

특허 요약문부터 화학 실험표까지, 인공지능 학습용 지식재산 자료 3종 무료 개방

- 지식재산처, 기업이 직접 구축하기 어려운 인공지능 학습 자료 3종 개방 -
- 특허요약·화학실험 자료로 민간 인공지능 개발·서비스 창출 지원 -

지식재산처(처장 김용선)는 9. 30.(수)부터 인공지능 학습용 지식재산 자료 3종, 약 961만 건을 지식재산정보 활용서비스(KIPRIS^{Plus})를 통해 무료로 개방한다고 밝혔다.

* 지식재산정보 활용서비스(KIPRIS^{plus}) : 국내·외 16개 국가·기관에서 발간된 지식재산 관련 공보와 행정정보 등 총 124종, 약 13.7억 건의 지식재산 정보를 제공하는 자료 개방 장

특허·상표 등 지식재산 자료는 전문용어가 많고 문서 구조도 복잡해 인공지능 학습에 적합한 형태로 가공하려면 상당한 시간과 비용이 필요하다. 특허 중소기업이나 새싹기업 등이 대규모 학습 자료를 직접 구축하기에는 부담이 크다.

이번에 새롭게 개방하는 자료는 지식재산처의 「인공지능(AI)기반 특허행정 혁신사업」 등을 통해 구축한 ▲특허 요약 학습 자료 ▲화학 및 금속분야 실험 학습 자료 ▲상품분류모형 학습 자료 등 3종이다. 이번 3종을 추가하면 개방되는 인공지능 학습 자료는 총 22종, 약 5,700만 건으로 늘어난다.

특허 요약 학습 자료(약 1만 6천 건)는 특허 명세서에서 ‘해결과제’, ‘대표 청구항’, ‘발명의 효과’ 등 핵심 내용을 전문가가 요약하고 원문과 요약문을 한 쌍으로 구성한 자료다. 인공지능 특허요약 모형 개발이나 요약 성능·품질 평가 등에 활용할 수 있다.

화학 및 금속분야 실험 학습 자료(약 300만 건)는 특허문헌에 사진이나 문자 형태로 포함된 실험결과표를 인공지능과 광학문자인식(OCR) 기술을 활용해 컴퓨터가 분석할 수 있는 자료로 변환·구조화한 것이다. 그동안 사진 형태로 존재해 활용하기 어려웠던 실험정보를 연구개발, 산업분석, 특허검색 등에 활용할 수 있게 된다.

* OCR(Optical Character Recognition): 사진 속 글자를 컴퓨터가 읽을 수 있는 문자로 바꾸는 기술

상품분류모형 학습 자료(약 660만 건)는 상표 출원서에 기재된 상품·서비스 명칭과 심사과정에서 활용되는 유사군분류 정보를 연결한 자료로, 상품 자동분류, 유사상품 검색, 시장·출원동향 분석 등을 위한 인공지능 모형 개발에 활용할 수 있다.

특히 이번 개방은 행정 과정에서 구축·축적한 자료를 민간이 인공지능 개발에 활용할 수 있는 학습 자료로 전환해 다시 개방한다는 데 의미가 있다. 기업이 많은 비용과 시간을 들여 자료를 직접 수집·가공하지 않고도 인공지능 연구개발과 서비스 개발에 활용할 수 있어 민간의 인공지능 개발 비용을 낮추고 새로운 지식재산 서비스 창출에도 도움이 될 것으로 기대된다.

지식재산청 정재환 지식재산정보국장은 “인공지능 경쟁력의 기반은 양질의 학습 자료를 얼마나 쉽게 확보하고 활용할 수 있느냐에 달려 있다”며, “기업이 자체적으로 구축하기 어려운 고품질 지식재산 자료를 지속적으로 발굴·개방해 민간의 인공지능 개발 부담을 낮추고, 지식재산 자료를 활용한 새로운 서비스와 산업이 성장할 수 있도록 지원하겠다”고 밝혔다.

한편, 이번에 개방되는 자료는 KIPRIS^{Plus} 누리집을 통해 무료로 제공되며, 자세한 사항은 KIPRIS^{Plus} 도움창구(02-6915-1553)로 문의하면 된다.

※ 붙임 1. 지식재산정보활용서비스(KIPRIS^{Plus}) 개요

2. 학습데이터 3종 개방 안내문

3. 지식재산정보활용서비스(KIPRIS^{Plus}) 개방 학습데이터 현황

담당부서	지식재산정보국 지식재산데이터관리과	책임자	과 장	신현철 (042-481-5134)
		담당자	사무관	김상훈 (042-481-5359)
			주무관	이민주 (042-481-5137)

□ 서비스 소개

- 지식재산처가 개방 중인 모든 IP 정보*를 실시간으로 활용할 수 있는
①벌크 데이터, ②Open API 방식의 정보제공 서비스

* 국내·외 16개 국가·기관에서 발간된 지식재산 관련 공보와 행정정보 등 총 124종, 약 13.7억 건의 지식재산 정보(AI 학습데이터 22종 포함)

□ 데이터 제공 형태

- (벌크데이터) 지식재산 정보를 온라인에서 대용량으로 일괄 제공하는 방식
- (Open API) 지식재산처 DB의 지식재산 정보를 네트워크 연계 기반으로 자체 DB 구축 없이 이용하는 방식

□ AI 학습데이터 개방 실적(총 22종)

'19~'24년 (12종)	'25년 (7종)	'26년 (3종)
· 특허 기술용어 연계표 · 특허 한-영 코퍼스 · AI이미지검색 학습데이터 등	· 특허 공개공보 학습데이터 · 의견제출통지서 학습데이터 · CPC 분류 학습데이터 등 7종	① 특허 요약 학습데이터 ② 화학금속분야 실험 학습데이터 ③ 상품분류모델 학습데이터

* 학습데이터 관련 세부사항은 [붙임3] KIPRIS Plus 개방 학습데이터 현황 참조

□ 활용 분야

구분	활용 기관	활용 목적
IP관리	특허법인, IP 관리기업, R&D 연구소	IP 고객관리 및 성과 관리에 활용
IP금융	은행, 신용보증기관, 금융 공사	기술 담보의 여신 및 금융심사에 활용
IP평가	기술 평가 전문기관, 평가 연구기관	(유망) 기술 평가 측정에 활용
IP거래	IP 거래 전문기관, 대학	권리 거래를 위한 중개에 활용
IP서비스	IP 번역기업, IP 정보서비스기업	번역엔진 성능 향상 및 신규 서비스 창출에 활용

1. 특허 요약 학습데이터

1. 개요

한국 특허공보(2024년 발간 공개·등록 공보 일부 대상) 내 명세서 주요 항목을 요약한 데이터로 특허 요약을 위한 AI 학습용 데이터를 제공합니다.

2. 세부정보

특허법에서 규정하는 요약문이 갖추어야 할 요건(「특허법 시행규칙」요약서 작성방법)을 충족하도록 특허 명세서 내 항목(해결과제, 대표청구항, 발명의 효과)을 선별하여 각 항목을 요약 후 최종 요약문(재요약)을 구축하였으며, 원문과 요약 쌍(Pair)을 제공합니다.

3. 활용 용도

AI 연구개발, AI 요약 학습, AI 요약 품질 평가 등

4. 특이 사항

휴먼 어노테이션(Human Annotation), 전문가 어노테이션(Expert Annotation) 방식으로 라벨링 되었으며, 생성형 AI 기술을 통해 특허정보의 핵심 내용을 효과적으로 요약하는 알고리즘 연구·실증 활용 목적으로 구축

메타 정보	분류체계	국내 IP데이터 > AI학습데이터 > 언어학습
	제공기관	지식재산처
	업데이트주기	미갱신
	상품개시일	'26. 9. 30.
	태그	특허, 요약, 학습, 데이터, 학습데이터, AI
세부 정보	서비스유형	벌크데이터
	데이터 상품명	특허 요약 학습데이터
	데이터 제공기간	2024년(2024년 발간 공개·등록 공보 대상)
	데이터 건수	16,000건
	데이터 제공방식	CSV
	데이터수수료	무료

데이터 예시			
구분	항목명	데이터 타입	예시
기본정보	① 출원번호	String(13)	1020220102604
	② 기술분야(대표IPC)	String	C12N 15/86
	③ 발명의 명칭	String	퇴행성 뇌질환 치료용 AAV 기반 유전자 치료제
요약정보	① 해결과제 원문	String	본 발명의 목적은 퇴행성 뇌질환의 예방 또는 치료용 아데노연관바이러스 벡터를 제공하는 것이다. 또한, 본 발명의 목적은 퇴행성 뇌질환의 예방 또는 치료용 조성물을 제공하는 것이다. 아울러, 본 발명의 목적은 퇴행성 뇌질환의 예방 또는 치료용 아데노연관 바이러스 벡터의 생산 방법을 제공하는 것이다.
	② 해결과제 요약	String	본 발명은 퇴행성 뇌질환의 예방 또는 치료용 아데노연관바이러스 벡터, 퇴행성 뇌질환의 예방 또는 치료용 조성물, 및 퇴행성 뇌질환의 예방 또는 치료용 아데노연관바이러스 벡터의 생산 방법을 제공한다.
	③ 대표 청구항 원문	String	인간 활성화 PAK1(hPAK1-CA)를 암호화하는 핵산 분자를 포함하는, 퇴행성 뇌질환의 예방 또는 치료용 아데노연관바이러스 벡터.
	④ 대표 청구항 요약	String	퇴행성 뇌질환의 예방 또는 치료용 아데노연관 바이러스 벡터는, 인간 활성화 PAK1(hPAK1-CA)를 암호화하는 핵산 분자를 포함한다.
	⑤ 발명의 효과 원문	String	본 발명의 세포사멸조절 단백질의 기능을 증진시켜 세포의 생존을 증진시킬 수 있는 단백질 인산화효소인 PAK1(p21-activated serine/threonine kinase 1)의 활성화(constitutively active, CA) 형을 발현하는 아데노연관바이러스 벡터는 신경퇴행성 질환, 특히, 약물 유도 파킨슨병 및 유전적 파킨슨병 동물모델에서 이들의 생존율을 현저히 증가시키므로, 이를 퇴행성 뇌질환의 예방 또는 치료 용도로 사용할 수 있다.
	⑥ 발명의 효과 요약	String	본 발명은 신경퇴행성 질환, 특히, 약물 유도 파킨슨병 및 유전적 파킨슨병 동물모델에서 이들의 생존율을 현저히 증가시키므로, 이를 퇴행성 뇌질환의 예방 또는 치료 용도로 사용할 수 있다.
	⑦ 재요약	String	본 발명은 퇴행성 뇌질환의 예방 또는 치료용 아데노연관바이러스 벡터, 퇴행성 뇌질환의 예방 또는 치료용 조성물, 및 퇴행성 뇌질환의 예방 또는 치료용 아데노연관바이러스 벡터의 생산 방법을 제공한다. 퇴행성 뇌질환의 예방 또는 치료용 아데노연관 바이러스 벡터는, 인간 활성화 PAK1(hPAK1-CA)를 암호화하는 핵산 분자를 포함한다. 본 발명은 신경퇴행성 질환, 특히, 약물 유도 파킨슨병 및 유전적 파킨슨병 동물모델에서 이들의 생존율을 현저히 증가시키므로, 이를 퇴행성 뇌질환의 예방 또는 치료 용도로 사용할 수 있다.

2. 화학 및 금속분야 실험 학습데이터

1. 개요

특허 문헌에 포함된 텍스트 및 이미지 형태의 화학 실험결과표를 데이터화하여, 검색 효율성을 높이고, DB 개방으로 산업·연구계 활용성 강화

2. 세부정보

AI를 활용하여 유기화학/합금분야 특허공보 내 존재하는 다양한 형태의 첨부 이미지 및 텍스트 중 실험결과표를 추출하고, 표 구조정보(학습데이터) 인식, OCR 텍스트 변환을 통해 실험데이터 구축

3. 활용 용도

연구 개발, 산업 분석, 특허 검색, 표 구조 학습 등

4. 특이 사항

이미지 형태의 실험결과표를 OCR 텍스트 변환을 통해 구축한 데이터로 오기재와 부분 누락이 존재할 수 있음

메타 정보	분류체계	국내 IP데이터 > AI학습데이터 > 이미지학습
	제공기관	지식재산처
	업데이트주기	미정신
	상품개시일	'26. 9. 30.
	태그	화학, 금속분야, 실험데이터, 학습, 데이터, 학습데이터, AI
세부 정보	서비스유형	벌크데이터
	데이터 상품명	화학 및 금속 분야 실험 학습데이터
	데이터 제공기간	1973년 ~ 2025년 12월
	데이터 건수	3,003,722건(이미지 및 텍스트 표 기준)
	데이터 제공방식	TXT
	데이터수수료	무료

데이터 예시

기본 정보	APPLNO	OPN_PUB_TPCD	CMDSK_KIND_CD	IPC_CD	APPL_DT	OPN_DT	PUB_DT	TBSTM_TP_TPCD	TBSTM_SEQ	IMG_FILE_NM
	1020007006010	공개	SGM	C07D 265/30	19981204	20010425	20060809	I	9	pct00049.tif
	1020007006010	공개	SGM	C07D 265/30	19981204	20010425	20060809	I	9	pct00049.tif
	1020007006010	공개	SGM	C07D 265/30	19981204	20010425	20060809	I	9	pct00049.tif
	1020007006010	공개	SGM	C07D 265/30	19981204	20010425	20060809	I	9	pct00049.tif
	1020007006010	공개	SGM	C07D 265/30	19981204	20010425	20060809	I	9	pct00049.tif
	1020007006010	공개	SGM	C07D 265/30	19981204	20010425	20060809	I	9	pct00049.tif
	1020007006010	공개	SGM	C07D 265/30	19981204	20010425	20060809	I	9	pct00049.tif
	1020007006010	공개	SGM	C07D 265/30	19981204	20010425	20060809	I	9	pct00049.tif
	1020007006010	공개	SGM	C07D 265/30	19981204	20010425	20060809	I	9	pct00049.tif
	1020007006010	공개	SGM	C07D 265/30	19981204	20010425	20060809	I	9	pct00049.tif
	TBSTM_ITM_NO	ITM_ROW_NO	ITM_COL_NO	ITM_ROW_ANXTN_CNT	ITM_COL_ANXTN_CNT	ITM_CONT	X_1_CRDNT_VAL	Y_1_CRDNT_VAL	X_2_CRDNT_VAL	Y_2_CRDNT_VAL
	1	8	4	1	1	단결합	1312	1095	1623	1228
	1	8	5	1	1	1	1623	1095	1730	1228
	1	8	6	1	1	0	1730	1095	1838	1228
	1	8	7	1	1	≡[CH ₂ CH ₂ CH ₂ ≡]	1838	1095	2153	1228
	1	8	8	1	1	지4	2153	1095	2405	1228
	1	9	0	1	1	117	26	1228	395	1361
	1	9	1	1	1	4-Plz	395	1228	731	1361
	1	9	2	1	1	3,4-dClPh	731	1228	1159	1361
	1	9	3	1	1	CO	1159	1228	1312	1361
	1	9	4	1	1	단결합	1312	1228	1623	1361

3. 상품분류모델 학습데이터

1. 개요

출원 상표에 기재된 지정상품/서비스 명칭(고시, 비고시 포함)과 심사 과정에서 활용된 유사군분류 정보를 매칭한 AI 학습용 데이터를 대용량(Bulk)으로 제공합니다.

2. 세부정보

지식재산처에서 지정상품/서비스의 유사판단 신속성 및 편의성을 위하여 활용 중인 유사군분류를 자동화 가능한 학습용 데이터로써 2008년* 이후 출원 상표에 존재하는 지정상품에 대하여 사용여부 분석, 중복 제거 등의 정제를 거쳐 텍스트 데이터로 제공합니다.

* 사람에 의한 유사군분류 부여 절차는 2008년 이후 적용

3. 활용 용도

연구 개발을 통한 상품 자동 분류, 유사상품 검색, 동향 분석 등

4. 특이 사항

유사군분류 정보는 고시명칭 개정 등의 이유로 매년 소급하여 변경되므로 검색서비스에 존재하는 분류와 다소 상이할 수 있음.

메타 정보	분류체계	국내 IP데이터 > AI학습데이터 > 언어학습
	제공기관	지식재산처
	업데이트주기	미갱신
	상품개시일	'26. 9. 30.
	태그	상품분류모델, 유사군분류, 학습, 데이터, 학습데이터, AI
세부 정보	서비스유형	벌크데이터
	데이터 상품명	AI 상품분류모델 학습데이터
	데이터 제공기간	2008년 1월 ~ 2023년 12월
	데이터 건수	660만건
	데이터 제공방식	TXT
	데이터수수료	무료

데이터 예시

기본 정보	상품/지정상품 명칭	상품류	유사군분류
	크로플전문 제과점업 진은선	43	G0301,G0502,S120602
	공팡이냄새제거용 공기청향제 주식회사엔큐랩	1	G1001
	기록된 또는 내려받기 가능한 스마트 파킹 관련 컴퓨터 소프트웨어 플랫폼	9	G390802
	계피를 주원료로 하는 차 남종현	30	G0501
	어묵 구매대행및 판매알선및 판매대행및 중개업 박종득	35	S2008
	기사식당업 박영숙	43	S120602
	곡물을 주원료로 하는 건강보조식품 유신환	29	G0206,G1009
	건설용 콘크리트 펌프카 임대업 주식회사거성중기	37	S040101
	건강기능식품 고려온단헬스케어주식회사	5	G1004,G1009
	소프트웨어 개발업 및 공급업 주식회사링플	42	G390802,S123301

□ KIPRIS^{PLUS} 학습데이터 개방 현황

○ 지식재산 분야 특화 학습데이터 22종 5,700만건 개방('19~'26)

순번	개방 상품명	세부 내용	데이터 건수(천건)
1	특허-기술용어 연계표 (전기, 기계, 화학)	특정 기술용어(전기전자, 기계, 화학)를 포함하고 있는 공보의 출원번호 정보 제공	90,393
2	특허 청구항 구조(전기전자)	전기전자분야 특허 청구항(독립항, 종속항)에 구조화 정보 태그를 매핑한 정보 제공	662,544
3	특허 분류코드 설명 한영코퍼스	분류코드(IPC, CPC) 설명을 영문으로 번역하여 문장쌍으로 제공	264,414
4	특허 초록 한영코퍼스	한국특허영문초록(KPA) 문장을 일부 발췌하여 문장쌍으로 제공	1,178,877
5	특허 중간서류철 한영 코퍼스	중간서류철(의견제출통지서, 등록결정서, 거절결정서)의 문장을 영문으로 번역하여 문장쌍으로 제공	500,000
6	AI 챗봇 학습데이터	출원 등록 수수료 등 분야별 특허상담 표준 질의 응답 정보 제공	25,000
7	특허 영한 코퍼스	미국 특허 공보의 문장을 한글로 번역하여 문장쌍으로 제공	152,791
8	특허 중한 코퍼스	중국 특허 공보 문장을 국문으로 번역하고 중한 문장의 문장쌍으로 제공	160,000
9	특허 공보 학습데이터	특허 공개공보 데이터를 AI 학습이 가능하도록 정제, 가공한 데이터	5,152,884
10	미국 특허공보 학습데이터	미국 특허공보 데이터를 AI 학습이 가능하도록 정제, 가공한 데이터	17,677,869
11	의견제출통지서 학습데이터	의견제출통지서 데이터를 AI 학습이 가능하도록 정제, 가공한 데이터	2,871,807
12	특허 패밀리 한영 기계번역용 학습데이터	특허 패밀리 건을 이용한 유사의미 문자 추출하여 매핑한 정보 제공	10,536,147
13	CPC 분류 학습데이터	발명의 명칭, 요약 정보, 기술분야 등 주요 문헌 정보를 CPC 분류 정보와 매핑한 정보 제공	4,987,269
14	특허 상담 기계독해용 학습데이터	특허 상담 내용(녹음 파일)의 구분 및 텍스트화한 정보 제공	5,871
15	특허 고객상담사례집 학습데이터	특허 고객 상담 사례집 데이터를 AI 학습이 가능하도록 정제, 가공한 데이터	868
16	특허 도면부호 학습데이터	국내 특허 도면의 좌표정보 및 이미지파일의 인식 기호 정보 제공	159,381
17	AI 도형상표 이미지 검색 학습데이터	국내 도형상표 건에 대한 출원번호 및 견본이미지-분류코드 매핑 정보 제공	1,384,000
18	AI 디자인도면 이미지 검색 학습데이터	국내 디자인 건에 대한 출원번호 및 도면, 분류코드 매핑 정보 제공	1,714,000
19	의약품 실험 학습데이터	의약품 분야 특허 명세서에 포함된 이미지(표) 실험데이터	447,000
20	특허 요약 학습데이터	한국 특허공보(2024년 발간 공개·등록 공보 일부 대상) 내 명세서 주요 항목을 요약한 데이터	16,000
21	화학 및 금속분야 실험 학습데이터	특허 문헌에 포함된 텍스트 및 이미지 형태의 화학 실험결과표를 데이터화하여 제공	3,003,722
22	상품분류모델 학습데이터	출원 상표에 기재된 지정상품/서비스 명칭과 심사 과정에서 활용된 유사군분류 정보를 매칭한 데이터	6,600,000
합 계			57,590,837