

# 검색 품질, 그 너머

Beyond retrieval quality (feat. Meilisearch)

Team 10 (매일리서치) / 김성수, 서동국, 안세호, 박제윤(Mentor)

# 목차

## I. 프로젝트 개요

### 1. 요구사항 정의

## II. 프로젝트 과정

### 1. 단어 사전 제작

### 2. Tokenizer Build (형태소 기반)

### 3. 단어 사전 적용

### 4. Build 후 문제점 개선

## III. 프로젝트 마무리

### 1. 성능 평가

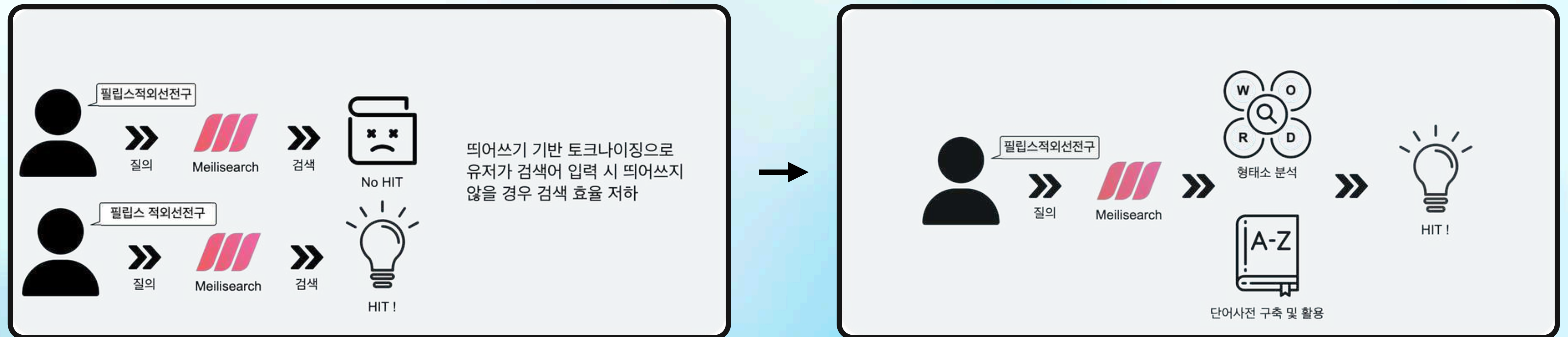
### 2. 부가적 문제

### 3. 개선 방안

# I. 프로젝트 개요

# I. 프로젝트 개요

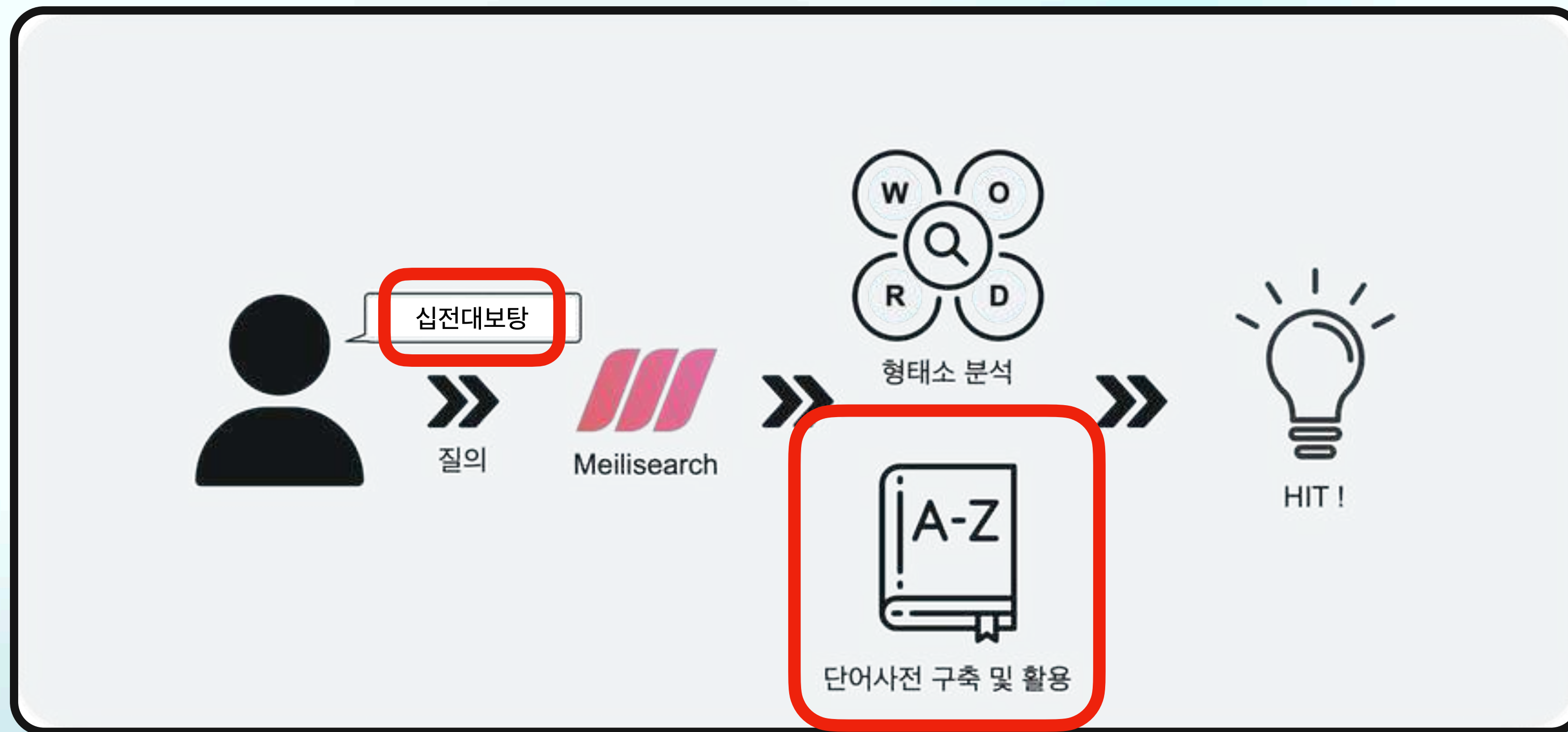
## 1. 요구사항 정의 - 형태소 분석 기반 검색엔진 Build





# I. 프로젝트 개요

## 1. 요구사항 정의 - 형태소 분석에 사용될 도메인 단어사전 구축

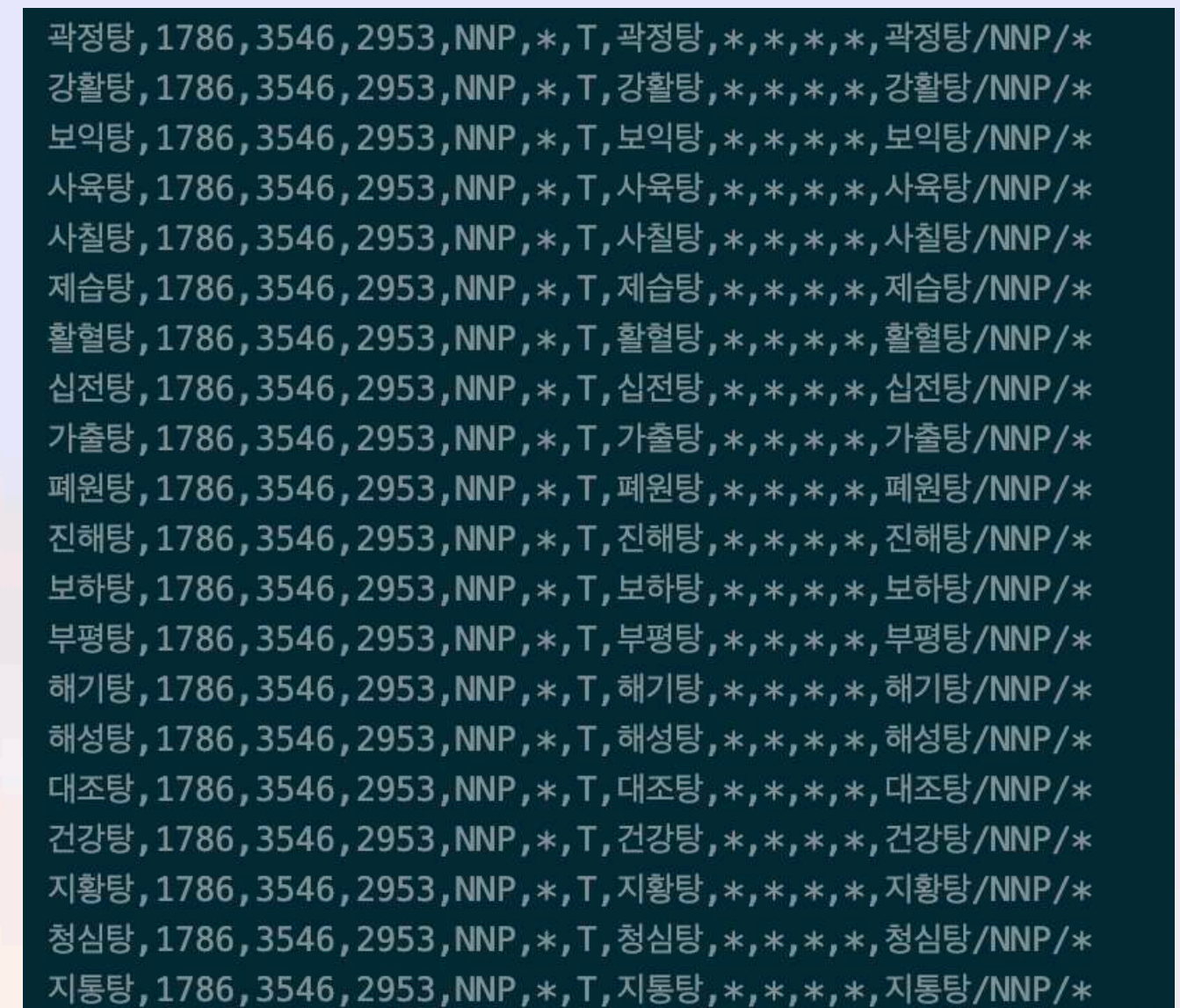
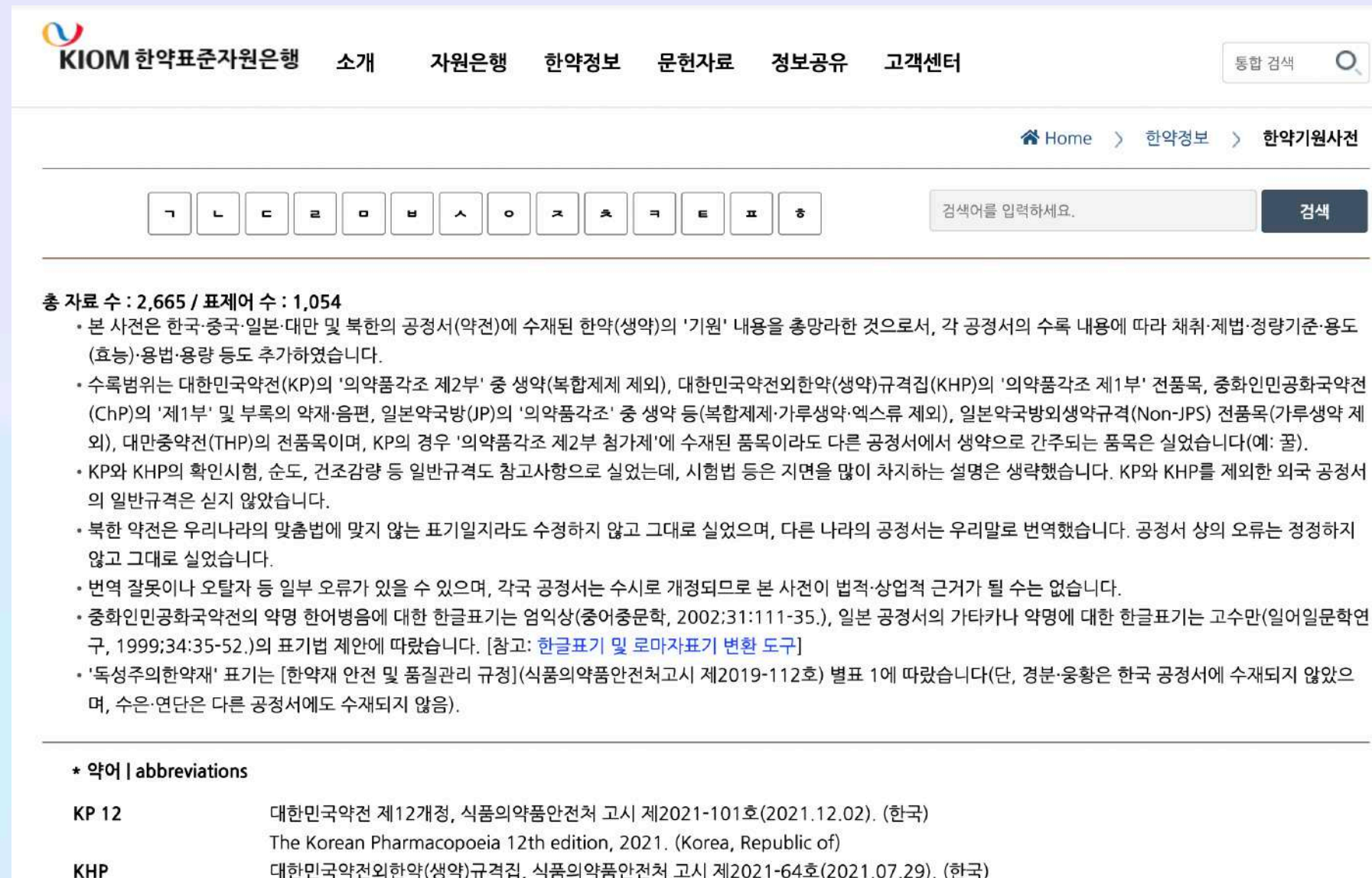
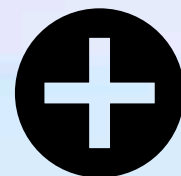
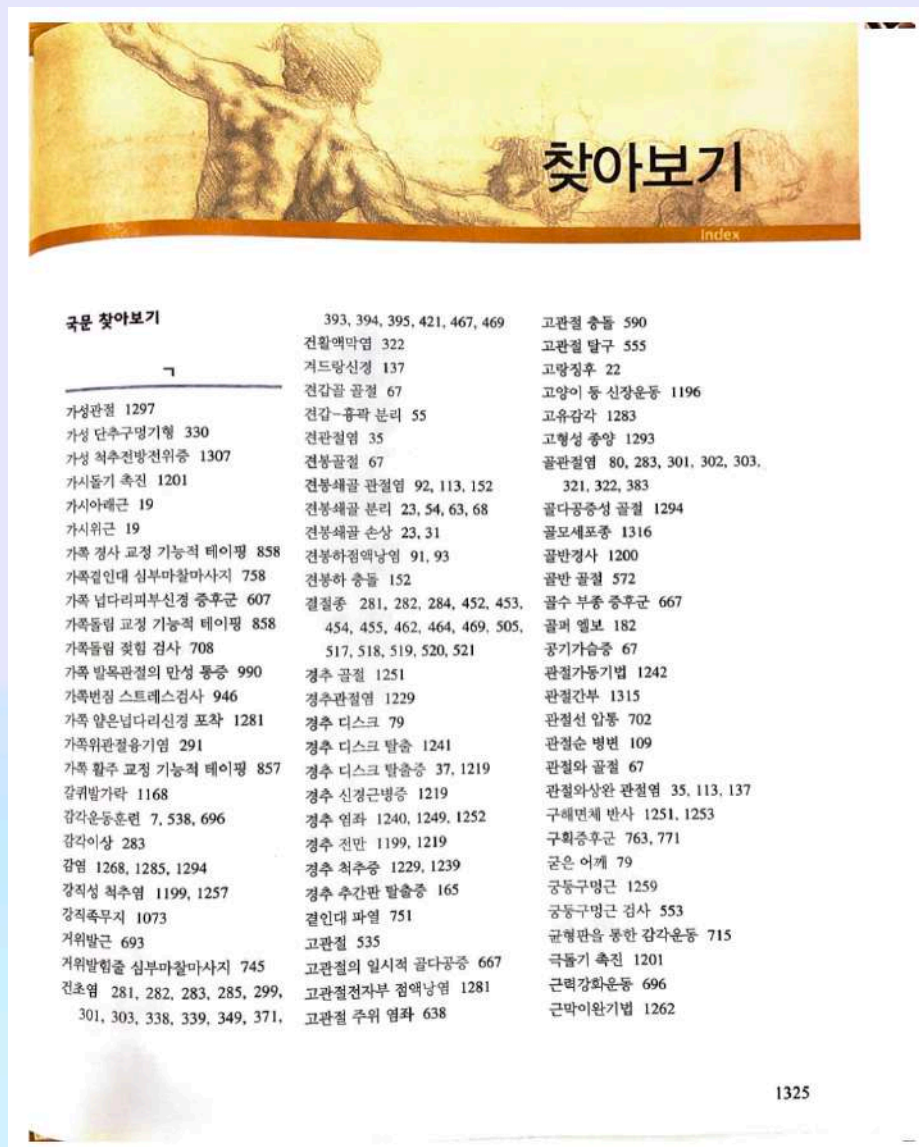


## II. 프로젝트 과정



# II. 프로젝트 과정

## 1. 단어사전(도메인) 구축



종이서적 OCR

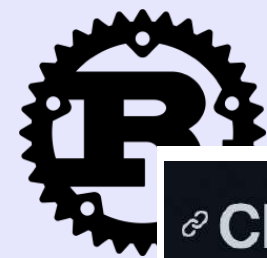
온라인 한의학 자료

도메인 단어사전



## II. 프로젝트 과정

### 2. Tokenizer(한국어용) Build



Charabia

Library used by Meilisearch to tokenize queries and documents

검색엔진 형태소 분석 코어 Library

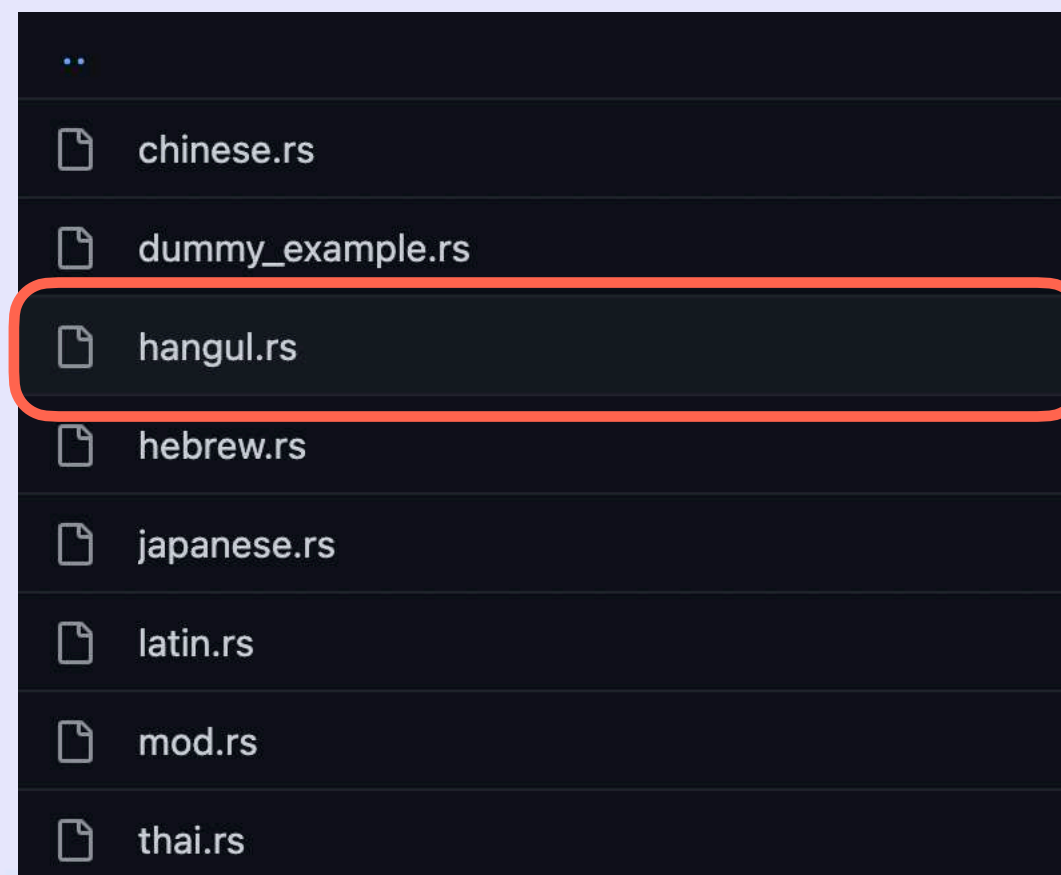
mosuka Upgrade lindra to 0.16.0		✓ 5439675 23 days ago History
..		
chinese.rs	Refactor script detection	5 months ago
dummy_example.rs	Rename example file	6 months ago
hebrew.rs	Include more changes by @ManyTheFish	3 months ago
japanese.rs	Upgrade lindra to 0.16.0	23 days ago
latin.rs	Patch latin segmenter to split words on quotes	5 months ago
mod.rs	Fixed formatting + put dictionary in root	2 months ago
thai.rs	Fix PR feedback	2 months ago

한국어 Tokenizer 미등재



# II. 프로젝트 과정

## 2. Tokenizer(한국어용) Build



한국어용 Tokenizer 파일 제작

```
use crate::segmenter::Segmenter;
use lindera::mode::{Mode, Penalty};
use lindera::tokenizer::{Tokenizer, TokenizerConfig, DictionaryConfig};
use once_cell::sync::Lazy;

/// Hangul specialized ['Segmenter'].
///
/// This Segmenter uses lindera internally to segment the provided text.
pub struct HangulSegmenter;

static LINDERA: Lazy<Tokenizer> = Lazy::new(|| {
    let config =
        TokenizerConfig { dictionary: DictionaryConfig {
            kind: "ko-dic",
            path: None,
        }, mode: Mode::Decompose(Penalty::default()) };
    Tokenizer::with_config(config).unwrap()
});

impl Segmenter for HangulSegmenter {
    fn segment_str<'o>(&self, to_segment: &'o str) -> Box<dyn Iterator<Item = &'o str> + 'o> {
        let segment_iterator = LINDERA.tokenize(to_segment).unwrap();
        Box::new(segment_iterator.into_iter().map(|token| token.text))
    }
}
```

Library 적용



Charabia

Library used by Meilisearch to tokenize queries and documents

형태소 분석 코어 Library

형태소 분석 적용



Lindera Morphology

A morphological analysis libraries and commands.

1 follower Tokyo, Japan

한국어/일본어 형태소 분석 Library



## II. 프로젝트 과정

### 3. 단어사전(도메인) 적용



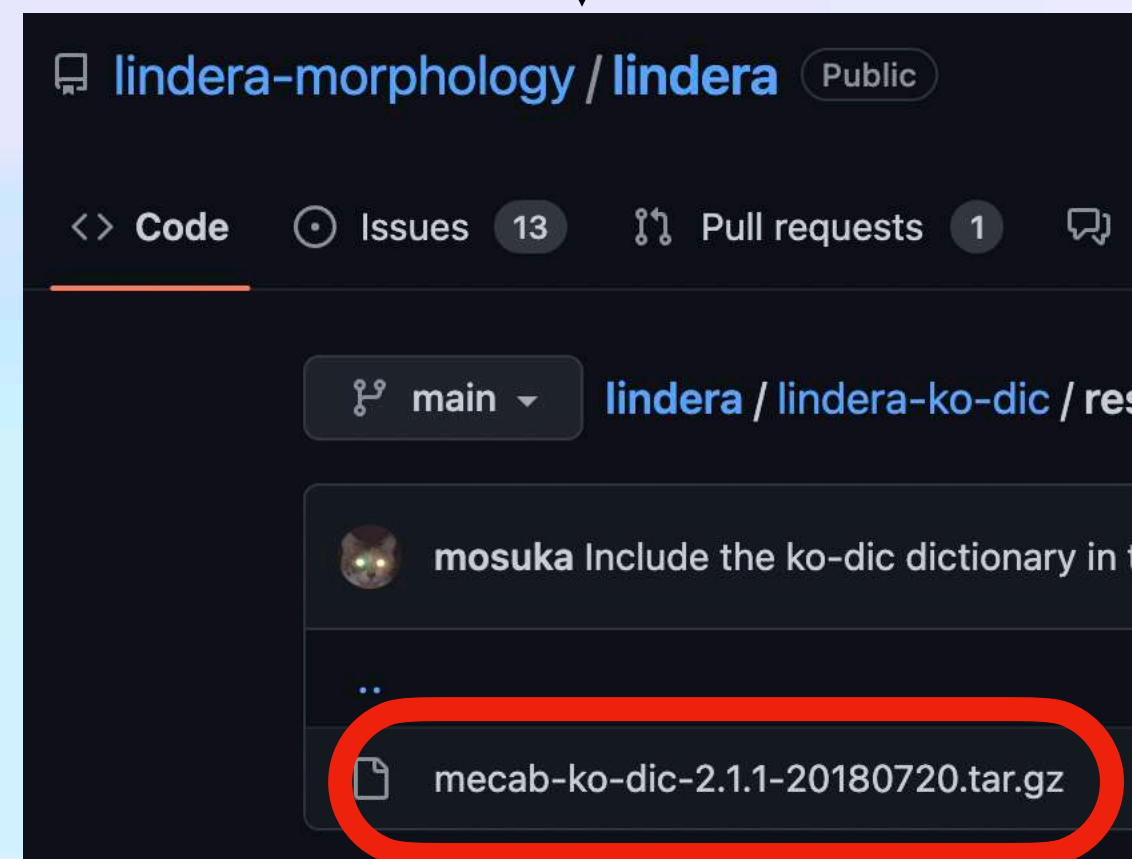
**Lindera Morphology**

A morphological analysis libraries and commands.

1 follower Tokyo, Japan

한국어/일본어 형태소 분석 Library

10



기존 단어사전

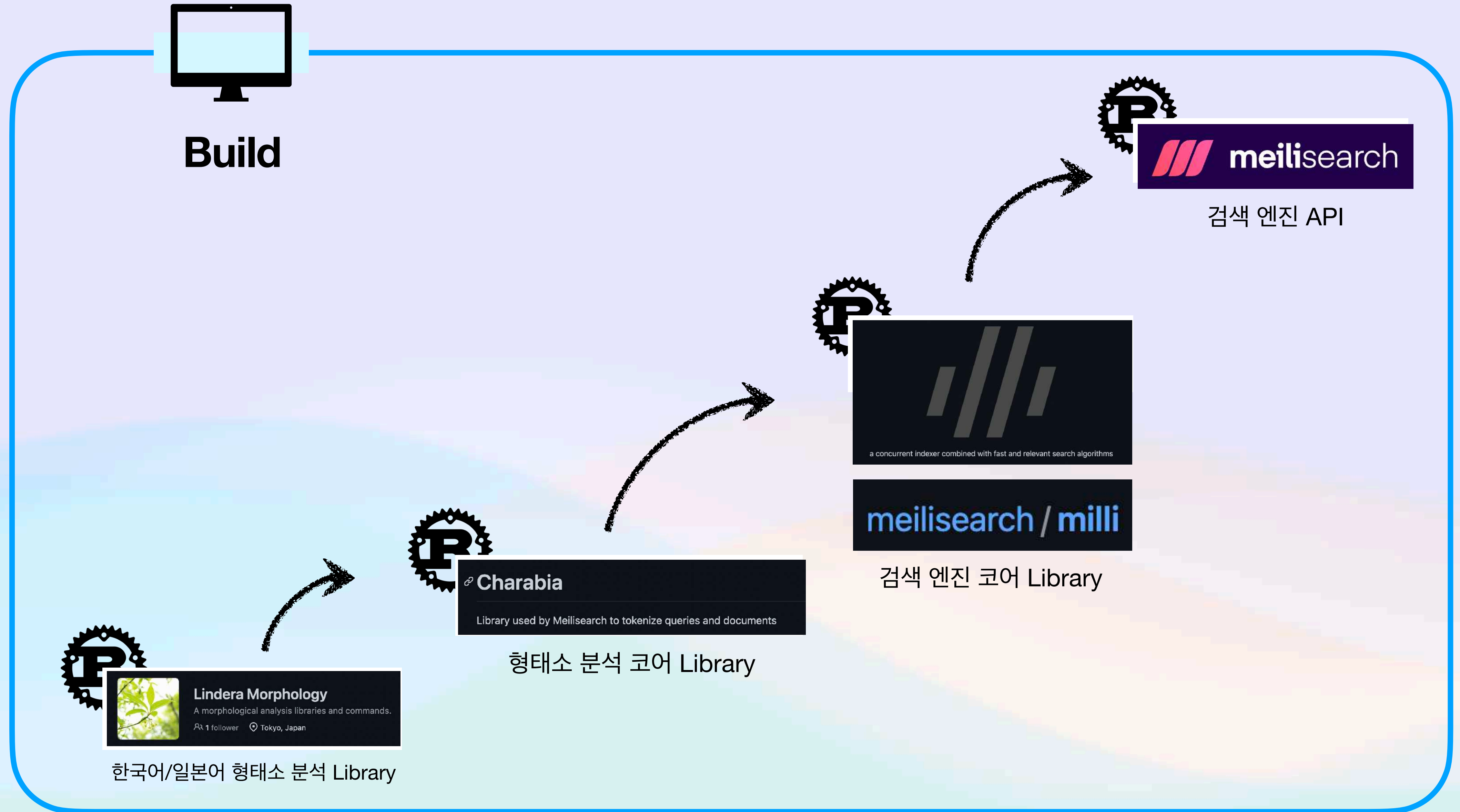


구축한 단어사전

곽정탕, 1786, 3546, 2953, NNP, \*, T, 곽정탕, \*, \*, \*, \*, 곽정탕/NNP/\*  
강활탕, 1786, 3546, 2953, NNP, \*, T, 강활탕, \*, \*, \*, \*, 강활탕/NNP/\*  
보익탕, 1786, 3546, 2953, NNP, \*, T, 보익탕, \*, \*, \*, \*, 보익탕/NNP/\*  
사육탕, 1786, 3546, 2953, NNP, \*, T, 사육탕, \*, \*, \*, \*, 사육탕/NNP/\*  
사철탕, 1786, 3546, 2953, NNP, \*, T, 사철탕, \*, \*, \*, \*, 사철탕/NNP/\*  
제습탕, 1786, 3546, 2953, NNP, \*, T, 제습탕, \*, \*, \*, \*, 제습탕/NNP/\*  
활혈탕, 1786, 3546, 2953, NNP, \*, T, 활혈탕, \*, \*, \*, \*, 활혈탕/NNP/\*  
십전탕, 1786, 3546, 2953, NNP, \*, T, 십전탕, \*, \*, \*, \*, 십전탕/NNP/\*  
가출탕, 1786, 3546, 2953, NNP, \*, T, 가출탕, \*, \*, \*, \*, 가출탕/NNP/\*  
폐원탕, 1786, 3546, 2953, NNP, \*, T, 폐원탕, \*, \*, \*, \*, 폐원탕/NNP/\*  
진해탕, 1786, 3546, 2953, NNP, \*, T, 진해탕, \*, \*, \*, \*, 진해탕/NNP/\*  
보하탕, 1786, 3546, 2953, NNP, \*, T, 보하탕, \*, \*, \*, \*, 보하탕/NNP/\*  
부평탕, 1786, 3546, 2953, NNP, \*, T, 부평탕, \*, \*, \*, \*, 부평탕/NNP/\*  
해기탕, 1786, 3546, 2953, NNP, \*, T, 해기탕, \*, \*, \*, \*, 해기탕/NNP/\*  
해성탕, 1786, 3546, 2953, NNP, \*, T, 해성탕, \*, \*, \*, \*, 해성탕/NNP/\*  
대조탕, 1786, 3546, 2953, NNP, \*, T, 대조탕, \*, \*, \*, \*, 대조탕/NNP/\*  
건강탕, 1786, 3546, 2953, NNP, \*, T, 건강탕, \*, \*, \*, \*, 건강탕/NNP/\*  
지황탕, 1786, 3546, 2953, NNP, \*, T, 지황탕, \*, \*, \*, \*, 지황탕/NNP/\*  
청심탕, 1786, 3546, 2953, NNP, \*, T, 청심탕, \*, \*, \*, \*, 청심탕/NNP/\*  
지통탕, 1786, 3546, 2953, NNP, \*, T, 지통탕, \*, \*, \*, \*, 지통탕/NNP/\*



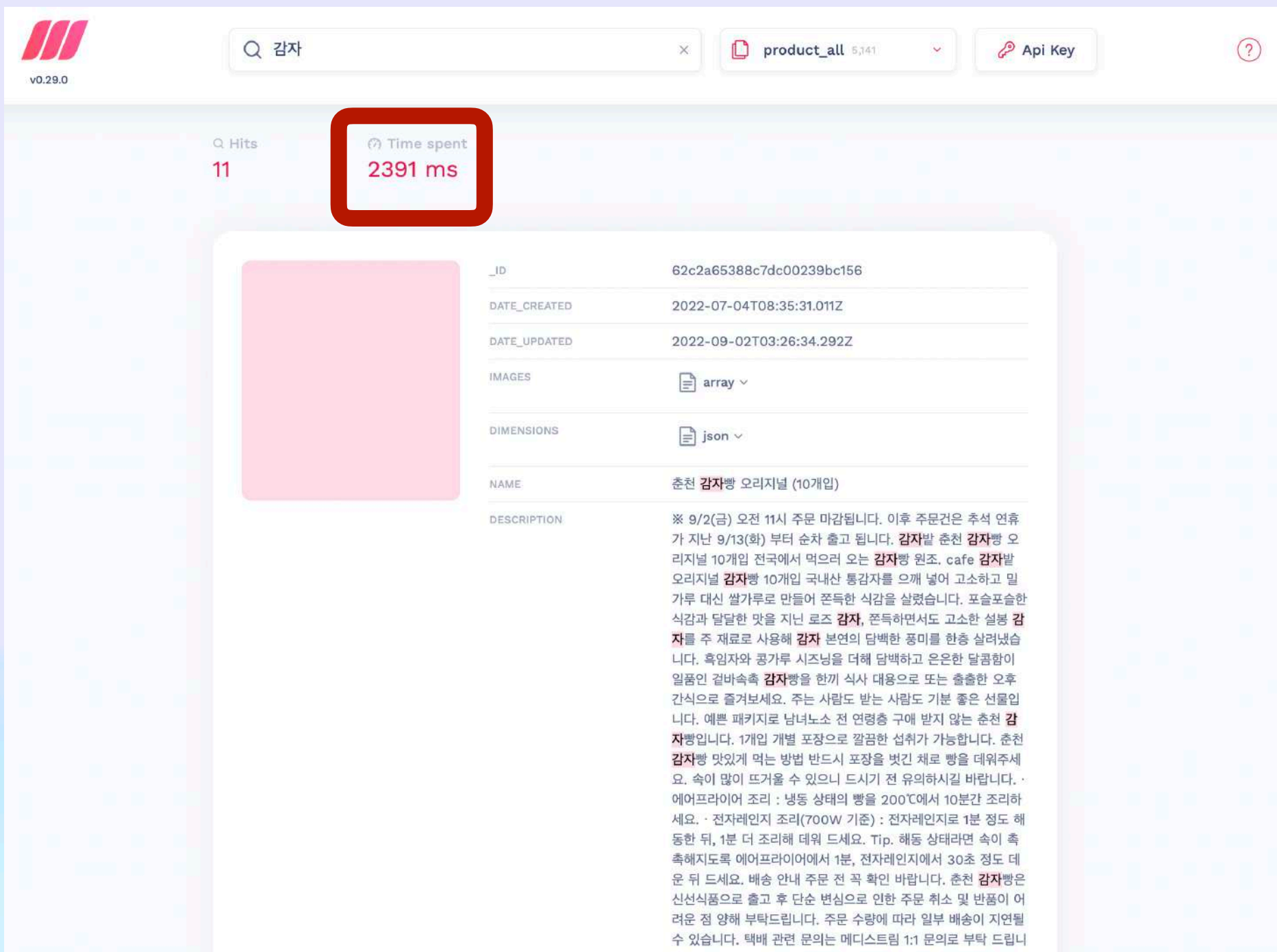
## II. 프로젝트 과정





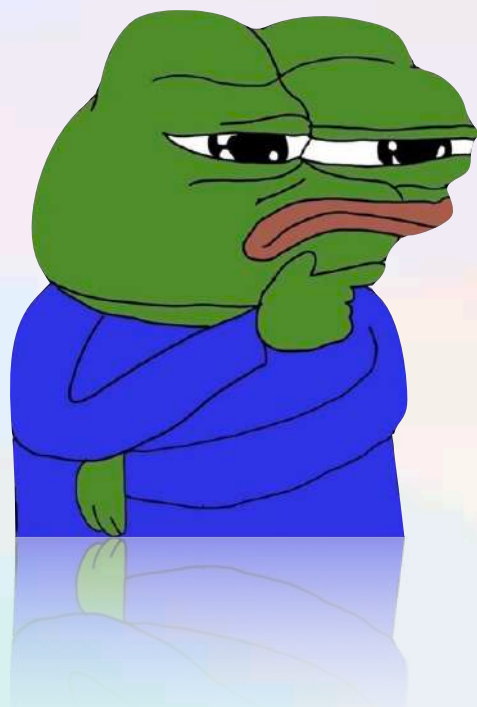
# II. 프로젝트 과정

## 4. Build 후 문제점 개선

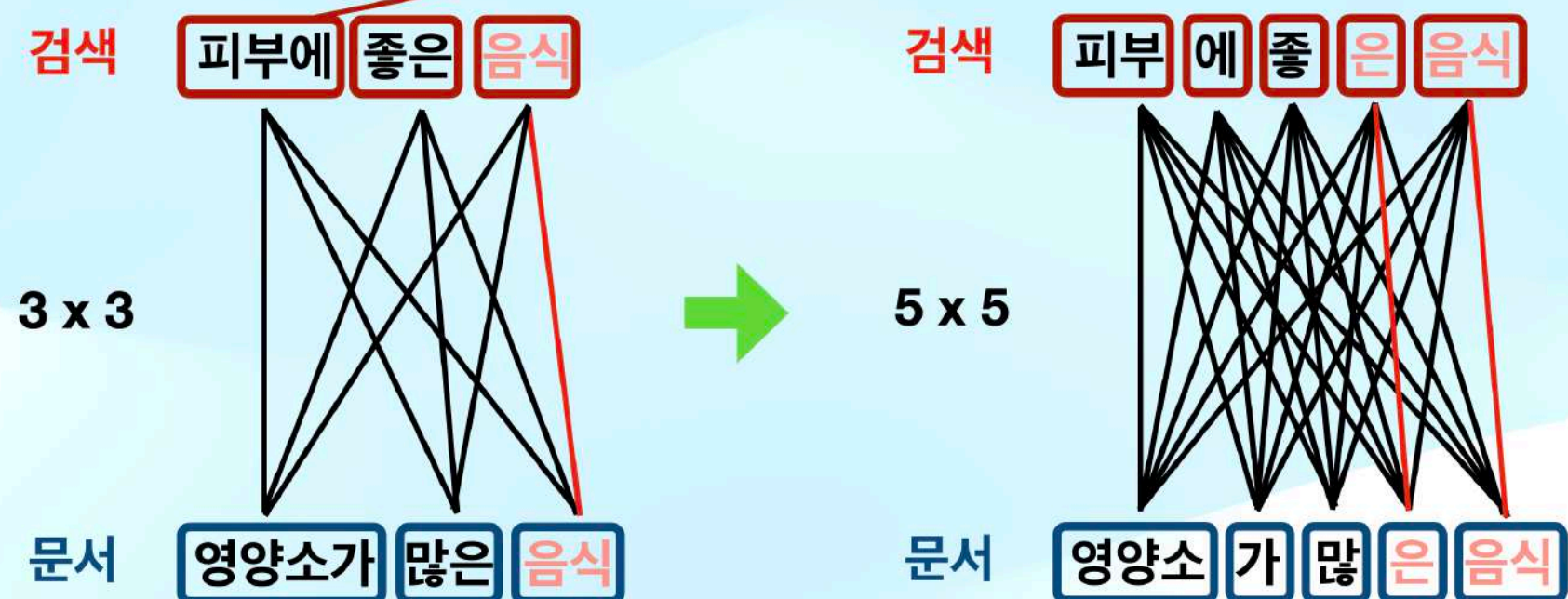


## 속도 저하 발생

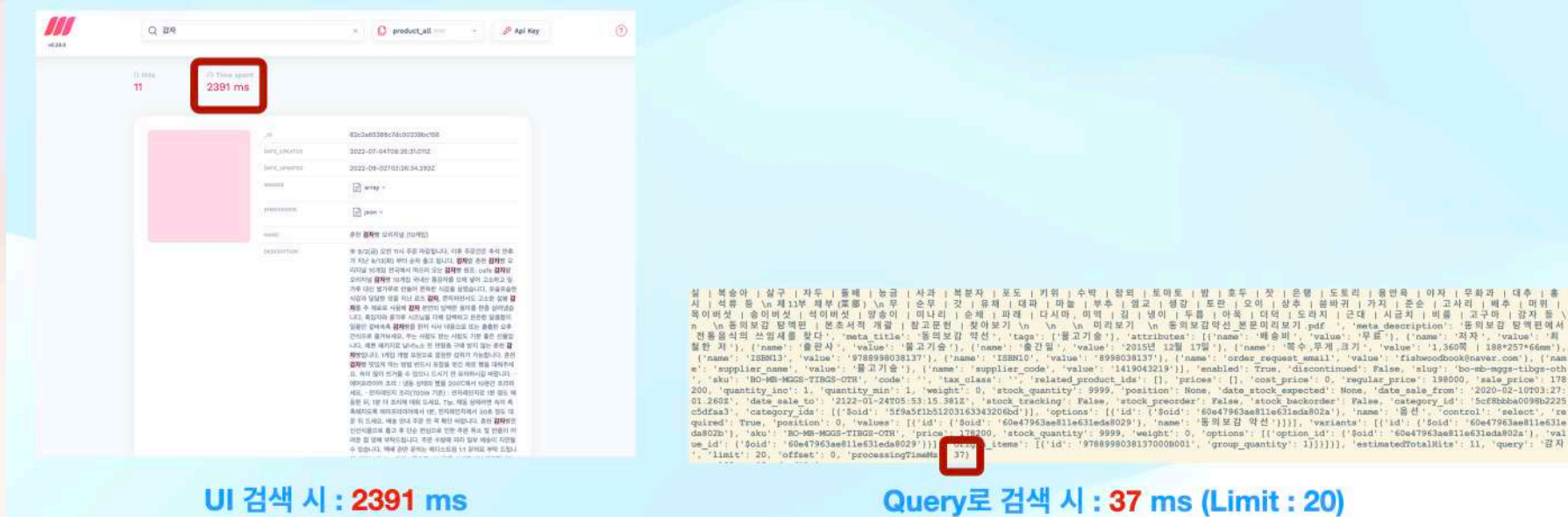
## 원인 분석



### 1-1) 속도 저하 원인 - 연산 ↑



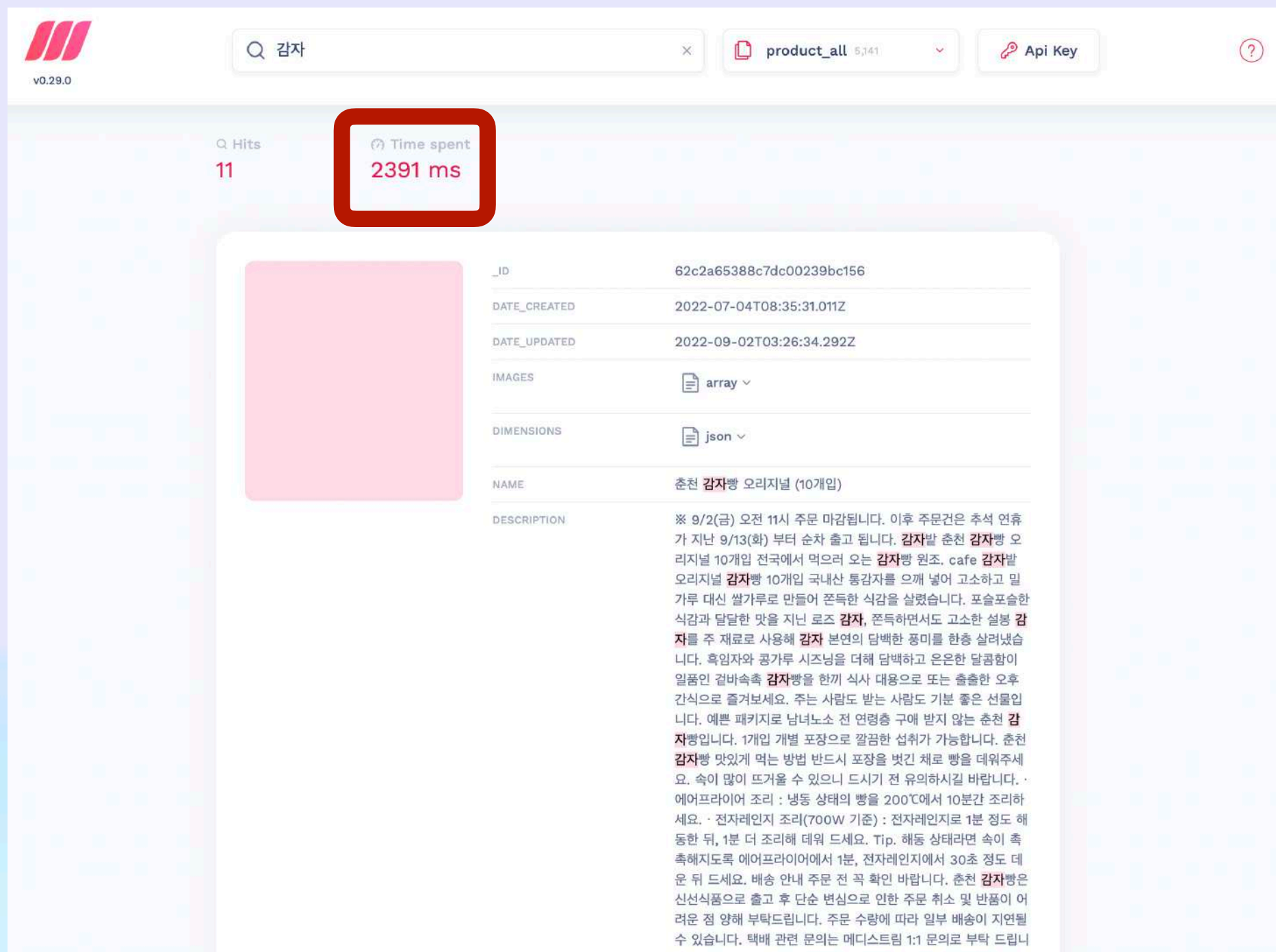
### 1-2) 속도 저하 원인 - UI 구현 문제





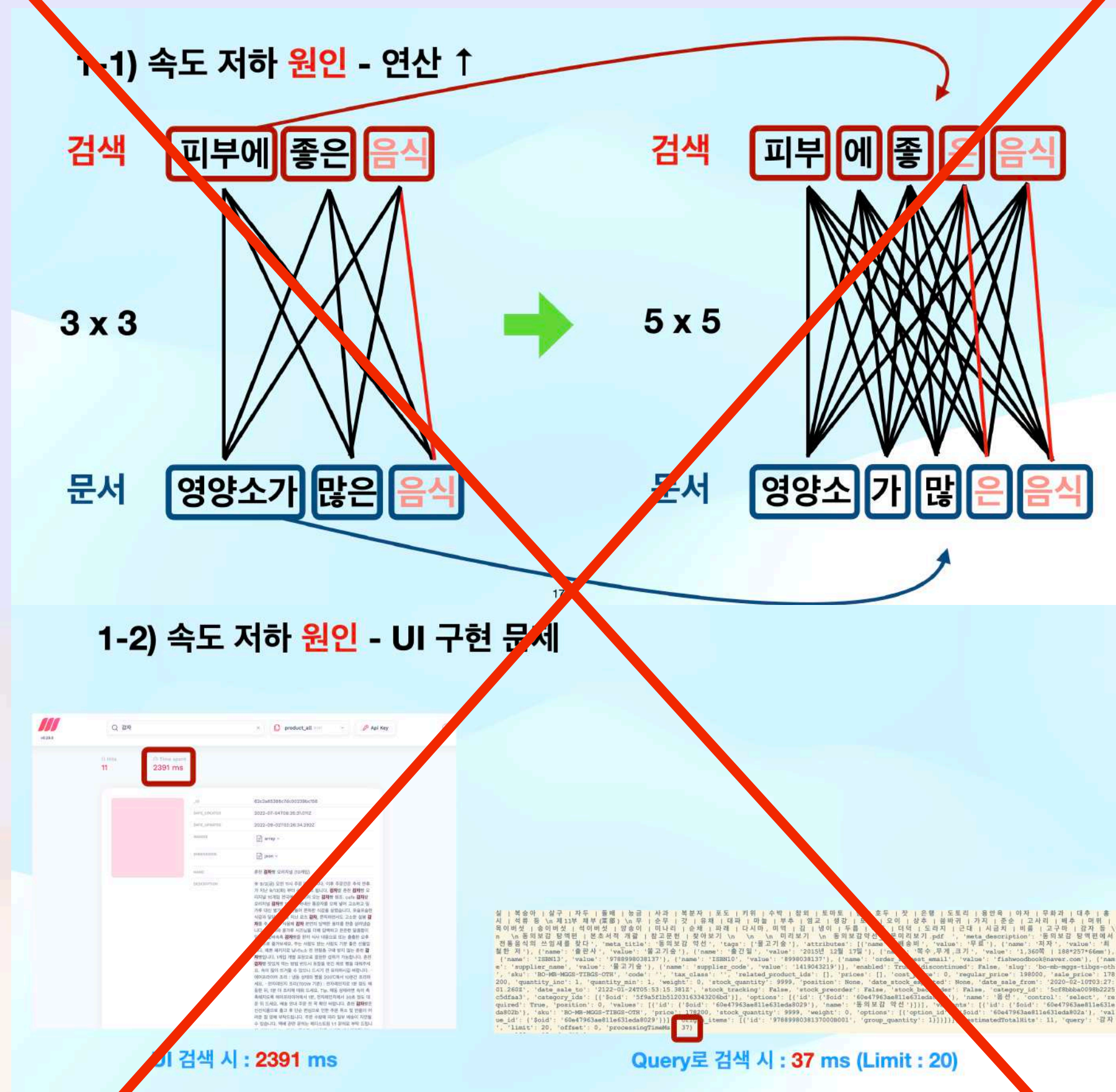
# II. 프로젝트 과정

## 4. Build 후 문제점 개선



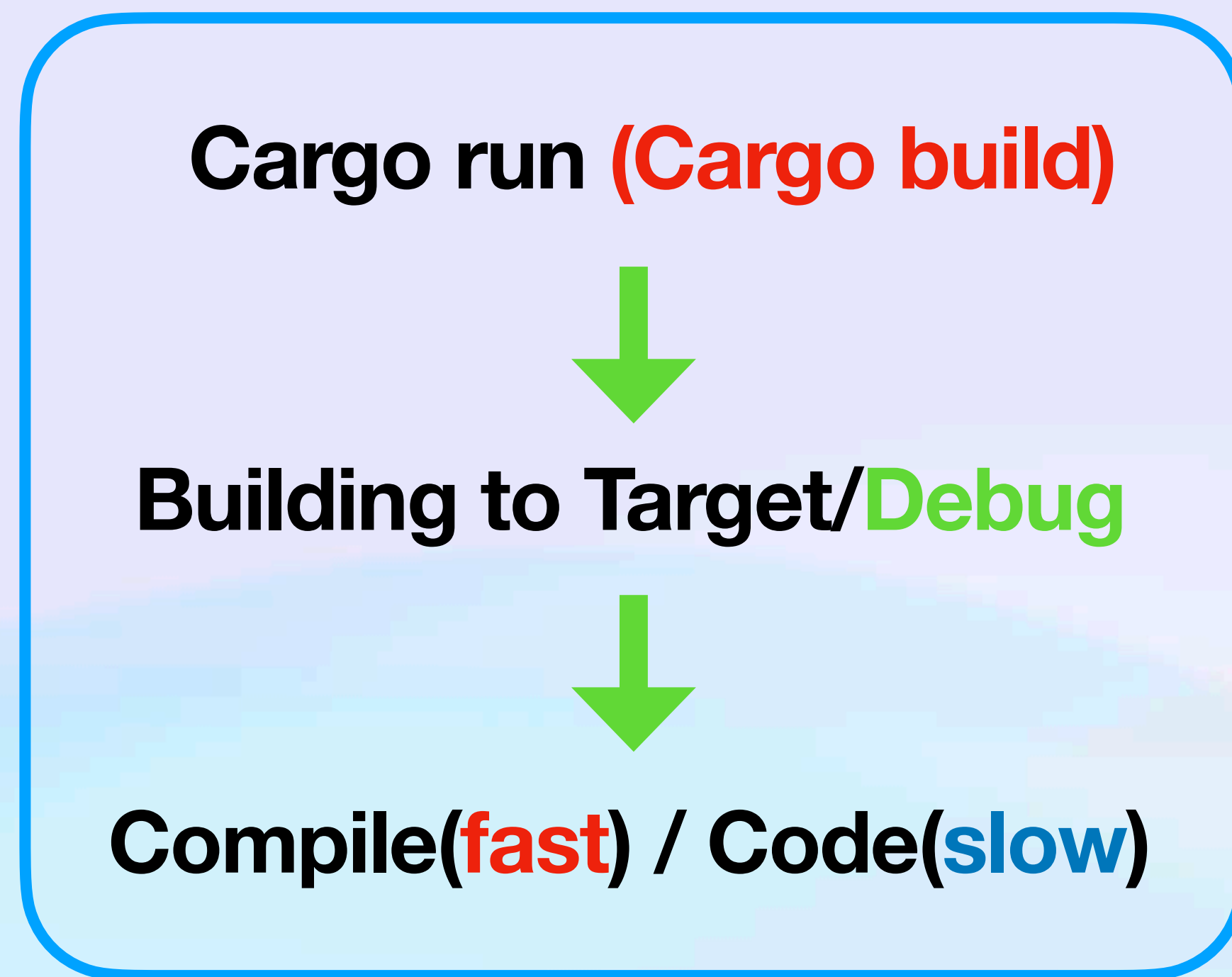
속도 저하 발생

원인 파악

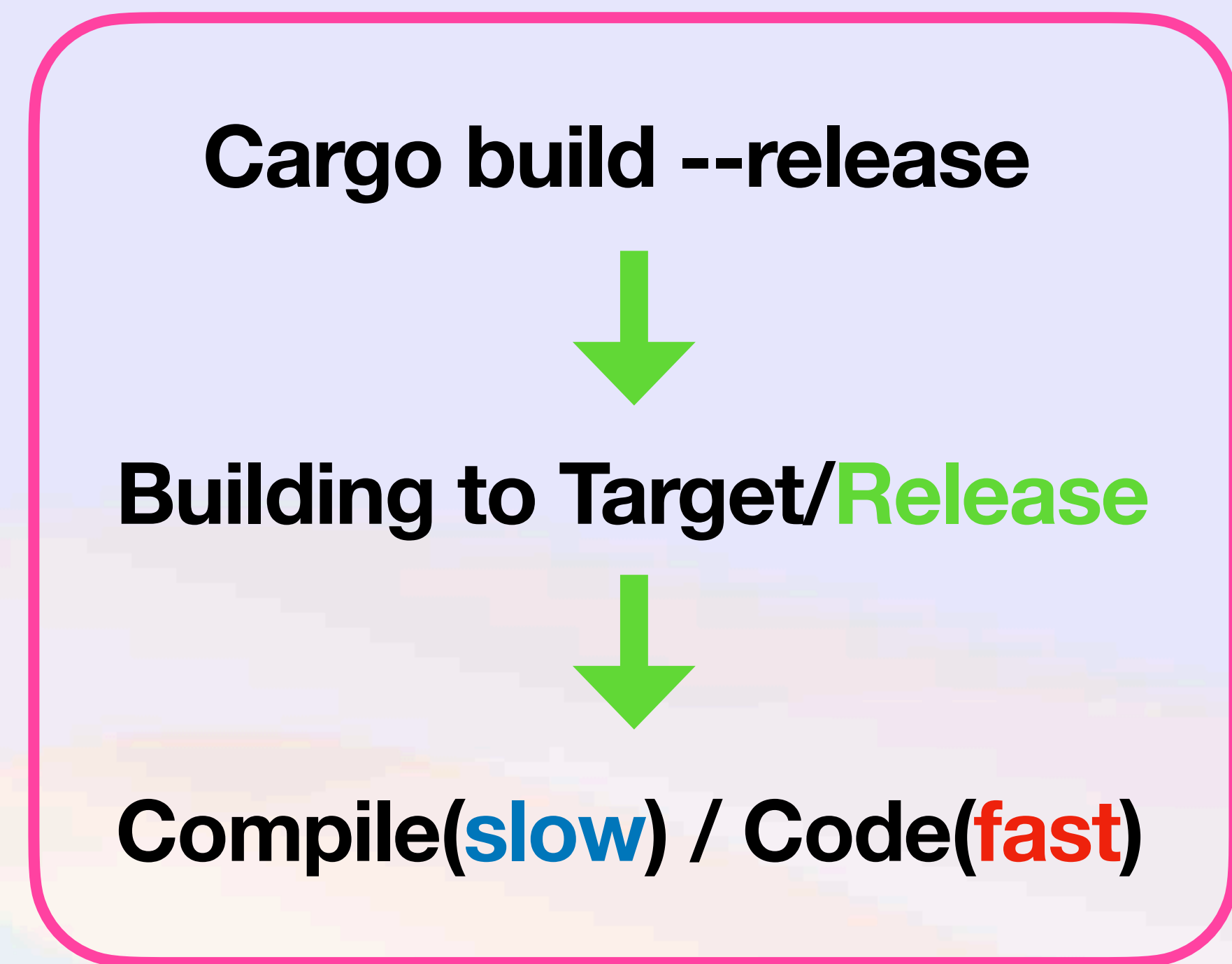


## II. 프로젝트 과정

### 4. Build 후 문제점 개선



Before




After



## II. 프로젝트 과정

### 4. Build 후 문제점 개선

  
v0.29.0


Q 감자

product 5,141

Api Key

Q Hits  
11

Time spent  
59 ms



NAME	춘천 감자빵 오리지널 (10개입)
META_TITLE	춘천 감자빵 오리지널 (10개입)
META_DESCRIPTION	춘천 감자발 카페의 원조 감자빵
_ID	62c2a65388c7dc00239bc156
DATE_CREATED	2022-07-04T08:35:31.011Z
DATE_UPDATED	2022-09-02T03:26:34.292Z
IMAGES	array
DIMENSIONS	json
DESCRIPTION	※ 9/2(금) 오전 11시 주문 마감됩니다. 이후 주문건은 추석 연휴가 지난 9/13(화) 부터 순차 출고 됩니다. 감자발 춘천 감자빵 오리지널 10개입 전국에서 먹으러 오는 감자빵 원조. cafe 감자발 오리지널 감자빵 10개입 국내산 통감자를 으깨 넣어 고소하고 밀가루 대신 쌀가루로 만들어 쫄득한 식감을 살렸습니다. 포슬포슬한

# III. 프로젝트 마무리

# III. 프로젝트 마무리

## 1. 성능 평가





# III. 프로젝트 마무리

## 1. 성능 평가

### Precision at K

#### Precision and recall at k: Definition

*Precision at k is the proportion of recommended items in the top-k set that are relevant*

Its interpretation is as follows. Suppose that my precision at 10 in a top-10 recommendation problem is 80%. This means that 80% of the recommendation I make are relevant to the user.

Mathematically precision@k is defined as follows:

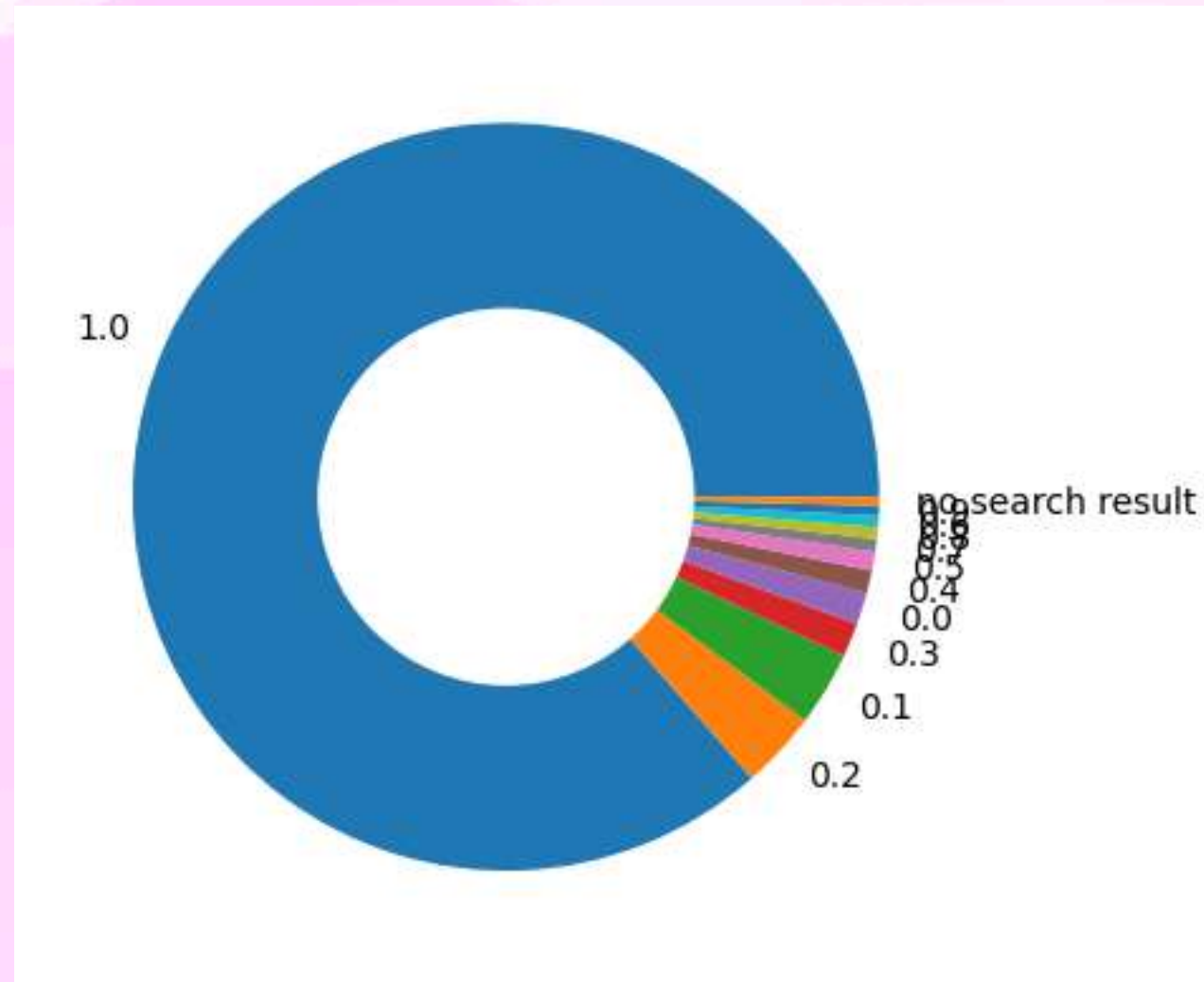
$$\text{Precision@k} = (\# \text{ of recommended items @k that are relevant}) / (\# \text{ of recommended items @k})$$

**P@K (Precision at k) : k개의 상위 검색 결과 중 정답과 연관된 결과가 몇 개가 있는 지 나타내는 지표**

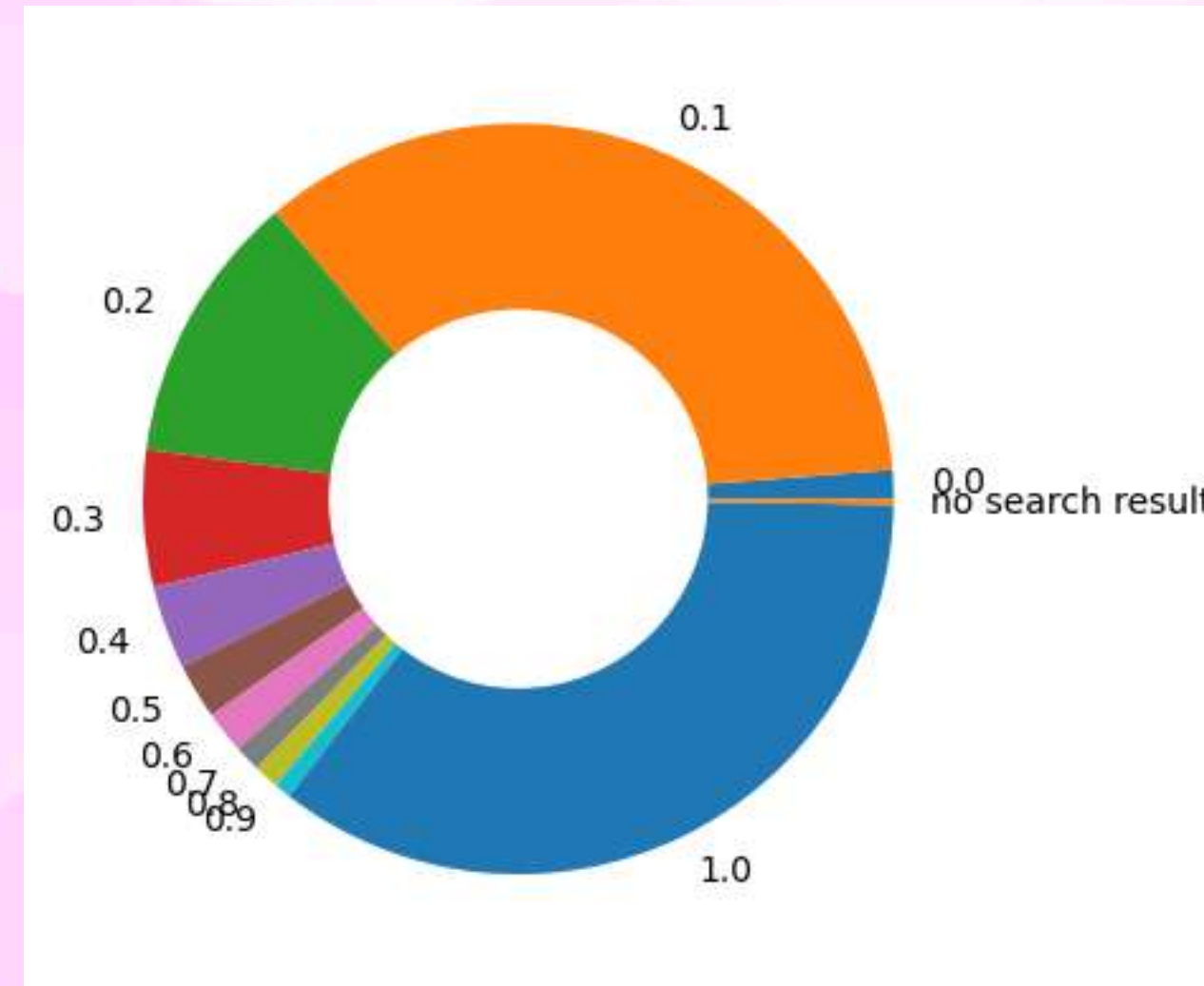
# III. 프로젝트 마무리

## 1. 성능 평가

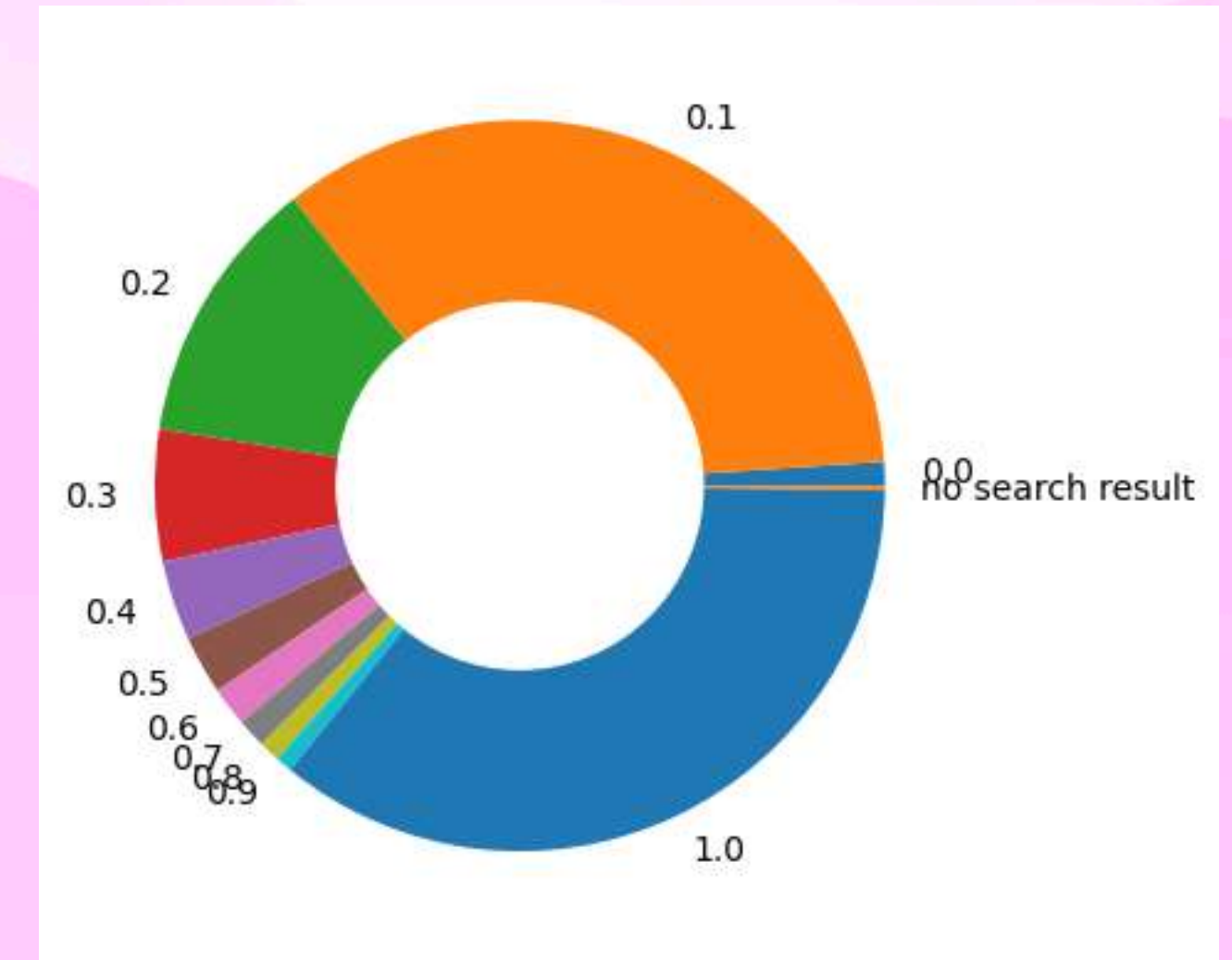
### Precision at K - Scores



기존 검색엔진



형태소 분석만 적용



형태소 분석 + 단어사전

# III. 프로젝트 마무리

## 1. 성능 평가

Q 경추통증

product 5,141

Api Key

Q Hits  
~ 43

Time spent  
139 ms

NAME	리메드 자기장 치료기 (Salus Talent)
META_TITLE	리메드 자기장 치료기 (Salus Talent)
META_DESCRIPTION	20년 이상의 임상과 기술적 노하우가 축적된 전자기장 장비의 표준
DESCRIPTION	20년 이상의 임상과 기술적 노하우가 축적된 전자기장 장비의 표준 리메드 자기장 치료기 (Salus Talent)의 장점 세계 최초로 레이저를 결합한 치료기 (자기장+레이저) 통증없이 심부 깊숙이 자기 자극을 통해 통증의 원인 치료 가능 세계 특허를 받은 오일 냉각 방식으로 소음 및 열이 현저히 감소 이런 증상에 좋아요 근골격계 손상에 의한 통증 퇴행성 관절염, 류마티스성 관절염에 의한 통증 근육이완 오십견 등 건부 통증 및 경추 통증 신경장애에 의한 통증 말초 신경 손상에 의한 통증 급성, 만성 요통 및 좌골 신경통 사양 자기장 출력 2 Tesla 20% 자기장 주파수 Symmetric Biphasic Pulse, 1~50 Hz 자기장 치료 모드 4 Manual Mode, 4 Auto Modes 치료 시간 1~60 Min 인터페이스 6 Buttons, 1 Jog Shuttle 장비 치수 40 x 53.5 x 119.3 cm 장비 중량 50 Kg



실질적 검색과 관련 X



# III. 프로젝트 마무리

## 1. 성능 평가

Q 경추통증

product 5,141

Api Key

Q Hits  
~ 43

Time spent  
139 ms

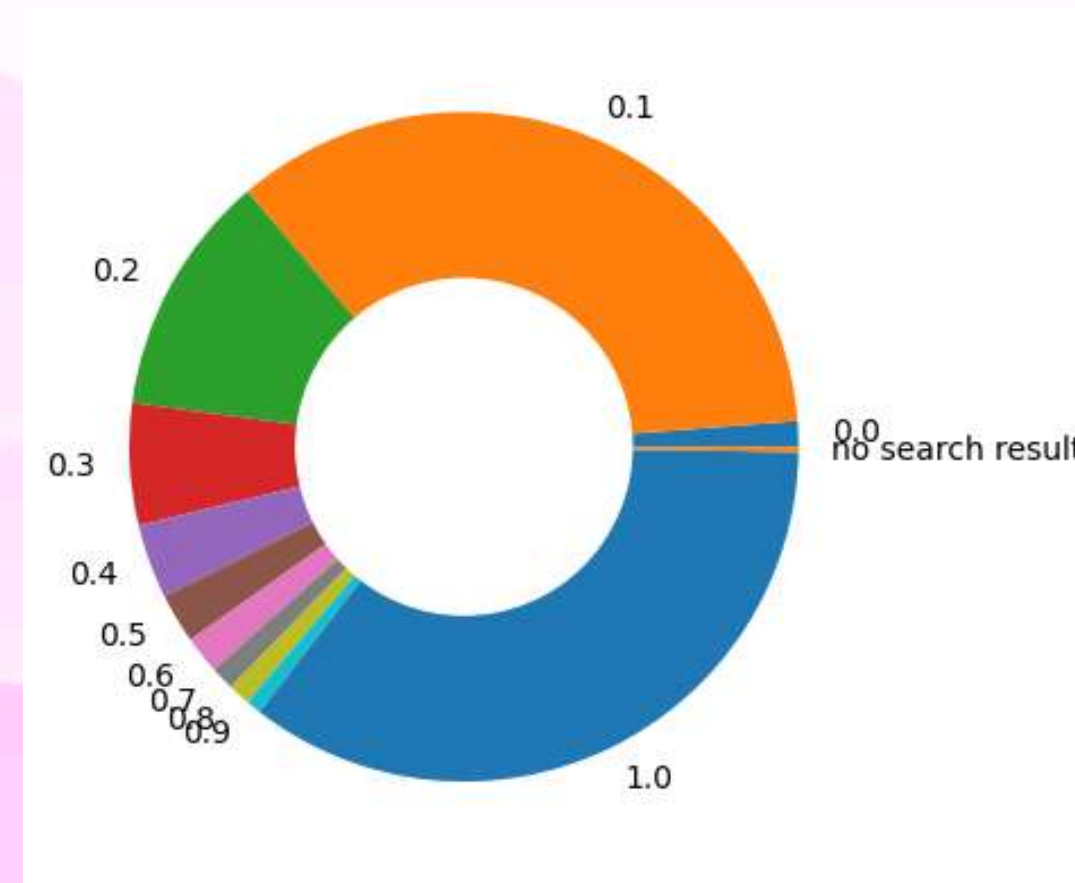
NAME	리메드 자기장 치료기 (Salus Talent)
META_TITLE	리메드 자기장 치료기 (Salus Talent)
META_DESCRIPTION	20년 이상의 임상과 기술적 노하우가 축적된 전자기장 장비의 표준
DESCRIPTION	20년 이상의 임상과 기술적 노하우가 축적된 전자기장 장비의 표준 리메드 자기장 치료기 (Salus Talent)의 장점 세계 최초로 레이저를 퓨전하여 치료 시너지 강화 Smart한 원터치 자동연결 치료모드(자기장+레이저) 통증없이 심부 깊숙이 자기 자극을 통해 통증의 원인 치료 가능 세계 특허를 받은 오일 냉각 방식으로 소음 및 열이 현저히 감소 이런 증상에 좋아요 근골격계 손상에 의한 통증 퇴행성 관절염, 류마티스성 관절염에 의한 통증 근육이완 오십견 등 건부 통증 및 경추 통증 신경장애에 의한 통증 말초 신경 손상에 의한 통증 급성, 만성 요통 및 좌골 신경통 사양 자기장 출력 2 Tesla 20% 자기장 주파수 Symmetric Biphasic Pulse, 1~50 Hz 자기장 치료 모드 4 Manual Mode, 4 Auto Modes 치료 시간 1~60 Min 인터페이스 6 Buttons, 1 Jog Shuttle 장비 치수 40 x 53.5 x 119.3 cm 장비 중량 50 Kg

특정 속성 검색 제외

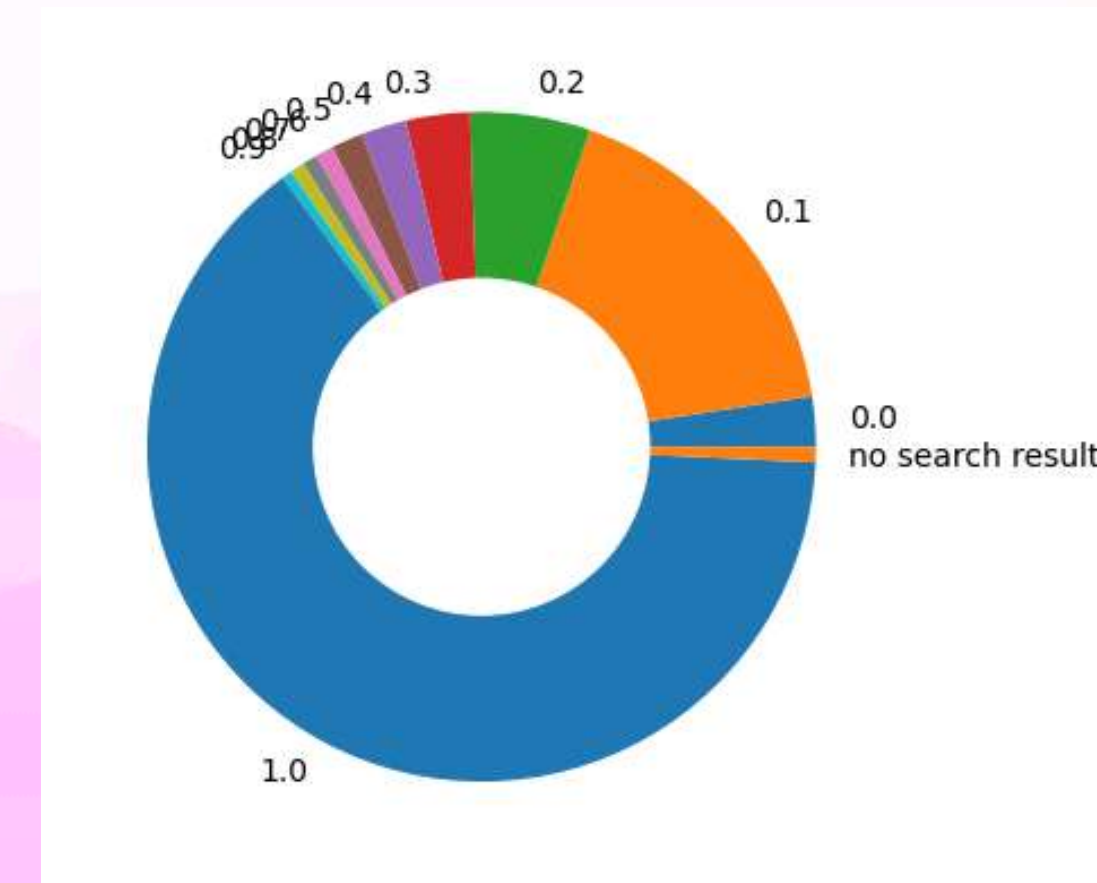
# III. 프로젝트 마무리

## 1. 성능 평가

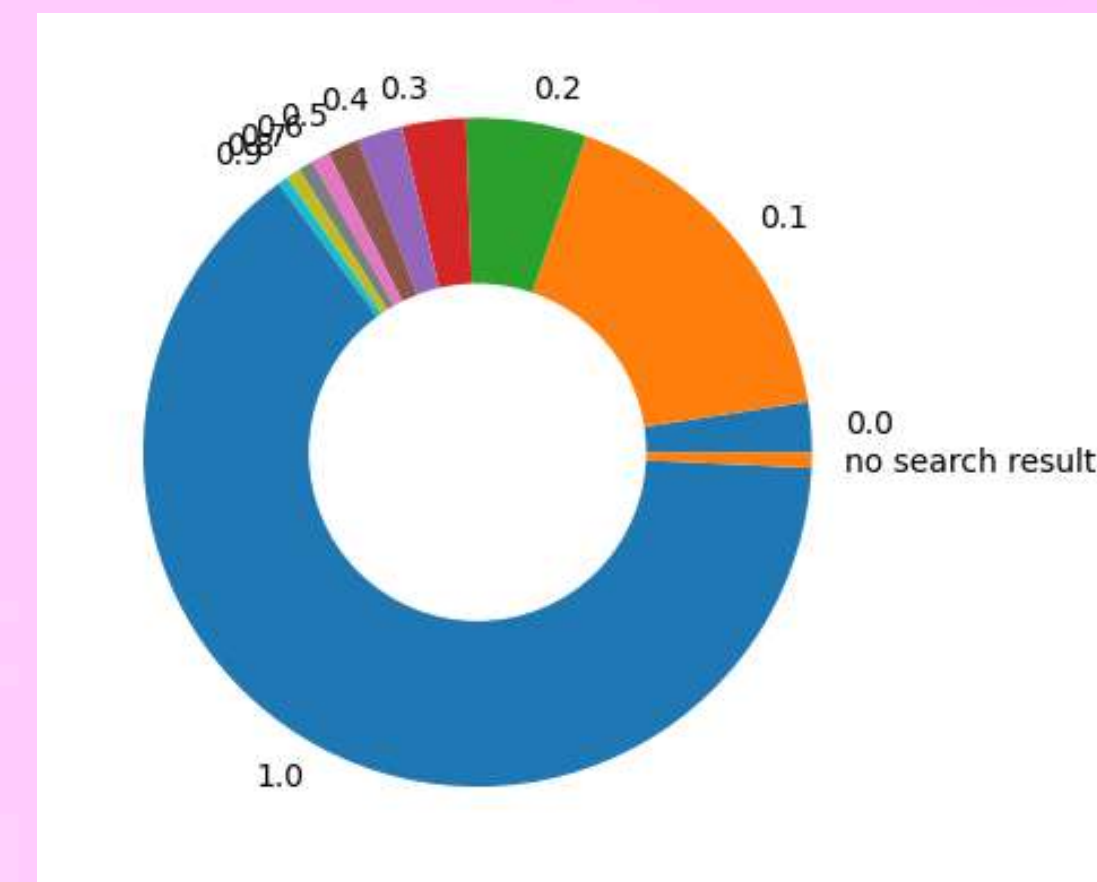
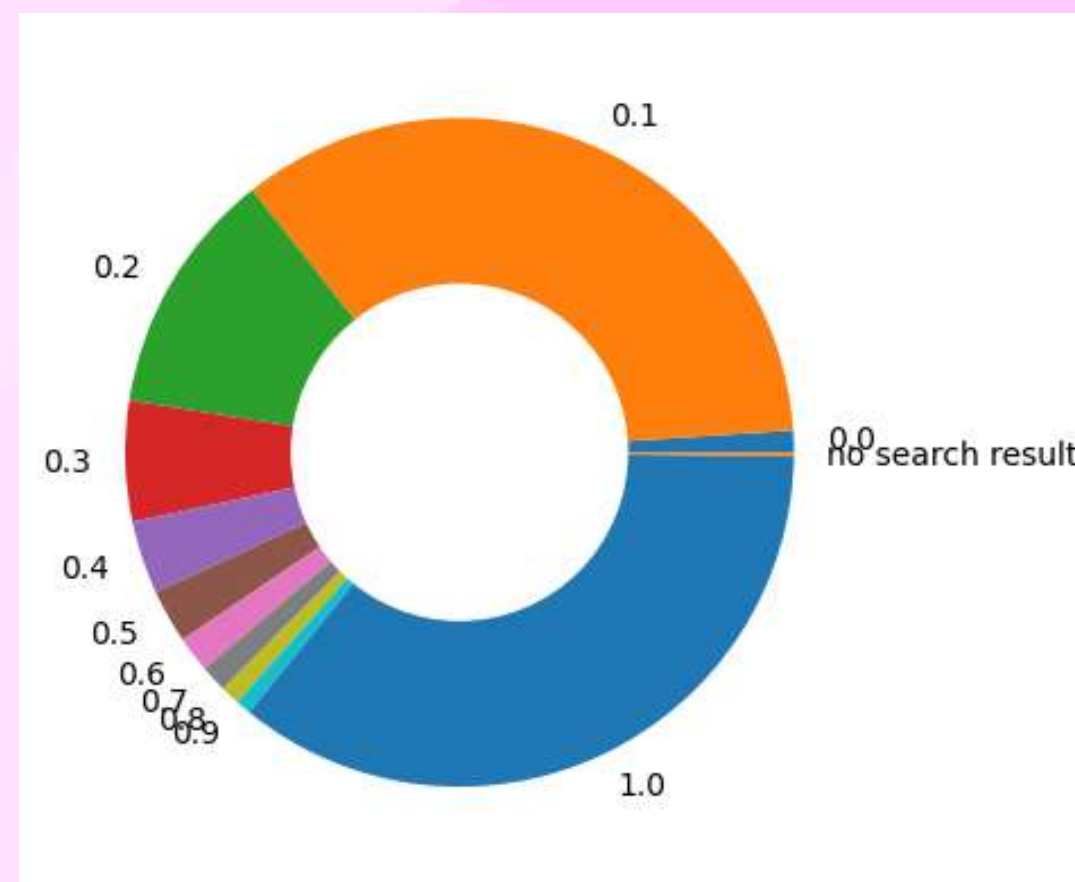
형태소 분석만 적용



Setting 후



형태소 분석 + 단어사전



# III. 프로젝트 마무리

## 1. 성능 평가

형태소 분석만 적용

```
<< P@k (precision at k) >>
1.0      569
0.1      565
0.2      189
0.3       95
0.4       58
0.5       39
0.6       29
0.0       20
0.7       19
0.8       17
0.9       12
no search result  4
```

Setting 후

```
<< P@k (precision at k) >
1.0      1036
0.1      280
0.2       94
0.3       50
0.0       39
0.4       35
0.5       25
0.6       14
0.7       12
no search result  12
0.8       11
0.9        8
```

형태소 분석 + 단어사전

```
<< P@k (precision at k) >>
1.0      576
0.1      559
0.2      192
0.3       94
0.4       57
0.5       42
0.6       28
0.7       21
0.0       17
0.8       15
0.9       12
no search result  3
```

```
<< P@k (precision at k) >>
1.0      1036
0.1      280
0.2       94
0.3       50
0.0       39
0.4       35
0.5       25
0.6       14
0.7       12
no search result  12
0.8       11
0.9        8
```



# III. 프로젝트 마무리

## 1. 성능 평가

### Precision: the first part of the F1 score

Precision is the first part of the F1 Score. It can also be used as an individual machine learning metric. It's formula is shown here:

$$Precision = \frac{\# \text{ of True Positives}}{\# \text{ of True Positives} + \# \text{ of False Positives}}$$

### Recall: the second part of the F1 score

Recall is the second component of the F1 Score, although recall can also be used as an individual machine learning metric. The formula for recall is shown here:

$$Recall = \frac{\# \text{ of True Positives}}{\# \text{ of True Positives} + \# \text{ of False Negatives}}$$

## F-1 Score

### F1 score formula

The F1 score is defined as the harmonic mean of precision and recall.

*As a short reminder, the harmonic mean is an alternative metric for the more common arithmetic mean. It is often useful when computing an average rate.*

In the F1 score, we compute the average of precision and recall. They are both rates, which makes it a logical choice to use the harmonic mean. The F1 score formula is shown here:

This makes that the formula for the F1 score is the following:

$$F1 \text{ score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Precision (정밀도) : Positive로 예상한 결과 중, 실제로도 Positive인 비율 =  $TP / (TP + FP)$

Recall (재현율) : Positive로 나온 결과 중, Positive로 예측했던 값의 비율 =  $TP / (TP + FN)$

검색엔진 성능평가에 사용하는 기준 :

- TP : 데이터셋의 검색어를 입력했을 때 실제로도 검색 결과로 제대로 나온 경우
- FP : 검색어가 들어있지 않지만 검색 결과로 나온 경우
- FN : 데이터셋의 검색어를 입력했을 때 검색 결과 나오지 않은 경우

$$F1 - \text{score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$$

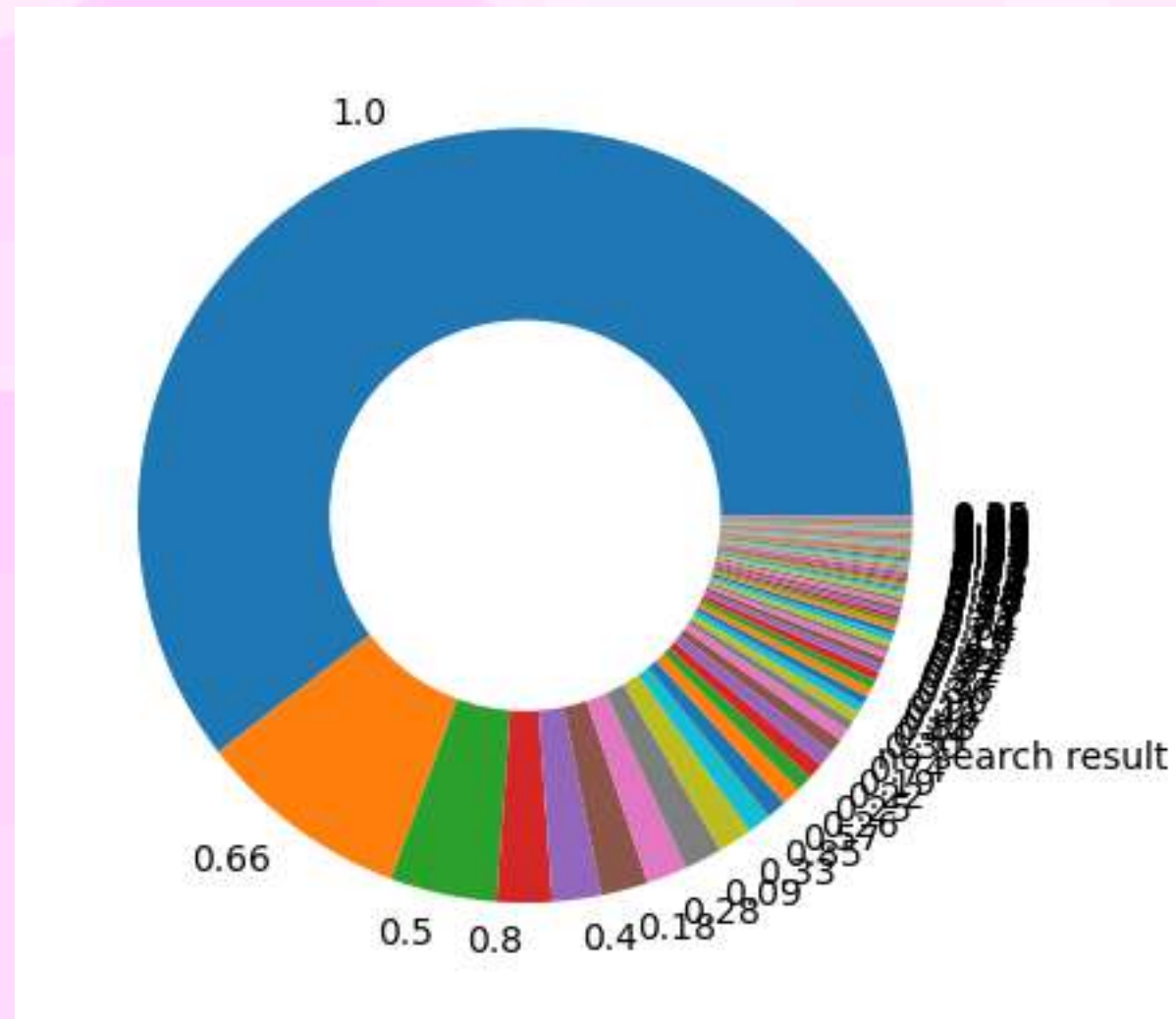
Precision(정밀도) & Recall(재현율)의 조화 평균

두 가지 수치를 모두 고려하는 지표로 정밀도와 재현율의 목표를 정하기 어려울 때 사용

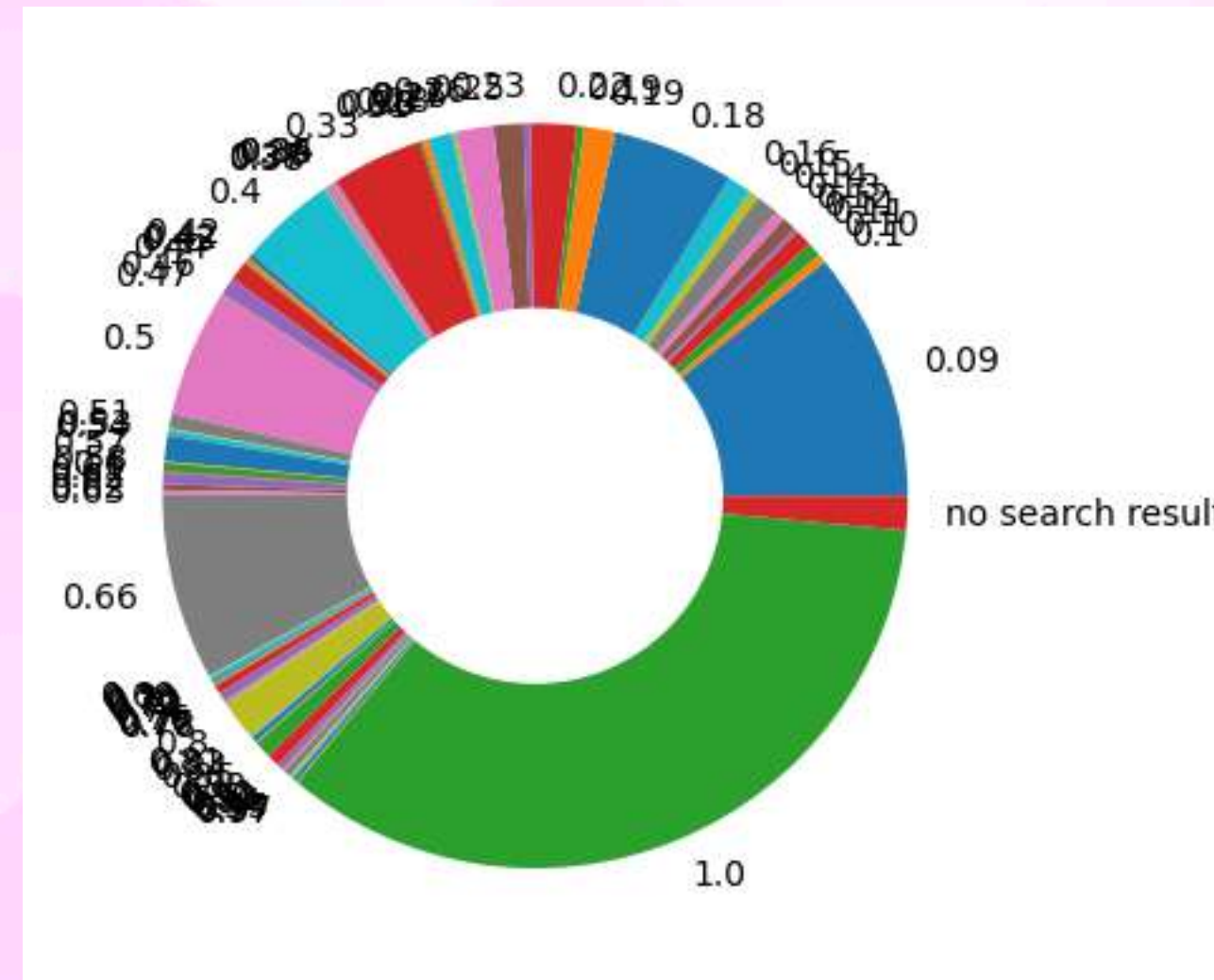
# III. 프로젝트 마무리

## 1. 성능 평가

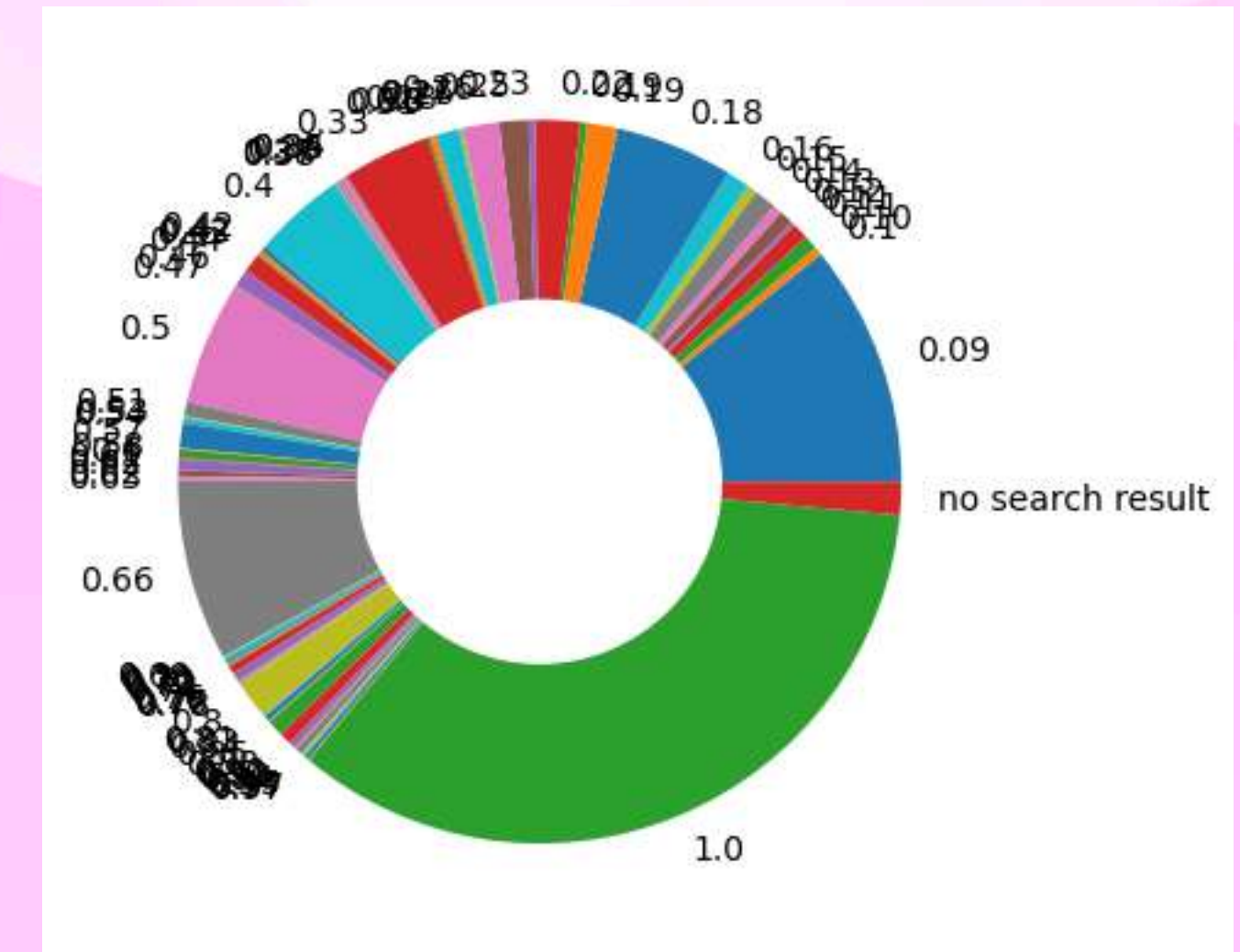
### F1-Score



기존 검색엔진



형태소 분석만 적용



형태소 분석 + 단어사전

# III. 프로젝트 마무리

## 1. 성능 평가

### F1-Score

```
<< F1 - score >>
1.000000    558
0.095238    175
0.666667    131
0.500000     90
0.181818     85

...
0.470588     1
0.266667     1
0.782609     1
0.916667     1
0.533333     1
```

형태소 분석만 적용

```
<< F1 - score >>
1.000000    558
0.095238    175
0.666667    131
0.500000     90
0.181818     85

...
0.470588     1
0.266667     1
0.782609     1
0.916667     1
0.533333     1
```

형태소 분석 + 단어사전



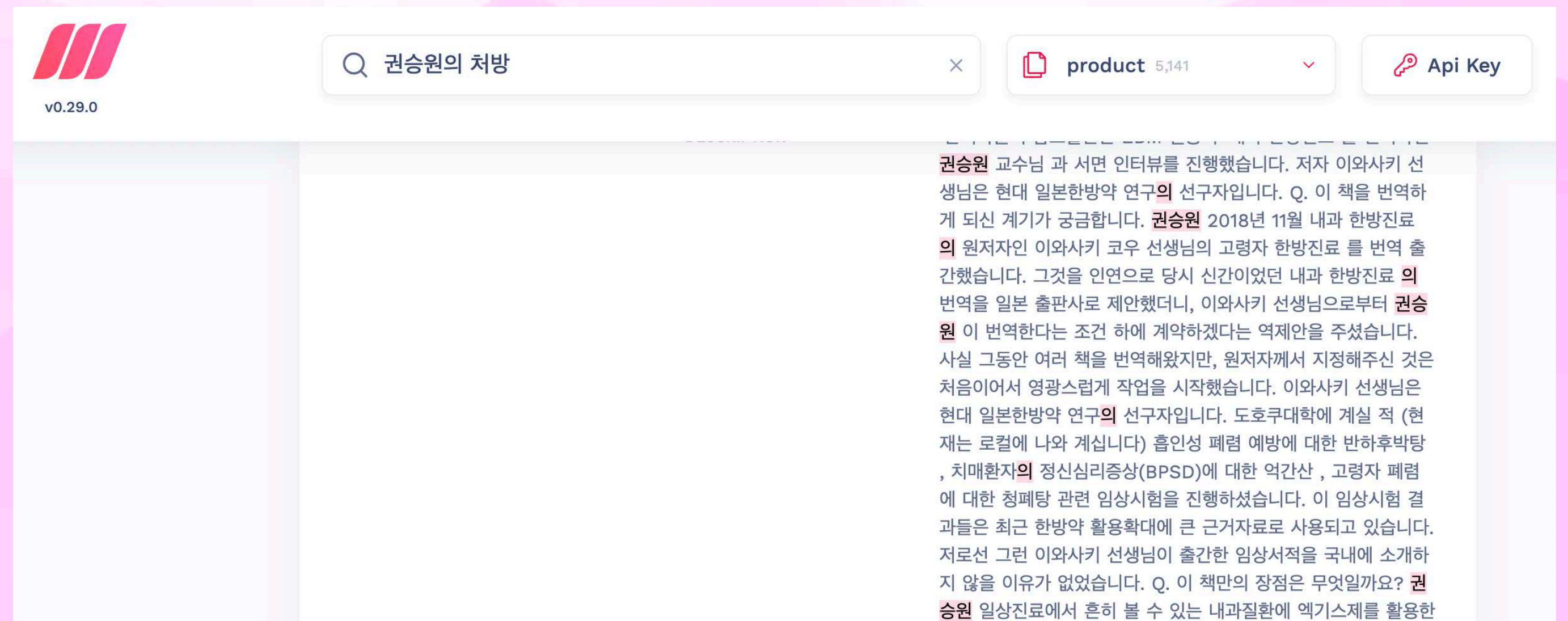
# III. 프로젝트 마무리

## 1. 성능 평가

기본통증진료학  
근육뼈대계통  
피부질환  
임상아틀라스  
사상의학  
본초정의  
간섭파자극기  
전이단계와  
수치료기  
힐세리온  
이카이로  
추나테이블  
폐기물통  
정기배송  
오양병원  
블랙에디션  
경방임증지남  
제피로스  
공기정화살균기  
크라시에  
어깨치료

← 차이 존재 →

Testset  
(단어 사전)



검색엔진 데이터

### III. 프로젝트 마무리

## 1. 성능 평가

# Testset

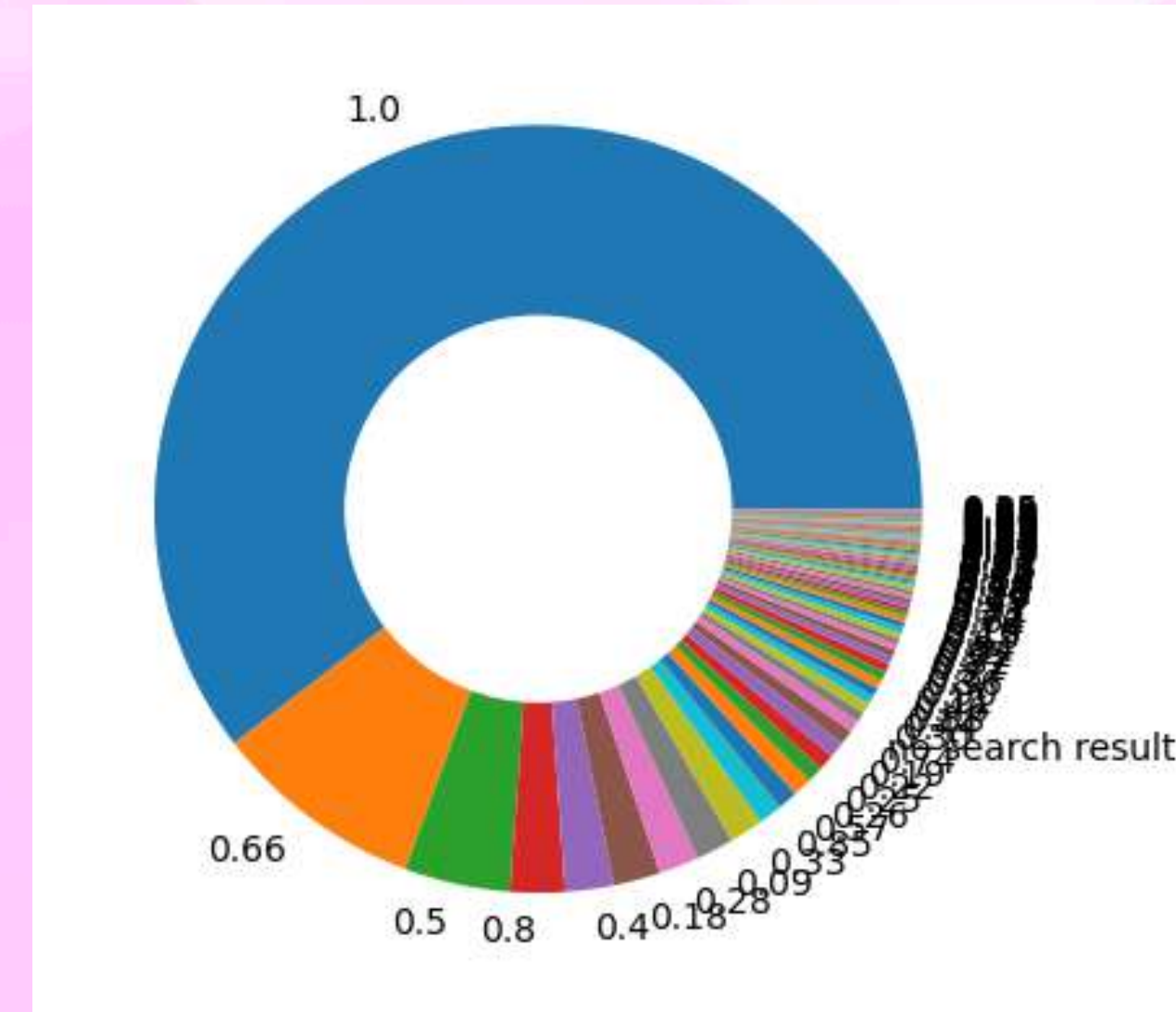
## 데이터 선별 작업



기본통증진료학  
근육뼈대계통  
피부질환  
임상아틀라스  
사상의학  
본초정의  
간섭파자극기  
전이단계와  
수치료기  
힐세리온  
이카이로  
추나테이블  
폐기물통  
정기배송  
요양병원  
블랙에디션  
경방임증지남  
제피로스  
공기정화살균기  
크라시에  
어깨치료



# Testset 재구축




## 재평가 필요



# III. 프로젝트 마무리

## 2-1. 부가적 문제 - 불용어

  
v0.29.0

Q 권승원의 처방

product 5,141


Api Key

권승원 교수님 과 서면 인터뷰를 진행했습니다. 저자 이와사키 선생님은 현대 일본한방약 연구의 선구자입니다. Q. 이 책을 번역하게 되신 계기가 궁금합니다. 권승원 2018년 11월 내과 한방진료의 원저자인 이와사키 코우 선생님의 고령자 한방진료 를 번역 출간했습니다. 그것을 인연으로 당시 신간이었던 내과 한방진료 의 번역을 일본 출판사로 제안했더니, 이와사키 선생님으로부터 권승원 이 번역한다는 조건 하에 계약하겠다는 역제안을 주셨습니다. 사실 그동안 여러 책을 번역해왔지만, 원저자께서 지정해주신 것은 처음이어서 영광스럽게 작업을 시작했습니다. 이와사키 선생님은 현대 일본한방약 연구의 선구자입니다. 도호쿠대학에 계실 적 (현재는 로컬에 나와 계십니다) 흡인성 폐렴 예방에 대한 반하후박탕 , 치매환자의 정신심리증상(BPSD)에 대한 억간산 , 고령자 폐렴에 대한 청폐탕 관련 임상시험을 진행하셨습니다. 이 임상시험 결과들은 최근 한방약 활용확대에 큰 근거자료로 사용되고 있습니다. 저로선 그런 이와사키 선생님이 출간한 임상서적을 국내에 소개하지 않을 이유가 없었습니다. Q. 이 책만의 장점은 무엇일까요? 권승원 일상진료에서 흔히 볼 수 있는 내과질환에 엑기스제를 활용한



# III. 프로젝트 마무리

## 2-2. 부가적 문제 - 형태소 분석의 본질적 문제

  
v0.29.0

product 5,141

Api Key

산四逆散 298 025-1 시호소간산柴胡疏肝散 303 026 소요산逍遙散 304 026-1 가미소요산加味逍遙散 315 026-2 흑소요산黑逍遙散 315 027 통사요방痛瀉要方 316 제3절 조화장위調和腸胃 319 028 반하사심탕半夏瀉心湯 320 028-1 생강사심탕生薑瀉心湯 324 028-2 감초사심탕甘草瀉心湯 325 028-3 황련탕黃連湯 326 제4장 청열제淸熱劑 327 제1절 청기분열淸氣分熱 328 029 백호탕白虎湯 329 029-1 백호가인삼탕白虎加人蔘湯 332 029-2 백호가창출탕白虎加蒼朮湯 333 030 죽엽석고탕竹葉石膏湯 334 제2절 청열량혈淸營涼血 338 031 청영탕淸營湯 338 032 서각지황탕犀角地黃湯 344 제3절 청열해독淸熱解毒 348 033 황련해독탕黃連解毒湯 349 034 양격산涼膈散 352 035 보제소독음普濟消毒飲 356 036 선방활명음仙方活命飲 362 제4절 청장부열淸臟腑熱 366 037 도적산導赤散 366 038 용담사간탕龍膽瀉肝湯 374 039 좌금환左金丸 380 040 위경탕葦莖湯 384 041 사백산瀉白散 387 042 청위산淸胃散 392 043 옥녀전玉女煎 397 044 갈근금련탕葛根芩連湯 401 045 작약탕芍藥湯 405 046 백두옹탕白頭翁湯 410 제5절 청허열淸虛熱 414 047 청호별갑탕靑蒿鱉甲湯 414 048 당귀육황탕當歸六黃湯 420 제5장 거서제祛暑劑 424 049 향유산香?散 427 049-1 신가향유음新加香?飲 431 050 청서익기탕淸暑益氣湯 433 제6장 온리제溫裏劑 436 제1절 온중거한溫中祛寒 437 051 이중환理中丸 437 052 소건중탕小建中湯 441 053 오수유탕吳茱萸湯 446 제2절 회양구역回陽救逆 450 054 사역탕四逆湯 450 054-1 통맥사역탕通脈四逆湯 454 054-2 사역가인삼탕四逆加人蔘湯 456 054-3 백통탕白通湯 457 054-4 삼부탕蔘附湯 458 제3절 온경산한溫經

# III. 프로젝트 마무리

## 2-3. 부가적 문제 - 단어 사전의 한계



Prefix-Search 사용

### Example

Given a set of words in a dataset:

film cinema movies show harry potter shine musical

query: s :

response:

- show
- shine

but not

- movies
- musical

예시

# III. 프로젝트 마무리

## 2-3. 부가적 문제 - 단어 사전의 한계



형태소 단위



단어사전 기준



# III. 프로젝트 마무리

## 2-3. 부가적 문제 - 단어 사전의 한계



형태소 단위(검색 가능)



단어사전 기준(검색 불가)

### III. 프로젝트 마무리

### 3-1. 개선 방안 - 불용어 (등재 및 지속 관리)

```
client.index(indexes[index_num-1]).update_stop_words(list(set([
    "ㄱ","ㄴ","ㄷ","ㄹ","ㅁ","ㅂ","ㅅ","ㅇ","ㅈ","ㅊ","ㅋ","ㅌ","ㅍ","ㅎ","ㅏ","ㅑ","ㅓ","ㅕ","ㅗ","ㅛ","ㅜ","ㅠ","ㅡ","ㅣ","애","에","어","여","요","우","유","ㅡ","예","의","이","피","페","헤","은",
    "진다",
    "가",
    "가형",
    "각",
    "각각",
    "각자",
    "각종",
    "갈다",
    "같이",
    "거니와",
    "거바",
    "거의",
    "것",
    "것들",
    "게다가",
    "계우다",
    "겨우",|
    "검사검사",
    "고려하면",
    "고로",
    "곧",
    "공통으로",
    "과",
    "과연"
```

## Stop words

A set of words defined for an index. Because some words neither add semantic value nor context, you may want to ignore them from your search. Stop words are **ignored during search**.

```
stopWords=[<String>, <String>, ...]
```

- [`<String>`, `<String>`, ...] (Array of strings, defaults to `[]` )

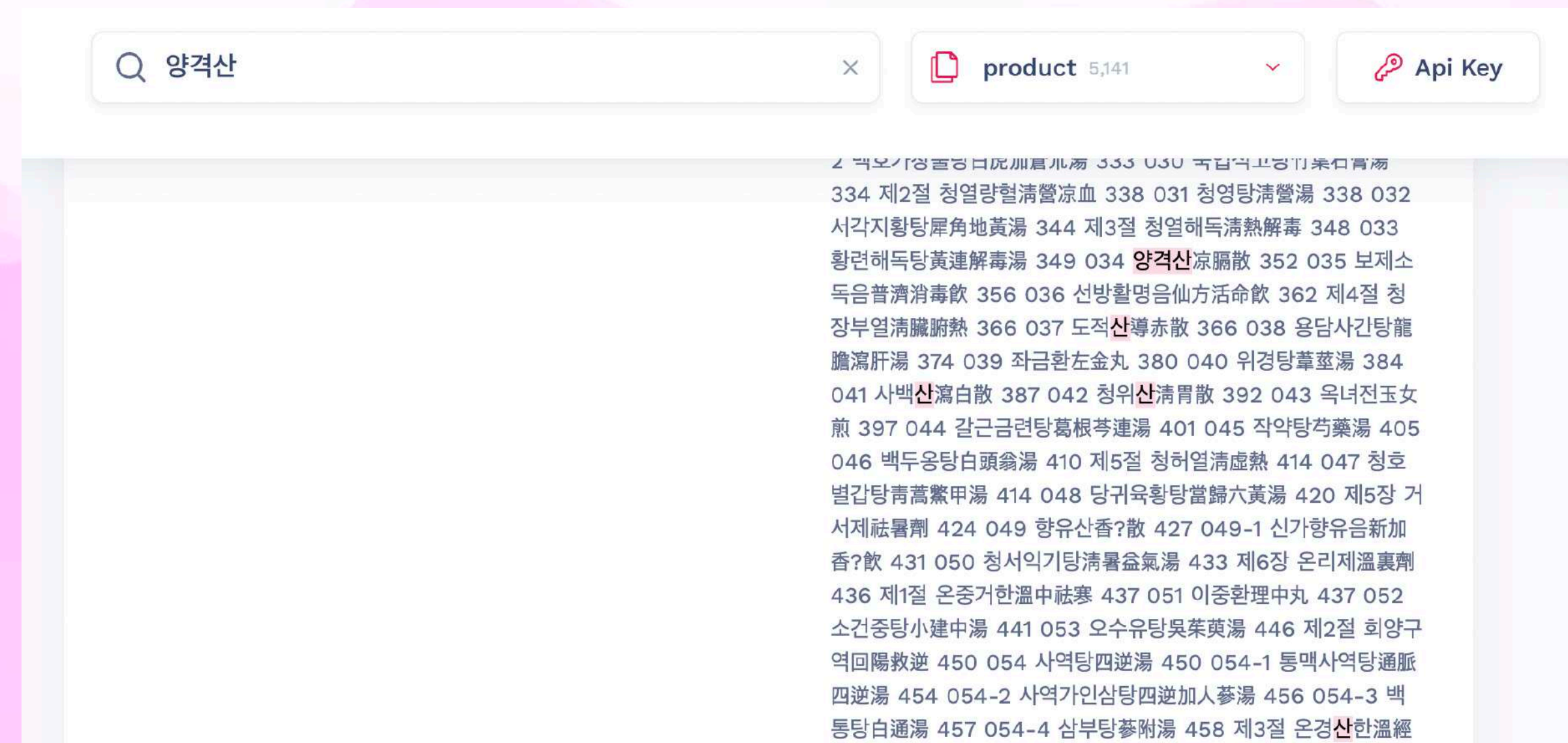
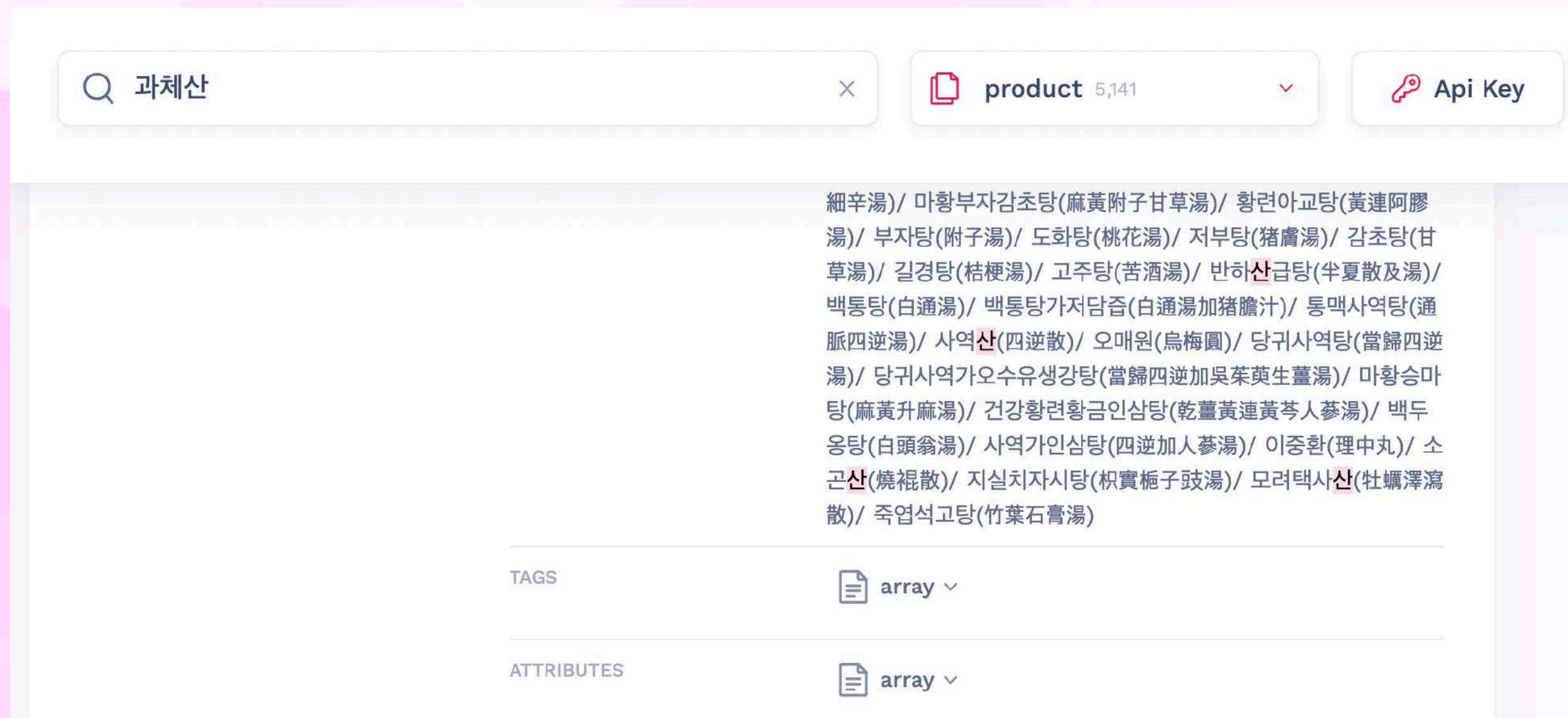
An array of strings that contains the stop words.

Learn more about stop words

## API 내 불용어 처리

### III. 프로젝트 마무리

### 3-2. 개선 방안 - 단어사전 (선별적 사전 등재)



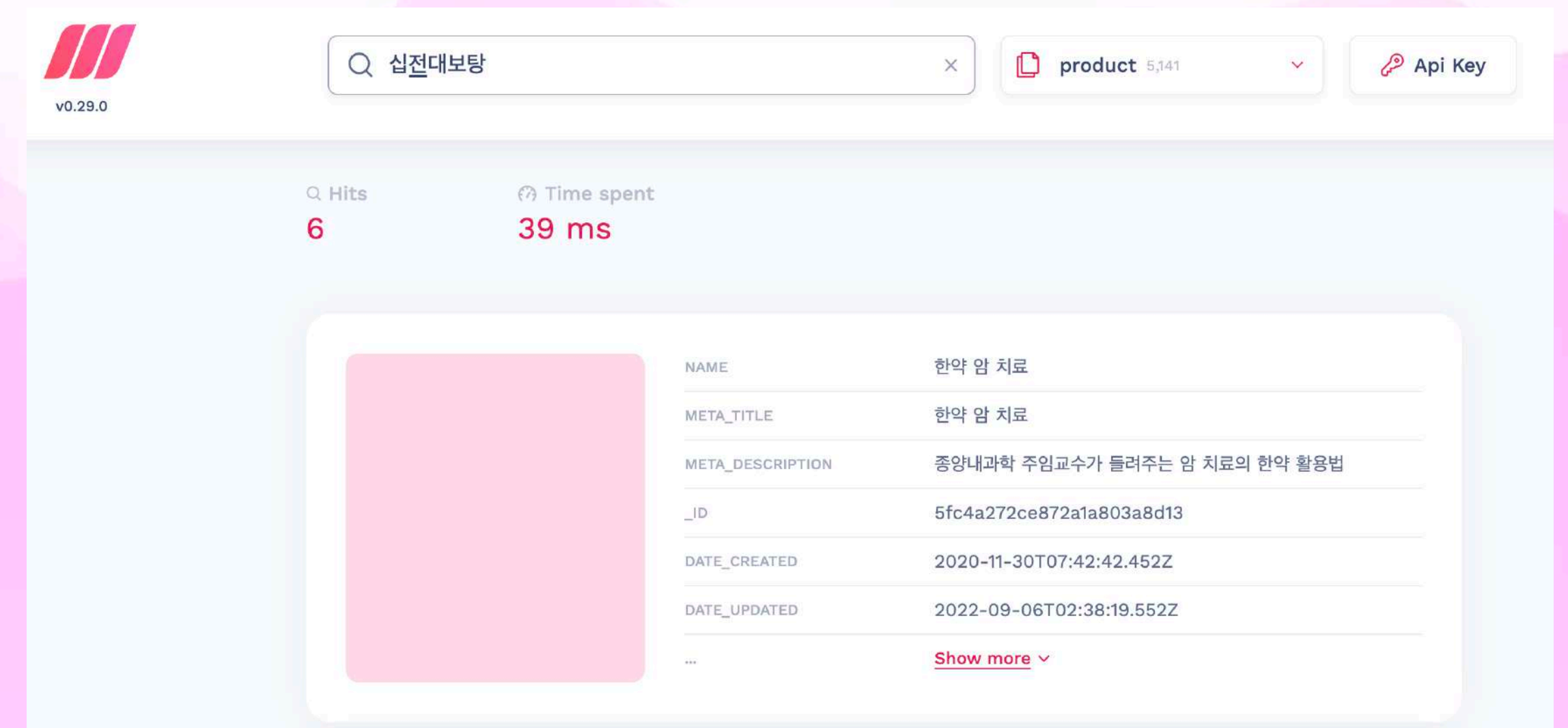
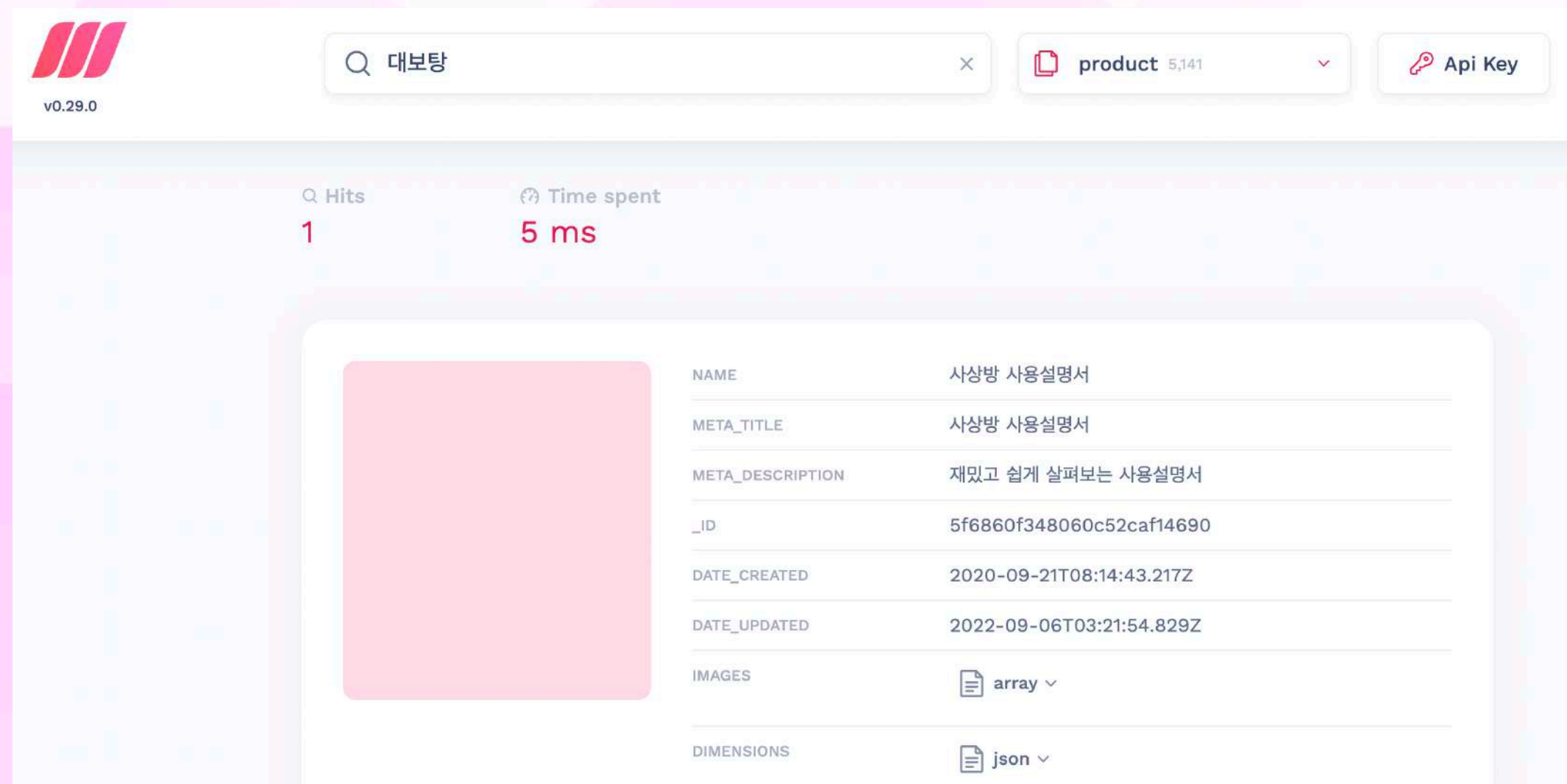
## 단순 복합 명사 (2개의 명사로 이루어진 복합 명사)

## 사전 등재 필요성 (~산/~탕/~환 등)



# III. 프로젝트 마무리

## 3-2. 개선 방안 - 단어사전 (선별적 사전 등재)



다수 복합 명사 (3개 이상의 명사)

사전 제외 필요성



김성수



서동국



안세호

Thank You!

박제윤  
(Mentor)