제 12회 산업 통상 자원부 공모전 과제 3번

산업통상자원부

Tri-Fusion 매핑 알고리즘

(LLM, 언어 모델, 추천 시스템에 기반한 HS CODE 10단위 매핑)

청랩: 김민경, 김유미, 민지현

Contents

CH01. 서론 및 분석 개요

제안 배경 및 필요성

KOTRA NEWS

07/01/2024

산업통상자원부

국내 기업 수출 증진 위한 HS 부호 매칭 시스템, 데이터 한계로 '삐걱'

2018, KOTRA 이미지 제공

"전세계 무역 지원기관, HS 부호 매칭 여전히 어려워..."

전 세계 무역 지원 기관들이 국내 기업의 수출 증진을 위해 디지털 매칭 프로그램을 운영하고 있지만, HS 부호 매칭의 어려움으로 여전히 난항을 겪고 있다.
HS 부호는 국가 간 상품 거래 시 통일된 기준을 제공하는 중요한 역할을 하지만, 데이터 부족 및 국가별 분류 체계 차이로 인해 정확한 매칭이 어려운 실정이다.

"기존 매칭 시스템 문제 원인은 데이터의 한계"

기존 매칭 시스템은 기업 정보의 변화를 따라가지 못하고, HS 부호 분류 체계의 국가별 차이로 인해 데이터 불일치 문제를 겪고 있다. 이는 시스템의 정확성과 효율성을 저해하여 기업의 수출 증진에 걸림돌이 되고 있다.
예를 들어, 기업의 기술 변화로 인한 원자재나 기계 수요 변화를 시스템에서 확인하기 어렵고, HS 부호 불일치로 인해 수작업 매칭에 의존하게 되어 시간과 비용 낭비를 초래한다. 이러한 문제는 결국 적절한 해외 파트너 발굴을 어렵게 만들고 거래 성사율 저하로 이어진다.

기존 매칭 시스템이 갖는 한계점을 보완하고자
생성형AI를 비롯하여 최신 빅데이터 기법을 접목한 매칭 모델을 제안

과제 목표

산업통상자원부

1. 국내기업 수출 증진을 위해, 국내기업의 품목을 수입할 수 있는
해외기업 발굴에 대한 필요성 상시 존재

STEP1 관세 문제가 아닌 국내기업과 해외기업 매칭을 통해 국내 기업 수출 증진을 목표로 함

2. 해외기업 발굴 시, 국내기업이 취급하고 있는 품목과 유사한 품목을
취급하는 해외기업을 발굴하고자 함

STEP2 LLM을 활용해 인간의 추론과정을 구현하여 해외 기업 설명(DSC)를 보고
바이어가 수입 가능한 품목을 도출할 수 있도록 함

사용 데이터 설명

산업통상자원부

1. 비식별된 해외기업별 영문 텍스트데이터.xlsx

ID	CODE	DSC
1	4520	automotive repair shops, nec specialized automotive repair, not elsewhere classified, ···
2	0149	general farms, primarily animals, nsk derives 50 percent or more of its total value ···
⋮	⋮	⋮
10645	4651	computers, peripherals, and software, nsk the wholesale distribution of computers, ···

2. 관세청_HS부호_240101.xlsx

HS부호	한글품목명	영문품목명	성질통합분류코드	성질통합분류코드명
0101211000	농가 사육용	For farm breeding	11020101	(말)
0101219000	기타	Other	11020101	(말)
⋮	⋮	⋮	⋮	⋮
9706903000	기타	Other	12050201	(수집품 및 골동품)

3. 통계청 국제표준산업분류 HSCODE 6단위 매핑.xlsx

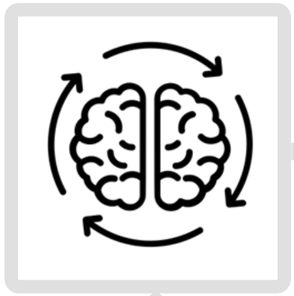
ISIC4 (국제표준산업분류, UN)	ISIC4 분류명(한글)	KSIC10 한국표준산업분류	KSIC10 분류명	HS2017 관세·통계통합품목분류 (WCO, 한국)
1061	곡물 가공품 제조업	10611	곡물 도정업	현미
1061	곡물 가공품 제조업	10611	곡물 도정업	기타 곡물로 만든 것
⋮	⋮	⋮	⋮	⋮
5223	항공 운송관련 서비스업	52931	공항 운영업	
5223	항공 운송관련 서비스업	52939	공항 운영업	

3번 데이터는 결측값이 존재하고,
최종 매칭 후보인 HS부호 데이터가 불완전
➡ 1, 2 데이터만 사용

서비스업은 HS부호 10자리 중 앞 6자리에
해당하는 HS2017이 없는 경우가 있음

Process 논리

산업통상자원부



해외 기업 설명
: 낙농 농장, 젖소 우유,
기타 유제품 생산,
젖소 암소 대체 사육
농장에서 우유(냉동 제외)를
가공 및 병에 담아
도매 또는 소매로 판매

해외 바이어 기업과
국내 기업 매칭

해외 기업이 구입할 만한
품목은?

→ LLM

기업 설명과 가장 유사한
HS10 분류 품목은?

→ 텍스트 임베딩
추천 시스템

이 중에서 냉동 제품은
제외한다면?

→ 예외 필터링

HS10 CODE 매핑

CH02. 분석과정 및 결과

HS CODE 데이터 전처리

산업통상자원부

STEP1 번역 모델(Marian MT)로 영문 번역 + 수작업 보완

관세청_HS부호_240101.xlsx

HS부호	한글품목명	영문품목명	성질통합분류코드	성질통합분류코드명
0102211000	젖소	For milk	11020102	소
8434100000	착유기	Milking machines	32030603	낙농기계
0402219000	기타	Other	11020490	기타 유제품
⋮	⋮	⋮	⋮	⋮

성질통합분류코드명_영문
Cow
Dairy machinery
Other dairy products
⋮

STEP2 '품목명' 변수 생성

영문품목명
For milk
Milking machines
Other

+ 'belonging to' +

성질통합분류코드명_영문
Cow
Dairy machinery
Other dairy products

품목명
For milk belonging to Cow
Milking machines belonging to dairy machinery
Other belonging to Other dairy products

LLM import 수입품목 추가

산업통상자원부

목표 : 해외 바이어 기업과 국내 기업 매칭

국내 기업 취급 유사품목 = 해외기업이 수입가능성 있는 품목 → 성약 극대화

DSC에 수입 가능성이 있는 품목 설명 추가 → LLM

LLM(초거대 언어 모델) : 생성형 AI

주어진 프롬프트에 대해 인간과 유사한 응답을 생성하기 위해 방대한 양의 텍스트 데이터로 훈련된 고급 AI 모델

뛰어난 유연성과 생성능력으로 다양한 영역에서 활용

기업에서도 적극적으로 도입하며 업무 혁신 도모

LLM import 수입품목 추가

산업통상자원부

LLM model Llama3

1. 고성능의 최신 LLM 모델
: Claude Sonnet, Mistral Medium, GPT-3.5보다 Llama3 70B Instruct 모델이 높은 성능

2. 우수한 보안성
: Ollama와 LangChain을 사용하여 모델을 로컬에서 실행하므로 클라우드 서비스에서 발생할 수 있는 데이터 유출 위험 최소화

3. 비용적 효율성
: Llama3 모델은 무료 배포로 유료인 GPT에 비하여 비용적 효율

높은 성능을 보이면서도
보안과 비용 측면에서의 강점

→ **Llama3 모델 채택**

Meta Llama 3 Instruct model performance

	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	82.0	81.9	79.0
GPQA 0-shot	39.5	41.5 CoT	38.5 CoT
HumanEval 0-shot	81.7	71.9	73.0
GSM-8K 8-shot, CoT	93.0	91.7 11-shot	92.3 0-shot
MATH 4-shot, CoT	50.4	58.5 Minerva prompt	40.5

LLM import 수입품목 추가

산업통상자원부

LLM prompt

This is a description of a company :
dairy farms, nsk the production of cows' milk and other dairy products and in raising dairy heifer replacements.
⋮
Examples of the domain-specific materials and parts that the company needs are (). Return only what will fit in the blank

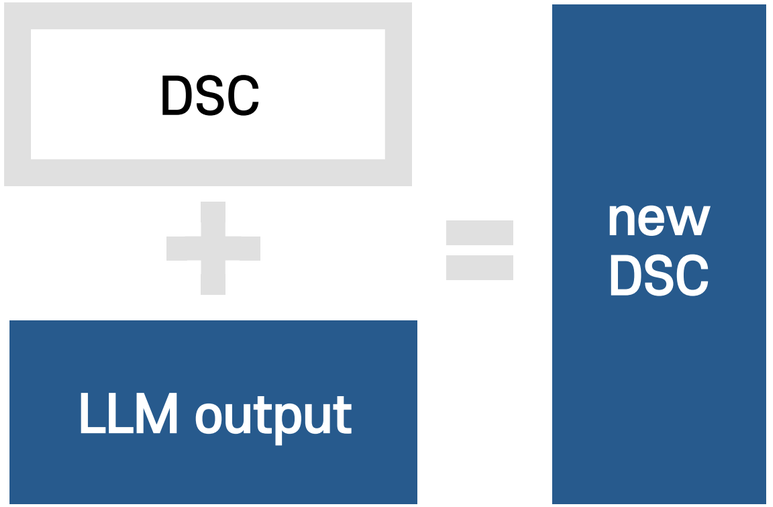
LLM

LLM output

This company may need Milking equipment (e.g. milking machines, milking parlors),
Cooling systems (e.g. refrigeration units, cooling tanks),
Feeding and nutrition supplies (e.g. hay, grain, mineral supplements),
⋮

생성형AI LLM 활용
: 내용을 생성하는 task를 수행하도록 프롬프트 설계

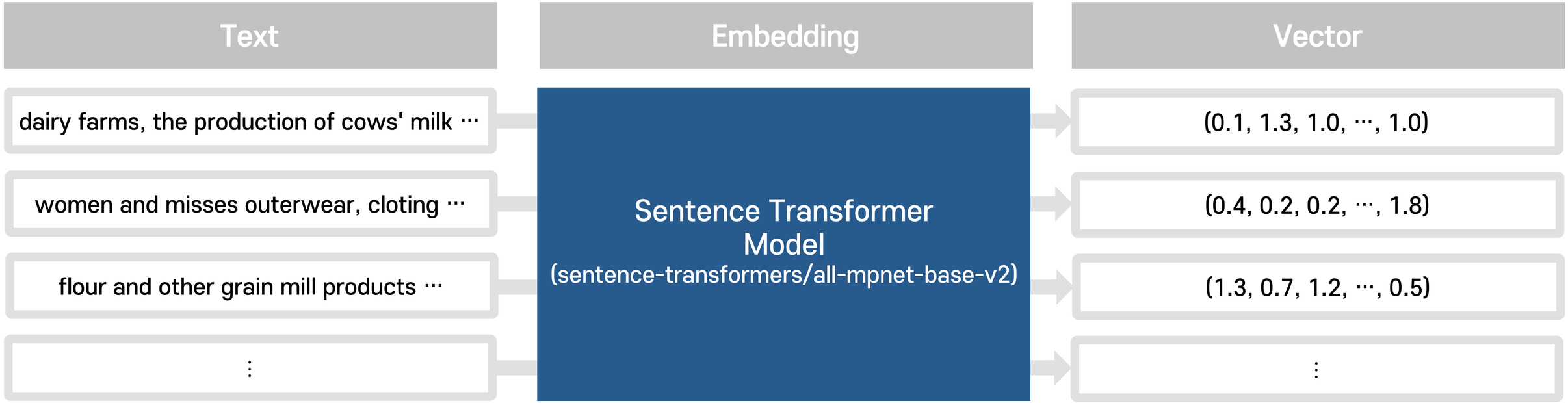
결과 반영



Embedding

산업통상자원부

단어나 문장 등의 텍스트를 기계가 이해할 수 있도록 숫자형태인 vector로 바꾸는 과정
임베딩을 통해 숫자 vector에 text의 의미를 부여



목표 : 기업 설명과 유사한 HS CODE 설명 찾기

Sentence Transformer 임베딩 모델을 사용
기업 설명 텍스트와 HS CODE 설명 텍스트를 숫자 형태로 임베딩
➡ 기업 설명과 HS CODE 설명의 의미를 담은 숫자 벡터 생성

Mapping

산업통상자원부

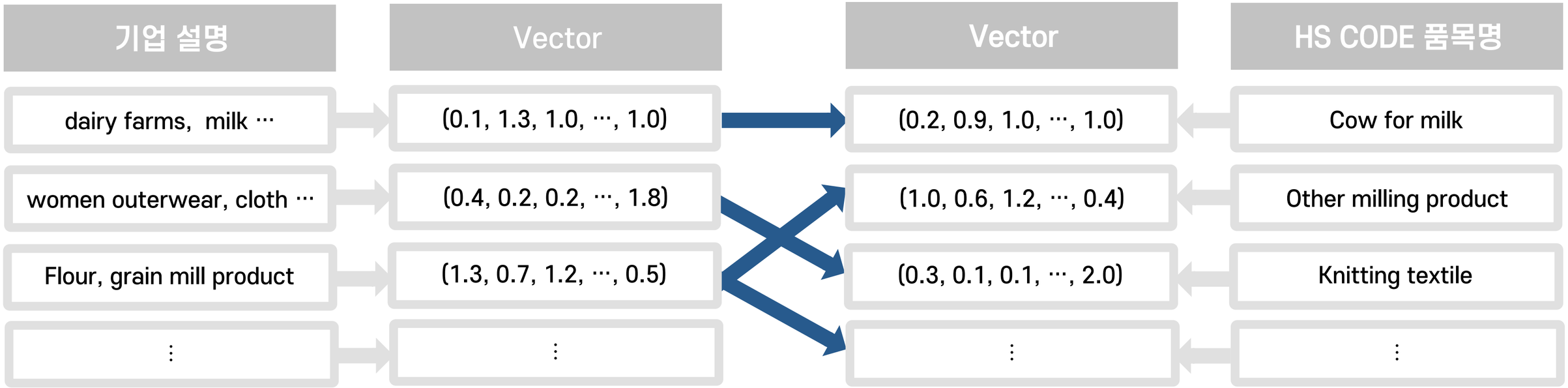
Cosine similarity (코사인 유사도)

벡터의 크기는 고려하지 않고 두 벡터 사이의 각도만을 고려하는 측정법
➡ 문서의 길이가 달라도 유사도 측정 가능

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Mapping by cosine similarity

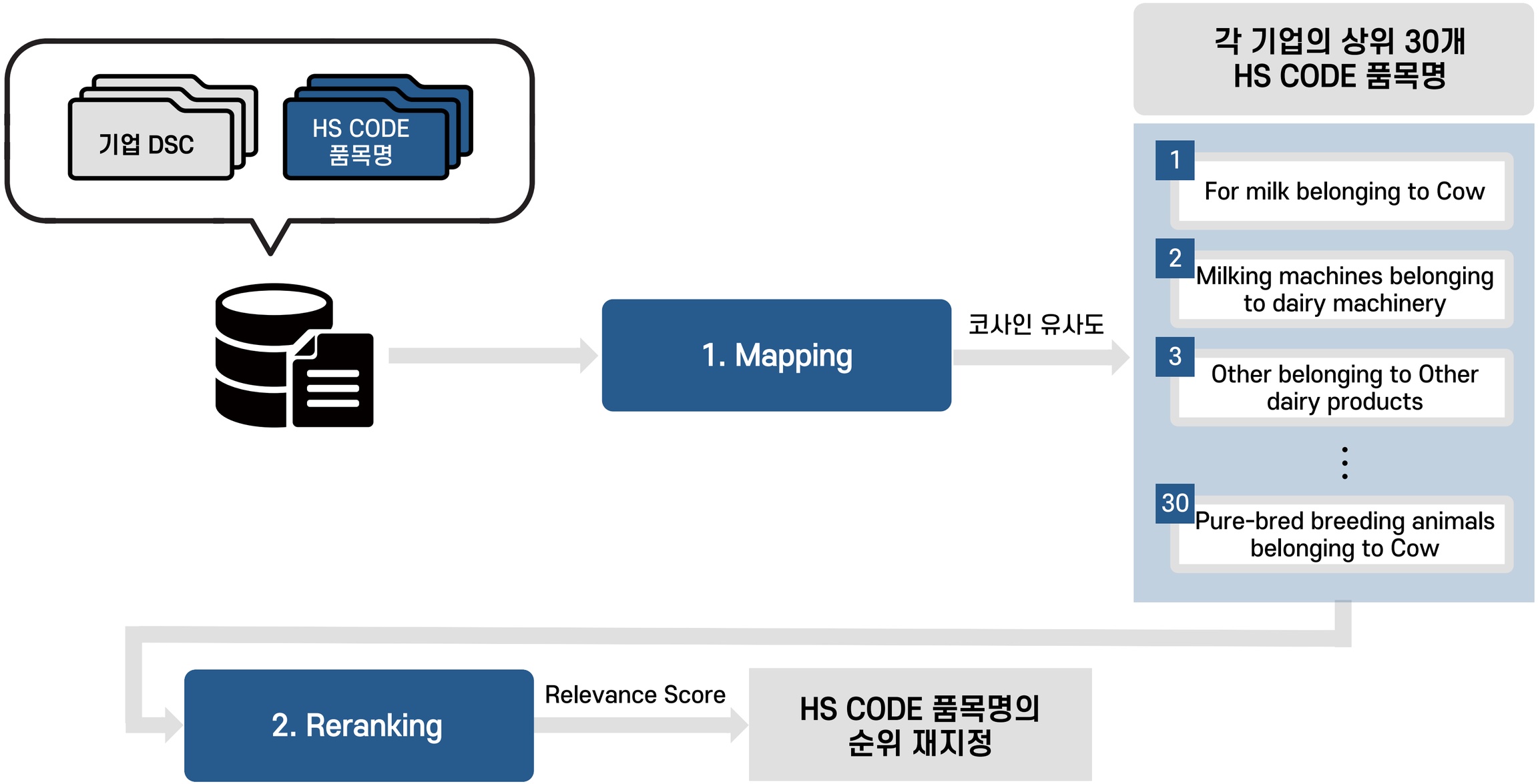
벡터 간의 거리가 가깝다 = 텍스트 간 의미가 유사하다



기업 설명 벡터와 HS CODE 품목명 설명 벡터간 코사인 유사도가 높은 경우를 매핑
길이가 긴 기업 설명과 길이가 짧은 품목명을 효과적으로 매핑

추천시스템1) Reranking

산업통상자원부

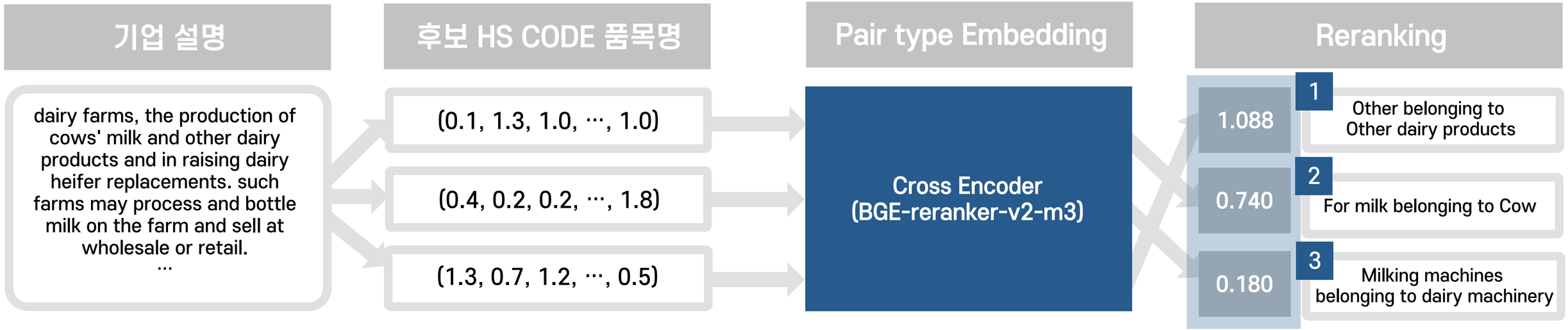


추천시스템1) Reranking

산업통상자원부

Reranking

- Cross encoder 임베딩을 통해 모든 HS CODE와 문맥적 관련성을 파악할 수 있음
- 비슷한 맥락이나 단어를 가진 문장 사이의 유사성을 잘 파악해 1차 mapping의 후보 HS CODE에 정확도 보조 역할을 수행
- 1차 mapping으로 도출한 후보 HS CODE를 대상으로 적용하므로, 계산 부담이 적음



Relevance score: 기업 설명과 후보 HS CODE의 관련성 점수

* 교차 엔트로피 손실을 최소화하는 방향으로 최적화될 때 '얼마나 관련있는지'를 나타내는 점수

Relevance score를 기준으로 순위 재정렬 가능

추천시스템2) VAE

산업통상자원부

VAE (Variational Autoencoder)

- 데이터를 저차원 벡터로 압축하고 복원하는 과정에서 데이터의 특징을 학습함
- 잠재 공간(latent space)을 확률적으로 모델링하여 데이터의 내재된 의미를 효과적으로 추출함
- 이러한 특성으로 인해 VAE는 추천시스템 분야에서 주로 쓰이는 모델 중 하나임

유사도 기반으로 추천하는 방식의 문제점

1. 서비스 품목 등의 경우 관련성이 낮은 품목을 추천하는 문제가 발생
2. 몇 개의 샘플에서 HS CODE 앞자리가 유사한 품목 위주로 추천하는 경향

VAE 추천시스템을 활용해 데이터 간의 깊은 관계를 학습하여 HS CODE를 추천하도록 함
앞서 도입한 방식에서 추천시스템 보완 가능

추천시스템2) VAE

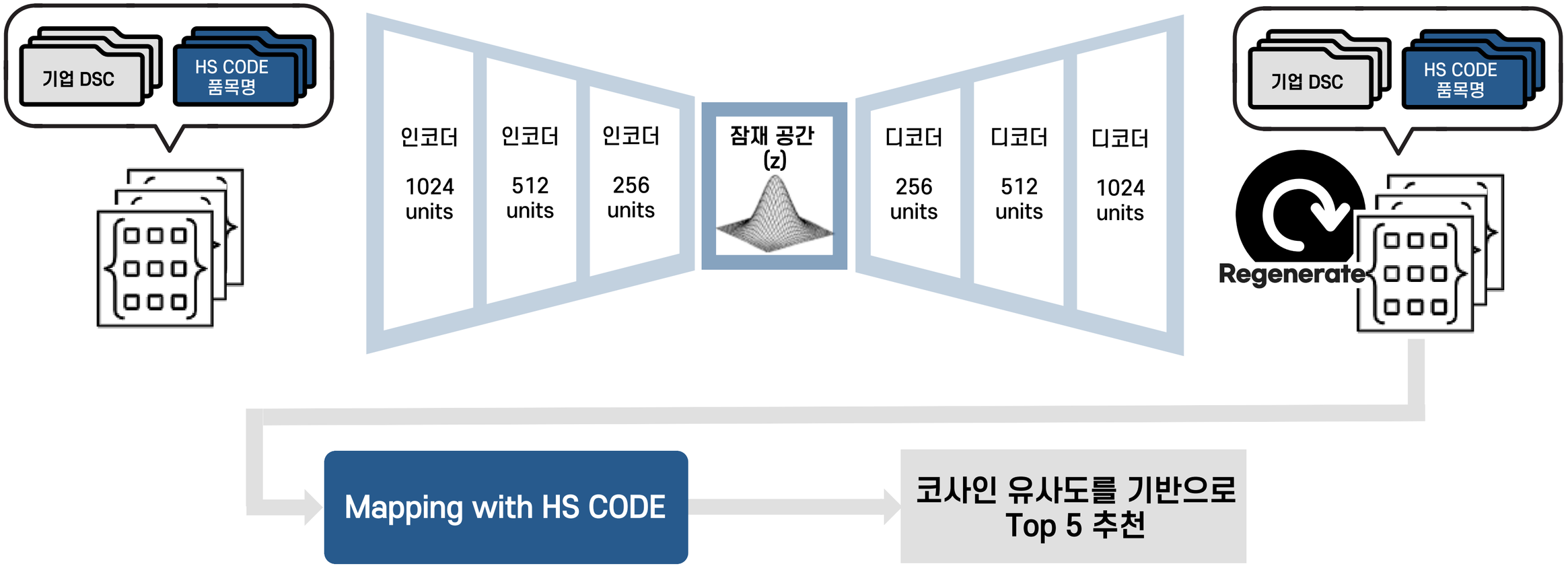
산업통상자원부

Input : 코사인 유사도

인코더 (Encoder)

디코더 (Decoder)

Output : 복원 데이터



Mapping with HS CODE

코사인 유사도를 기반으로
Top 5 추천

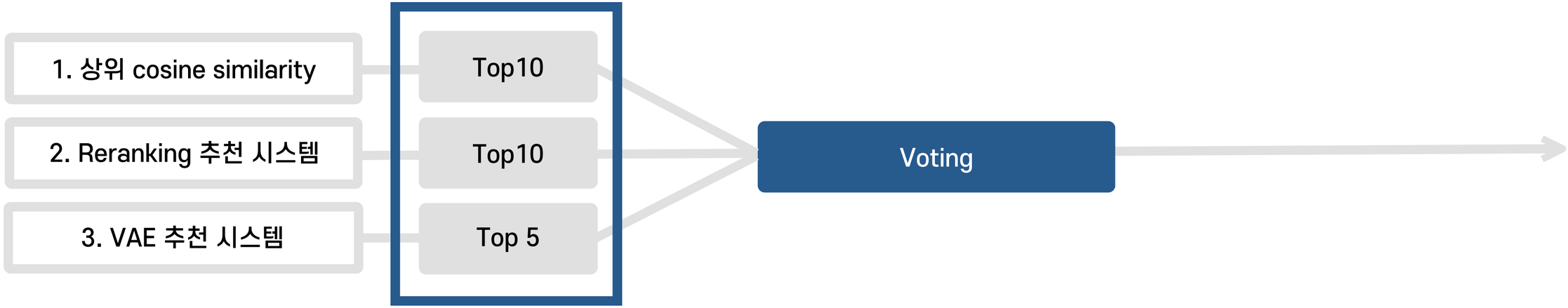
데이터의 내재된 의미와 전체적인 관계를 효과적으로 학습하여
서비스 품목 같이 추론이 필요한 해외기업 설명(DSC)에 더 정확한 HS CODE를 추천

Ensemble Mapping

산업통상자원부

3가지 기법을 Voting 방식으로 앙상블

1. 다양성 확보 : 코사인 유사도 외의 다양한 추천 기법을 활용
2. 정확도 향상 : 2개 이상의 기법이 추천한 코드를 최종적으로 선정



앙상블 가중치 Tuning by Human Labeling

1. 비지도 학습의 한계 극복 : 성능 평가를 위해 수작업으로 약 100개 기업에 대하여 가상 답안을 직접 작성
2. Test 성능을 기준으로 가중치 부여 : Top10, Top10, Top5 로 추천 시 가장 높은 성능

➡ 휴리스틱에 객관성을 부여

Filtering

산업통상자원부

예외 사항 처리

: 기업 설명에 명시된 제외 품목(except, other than) 을 추천에서 제외

Filtering → Final prediction

기업 설명	추천 코드	Filtering
: 낙농 농장, 우유, 유제품 생산, 젖소 사육, 우유(**냉동 제외**)를 가공해 판매	1. 기타 유제품 2. 사료용 유제품 3. 젖소 4. 연유 5. **냉동 우유** 6. 낙농용 착유기 7. 축우용 사료	1. 기타 유제품 2. 사료용 유제품 3. 젖소 4. 연유 ~~5. 냉동 우유~~ 6. 낙농용 착유기 7. 축우용 사료

: 제외할 품목 벡터와
HS CODE 품목명 벡터간
코사인 유사도가 높은 경우를
추천에서 제외

➡ 세부 사항을 효과적으로 고려
➡ 기업 맞춤형 추천을 실현

CH03. 사업화 방안 및 기대효과

강점

산업통상자원부

1. 모든 프로세스 자동화 가능

- 기업 설명 추가 보완, Mapping부터 Reranking, VAE 추천 시스템까지 자동화를 통해 인간 개입 최소화
- 언어모델과 수리적 자동 계산을 통해 객관적 기업 매칭 가능

2. 비용 효율화

- Open source인 Llama3를 무료로 사용함으로써 사업 비용 절감
- 추론 과정을 LLM이 함으로써 수동 작업에 드는 시간과 노력을 줄일 수 있음

3. 일괄 적용 가능

- Llama3으로 일관된 형식과 품질의 문장을 빠르게 대량 생산 가능
- 모든 프로세스 코드를 구축함으로써, 데이터 입력 후 일괄 적용 및 결과 도출 가능

보완(제언)

산업통상자원부

1. ISIC4 & HS6 CODE 필터링

2. LLM을 통한 service 제공 기업 분류

IDEA

통계청 국제표준산업분류 HSCODE 6단위 매핑 데이터를 활용하여 필터링 방식을 강화

BACKGROUND

1. 기업 설명과 ISIC4 코드는 매핑된 상태
2. 우리나라의 HSCODE 10단위의 앞 6자리는 국제 HSCODE 6단위와 동일, 뒤 4자리는 국제 HSCODE 6단위 이상의 세분류

DETAIL

해외 기업이 해당하는 ISIC4 코드에 매핑된 HSCODE 6단위와 추천 HSCODE 10단위의 앞 6자리가 일치하는 경우로 필터링

➡ **추천 정확도 향상**

보완(제언)

산업통상자원부

1. ISIC4 & HS6 CODE 필터링

2. LLM을 통한 service 제공 기업 분류

IDEA

LLM에게 해외 기업 설명을 제공한 후
해당 기업이 무형의 서비스를 다루는지를 질문

BACKGROUND

1. 서비스의 수입 및 수출에 적용되지 않는 관세
2. HS CODE 기준의 분류는 재화 품목이 과반수로
서비스 제공 기업에 대한 맞춤 추천의 한계

DETAIL

실제 활용한 LLM 프롬프트 내용

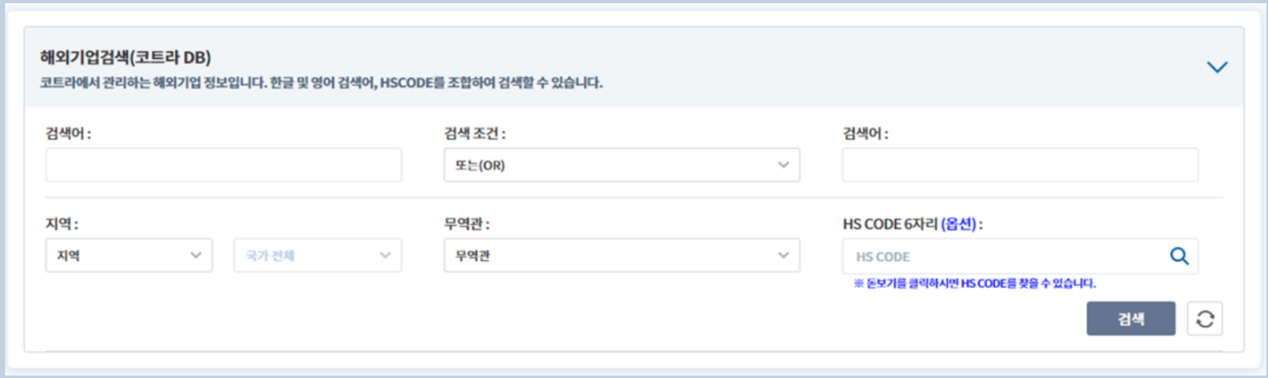
Does this company provide intangible services(not tangible goods)?
Only return true or false. The following is description of a business :

➡ **추천에서 예외적으로 고려할 기업을 손쉽게 탐지**

사업화 방안

산업통상자원부

KOTRA의 유료 서비스인 'TriBIG 프리미엄 서비스 ' 강화



1. 기존 서비스의 정확도를 개선하고 높은 품질의 정보를 제공

- 해외 바이어에 대한 풍부한 정보 획득 가능
- LLM을 활용해 수출 관련 품목을 자동으로 생성할 수 있어 사람의 추론을 대신할 수 있음
- 추천 시스템에 적용하여 해외 바이어의 니즈에 맞는 품목을 다양하게 제시 가능

➡ 두 가지 추천 시스템을 결합해 서비스 정확도 향상

2. 기존 서비스의 정확도를 개선하고 높은 품질의 정보를 제공

- 해외 바이어 정보 생성부터 추천 시스템까지 자동화가 가능
- 즉시 정보를 업데이트하고 추천 시스템에 적용하여 신뢰할 수 있는 서비스를 지속적으로 유지

기대효과

산업통상자원부

중소, 중견기업에 비교적 저렴한 가격으로 서비스 제공

- 최소의 비용으로 Tri-Fusion 매핑 추천 시스템 구축 가능

기업 경쟁력 강화

- 최신 시장 동향과 바이어 정보를 신속하게 획득 가능
- 국내 시장을 넘어 글로벌 시장에서 경쟁력을 갖출 수 있음

최적의 바이어 매칭으로 시간, 비용 지출 감소

- 적절한 바이어를 매칭하여 거래 성사율을 높이고 기업의 매출 증대를 기대할 수 있음
- 무역 활동의 효율성 증가
- 글로벌 시장에서의 우위 선점 가능

기업 업무 프로세스 혁신

- 수작업을 AI가 함으로써 일관성 확보 및 자동화를 통해 업무 프로세스의 혁신적 변화를 도모
- 인적 자원의 효율적 배치와 기업 운영의 생산성 향상

참고문헌

- Yuichi Sasazawa, Kenichi Yokote, Osamu Imaichi, Yasuhiro Sogawa. 2023. “Text Retrieval with Multi-Stage Re-Ranking Models”

- 오동석, 김수완, 박기남, 임희석, 2022. “문장 임베딩을 위한 Cross-Encoder의 Re-Ranker를 적용한 의미 검색 기반 대조적 학습”, 한국정보과학회 언어공학연구회

- Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, Tony Jebara. 2018, "Variational Autoencoders for Collaborative Filtering"

Thank you