미래기술마당 개선 프로젝트

공공빅데이터 일경험 수련생

분석 결과 보고서

01: S-BERT 기반 수요기술 매칭 서비스

기 관 명 과학기술일자리진흥원
수행기간 2022년 10월 ~ 2023년 01월
수 련 생 김인수 황양하

행정안전부

NIA 한국지능정보사회진흥원

CSLEE
㈜씨이에스리

목차

S-BERT기반 수요기술 매칭 서비스

01

프로젝트 개요

전체프로젝트 개요

미래기술마당 **서비스 개선**을 위한 프로젝트 진행

프로젝트 1:
S-BERT 기반 수요기술 매칭 서비스

수요자가 신청한 수요기술의 내용을 기반으로 유사한 특허를 찾아내어 수요자에게 유사 특허 리스트를 제공

프로젝트 2:
BI 기반 통계 서비스

미래기술마당 관리자 통계 화면의 문제점을 개선하고, 사용자를 위한 검색어 통계 시각화 서비스를 제공

프로젝트 3:
자연어 처리 기반 특허 기술 분류

자연어 처리 기술을 활용하여 특허 데이터를 미래유망기술(6T) 분류 기준으로 분류하는 모델을 제작

프로젝트 진행 과정


9월
10월
11월
12월
1월
2월
사업화유망기술 DB정제
SBERT 기반 수요기술 매칭 서비스
BI기반 통계 서비스
모델 개선 & 자연어 처리 기반 특허 분류

02

분석 배경

수요기술매칭이란?

문제점

해결방안

PROBLEM

내가 찾고 싶은 **수요기술**,
바로 **추천** 받을 수 없을까?

미래기술마당의 서비스 중 하나인 **수요기술 매칭 서비스**에서 사용자 불편 요소를 찾음.
자연어 처리 알고리즘 중 하나인 **Sentence-BERT**를 이용해 사용자의 불편을 개선하고자 함.

미래기술마당 **수요기술매칭** 서비스란?

미래기술마당 수요기술매칭 서비스란,
대학 및 출연(연)이 보유한 **사업화 유망기술**과 **기술 수요기업의 매칭**을 통해
공공기술의 **기술이전 및 창업 활동**을 지원하는 서비스

사업화 유망기술

미래기술마당에서
유망기술과 수요기술을 매칭

수요자 – 공급자 매칭 Facilitator

기술 수요자

수요기술 매칭 서비스에서 발견한 **문제점**

Pain Point 1

수요기술 매칭이 수동으로 이루어지는 시스템

수요기술 접수부터 기술 매칭, 신청자 수신까지의 과정이 수동으로 이루어지고 있어 진행 과정이 매우 느림.

Pain Point 2

수요기술의 매칭 여부를 알 수 없는 시스템

수요자와 기술 권리자의 매칭 여부를 알 수 없어 수요자가 기술 이전 진행 과정에 대해 빠르게 알기 어려움

기술 수요자

S-BERT 기반
수요기술 매칭 서비스

기술 권리자

Pain Point 1

수요기술 매칭이
수동으로 이루어지는 시스템

Solution1

수요자가 등록한 **수요기술 상세정보**를 기반으로 AI 기술을 통해
상세정보와 유사한 내용의 **특허를 추천**해주는 서비스를 제공.

Pain Point 2

수요기술의 매칭 여부를
알 수 없는 시스템

Solution2

수요자에게 추천된 특허 리스트에서 매칭에 성공한 경우,
자동으로 특허 권리자에게 메일을 보내 **기술이전을 진행**할 수 있도록 함.

03

모델 프로세스

프로젝트 개요

모델 프로세스

모델 시연

DATA

프로젝트 활용 데이터

데이터	건수
KIPRIS 특허 서지 정보	1,588,185건
KPEG 특허 평가 정보	1,262,211건

TOOL

프로젝트에서 사용한 툴

PyTorch

TensorFlow

Flask
web development,
one drop at a time

Hugging Face

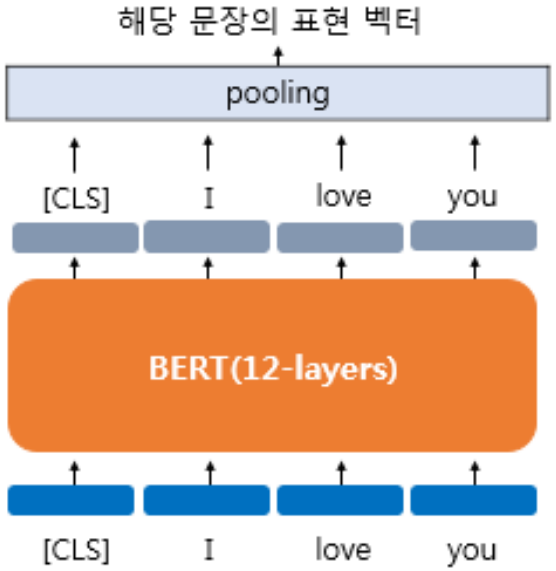
수요기술 추천 서비스 **프로세스**

S-BERT 기반 수요기술 매칭 서비스를 통해 **수요기술 매칭이 활발**해 질 수 있음을 기대

수요자가 등록한
수요기술정보

수요기술명과
특허 서지정보 사이
유사도 측정

수요기술과 **유사한**
추천 특허 리스트

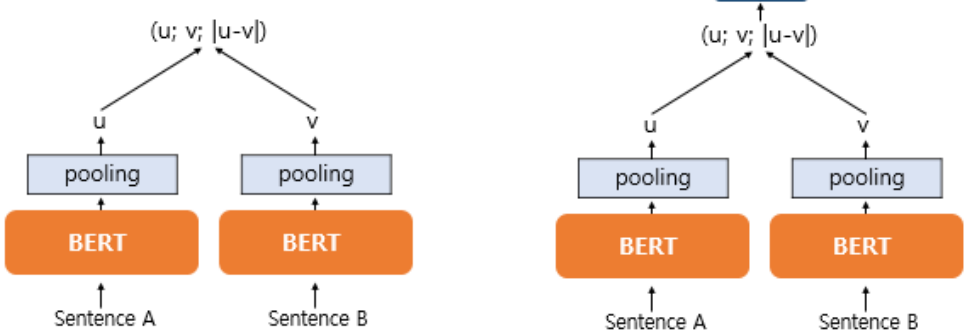


수요기술정보와 특허 서지정보 사이의 유사도를 측정하기 위해서는
사람이 쓰는 **자연어**를 **기계가 이해**할 수 있도록 변화시켜 주어야 함.

즉, 자연어를 숫자 집합인 벡터로 바꾸어주는 **임베딩** 과정을 거쳐야 함.

특허 서지 정보를 임베딩하기 위해 **Sentence-BERT** 를 활용함.

왜 **Sentence BERT**를 사용했을까?



다른 방식으로 임베딩을 진행하고 추천 결과를 보았을 때
단어의 문맥을 잘 파악하지 못하는 경우가 발생했음.
이러한 문제들을 해결하기 위해 **S-BERT 모델**을 사용함.

S-BERT는 **의미론적으로 의미 있는 문장 임베딩**을 도출할 수 있도록
BERT를 Fine-Tuning하여 만들어진 모델.

특허 서지정보의 문장들의 **문맥을 충분히 반영**하여
추천 시스템의 성능을 향상시키기 위해 S-BERT모델을 선택했음

임베딩을 위한 **S-BERT** 모델 제작 과정

사전학습 모델 선정
BERT 다국어 지원 모델 선정

▶

한국어 추가 사전 학습
수집된 말뭉치로 Domain Token
생성 후 기존 모델에 추가 학습

▶

학습된 BERT를 통해
S-BERT 모델 제작

▶

Sementic 검색 모델 구축

STEP 1

사전학습 모델 설정

학습 대상이 될 BERT 다국어 지원 모델을 선정

KLUE

KLUE(Korean Language Understanding Evaluation)-BERT

대표적인 한국어 BERT모델인 KLUE는 64GB 상당의 한국어 학습데이터를 통해 학습된 모델이며,
Morpheme-based subword 토크나이저를 사용하여 학습되었음.
한국어 데이터에 특화된 BERT 모델이기 때문에 학습 대상 모델로 선택했음.

STEP 2

한국어 추가 사전 학습

Domain에 대해 수집된 말뭉치로 기존 BERT모델에 추가 학습 진행

Step 1

KIPRIS

카바메이트 펩타이드 이산화탄소 데쿠르시놀 뉴럴네트워크 앱 ##타머

KIPRIS의 특허 서지 정보 데이터를
Special domain 말뭉치로 활용하여
기존 모델 서브 토크나이저에 추가 학습함.

Step 2

한국특허정보원
Korea Institute of Patent Information

Special domain 말뭉치와 KLUE,
한국특허정보원에서 개발한
특허 분야 언어 모델인 KoPatElectra의
단어 사전을 모두 통합하여 모델에 학습.

Step 3

[CLS] 앱타머, 를, 활용한, 검출, 키트 [SEP]

↑

Masked Language Model

↑

[CLS] 앱타머, ##를, [MASK] 검출, 키트 [SEP]

↑

앱타머를 이용한 검출 키트

Special Domain이 학습된 말뭉치로
기존 BERT모델에
Masked Language Model 추가 학습.

STEP 3

S - BERT

앞서 학습한 BERT로 S - BERT 모델 제작

Special Domain으로
학습된 BERT 모델

자연어 추론 학습
Natural Language Inference

두 문장이 주어졌을 때,
하나의 문장이 다른 문장과
논리적으로 어떤 관계에
있는지 분류하는 것.

STS 학습
Semantic Textual Similarity

서로 다른 두 문장이 주어졌을 때,
두 문장이 얼마나 유사한 지
그 유사도를 판별하는 것.

S – BERT
모델 완성

Special Domain으로 학습된 BERT모델에 NLI학습과 STS학습을 진행하여 **S - BERT 모델**을 제작함.

STEP 4

Semantic 검색 모델 구축

S - BERT와 Faiss를 이용하여 Semantic 검색 모델 구축

시맨틱 검색(Semantic search)은 기존의 키워드 매칭이 아닌 **문장의 의미에 초점**을 맞춘 정보 검색.
Faiss를 이용하여 보다 **빠르고 의미론적인 검색**에 초점을 맞춘 모델을 만들었음.

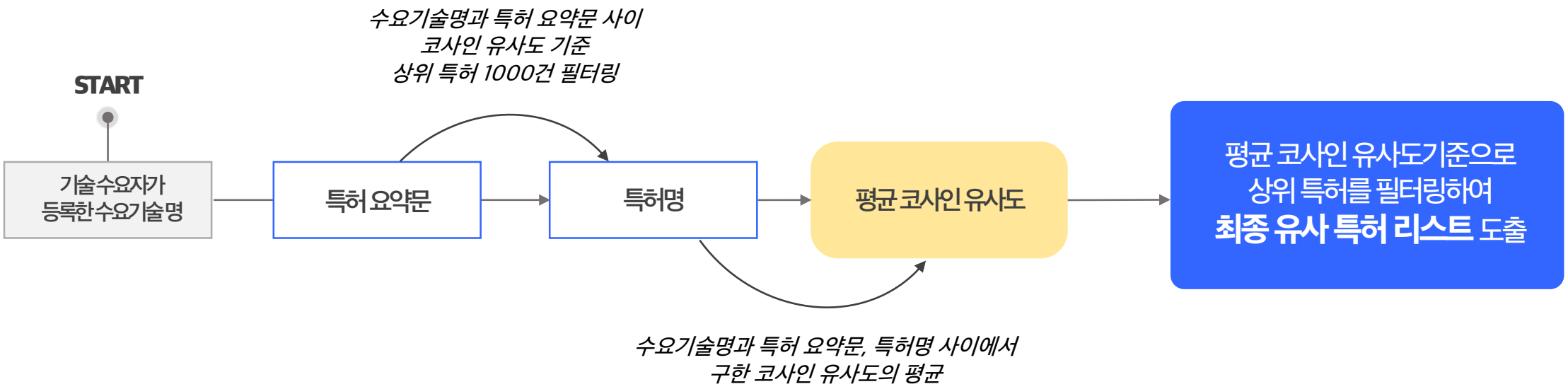
질의문 → 학습시킨 S-BERT로 질의문 인코딩 → 벡터화 된 질의문 → Faiss를 통해 서지정보 임베딩 값을 인덱싱 후 비슷한 문서를 찾음 → *가장 유사한 문서 도출* → 유사 문서 결과

서지정보 임베딩 값 ← 학습시킨 S-BERT로 특허 서지 정보 인코딩

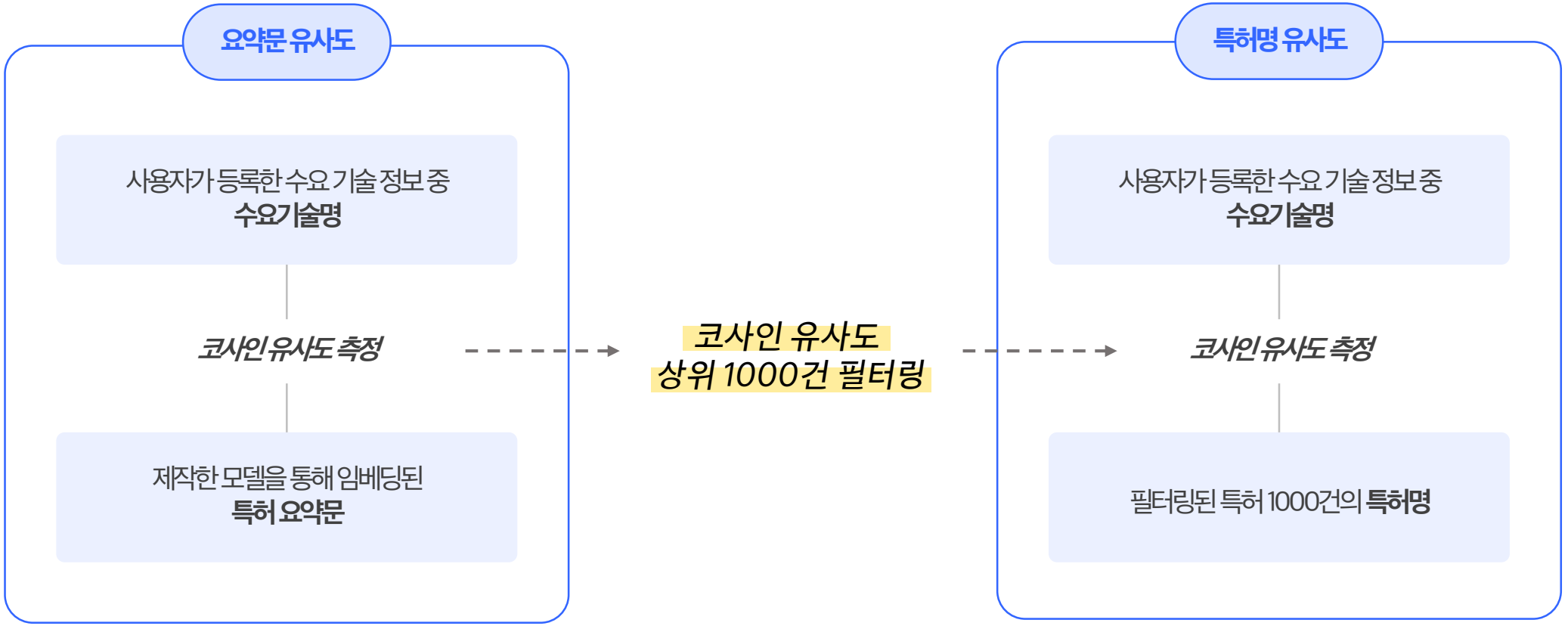
Faiss는 GPU를 지원하는 라이브러리로
sklearn보다 빠르고 강력하게 유사도를 측정할 수 있음

유사 특허 추천 서비스 **프로세스**

다음과같은 프로세스를 거쳐 기술수요자가원하는기술과유사한특허를 찾아냄



수요기술명과 **특허 요약문** 사이의 코사인 유사도를 구해 상위 1000건의 유사 특허를 필터링 하고,
필터링 된 특허의 **특허명**과 수요기술명의 코사인 유사도를 다시 측정하여 최종 유사 특허를 결정.



왜 **특허명**과 **특허 요약문**을 나누어서 **유사도**를 측정할까?

특허명만으로는 특허의
정확한 내용을 파악할 수 없음.

특허 요약문까지 모두 고려하여
유사도 측정을 진행.

두 유사도 값의 평균으로
최종 유사도 평가 지수 산출.

특허명만으로는 특허의 **정확한 내용**을 파악할 수 없음.

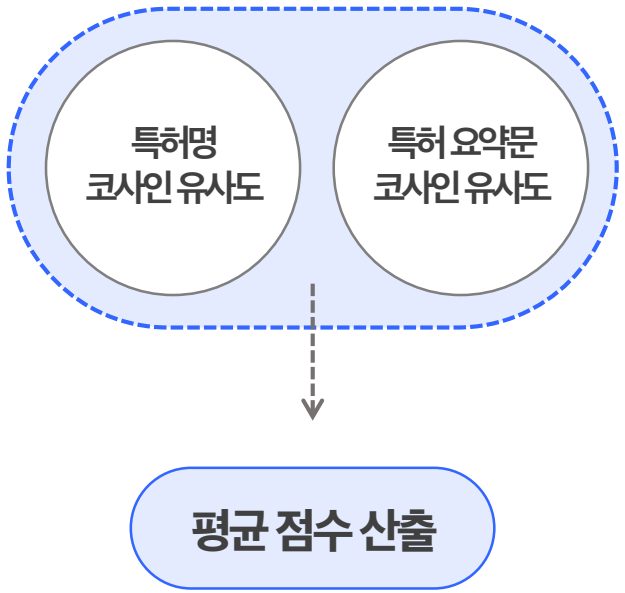
특허명	*특허 요약*
공기정화장치	본 발명은 공기정화장치에 관한 것으로, 더욱 상세하게는 분진이 포함된 고온의 공기를 냉각된 처리수를 통해 분진을 흡착 제거하고, 고온의 공기와 열 교환한 가열된 처리수를 열 교환하여 냉각하고, 열 교환된 열은 재활용할 수 있는 공기정화장치에 관한 것이다
공기정화장치	본 발명은 산소의 공급을 위한 공기정화장치에 관한 것으로, 종래의 연탄가스중독 방지시스템에 있어 예방은 되나 이미 연탄가스에 중독되었거나 공해로 인한 천식환자 또는 응급환자에게 짧은 순간에 산소를 다량으로 공급하지 못하는 장치의 한정적인 문제를 해결하고, 종래의 가습 수단과, 공기중의 산소함량을 디지탈 정보 디스플레이 하여 주며, 자동적으로 산소함량이 낮을 때는 경보를 울려주고, 높을 때는 산소 공급을 중단하며, 동시에 배기송풍기를 자동으로 온, 오프 시켜 실내에 산소를 공급하여 주는데 있어서, 상기 배기 송풍기는 변압기 없는 직류 전원에 의해 작동되어 오염된 공간에 항상 맑은 공기를 제공할 수 있다

➡ 분진이 포함된 공기를 **냉각된 처리수를 통해 정화**하는 장치

➡ 공기중의 산소 함량을 감지하여 **산소를 공급해주는** 장치

위의 경우처럼, **특허명은 동일**하지만 특허 요약문을 살펴 보았을 때 **전혀 다른 특허**인 경우가 있음.
이러한 경우를 고려하여 특허명과 특허 요약문을 나누어서 유사도를 측정했음.

특허명 유사도와 특허 요약문 유사도의 평균을 산출해 값이 큰 순서대로 정렬하여 최종 추천 리스트 완성



추천 유사 특허 리스트

특허명	Mean Score
딥러닝기반 영상내 객체 인식 시스템	0.90
GPU장치를 기반으로하는 딥러닝분석을 이용한 영상 보정방법	0.89
딥러닝 인공신경망 기반 영상 인식방법 및 시스템	0.856
인공지능 기반 사물인식을 활용한 가상 증강 정보제공 시스템 및 그 방법	0.851
⋮	⋮

유사도 점수의 평균값이 큰 순서대로 정렬하여 특정한 수(10~20개)의 상위 값을 출력해
출원 번호, 특허명, 특허 요약, 평균 유사도 점수를 제공

모델을 통해 찾은 최종 유사 특허 리스트를 사용자에게 제공하기 위해 **대시보드**를 제작함.

최종 추천 특허 리스트

Flask
web development,
one drop at a time

Flask를 이용해
추천 결과를 웹 페이지에
JSON형식으로 넘겨줌

Dash
by plotly

JSON형식으로 넘겨받은
추천 특허에 대한 특허 평가 정보를
시각화하여 대시보드 형태로
사용자에게 제공

User Interface

04

프로젝트 결과

프로젝트 결과

고도화 방안

미래기술마당 **유사 특허 탐색** 대시보드

미래기술마당 수요기술 검색 페이지에서 유사 특허를 탐색할 수 있는 **API**와 **시각화 대시보드** 개발

유사특허 탐색

initial value

	출원번호	특허명	pqi지수	권리성평가지수	기술성평가지수	상업성평가지수	유사도	링크
☐	1020120155653	IE형 트렌치 게이트 IGBT	1.94	66	78.2	72.6	0.57	Link
☐	1020150089927	보호 IED의 Live 시험 시스템 및 그 방법	0.58	87	80.2	55.8	0.57	Link
☐	1020060087451	TII 디코더 및 디코딩 방법	0.52	84	82.2	66.6	0.56	Link
☐	1020120126208	상태 기반의 테스트 시나리오 모델을 이용한 GUI 테스트 장치 및 방법	0.48	72	78.2	55.8	0.55	Link
☐	1020160006528	EMI 저항 및 통전 검사 장치	0.37	69	78.2	53.4	0.55	Link
☐	1020150138985	MRI용 다채널 RF 코일 어레이	0.52	75	80.2	54.6	0.54	Link
☐	1020180005855	니모닉 기반의 GUI 테스트 자동화 방법 및 이를 이용하는 장치	1.61	90	76.2	55.8	0.54	Link
☐	1020150179963	BCI(Bulk Current Injection) 테스트 장치 및 BCI 테스트 방법	0.64	81	78.2	67.8	0.54	Link
☐	1020140109627	IEEE 11073 서비스 제공 방법 및 시스템	0.41	69	78.2	54.6	0.53	Link
☐	1020040065449	DDI의 검증방법	0.43	69	78.2	64.2	0.53	Link

« ‹ 1 / ›

유사 특허 결과와 함께 다양한 **특허 관련 정보**를 볼 수 있음.

유사 특허 검색

검색창에 수요기술명을 입력하면
S-BERT 기반 유사 특허 결과 출력

유사 특허 평가 지수

출력된 유사 특허 중 사용자가 원하는
특허를 선택할 수 있도록
각 특허 별 평가 지수를 함께 출력

유사 특허 선택

선택한 특허의 평가 지수를
그래프로 확인할 수 있음

딥러닝을 활용한 영상인식

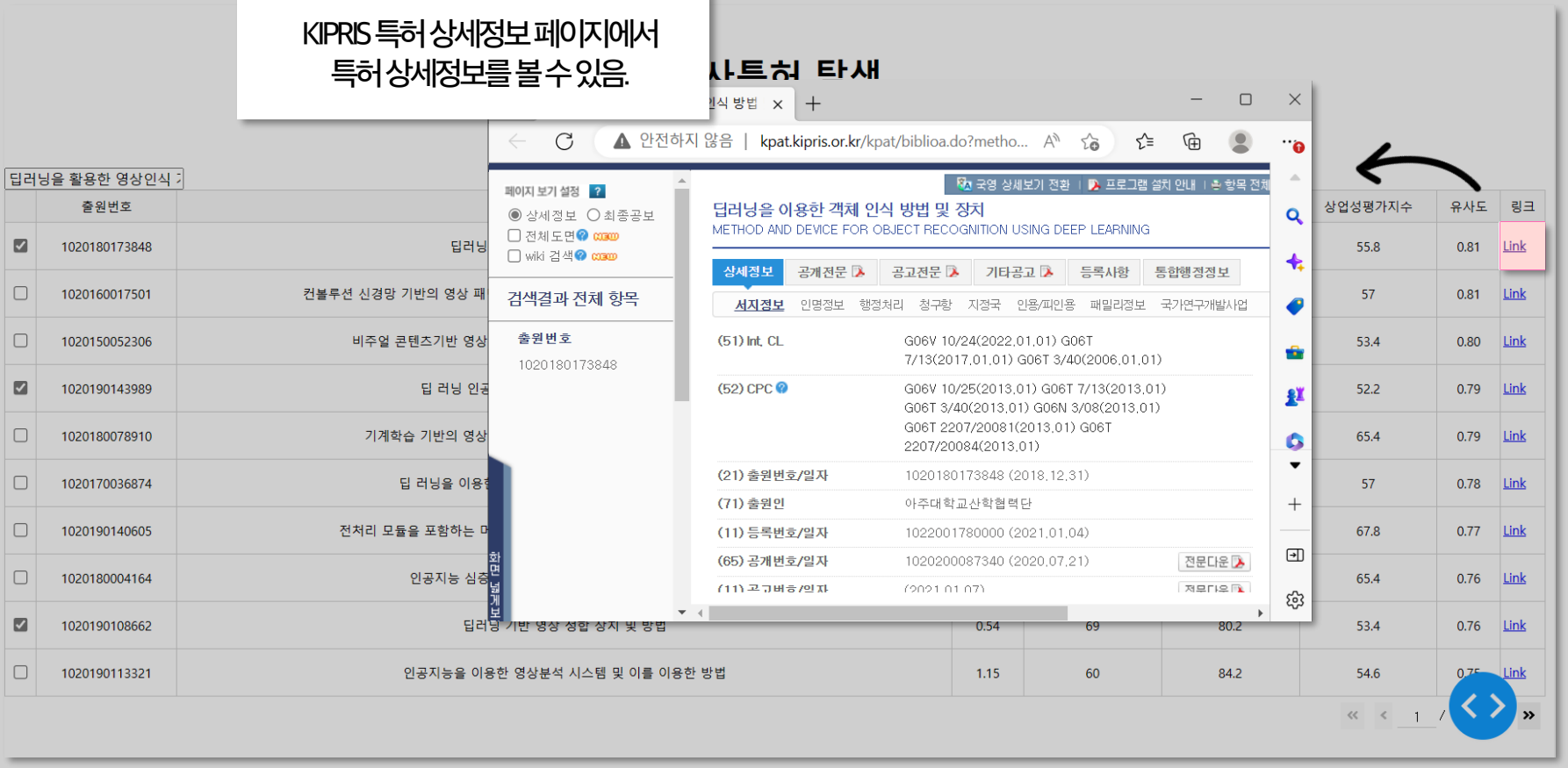
	출원번호	특허명	pqi지수	권리성평가지수	기술성평가지수	상업성평가지수	유사도	링크
☑	1020180173848	딥러닝을 이용한 객체 인식 방법 및 장치	1.11	72	80.2	55.8	0.81	Link
☐	1020160017501	컨볼루션 신경망 기반의 영상 패턴화를 이용한 딥러닝 시스템 및 이를 이용한 영상 학습방법	0.77	66	94.1	57	0.81	Link
☐	1020150052306	비주얼 콘텐츠기반 영상 인식을 위한 딥러닝 프레임워크 및 영상 인식 방법	0.68	78	78.2	53.4	0.80	Link
☑	1020190143989	딥 러닝 인공신경망 기반 영상 인식 방법 및 시스템	0.99	60	82.2	52.2	0.79	Link
☐	1020180078910	기계학습 기반의 영상 인식 방법 및 기계학습 기반의 영상 인식 시스템	1.31	75	80.2	65.4	0.79	Link
☐	1020170036874	딥 러닝을 이용한 인공지능 기반 영상 감시 방법 및 시스템	1	72	82.2	57	0.78	Link
☐	1020190140605	전처리 모듈을 포함하는 머신 러닝 기반의 인공지능을 이용하는 영상 분석 장치	1.54	75	80.2	67.8	0.77	Link
☐	1020180004164	인공지능 심층학습 기반의 영상물 인식 시스템 및 방법	1.13	75	80.2	65.4	0.76	Link
☑	1020190108662	딥러닝 기반 영상 정합 장치 및 방법	0.54	69	80.2	53.4	0.76	Link
☐	1020190113321	인공지능을 이용한 영상분석 시스템 및 이를 이용한 방법	1.15	60	84.2	54.6	0.75	Link

« ‹ 1 / »

특허 상세 정보

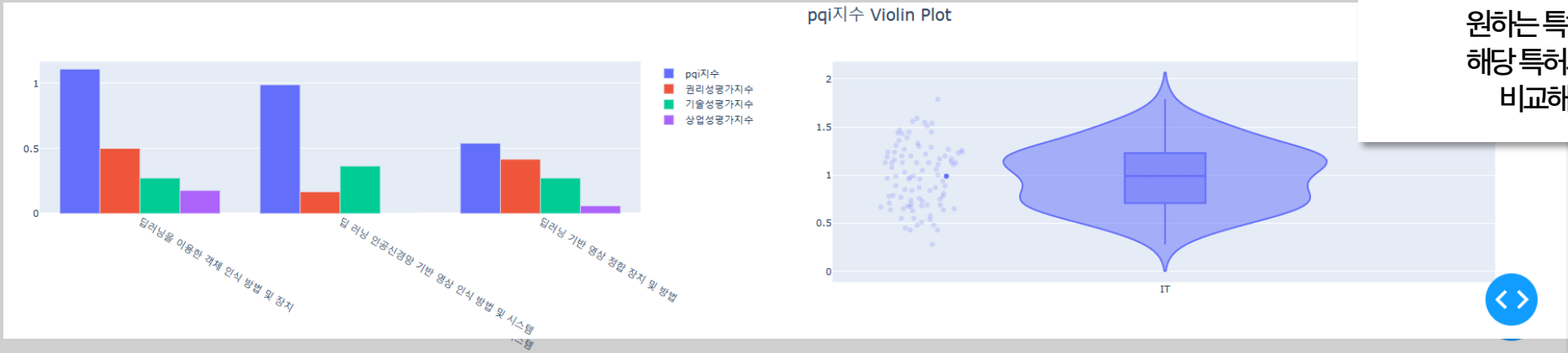
링크를 클릭하면
KIPRIS의 특허 상세정보
페이지로 접속할 수 있음

KIPRIS 특허 상세정보 페이지에서
특허 상세정보를 볼 수 있음.



	출원번호	특허명	pqi지수	권리성평가지수	기술성평가지수	상업성평가지수	유사도	링크
☑	1020180173848	딥러닝을 이용한 객체 인식 방법 및 장치	1.11	72	80.2	55.8	0.81	Link
☐	1020160017501	컨볼루션 신경망 기반의 영상 패턴화를 이용한 딥러닝 시스템 및 이를 이용한 영상 학습방법	0.77	66	94.1	57	0.81	Link
☐	1020150052306	비주얼 콘텐츠기반 영상 인식을 위한 딥러닝 프레임워크 및 영상 인식 방법	0.68	78	78.2	53.4	0.80	Link
☑	1020190143989	딥 러닝 인공신경망 기반 영상 인식 방법 및 시스템	0.99	60	82.2	52.2	0.79	Link
☐	1020180078910	기계학습 기반의 영상 인식 방법 및 기계학습 기반의 영상 인식 시스템	1.31	75	80.2	65.4	0.79	Link
☐	1020170036874	딥 러닝을 이용한 인공지능 기반 영상 감시 방법 및 시스템	1	72	82.2	57	0.78	Link
☐	1020190140605	전처리 모듈을 포함하는 머신 러닝 기반의 인공지능을 이용하는 영상 분석 장치	1.54	75	80.2	67.8	0.77	Link
☐	1020180004164	인공지능 심층학습 기반의 영상물 인식 시스템 및 방법	1.13	75	80.2	65.4	0.76	Link
☑	1020190108662	딥러닝 기반 영상 정합 장치 및 방법	0.54	69	80.2	53.4	0.76	Link
☐	1020190113321	인공지능을 이용한 영상분석 시스템 및 이를 이용한 방법	1.15	60	84.2	54.6	0.75	Link

« ‹ 1 / 10 › »

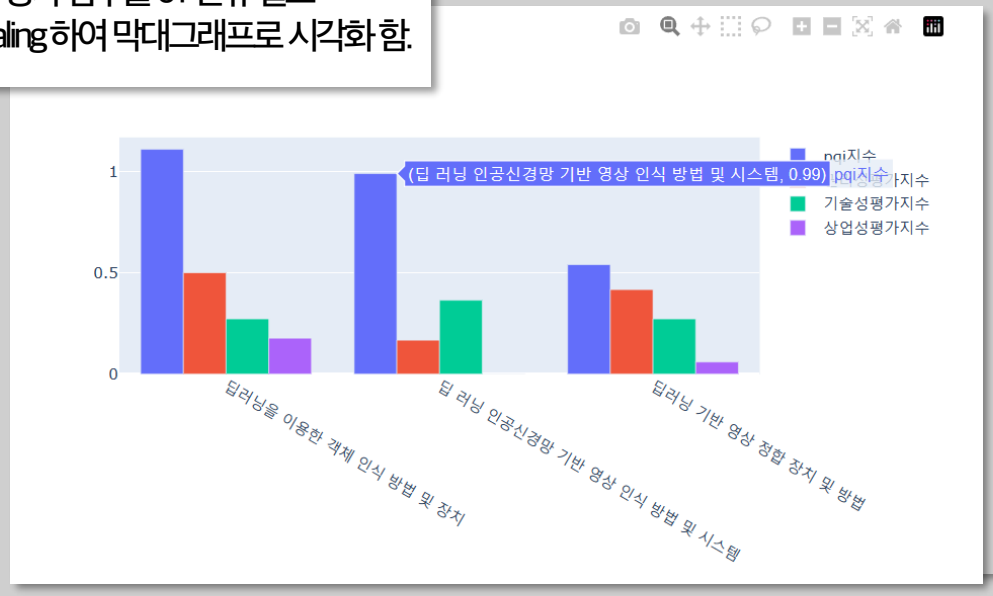


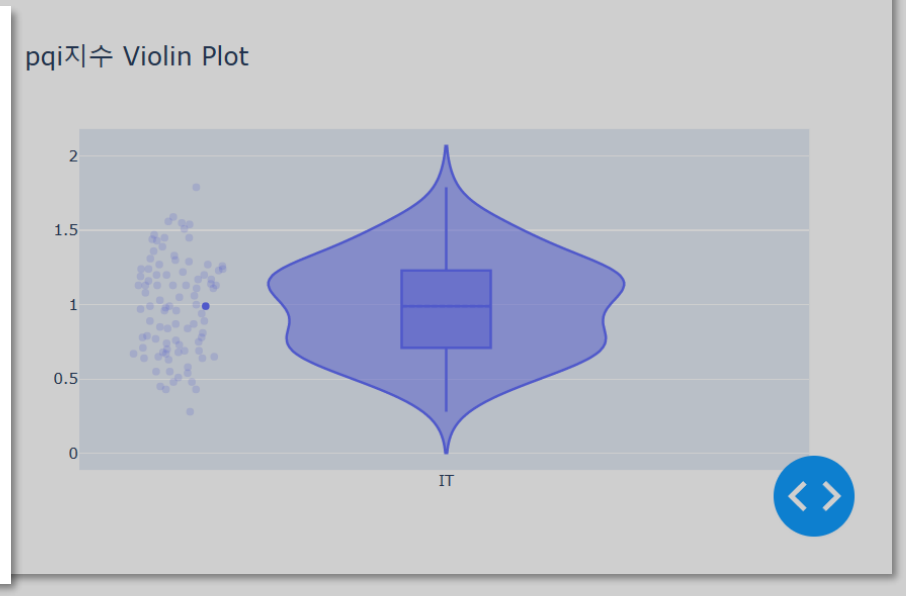
원하는 특허를 선택하면
해당 특허의 평가정보를
비교해 볼 수 있음

☐	1020170036874	딥 러닝을 이용한 인공지능 기반 영상 감시 방법 및 시스템	1	72	82.2	57	0.78	Link
☐	1020190140605	전처리 모듈을 포함하는 머신 러닝 기반의 인공지능을 이용하는 영상 분석 장치	1.54	75	80.2	67.8	0.77	Link
☐	1020180004164	인공지능 심층학습 기반의 영상물 인식 시스템 및 방법	1.13	75	80.2	65.4	0.76	Link
☑	1020190108662	딥러닝 기반 영상 정합 장치 및 방법	0.54	69	80.2	53.4	0.76	Link
☐		인공지능을 이용한 영상분석 시스템 및 이를 이용한 방법	1.15	60	84.2	54.6	0.75	Link

« ‹ 1 / 10 › »

특허 평가정보를 한눈에 볼 수 있도록
모든 평가 점수를 6T 분류 별로
Min-Max Scaling 하여 막대그래프로 시각화 함.

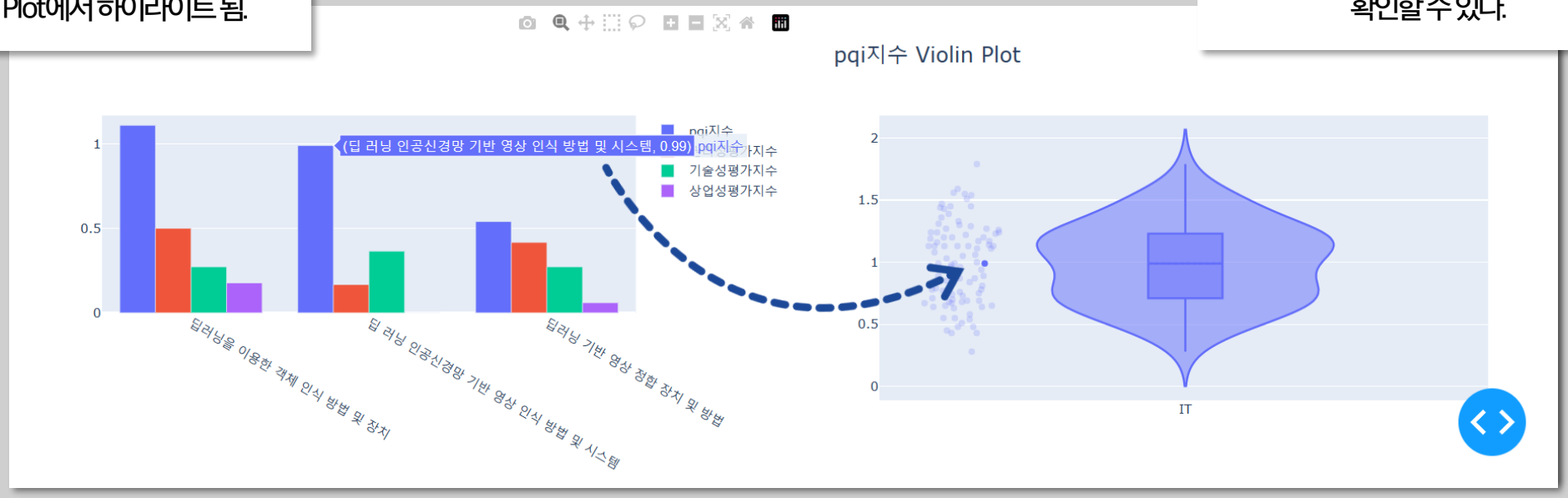




☐	1020170036874	딥 러닝을 이용한 인공지능 기반 영상 감시 방법 및 시스템	1	72	82.2	57	0.78	Link
☐	1020190140605	전처리 모듈을 포함하는 머신 러닝 기반의 인공지능을 이용하는 영상 분석 장치	1.54	75	80.2	67.8	0.77	Link
☐	1020180004164	인공지능 심층학습 기반의 영상물 인식 시스템 및 방법	1.13	75	80.2	65.4	0.76	Link
☑	1020190108662	딥러닝 기반 영상 정합 장치 및 방법	0.54	69	80.2	53.4	0.76	Link
☐		인공지능을 이용한 영상분석 시스템 및 이를 이용한 방법	1.15	60				

그래프에 마우스 오버하면
해당 데이터의 값이
Violin Plot에서 하이라이트 됨.

검색결과 중 선택한 특허와 동일한
6T분류 데이터 내 평가 지수별 분포를
확인할 수 있다.



방안 1:
특허 데이터

수요자가 작성한 수요기술과 관련된 특허가 존재하지 않는 경우 추천 성능이 약간 떨어짐.
특허 서지 데이터를 보충해 추천 성능 개선.

방안 2:
오타, 번역문제

수요기술명과 특허 서지정보에 존재하는
영어 단어와 **오타**는 모델에서 인식하지 못함.
전처리 과정을 거쳐 모델이 인식할 수 있도록
모델 프로세스 개선.

부록 참고문헌

특허 분야 언어모델 github: https://github.com/kipi-ai/korpatelectra

BERT 한국어 지원 모델 KLUE: https://huggingface.co/klue/bert-base

Transformer: https://huggingface.co/docs/transformers/index

Sentence BERT model: https://huggingface.co/bongsoo/moco-sentencedistilbertV2.1

Sentence Transformer: https://www.sbert.net/

faiss 라이브러리: https://github.com/facebookresearch/faiss

감사합니다