[오픈소스 프로젝트 소개] AI-ready 데이터 추출 PDF 파서 'OpenDataLoader PDF', GitHub 오픈소스 트렌딩 1위 달성

(`26.4.24, 작성자 : 오픈데이터로더 PDF PM 조수지)

OpenDataLoader PDF는 한글과컴퓨터가 개발한 오픈소스 PDF 데이터 추출 엔진입니다. PDF 문서의 표, 제목, 이미지, 수식 등 구조를 분석하여 JSON, Markdown, HTML로 변환하며, LLM 및 RAG 파이프라인에 최적화되어 있습니다. Apache 2.0 라이선스로 무료 제공되며, GitHub 전체 트렌딩 1위, 글로벌 오픈소스 벤치마크 종합 1위(0.907, 2026년 4월 기준)를 달성했습니다. Rule-based 모드(AI 없이 0.015s/page)와 하이브리드 모드(AI 모델 결합, 종합 1위)를 제공하여 속도와 품질을 용도에 따라 선택할 수 있습니다. PDF Association의 멤버로서 PDF 표준을 준수하며 개발하고, veraPDF 개발팀(Dual Lab)과 공동 개발하고 있습니다. OpenDataLoader PDF는 LangChain 공식 통합이 완료되어 있습니다. 이후 Langflow, Llamaindex 등과의 통합도 예정되어 있습니다. X와 Medium을 공식 채널로 운영하고 있으며, GitHub Discussions와 Issues가 활성화되어 있어 누구나 자유롭게 참여할 수 있습니다

▶ 오픈소스 프로젝트(OpenSource Project) – OpenDataLoader PDF

구분	세부 항목	설명
프로젝트 개요	프로젝트 저장소	https://github.com/opendataloader-project/opendataloader-pdf
	홈페이지	https://opendataloader.org
	라이선스	Apache License 2.0
	프로젝트 분야	AI, Document Processing, Developer Tools
	프로젝트 소개	OpenDataLoader PDF는 PDF 문서 내의 표, 제목, 이미지, 수식, 목록 등 모든 구성 요소를 정확하게 인식하고, JSON, Markdown, HTML, Tagged PDF 형식으로 변환하는 오픈소스 PDF 데이터 추출 엔진입니다. LLM, RAG, 벡터 검색 등 AI 파이프라인에 최적화된 고품질 구조 데이터를 제공합니다. Java, Python, Node.js SDK를 지원하며, Docker 이미지로도 제공됩니다.
핵심 가치	AI 시대에 PDF는 가장 방대한 비정형 데이터 소스입니다. 전 세계에 약 2.5조 개의 PDF 문서가 존재하며, 기업 데이터의 상당 부분이 PDF 형태로 저장되어 있습니다. 그러나 기존 도구로는 다단 레이아웃의 읽기 순서, 복잡한 표 구조, 이미지 내 정보를 정확히 추출하기 어렵습니다. OpenDataLoader PDF는 이 문제를 오픈소스로 해결합니다. 글로벌 벤치마크 종합 1위(0.907, 2026년 4월 기준)를 달성하여 PDF 데이터 추출 기술의 새로운 기준을 제시했으며, Apache 2.0 라이선스로 전 세계 개발자가 제한 없이 활용할 수 있습니다.	
주요 특징	등장배경	AI/LLM 시대가 도래하면서 학습, 검색, 분석에 활용할 고품질 문서 데이터의 수요가 폭증했습니다. 그러나 PDF는 "보기 위한"

구분	세부 항목	설명
		포맷이지 "데이터로 활용하기 위한" 포맷이 아닙니다. 기존 방식의 한계는 명확합니다. 다단 레이아웃에서 읽기 순서가 뒤섞이고, 표가 단순 텍스트로 풀어지며, 이미지와 차트의 정보가 손실됩니다. 바운딩 박스 좌표가 없어 출처를 추적할 수 없고, 접근성 태그가 없어 장애인이 문서에 접근할 수 없습니다. OpenDataLoader PDF는 이러한 구조적 한계를 해결하기 위해 개발되었습니다. 한컴이 30년간 축적한 문서 처리 기술을 기반으로, PDF Association의 멤버로서 표준 지침을 참고하여, veraPDF 개발팀(Dual Lab)과 공동 개발을 진행하고 있습니다.
	핵심 기능	**- XY-Cut++ 기반 읽기 순서 정렬.** 다단 레이아웃, 사이드바, 각주 등 복잡한 구조에서도 사람이 읽는 순서 그대로 텍스트를 정렬합니다. 벤치마크 점수 0.934로 12개 경쟁 파서 중 1위입니다. **- 표 구조 추출.** 병합 셀, 테두리 없는 표, 중첩 표를 포함한 복잡한 표 구조를 인식합니다. 하이브리드 모드에서 벤치마크 0.928로 1위를 기록했으며, 2위(0.887)와 의미 있는 격차를 보입니다. **- 제목 인식.** 문서의 제목 계층(H1-H6)을 자동으로 판별합니다. 하이브리드 모드에서 벤치마크 0.821로 2위를 기록했으며, 지속적으로 개선 중입니다. **- 바운딩 박스 좌표 제공.** 모든 추출 요소에 [x1, y1, x2, y2] 좌표를 포함합니다. RAG 시스템에서 답변의 출처를 PDF 원문 위치로 정확히 추적할 수 있습니다. **- AI 모델 연동 하이브리드 모드.** Rule-based 엔진에 AI 모델을 결합하여 OCR(80개 이상 언어), 이미지 설명, 차트 인식, 수식(LaTeX) 변환 등 고급 기능을 제공합니다. **- 접근성을 위한 Tagged PDF 자동 생성(Auto-Tagging).** H1-H6, Table, List, Figure, Link, Caption 등 구조 태그를 자동으로 4월 23일 패치 예정입니다. **- AI-Safety 내장** 히든 텍스트, 페이지 외 콘텐츠, 프롬프트 인젝션 시도를 자동으로 필터링합니다.
	차별화 요소	- GitHub 전체 개발 언어 대상 트렌딩 1위를 달성하고 트렌딩 배지를 획득했습니다 (2026.03.20.). - 13,000개 이상의 Stars와 1,060개 이상의 Forks를 기록하고 있습니다. 글로벌 오픈소스 벤치마크에서 종합 1위(0.907)를 달성했습니다 (2026년 4월 기준). - docling(0.882), nutrient(0.880), marker(0.861), unstructured(0.841) 등 12개 파서와 비교한 결과이며, 200개 실제 PDF를 대상으로

구분	세부 항목	설명
		측정한 수치입니다. 벤치마크 저장소(opendataloader-bench)를 공개하여 누구나 직접 검증할 수 있습니다. - OpenDataLoader PDF는 두 가지 모드를 제공합니다. Rule-based 모드는 AI 없이 0.015s/page로 동작하며, 속도와 메모리 효율 모두 1위입니다. 하이브리드 모드는 AI 모델을 결합하여 종합 품질 1위(0.907)를 달성합니다. 용도에 따라 선택할 수 있습니다. - 엔진 자체가 오픈소스(Apache 2.0)이므로 vendor lock-in이 없습니다. veraPDF 개발팀과 공동 개발하여 태깅 엔진과 검증 도구가 같은 팀에서 나옵니다. - LangChain 공식 통합이 완료되어 RAG 파이프라인에 즉시 연동할 수 있습니다.
	대상 사용자	RAG/LLM 파이프라인을 구축하는 AI 개발자에게 적합합니다. 대량 PDF 문서를 AI 학습 데이터로 변환해야 하는 기업과 기관, PDF 접근성(Section 508, ADA Title II, EAA) 준수가 필요한 공공·교육기관에서 활용할 수 있습니다. 금융, 법률, 의료 등 정형화된 문서를 다루는 모든 산업에서 사용 가능합니다.
	운영 환경	Windows, macOS, Linux를 지원합니다. Java, Python, Node.js SDK를 제공하며, Docker 이미지로도 배포됩니다. Rule-based 모드는 CPU만으로 동작하며 GPU가 불필요합니다. 하이브리드 모드에서는 GPU(CUDA 지원)를 권장합니다. 완전 온프레미스 실행이 가능하여 FERPA, HIPAA 등 데이터 정책 제약이 있는 환경에서도 사용할 수 있습니다.
	활용 분야	RAG 기반 검색 시스템 구축, 기업 문서 AI 학습 데이터 변환, 공공기관 PDF 접근성 자동화(EAA / Section 508 / ADA Title II), 금융·법률 문서 자동 분석, 교육기관 학습자료 디지털 전환 등에 활용됩니다.
프로젝트 생태계	**커뮤니티 현황**	현재 X(@opendatalo51205)와 Medium(@opendataloader)을 공식 채널로 운영하고 있습니다. Reddit, Hacker News, Discord 채널은 곧 오픈 예정입니다. 문의사항이나 버그 리포트는 GitHub Discussions 또는 Issues에 남겨주시면 됩니다. LangChain 공식 통합(langchain-opendataloader-pdf)이 완료되었으며, PyPI와 npm을 통해 패키지가 배포됩니다.
오픈소스 허브	https://www.oss.kr/opensource/hub/62651	

▶ 주요 개발자 (Main Developers) – 이분도, 조현희

구분	세부 항목	내용
개발자 개요	개발자 소개	AI Agent를 하나의 팀처럼 운영하며, 오픈소스 프로젝트의 기획, 개발, 마케팅, 고객 유입까지 End-to-End로 직접 만들어가는 개발자입니다. 핵심 역량은 AI Agent가 모든 영역에서 프로덕션 수준의 결과물을 만들어낼 수 있도록 컨텍스트를 설계하고 관리하는 데 있습니다.
	개발자 이름	이분도
	이메일 또는 SNS	https://www.linkedin.com/in/bundo-lee-90670925b
	전문 분야	- AI Agent & Product Engineer
	경력	- 한글과컴퓨터 (2014.02 ~ 현재)
참여 프로젝트	https://github.com/opendataloader-project/opendataloader-pdf	
기여 내용	- 기술 설계 및 구현 (하이브리드, XY-Cut 알고리즘, 검증/배포 자동화) - 기술 문서 작성 - 커뮤니티 운영 - 마케팅 - 고객 유입 구조의 기술적 설계	
오픈소스 참여 계기	회사에서 오픈소스 프로젝트를 운영할 기회를 받았습니다. 한글과컴퓨터는 전 세계 개발자 커뮤니티에 의미 있는 기여를 하기를 바랐고, 저 역시 AI로 문서를 처리하려는 개발자들이 라이선스 걱정 없이 쓸 수 있는 좋은 기술을 많은 곳에서 활용하게 하고 싶었습니다.	
기여 활동	오픈소스의 성능과 품질을 세계 최고 수준으로 끌어올리기 위해, AI Agent를 활용해 프로덕션 등급 상용 제품의 아키텍처를 리서치하고, 하이브리드 구조를 설계·구현하여 공개된 벤치마크 기준 글로벌 오픈소스 대비 1위를 달성했습니다. 또한 AI 시대에 발맞춰, 개발자가 아닌 개발자의 AI Agent를 대상으로 집중적으로 알리려 노력했습니다. AI Agent를 활용해 논문 및 연구자료를 리서치하고, AI가 인용할 확률이 높은 콘텐츠를 설계했습니다. 현재 AI Agent들이 해당 콘텐츠를 적극적으로 소비하면서, 단기간에 1만 스타를 달성하는 등 오픈소스 커뮤니티가 폭발적으로 성장했습니다.	
성장 포인트	오픈소스를 성장시키는 과정은 개발자가 제품의 전체 생애주기를 직접 경험할 수 있는 좋은 기회입니다. GitHub 생태계는 AI Agent를 운영하기에 최적화된 환경을 제공합니다. 최신 AI 기술을 활용해 제품의 기획부터 출시·운영까지 전 과정을 직접 경험하며, 한 명의 개발자로서 더 넓은 역할을 해낼 수 있다는 것을 체감했습니다.	
후배들에게 조언	아직 저도 인생과 기술 모두 탐구 중인 입장이라 조언은 어렵지만, 경험은 나눌 수 있을 것 같습니다. 오픈소스나 AI 관련 커피챗은 언제나 환영합니다.	

구분	세부 항목	내용
개발자 개요	개발자 소개	시스템 소프트웨어 기반의 개발 경험을 바탕으로, 오픈소스 프로젝트의 기술 통합부터 전략, 콘텐츠, 운영까지 반복 가능한 구조로 만들어 가는 개발자입니다. 기술을 만드는 것만큼, 그 기술이 계속 쓰일 수 있는 기반을 다지는 데 같은 무게를 두고 있습니다.
	개발자 이름	조현희
	이메일 또는 SNS	https://github.com/hyunhee-jo
	전문 분야	System Software Development
	경력	8년차 시스템 소프트웨어 개발자 - 방산 SW 개발 4년 - 한글과컴퓨터 (2022~현재)
참여 프로젝트	https://github.com/opendataloader-project/opendataloader-pdf	
기여 내용	- AI 프레임워크 생태계 통합 - 프로젝트 전략 수립 - 프로젝트 운영 체계 구축 - 기술 콘텐츠 작성 - 커뮤니티 활동 - 사용자 도구 개발	
오픈소스 참여 계기	회사에서 오픈소스 프로젝트에 참여할 기회를 받았습니다. 그동안 시스템 소프트웨어를 개발하면서 만든 기술은 사내 제품 안에서만 쓰였는데, 오픈소스를 통해 전 세계 개발자에게 직접 닿을 수 있다는 점이 가장 큰 동기였습니다. 좋은 기술을 공유하고, 다양한 개발자들과 함께 발전시켜 나갈 수 있는 환경에서 일하고 싶었습니다.	
기여 활동	OpenDataLoader PDF가 개발자들의 AI 워크플로우 안에서 자연스럽게 쓰일 수 있는 환경을 만드는 데 집중했습니다. LangChain, LangFlow, LlamaIndex 각각에 대해 생태계 조사부터 설계, 개발, 문서 작성, 검증, 코드 리뷰, 배포 자동화까지 전 과정을 직접 주도했습니다. PR을 제출할 때는 개발자, 사용자, 문서 독자 등 여러 관점에서 사전 검증을 거치는 방식을 적용했습니다. 기술 구현과 함께 AI 시대에 맞는 확산 전략을 수립하고, 기술 콘텐츠를 운영하며, 커뮤니티에 직접 참여하여 프로젝트를 알렸습니다. 릴리스 자동화, 보안 업데이트 대응, 사용자 안내 도구 개발 등 운영 기반도 함께 만들어 가고 있습니다.	
성장 포인트	가장 큰 변화는 시야입니다. 사내 제품만 바라보던 엔지니어에서, 전 세계 개발자가 실제로 쓰는 코드를 직접 만들고 대화하는 경험을 하게 되었습니다. 폐쇄적인 환경에서는 발견하지 못했던 문제를 외부 개발자들이 찾아주고, 함께 해결해 나가는 과정을 통해 기술이 더 빠르게 성장한다는 것을 체감했습니다. 업무 방식도 달라졌습니다. 이전에는 하나의 관점으로 일했지만, 오픈소스를 운영하면서 결과물을 다각도로 검증하고, 기술 스택의 경계를 넘나드는 유연함을 얻었습니다.	
후배들에게 조언	저도 아직 배우는 중이라 조언보다는 경험 공유입니다. 익숙한 영역 밖으로 나가는 게 처음에는 부담스럽지만, 오픈소스는 그 과정을 가장 자연스럽게 만들어 주는 환경이라고 느꼈습니다. 관심이 있다면 편하게 연락주세요.	