

Wello x DSL

Policy Text Generation

최종 발표

프로젝트2 -C팀

김하람 · 남영욱 · 조신형 · 최연수 · 편시현 · 허유진

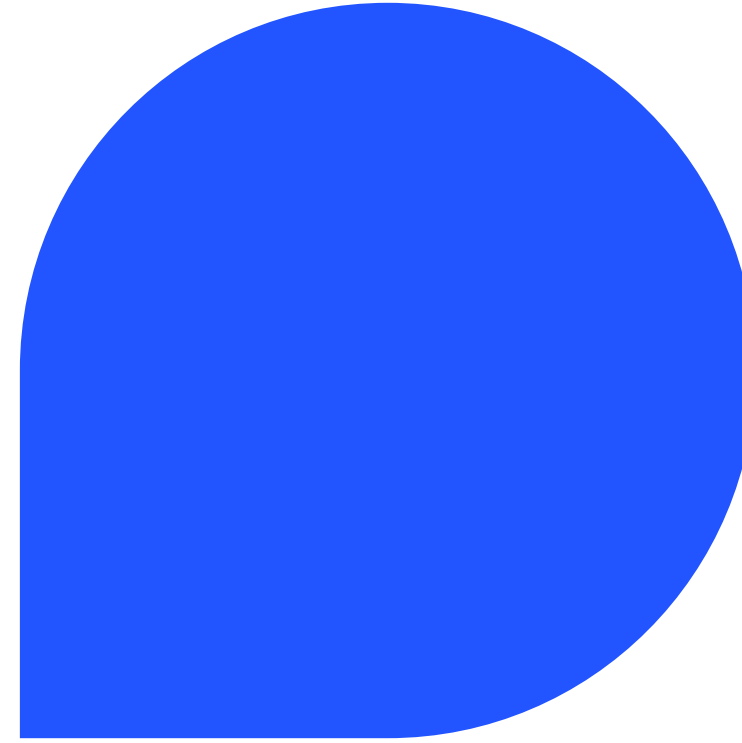
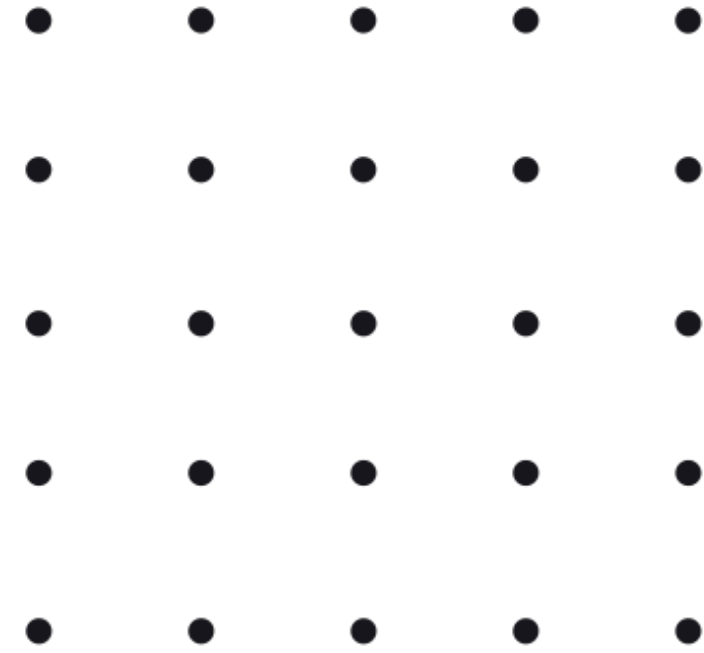


Final Project Wello x DSL

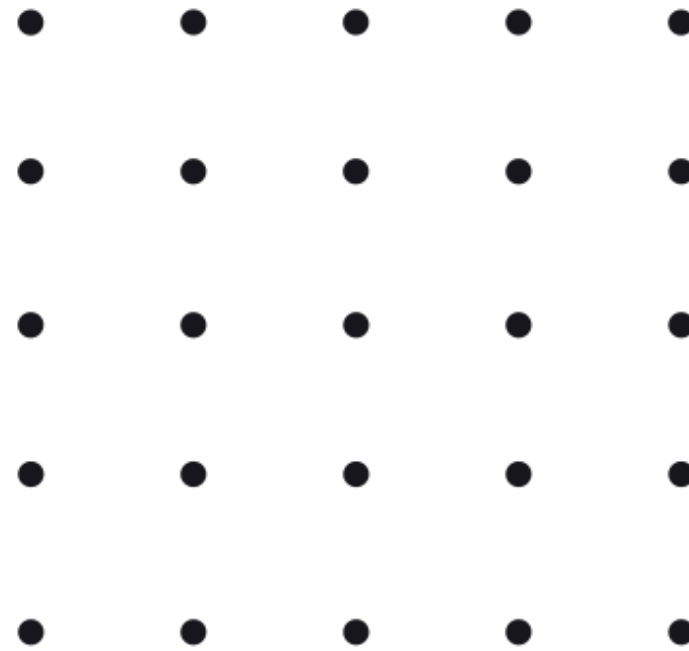
키워드 추출 With Word2vec

Contents

- 1 Project Ideation
- 2 Modeling
- 3 Modeling result
- 4 개선방향 및 한계



1 . Project Ideation



1. Project Ideation

정책의 핵심 키워드를 뽑자

정책 상세내용

청년우대형청약통장

저소득 무주택 청년의 주택구매 및 임차자금
마련 지원을 위해 재형 기능을 강화한
청약통장 도입

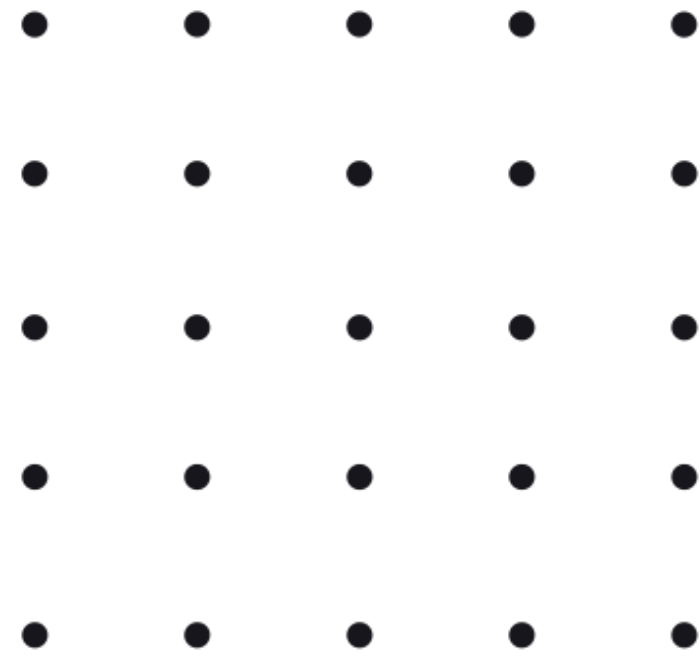
✓ 바로 신청하기

정책 상세내용

청년 취업교육서비스 제공

청년의 성공적인 자립에 필요한 다양한
직무교육과 기업연계, 멘토링 탐방 등의
취업연계 지원으로 실질적인 취.창업의 빠른
사회진입을 돕고 사회주체로의 성장을
지원하고자 함

✓ 바로 신청하기



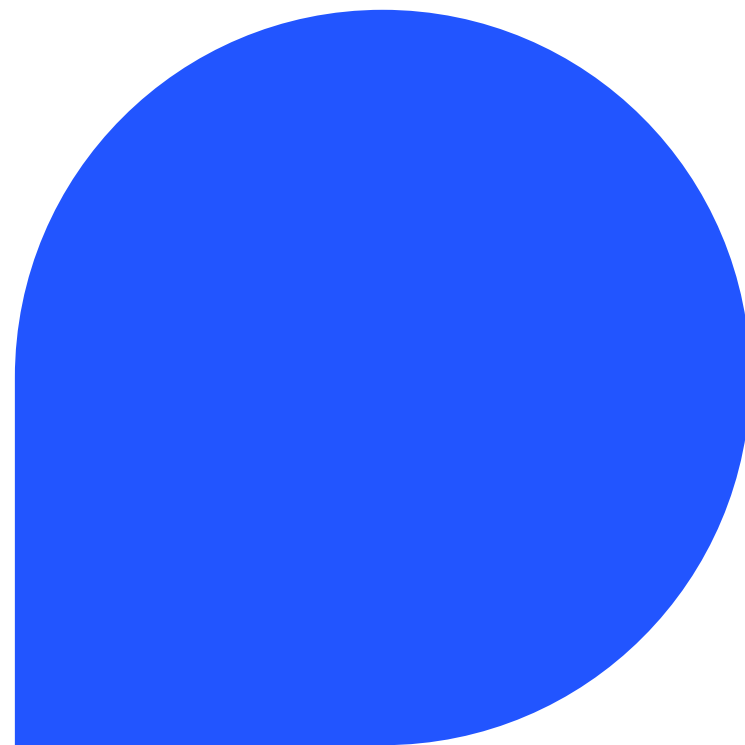
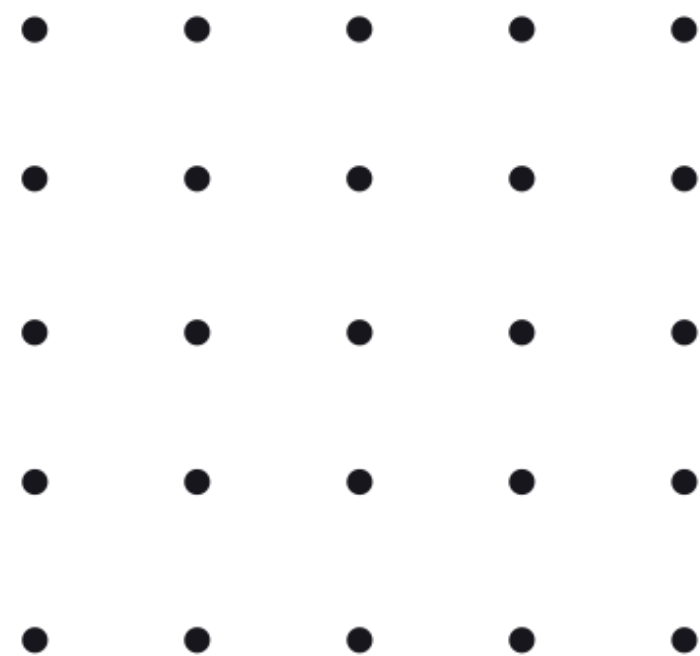
2. Modeling

2.1 Word2vec

2.2 Modeling

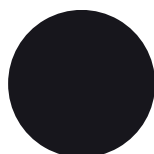
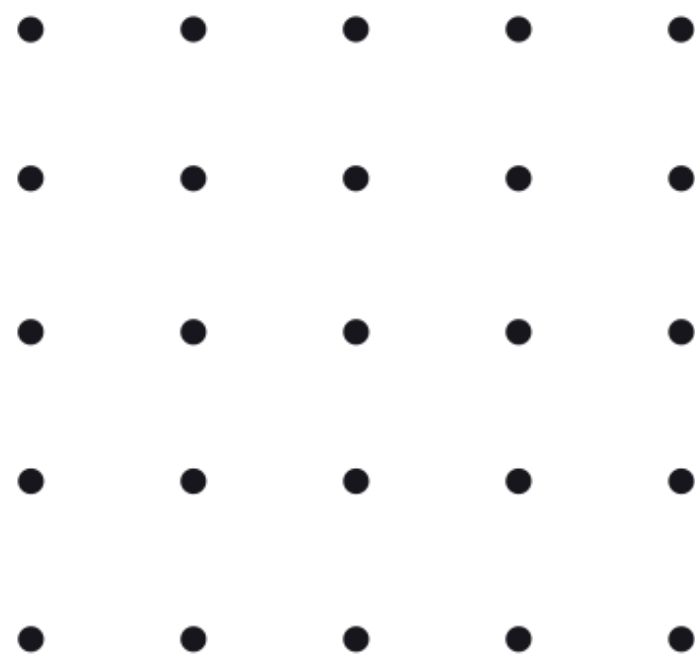
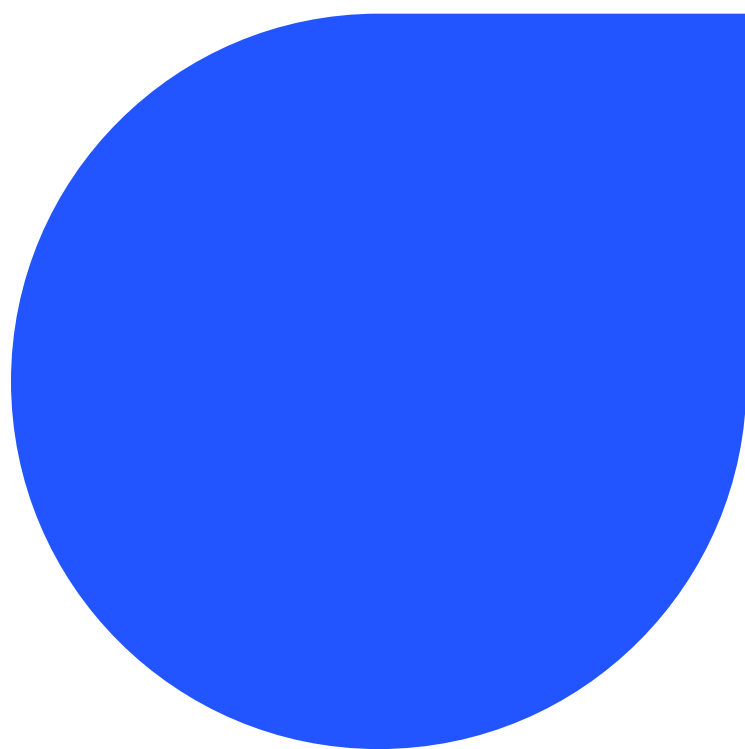
2.3 모델 개선





2-1 Word2vec

선택이유 및 소개



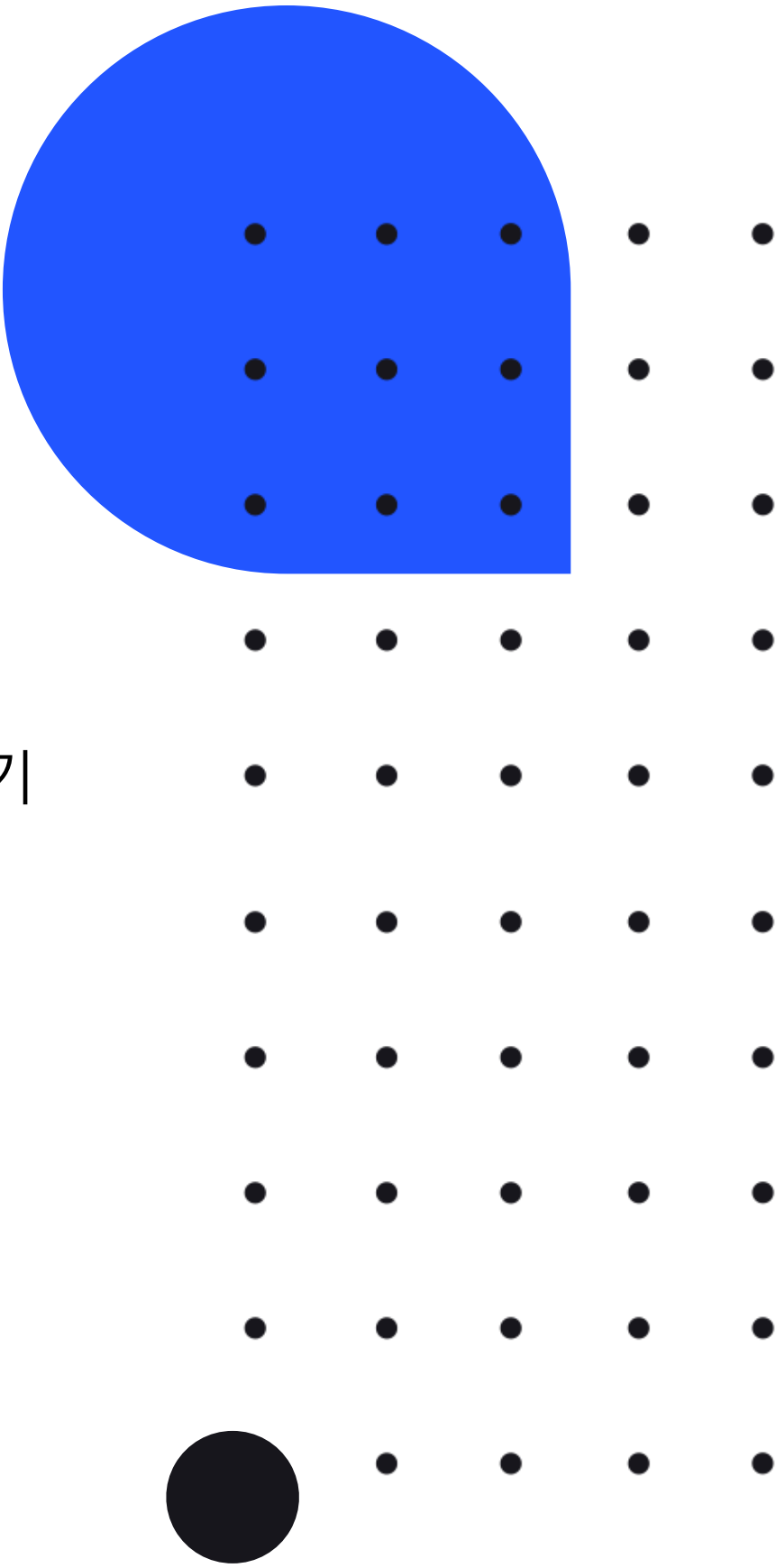
Word2Vec

단어의 '의미'를 함께 벡터화

“비슷한 위치의 단어는 비슷한 의미를 가진다”는 가정을 바탕으로
단어의 의미를 여러 차원에 분산하여 표현

C-BOW vs Skip-Gram

: 주변에 있는 단어로中间的의 단어 예측하기 vs 中间에 있는 단어로 주변의 단어 예측하기



중심 단어 주변 단어
↓ ↓
The fat cat sat on the mat
The fat cat sat on the mat
The fat cat sat on the mat
The fat cat sat on the mat
The fat cat sat on the mat
The fat cat sat on the mat
The fat cat sat on the mat

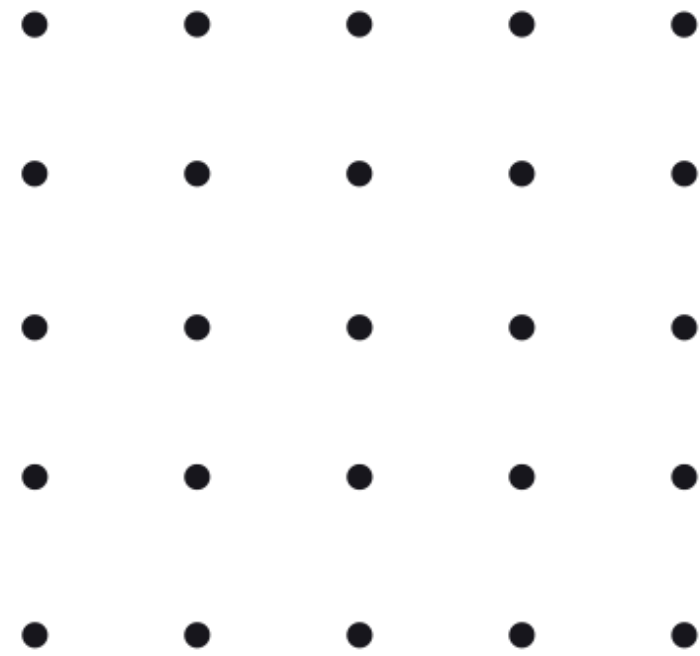
중심 단어	주변 단어
[1, 0, 0, 0, 0, 0, 0]	[0, 1, 0, 0, 0, 0, 0], [0, 0, 1, 0, 0, 0, 0]
[0, 1, 0, 0, 0, 0, 0]	[1, 0, 0, 0, 0, 0, 0], [0, 0, 1, 0, 0, 0, 0], [0, 0, 0, 1, 0, 0, 0]
[0, 0, 1, 0, 0, 0, 0]	[1, 0, 0, 0, 0, 0, 0], [0, 1, 0, 0, 0, 0, 0], [0, 0, 0, 1, 0, 0, 0], [0, 0, 0, 0, 1, 0, 0]
[0, 0, 0, 1, 0, 0, 0]	[0, 1, 0, 0, 0, 0, 0], [0, 0, 1, 0, 0, 0, 0], [0, 0, 0, 1, 0, 0, 0], [0, 0, 0, 0, 1, 0, 0]
[0, 0, 0, 0, 1, 0, 0]	[0, 0, 1, 0, 0, 0, 0], [0, 0, 0, 1, 0, 0, 0], [0, 0, 0, 0, 1, 0, 0], [0, 0, 0, 0, 0, 1, 0]
[0, 0, 0, 0, 0, 1, 0]	[0, 0, 0, 1, 0, 0, 0], [0, 0, 0, 0, 1, 0, 0], [0, 0, 0, 0, 0, 1, 0], [0, 0, 0, 0, 0, 0, 1]
[0, 0, 0, 0, 0, 0, 1]	[0, 0, 0, 1, 0, 0, 0], [0, 0, 0, 0, 1, 0, 0], [0, 0, 0, 0, 0, 1, 0]

C-BOW

중심 단어 주변 단어
↓ ↓
The fat cat sat on the mat
The fat cat sat on the mat
The fat cat sat on the mat
The fat cat sat on the mat
The fat cat sat on the mat
The fat cat sat on the mat
The fat cat sat on the mat

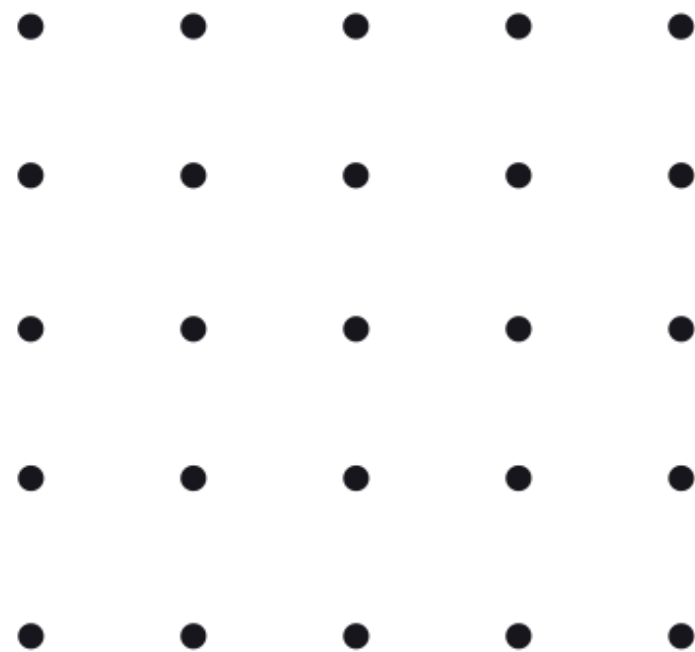
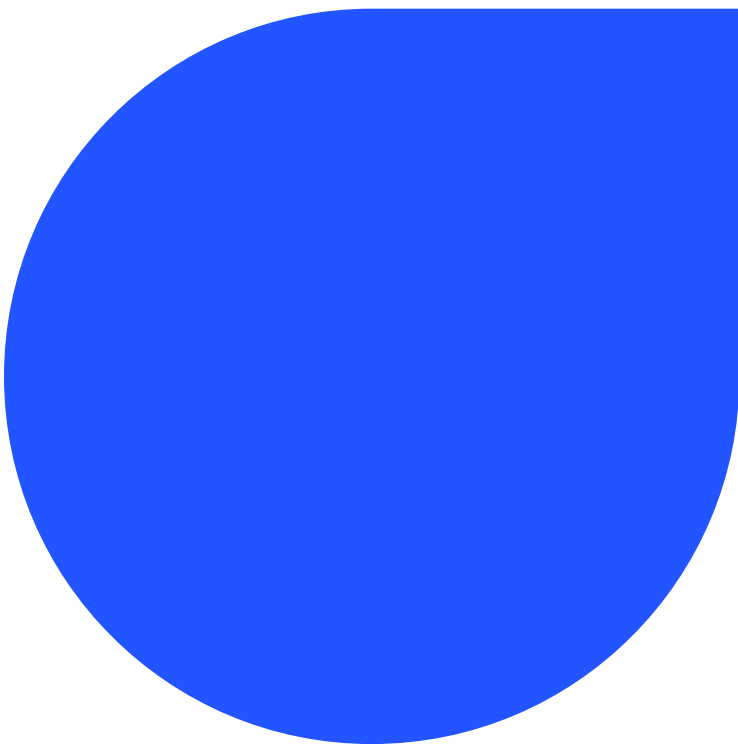
중심 단어	주변 단어
cat	The
cat	Fat
cat	sat
cat	on
sat	fat
sat	cat
sat	on
sat	the

Skip-Gram



2-2 Modeling

모델링 및 코드 흐름 소개



Modeling

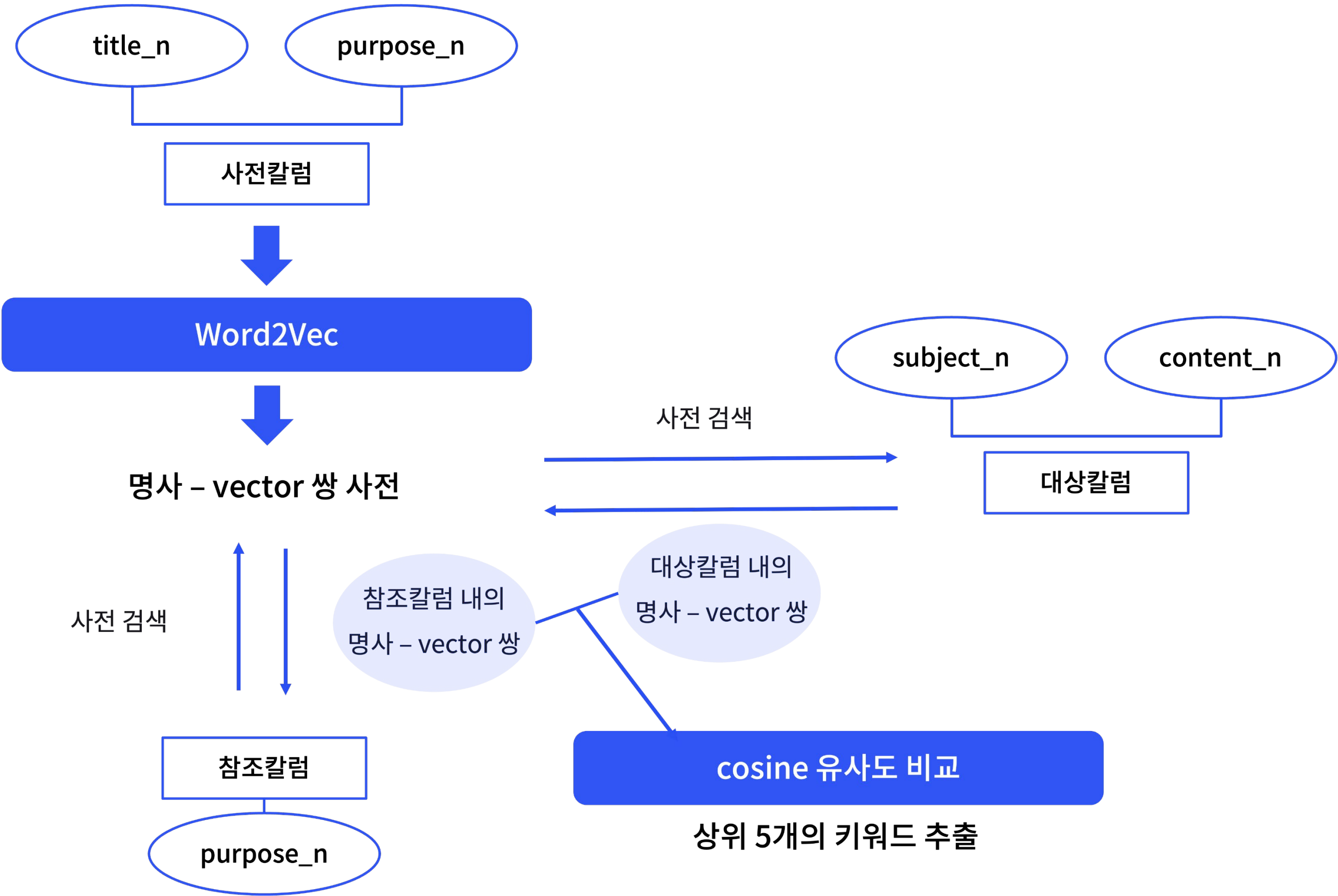
Word2Vec을 이용한 키워드 추출

0. 데이터: admin 페이지의 '서비스명', '서비스목적', '정책내용', '정책대상' 컬럼
1. POS(품사) 태깅: 각 컬럼에서 명사만 추출
2. 사전(dict)컬럼, 대상컬럼, 참조컬럼 지정
3. 사전컬럼으로부터 명사 - vector 쌍 사전 생성
4. 대상컬럼 명사 - vector 쌍 생성 & 참조컬럼 명사 - vector 쌍 생성
5. 코사인 유사도 기반 키워드 추출



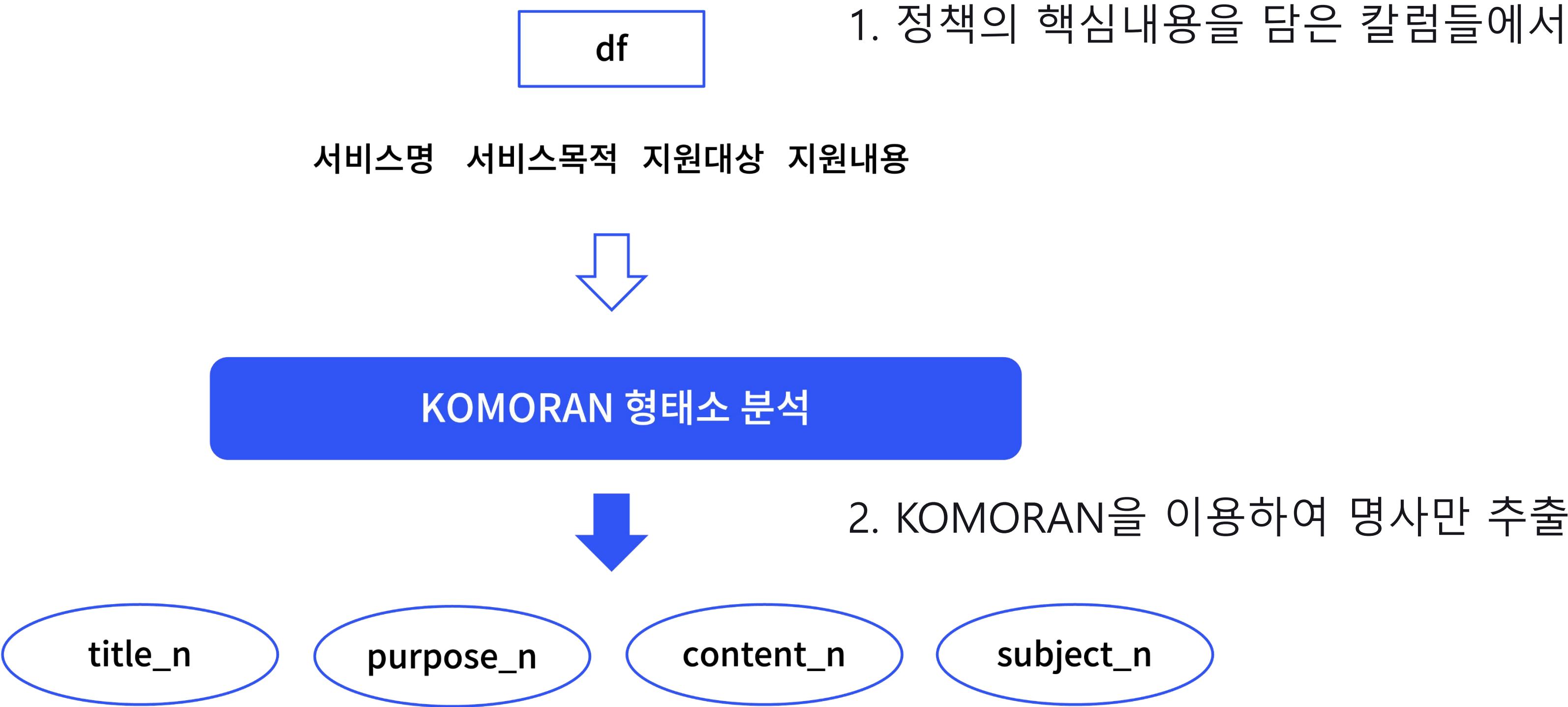
모델 개요

전체 흐름도

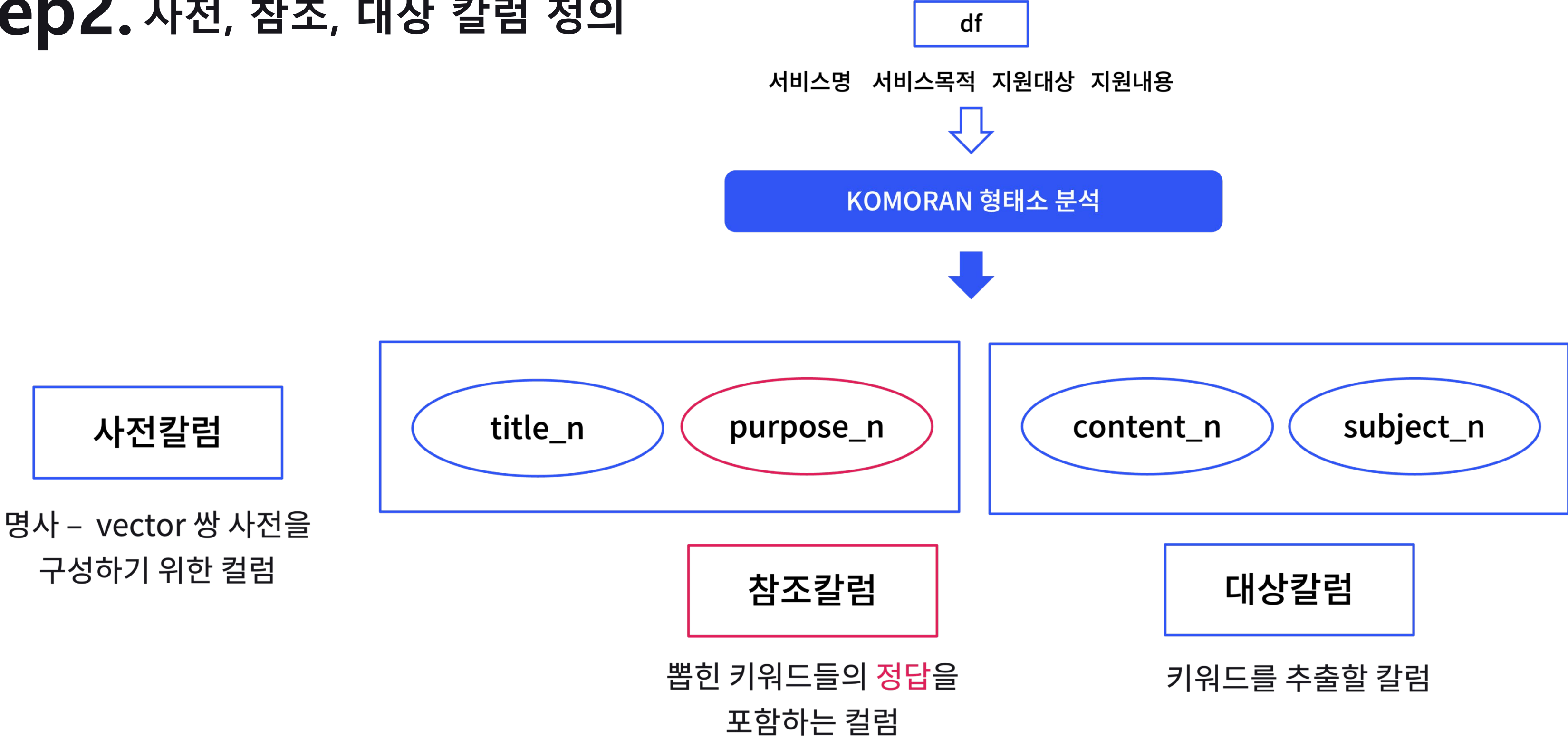


*title_n, purpose_n, subject_n, content_n은 각각 original data에서 명사를 추출한 컬럼

Step1. 명사 추출

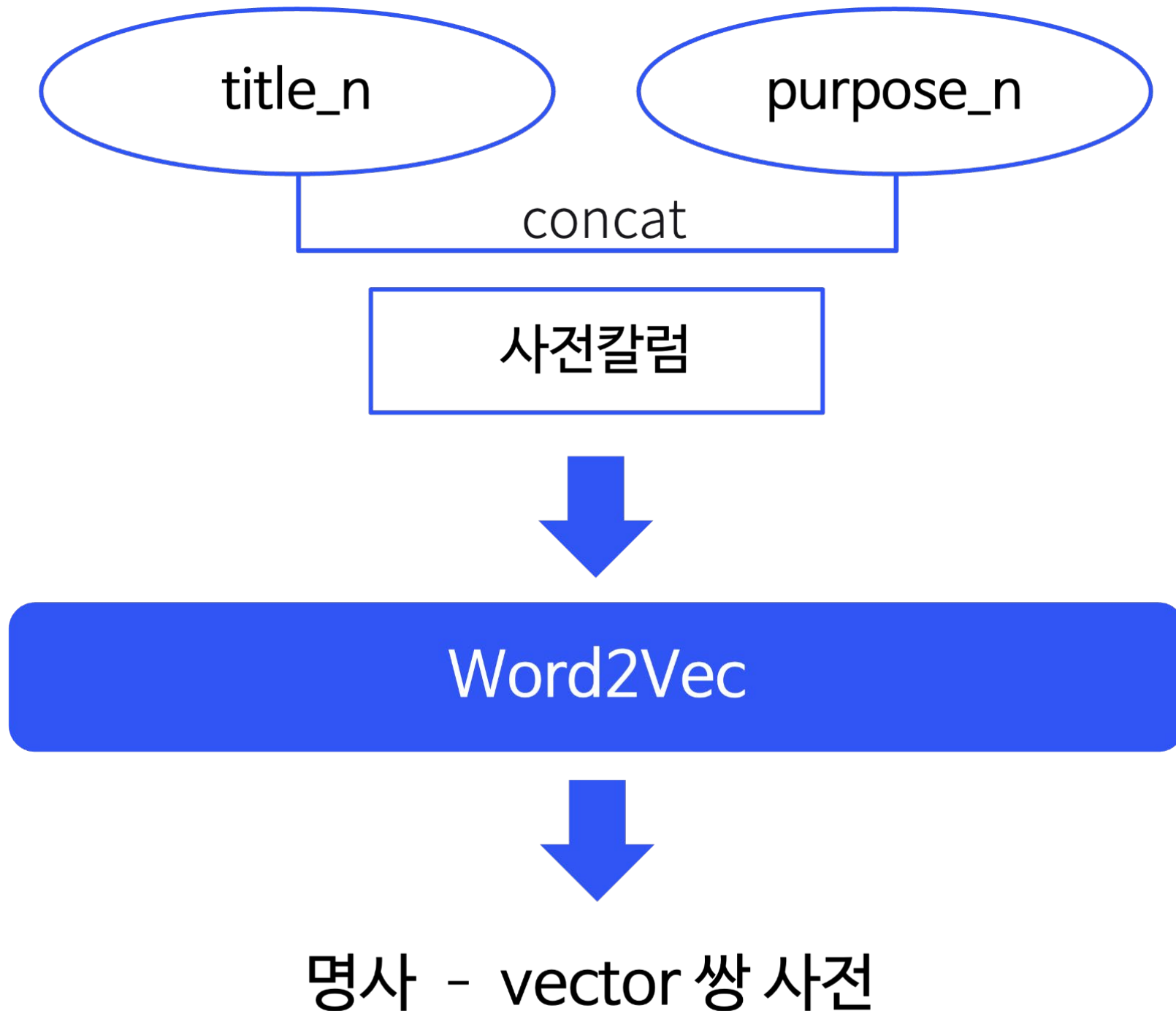


Step2. 사전, 참조, 대상 칼럼 정의



2. Modeling

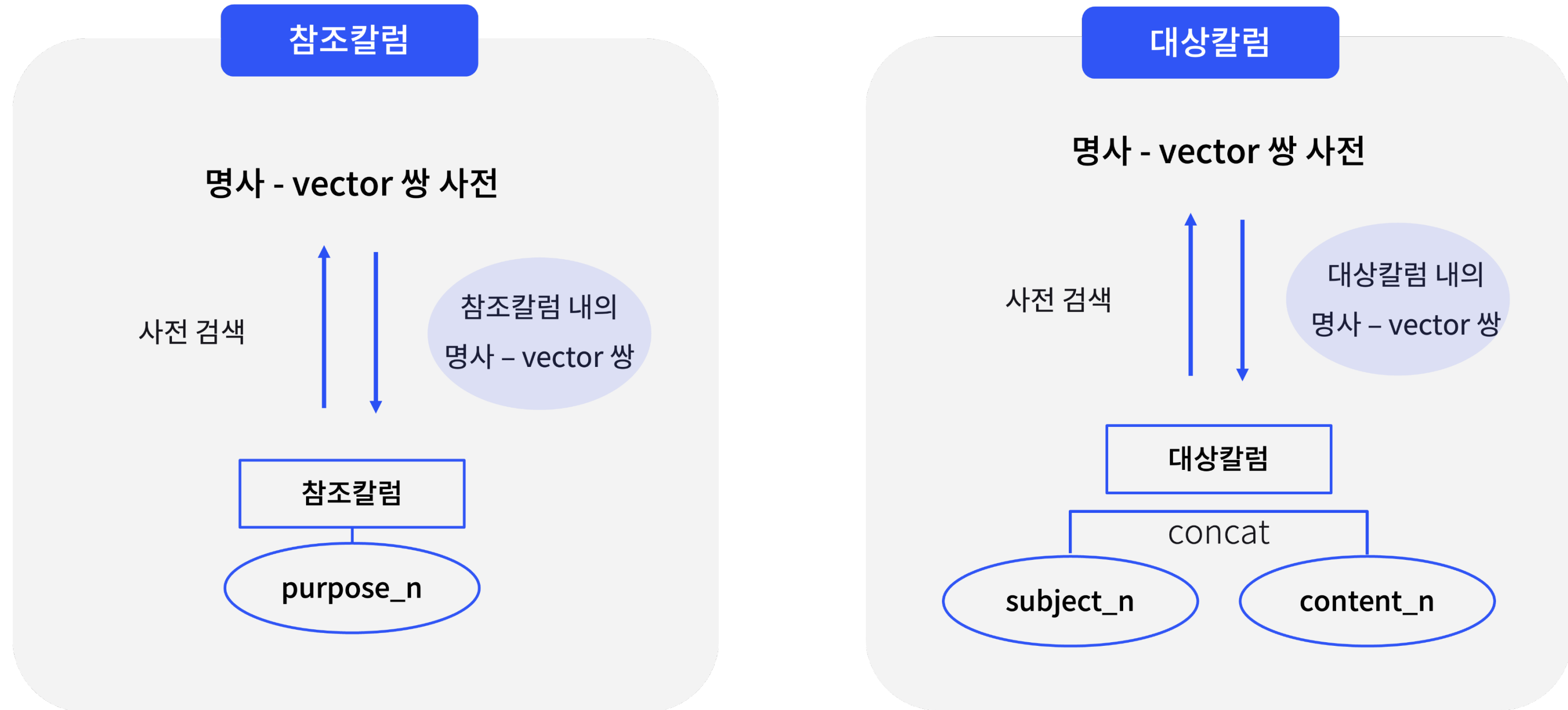
Step3. (단어-벡터) 사전 생성



`title_n`과 `purpose_n`을 concat하여
사전칼럼으로 지정,
Word2Vec을 이용하여 단어 embedding

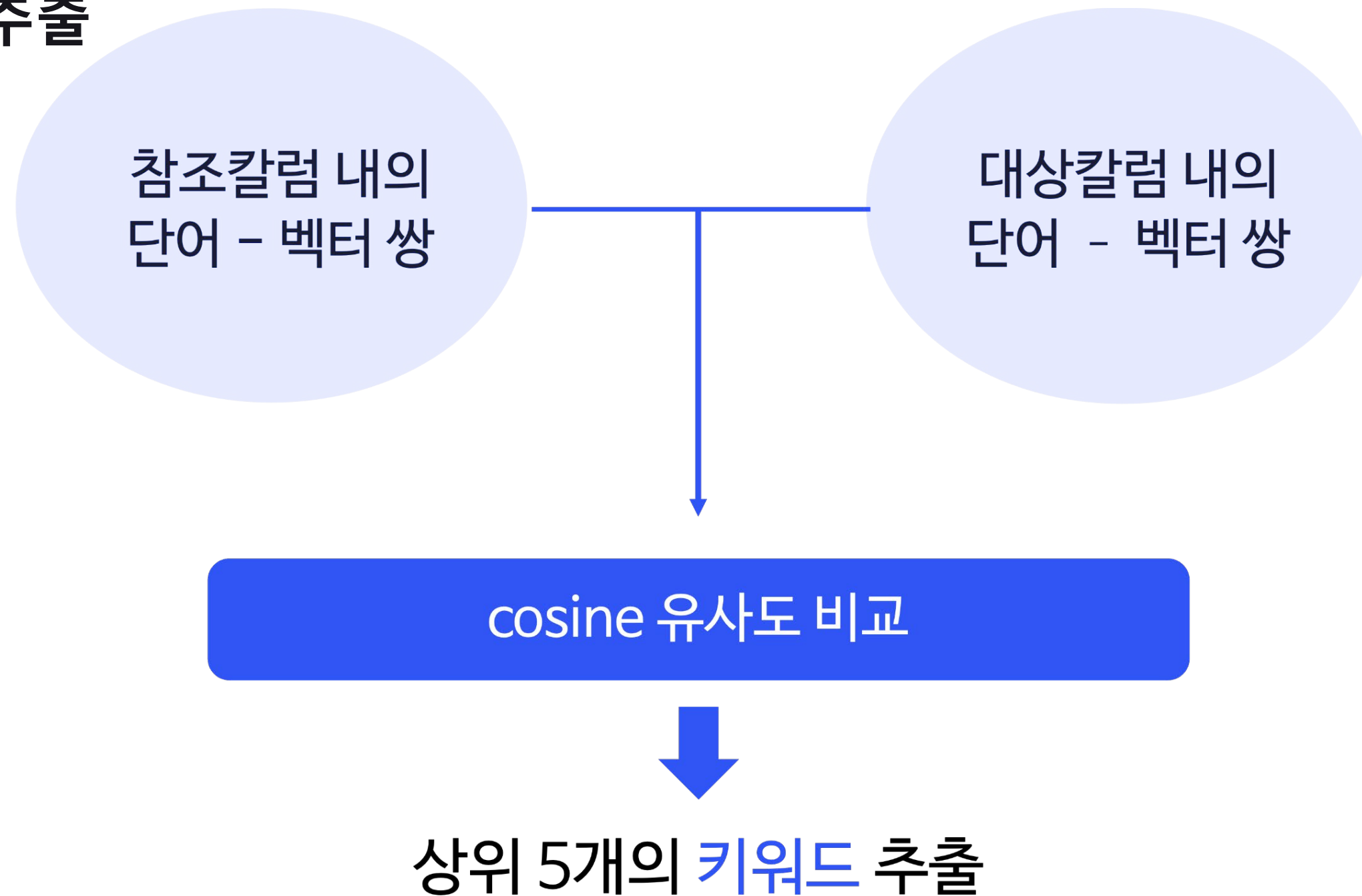
2. Modeling

Step4. 참조칼럼, 대상칼럼 내의 (단어-벡터) 쌍 생성



Step3에서 생성한 명사 - vector 쌍 사전에서
참조칼럼, 대상칼럼 내의 단어에 해당하는 embedding vector를 찾는다

Step5. 키워드 추출



대상 embedding vector 중 참조 embedding vector와의 cosine 유사도가 높은
상위 5개의 단어를 키워드로 추출한다

2. Modeling

Eg. 추출 예시

단어가 벡터화 된 사전컬럼

중소기업 -> (0,1,2)
기업부담 -> (0,2,0)
안정 -> (1,0,0)
고용안정 -> (2,0,0)
....

←
참고

참조컬럼(정답)
[서비스 목적]

대상컬럼
[지원 내용 & 지원
대상]

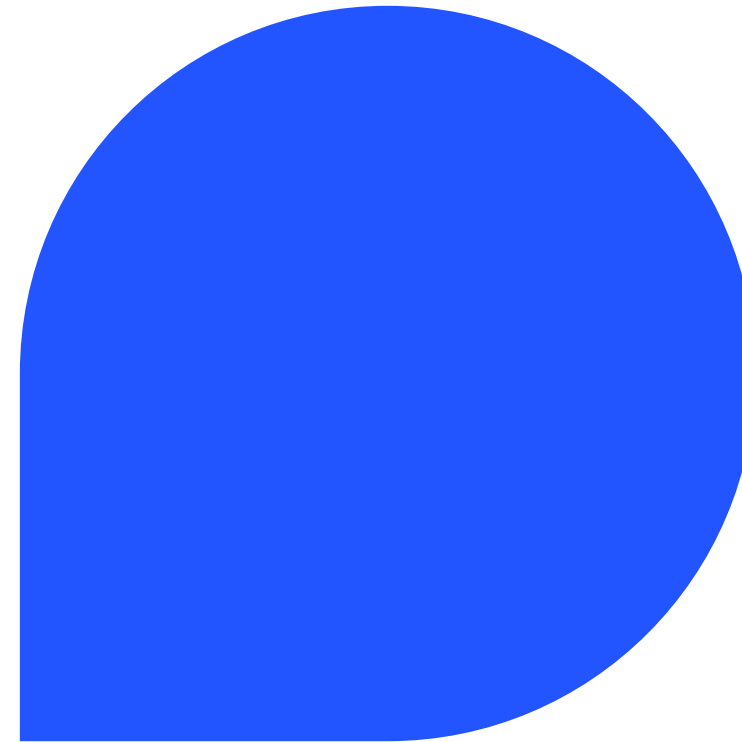
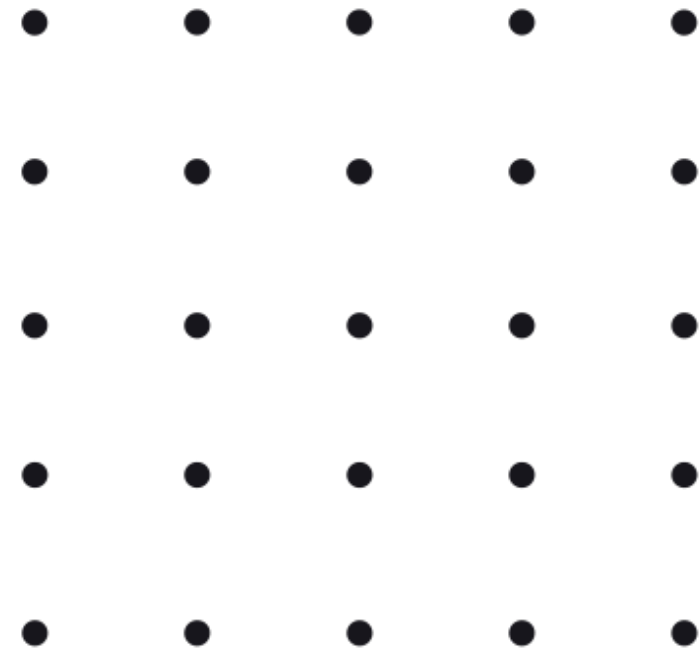
컬럼별 명사들

정규직, 기업부담, 생산성, 안정, 경쟁력

중소기업, 업무, 지원, 정규직전환, 제출, 간접,
고용안정



키워드 추출: 중소기업, 정규직전환, 고용안정



2-3 Improvements

사용자사전, Stopwords



모델 개선

문제점

KOMORAN

‘북한이탈주민’과 같이
하나의 뜻을 가지는 명사를
북한/이탈/주민으로 나누어 명사 추출



사용자 사전

‘북한이탈주민’과 같은 단어들을 하나의
명사로 인식할 수 있도록
전처리 및 ‘사용자 사전’ 구성하여
명사 태깅과 임베딩 성능 개선

키워드 추출

동사형 명사(ex. 처리, 해소, 개선 ...),
한 단어 명사 등
키워드로 적합하지 않은 불용어 존재



불용어 제거

후처리로 불용어 제거하여
키워드 추출 결과물 개선

모델 개선

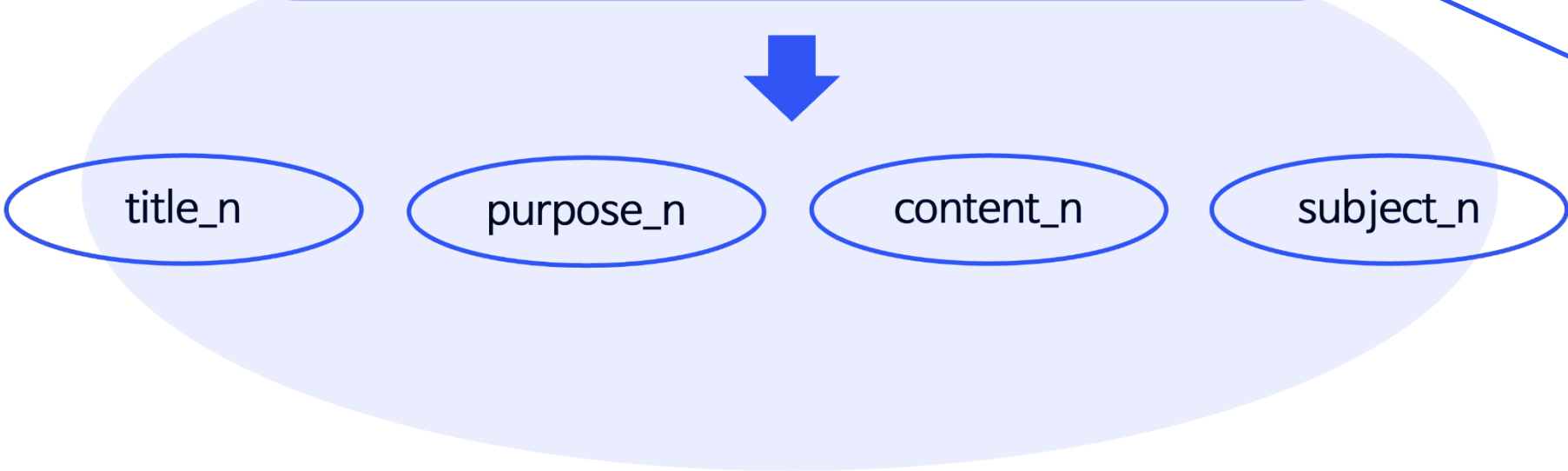
사용자 사전 등록

df

서비스명 서비스목적 지원대상 지원내용



KOMORAN 형태소 분석



전처리

re.sub

북한 이탈 주민	북한이탈주민
독립 유공자	독립유공자
기초 생활 수급자	기초생활수급자

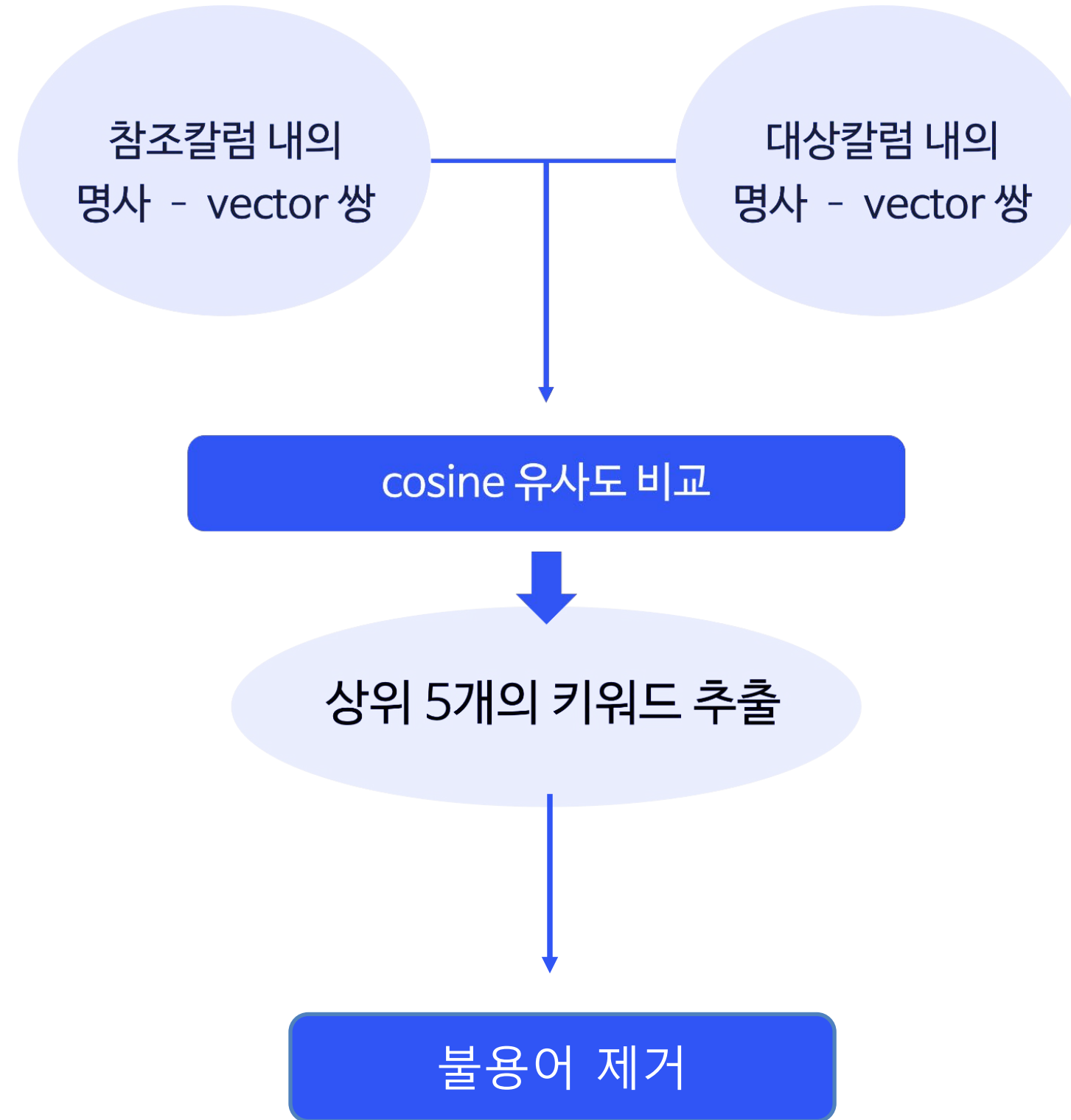
사용자 사전

정책의 내용을 잘 나타내주는 핵심적인 단어들이
단어 - vector사전에 더 많이 포함될 수 있도록
태그 데이터 + KOMORAN 태깅이 잘못된 단어들로
'사용자 사전' 구성

ex) 근로자/녀로 잘못 태깅된 경우 → '근로자녀' 사용자 사전에 등록

모델 개선

불용어 제거



ex) 불용어 예시

수립

고취

고려

확인

평가

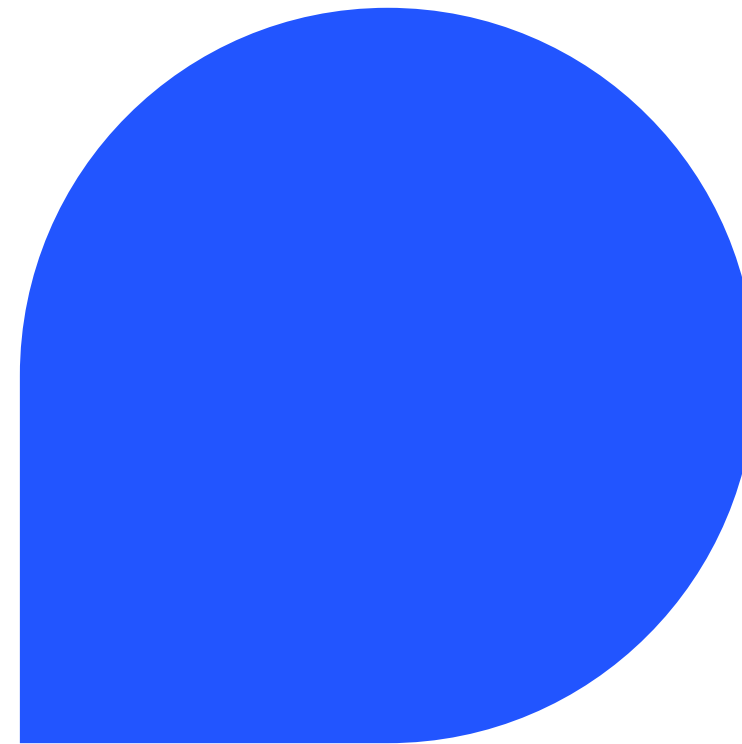
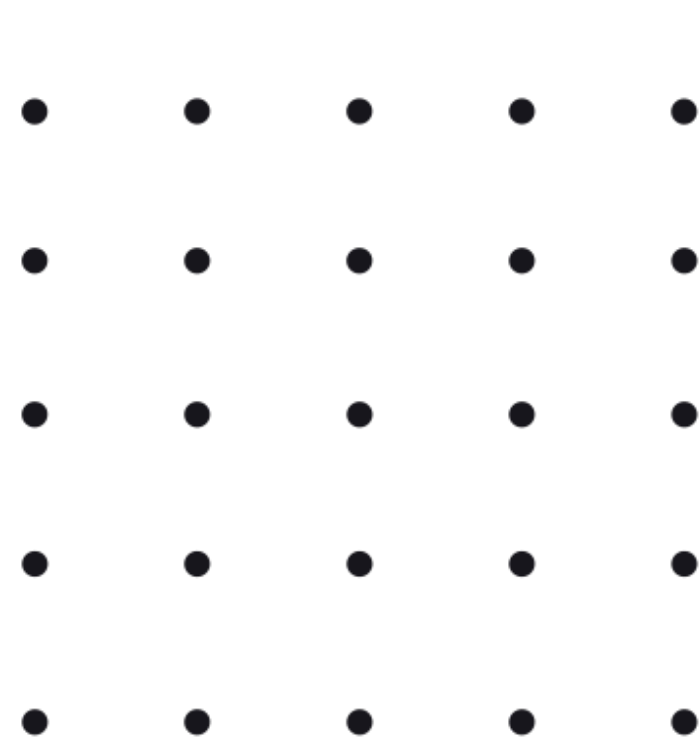
발굴

...

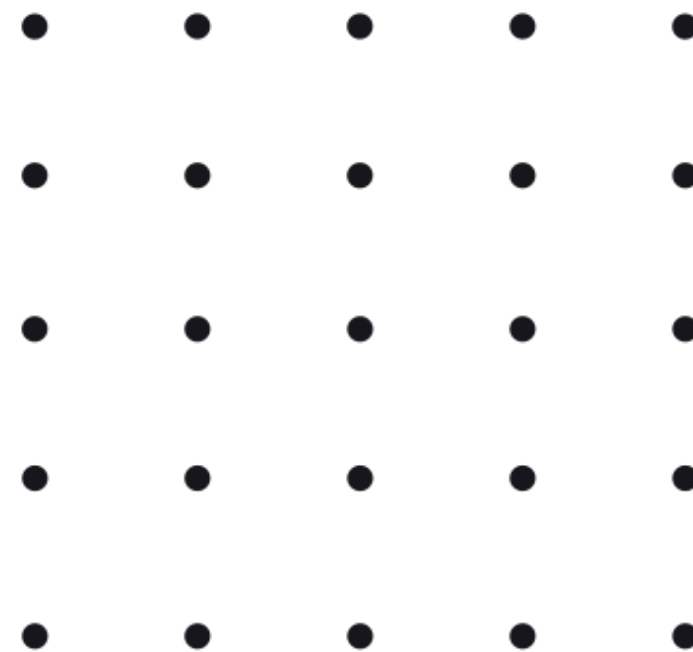
```
if len(word)>1 and word not in stopwords:
```

키워드로 부적합하다 판단되는

동사형 명사, 한 글자 단어 등의 불용어를 후처리로 제거



3 . Modeling Result

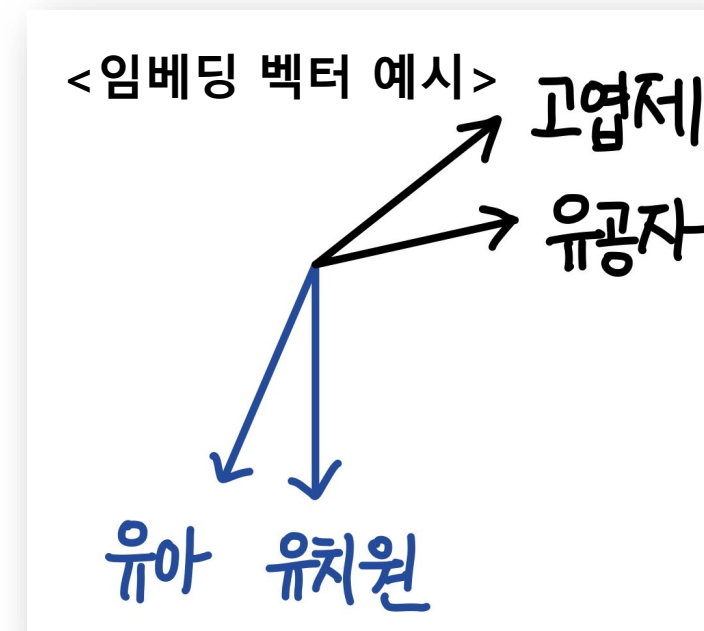


3. Modeling Result

임베딩 결과 - 코사인유사도 확인

```
[67] word1 = ["고엽제"]  
word2 = ["유공자"]  
  
example1_vec = get_embedding_vector(word1)  
example2_vec = get_embedding_vector(word2)  
  
print(np.isnan(example1_vec).sum())  
print(np.isnan(example2_vec).sum())  
  
cosine_similarity(example1_vec, example2_vec)  
  
0  
0  
array([[0.46699667]], dtype=float32)
```

단어	코사인 유사도
(“고엽제,”유공자”)	+ 0.466
(“유치원,”유아”)	+ 0.571
(“고엽제,”유아”)	- 0.0896



키워드 추출 결과

(참조컬럼)

서비스명: 농번기 아이돌봄방 운영 지원

서비스 목적: 돌봄시설이 부족한 농촌에서 농번기 주말동안 영유아를 안심하고 맡기고, 영농에 종사할 수 있도록 아이돌봄방을 설치·운영하여 일·가정 양립지원 및 농촌 돌봄사각지대 해소



(대상컬럼)

지원 대상:

- 농촌지역에서 보육에 필요한 전문성, 시설 및 인력을 갖추고 농번기 주말동안 아이돌봄방을 운영하고자 하는 법인·단체
- 어린이집, 지역농협, 여성농업인센터, 여성농업인종합지원센터, 지역아동센터, 사회복지법인 등

지원 내용:

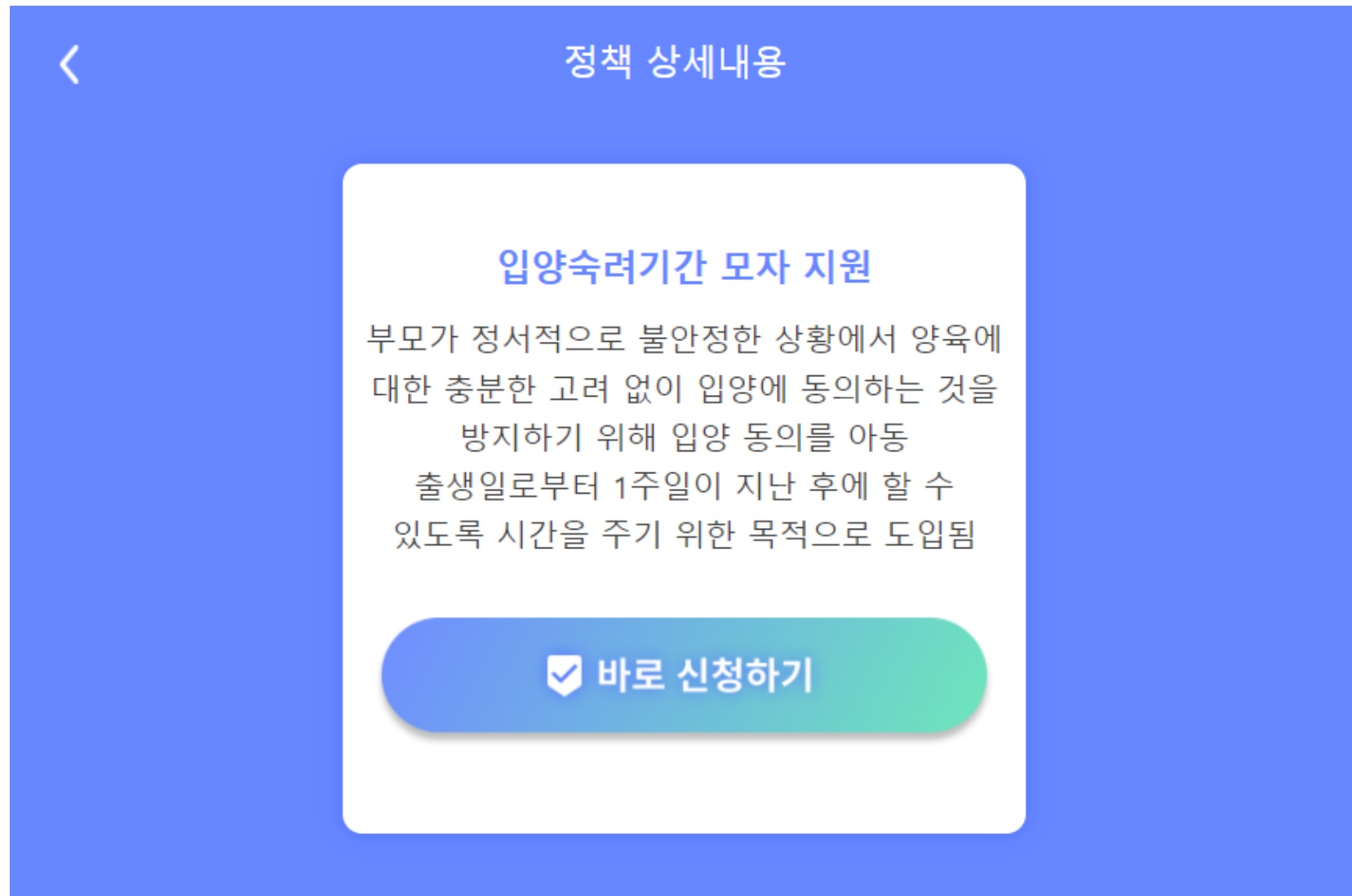
- 시설비: 개소당 최대 20백만원 지원
- 리모델링 및 장비, 기자재 구입비

- 운영비: 개소당 17~26백만원 지원
- 운영기간(4개월~6개월)에 따라 차등지원
- 인건비, 급간식비, 교재교구비 등

키워드: 돌봄, 교구비, 구입비, 어린이집, 간식

3. Modeling Result

모델 개선 이전 vs 이후



Before

아래, 아동, 방문, 복지, 여성 →

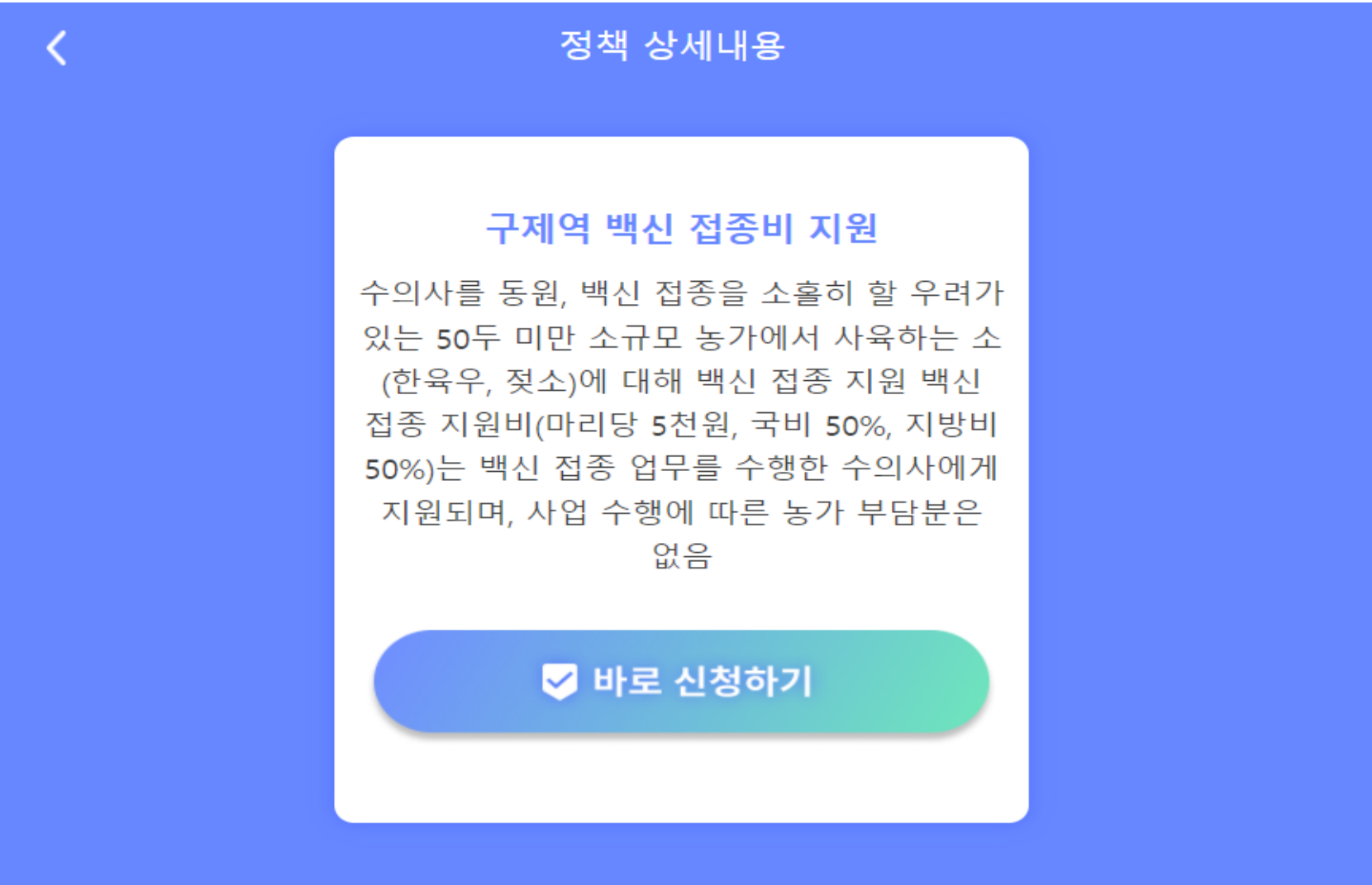
무엇을 받나요?

- 미혼이혼 한 부모가 선택하는 서비스 이용 비용 지원
 - 출산(예정) 여부 : 출산(예정) 일 전 40일 또는 후 7일 이내에 있는 자
 - 혼인 여부 : 혼인관계 증명서상 혼인관계에 있지 아니한 자
- 가정 내 보호 지원
 - 산후 지원인력 가정방문서비스 지원(1주), 500,000원(지원 단가)
 - * 산후 지원인력 서비스 이용료 (40만 원 한도)
 - * 아동 생필품비 포함 (10만 원)
 - 가족 또는 친구 등 지인의 도움을 받기 원할 경우(1주), 350,000원(지원 단가)
 - * 아동 생필품비 포함
- 미혼모 가족복지시설 내 입소자 지원
 - 미혼모자가족시설 입소 시, 산후 지원인력 인건비 지원(1주), 400,000원(지원 단가)
- 산후조리원 보호 지원
 - 1주 산후조리원 이용료 지원 최대 700,000원
 - * 1주 이용료가 700천 원 미만인 산후조리원의 경우, 실비 지원
 - * 아동 생필품비 및 생모 식료품비 등 포함

After

방문, 산후, 산후조리원, 여성, 한부모

모델 개선 이전 vs 이후



누가 받나요?

0-10

10-20

20-40

40-60

60-

청년, 중장년, 노년

매일 뜨거운 삶을 살아가는 당신을 위해
새로운 인생 2막을 즐길 당신을 위해
황혼의 여유를 즐길 당신을 위해

50두 미만 소 사육농가

Before

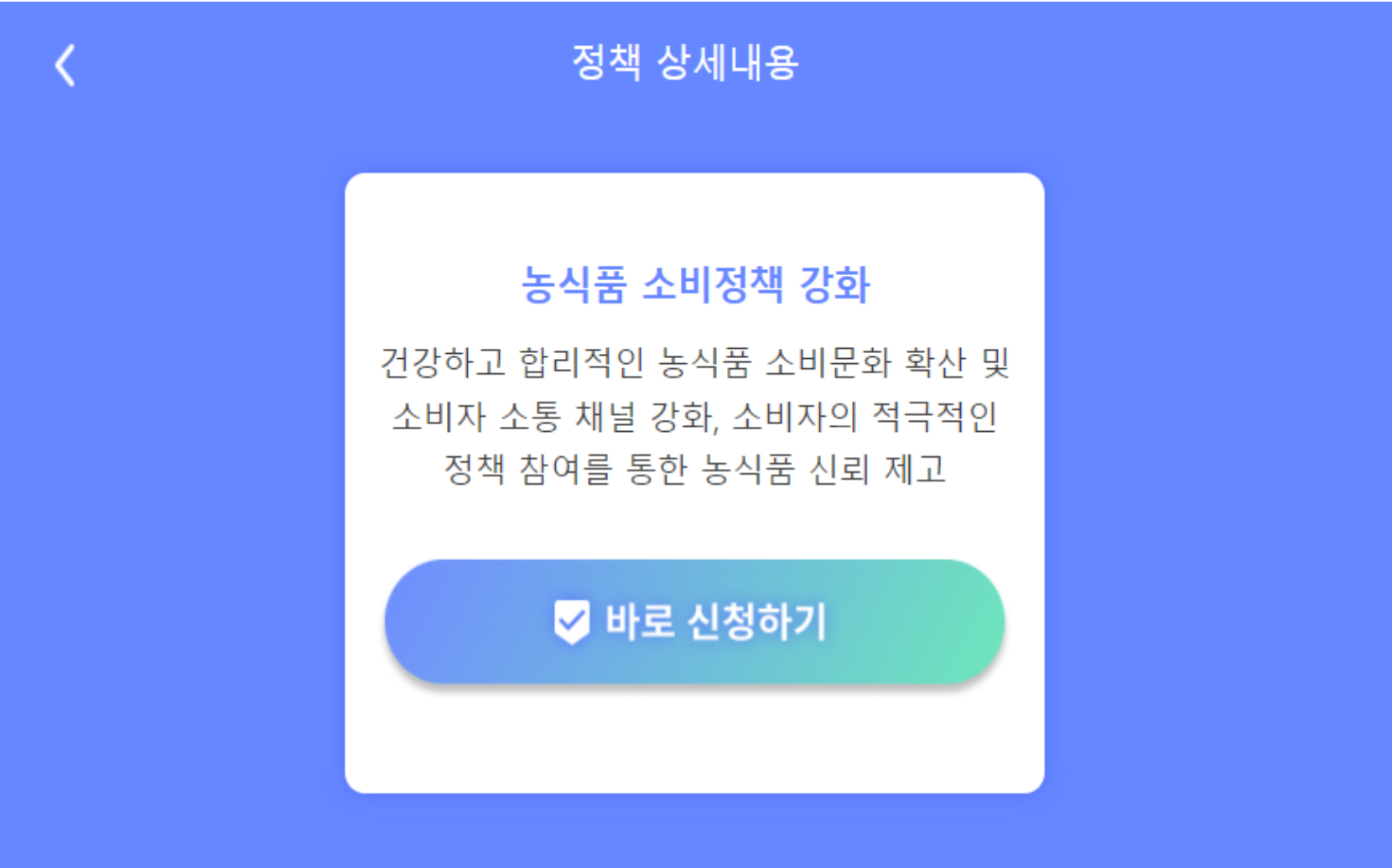
젖소, 지원비, 우려, 국비, 미만

→

After

젖소, 소규모, 운영지원, 국비, 사업자금

모델 개선 이전 vs 이후



- (공통 필수사항) 소비자 대상 농식품 교육·홍보 프로그램 기획·운영 및 소비자 정책제언·모니터링 등 수행 실적이 있는 기관 및 단체 등
 - (소비자단체 부문) : 소비자 기본법 제29조의 소비자단체
 - (비소비자단체 부문) : 비영리민간단체 (비영리민간단체지원법 제4조), 비영리법인 (민법 제32조에 따라 허가받은 법인)

Before

 관련, 단체, 실적, 영리, 기관 →

After

 홍보, 영리, 교육, 민간단체, 단체

3. Modeling Result

활용방안

유저들의 효율적인 정책 이용을 위해, 키워드 몇 개로 정책을 한 눈에!

중앙행정기관 국토교통부
청년우대형청약통장
D-36
중앙행정기관 고용노동부
국민내일배움카드 훈련과정
상시
중앙행정기관 고용노동부
청년내일채움공제 2021
상시
중앙행정기관 고용노동부
워라밸일자리 장려금 지원 사업
상시
중앙행정기관 고용노동부
무급휴업휴직 근로자지원
상시

맨 위 정책 클릭

<

정책 상세내용

청년우대형청약통장

저소득 무주택 청년의 주택구매 및 임차자금 마련 지원을 위해 재형 기능을 강화한 청약통장 도입

✓ 바로 신청하기

✓ 2021년 12월 31일 마감

활용방안

정책 목록에 키워드 노출시,
유저들이 정책 내용을 직관적으로 파악 가능!

중앙행정기관 국토교통부

청년우대형청약통장

D-36 #금리우대 #비과세 #소득공제 #무주택자 #청약

중앙행정기관 고용노동부

국민내일배움카드 훈련과정

상시 #실업자 #재직자 #자영업자 #훈련비 #훈련장려금

중앙행정기관 고용노동부

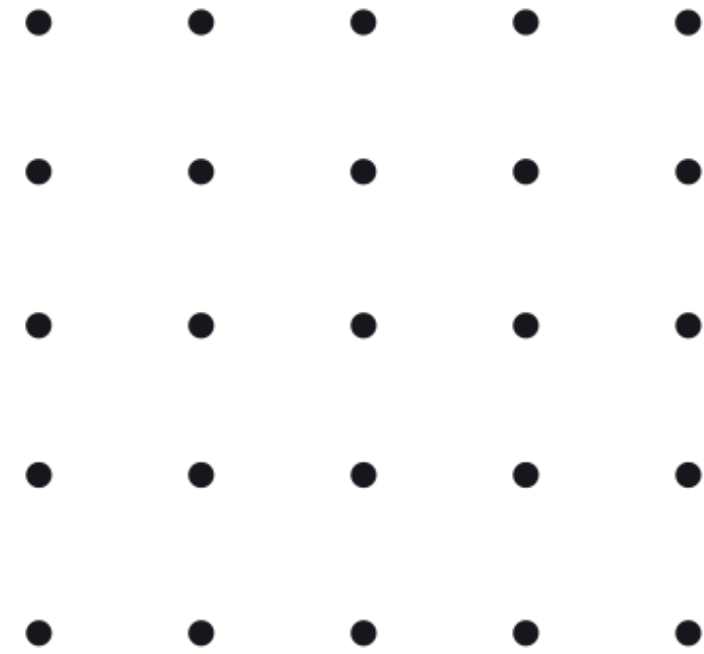
청년내일채움공제 2021

상시 #취업 #청년 #중소기업 #정규직 #채용기업

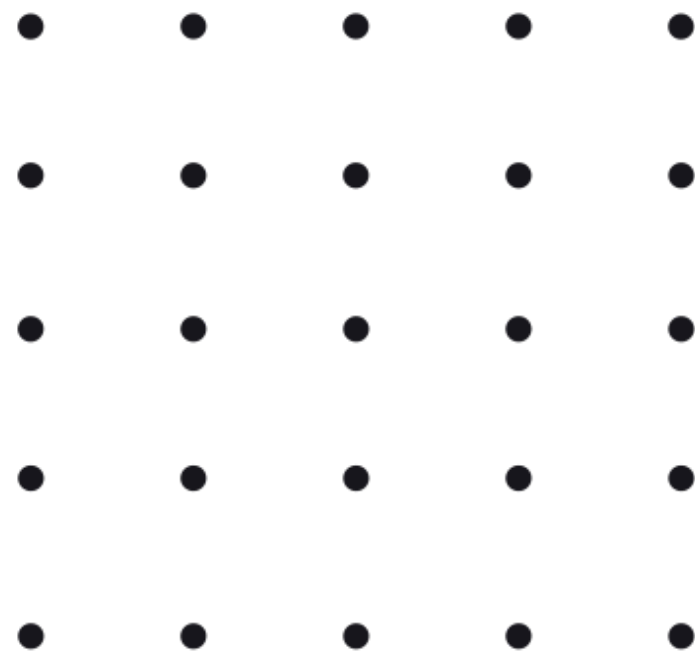
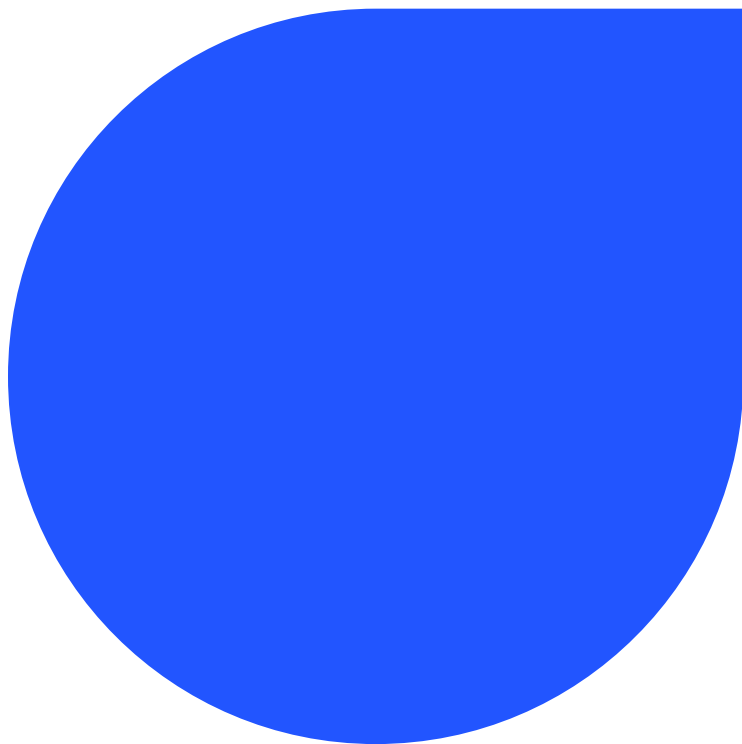
중앙행정기관 고용노동부

워라밸일자리 장려금 지원 사업

상시 #근로시간단축 #정액지원 #사업주 #근로시간 #필요




4. 개선방향 및 한계



4. 개선방향 및 한계



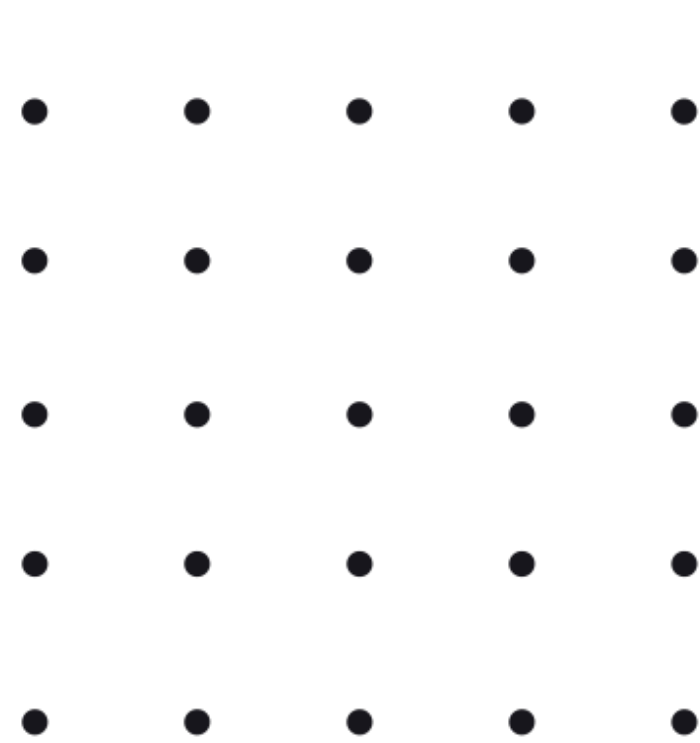
1. 구체적인 정책 용어를 사용자 사전에 입력
=> 모델 성능 
2. 키워드의 범주를 분류할 수 있다면 더 높은 퀄리티 기대 가능



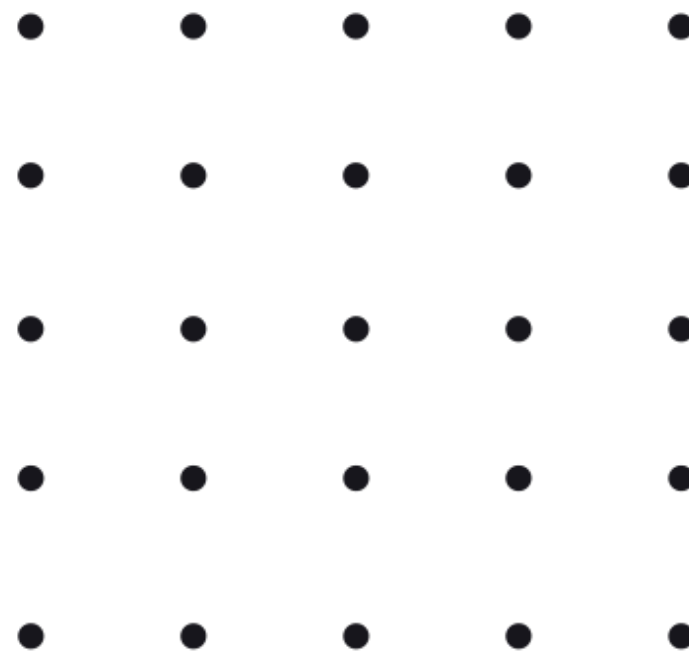
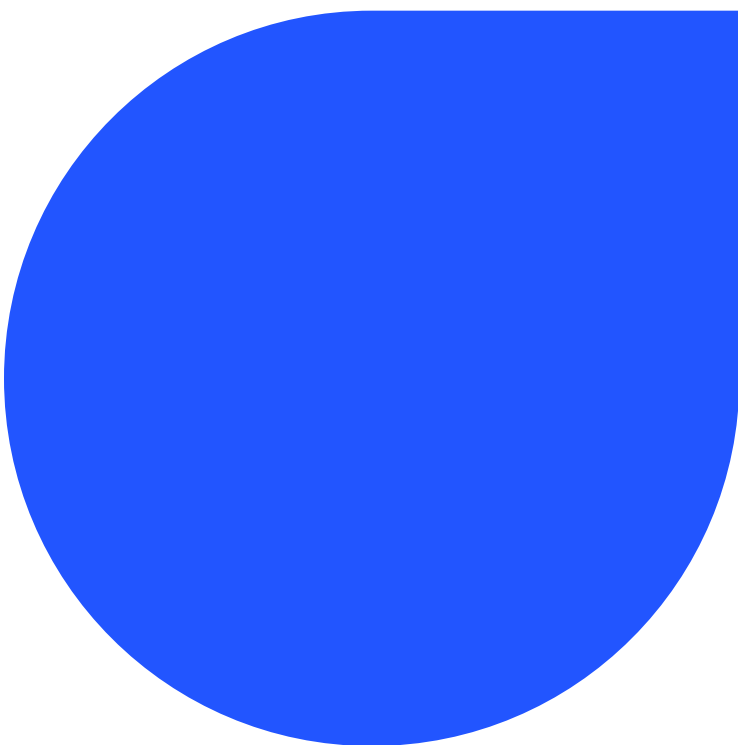
1. 사용자 사전, 불용어 처리와 같은 수작업에 투자한 시간 부족
2. 단어 입력 및 삭제 방법으로 모델 성능을 개선하는 방식 ,
다양한 방법을 사용하지 못해 아쉬움



Q&A
감사합니다!!

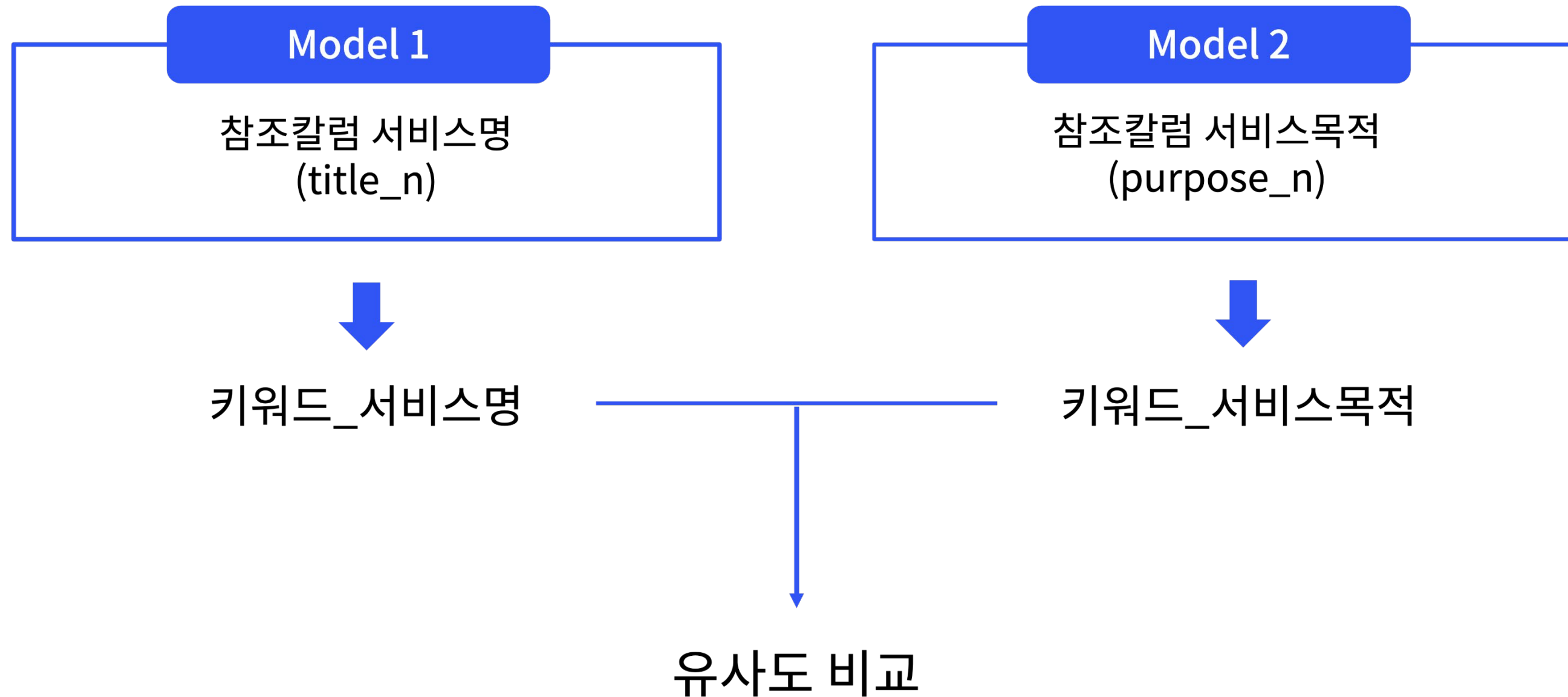


Evaluation Metrics



모델 개선

개선 전후 모델 성능 비교



개선 전보다 개선 후에 모델1과 2에서 추출된 키워드들 간의 유사도가 높아졌으면
모델 성능의 개선이 이뤄진 것으로 판단할 수 있음