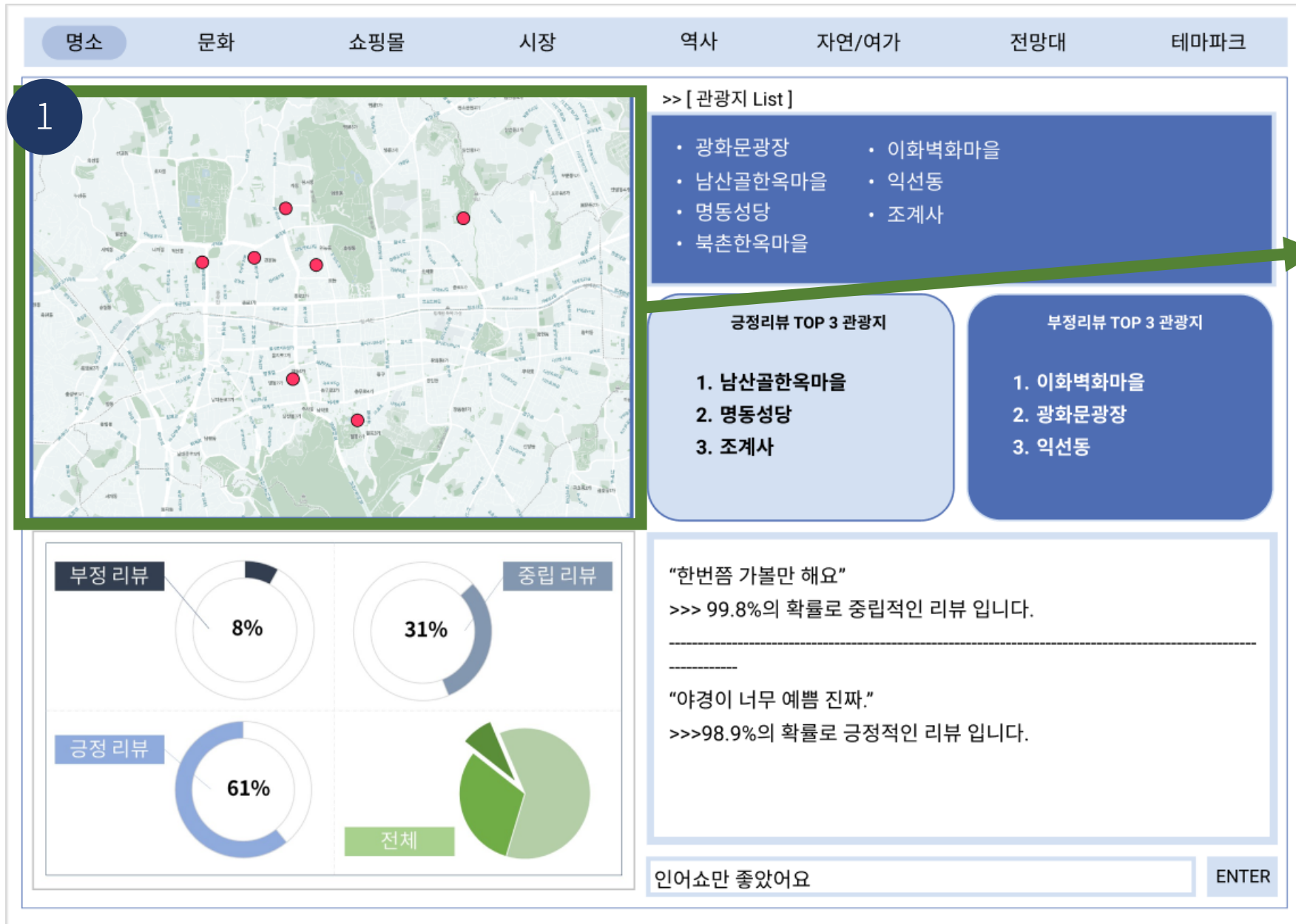
Django 로 웹 구축

# Dash board\_Theme tool

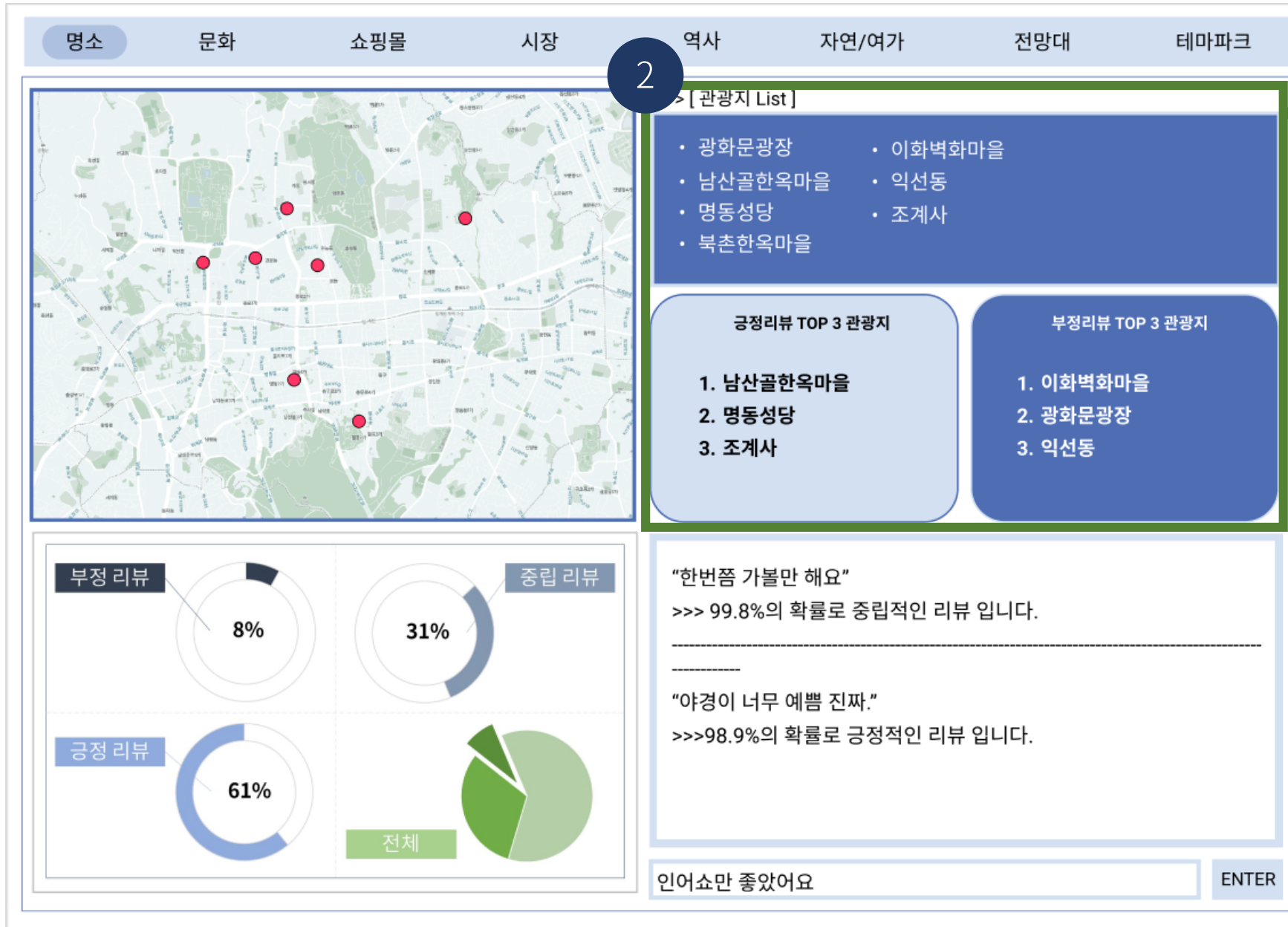




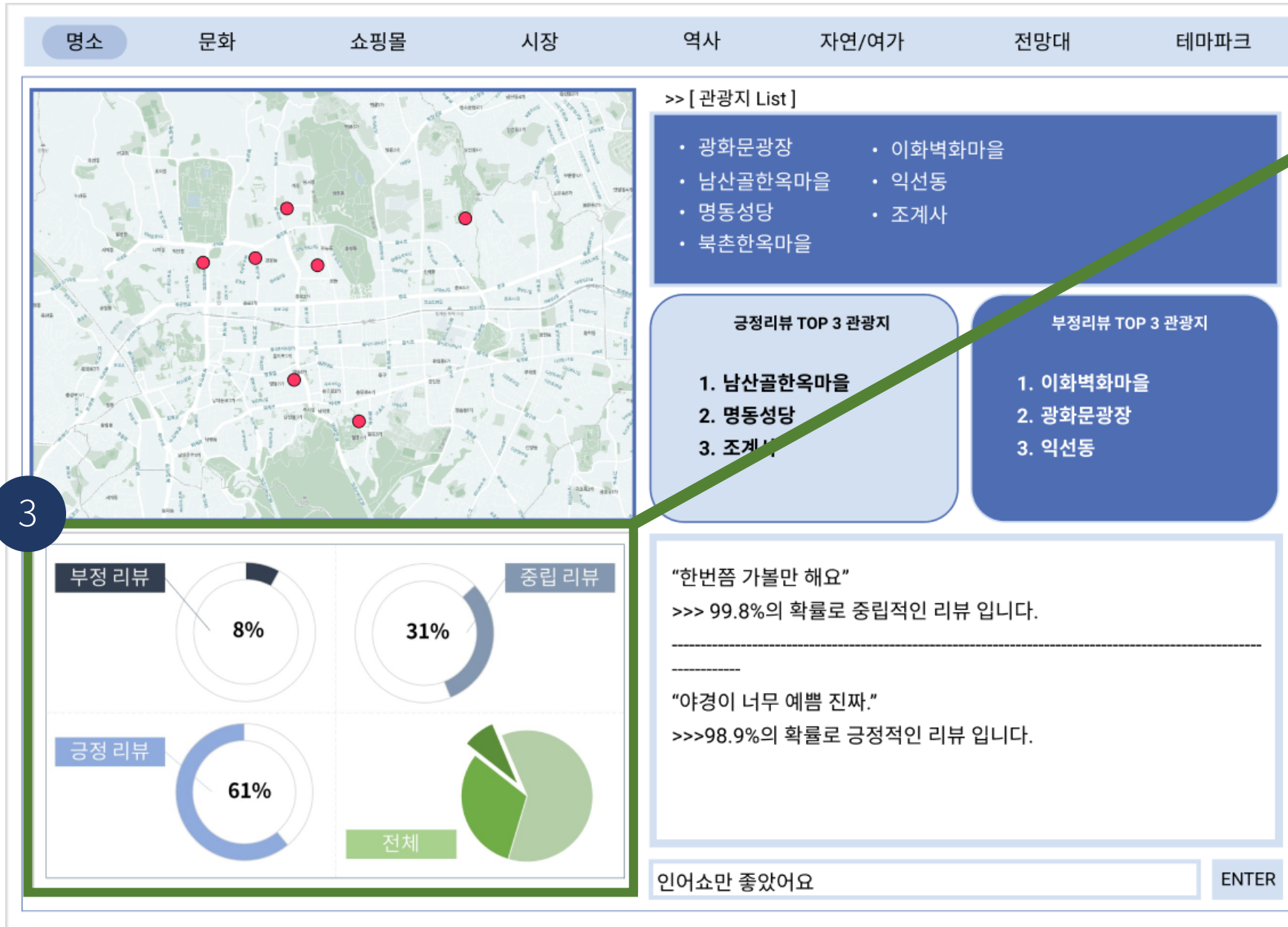
# Dash board\_Theme tool



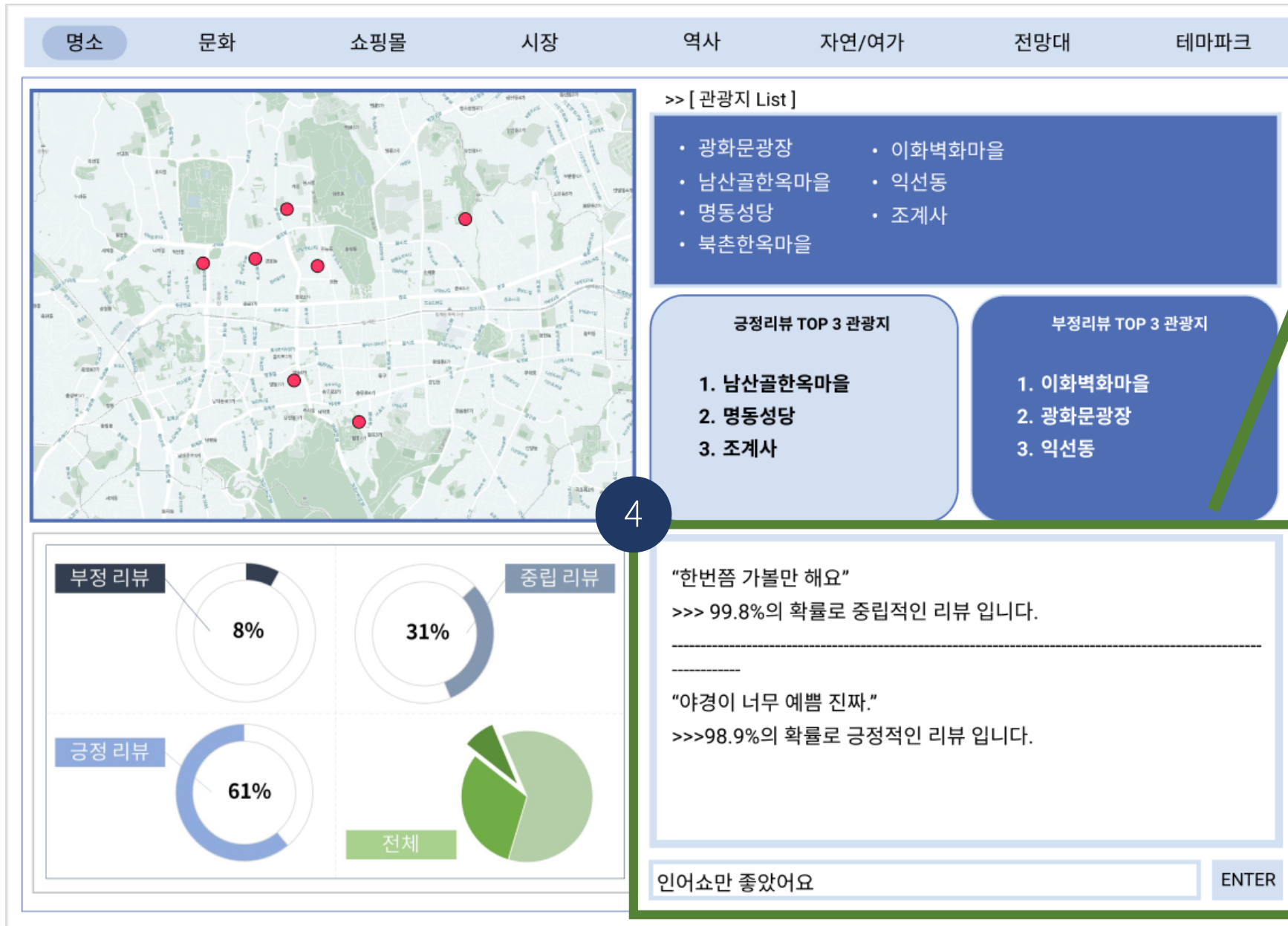
- Google Maps API를 통해 각 테마 별 관광지의 위치 정보 수집
- 수집한 위치 정보를 QGIS를 이용하여 mapping
- 테마 별 지도 이미지를 저장하고 대시보드에 게시



- 각 테마 별 관광지 목록 게시
- 관광지 선택 시 관광지에 대한 페이지로 이동
- 테마 별로 리뷰에 긍정 비율 및 부정 비율이 높은 관광지를 각각 게시



- 테마 별 긍정, 중립, 부정 리뷰 비율 그래프 게시





# Dash board\_Tourist attraction tool



- 관광지 별 부정 리뷰에 대해 워드 클라우드 제작
- 부정리뷰에 대한 솔루션 작성 및 대시보드 게시



# BERT code Demo

 **Model\_test.ipynb** ☆

파일 수정 보기 삽입 런타임 도구 도움말 모든 변경사항이 저장됨

RAM 디스크 수정 가능

Mount

```
from google.colab import drive
import os

drive.mount("/content/drive", force_remount=True)

Mounted at /content/drive
```

Check directories and Change directory

```
[2] !ls

drive sample_data
```

```
[3] # 각자에게 맞는 directory를 설정해주세요.
    eojeol = "/content/drive/MyDrive/bert_classification_kor/eojeol"
    os.chdir(eojeol)
```

3초 오전 12:54에 완료됨



리뷰에 대한 정확한 정보 제공



관광지의 보완점 빠르게 파악 후 개선



다양한 리뷰를 수집하여 전국 관광지를  
평가할 수 있는 모델로 확대



**End of document.**

