AI를 활용한 오픈소스 자동화 경험

오픈소스프로젝트팀 이분도

HANCOM

왜 PDF에 주목해야 할까?

HANCOM

PDF는 가장 많이 활용되는 포맷이자, AI가 학습하는 데이터 생태계의 출발점

데이터로 보는 PDF 생태계 규모

전 세계
2.5조 개
PDF 문서 존재

매년
2900억 개+
신규 PDF 문서 생성

글로벌 기업의
98%
배포 문서 형식 = PDF

디지털 계약 PDF 사용 현황

산업 별 PDF 사용 현황

* https://smallpdf.com/pdf-statistics (25년)

AX 시대, PDF의 중요성

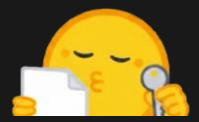
HANCOM

고품질 PDF 데이터셋, AI 성능을 극대화

2025년 9월, 허깅페이스 고품질 PDF 데이터셋 공개

FinePDFs

Liberating 3T of the finest tokens from PDFs



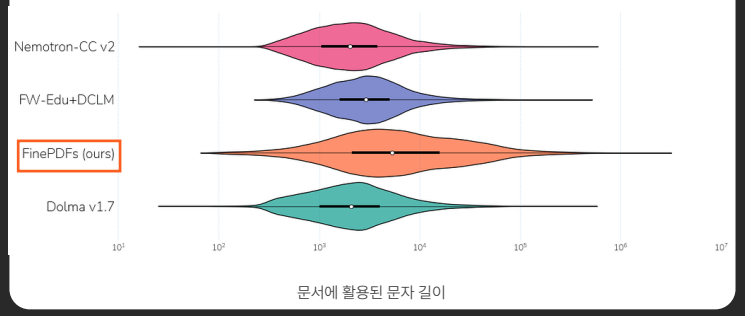
- 2013년부터 2025년까지의 데이터 수록
- Common Crawl 스냅샷 기반 수집
- 총 3조 토큰, 3.65TB 규모
- 4억 7,500 만 개의 PDF 문서 포함
- 1,733개의 언어 지원

* 파이토치, FinePDFs: Hugging Face가 공개한 3T 규모의 공개 PDF 데이터셋 (25년)

Fine PDFs로 입증된 PDF 데이터의 가치

PDF, 더 길고 구조화된 문서 제공

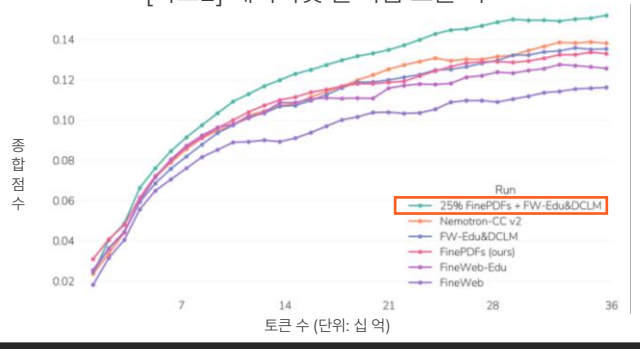
[자료1] 문서 포맷 별 문자 길이 분포 비교



PDF 데이터는 AI 학습에 필요한 문자이 더 많이 포함되어 있어 일반 데이터보다 학습 효율이 높음

PDF 포함 데이터, LLM 성능 극대화

[자료2] 데이터셋 별 학습 효율 비교



LLM 학습 시 PDF 데이터를 활용한 경우 다른 데이터셋 대비 뛰어난 성능을 보임

* https://huggingface.co/datasets/HuggingFaceFW/finepdfs (25년)

PDF 데이터화를 위한 다양한 도구

HANCOM

요청 데이터

복잡한 표

다단 구조

이미지 속 텍스트

1

PDF
데이터 입력

PDF 데이터화 도구

Adobe Document Cloud

Microsoft Azure
Form Recognizer

Google AI
Document AI

Amazon Textract

upstage
Document Parser

OpenDataloader PDF

NAVER CLOVA OCR

MuPDF

IBM DATACAP

SYNAP DocuAnalyzer

PyPDF

Xtracta

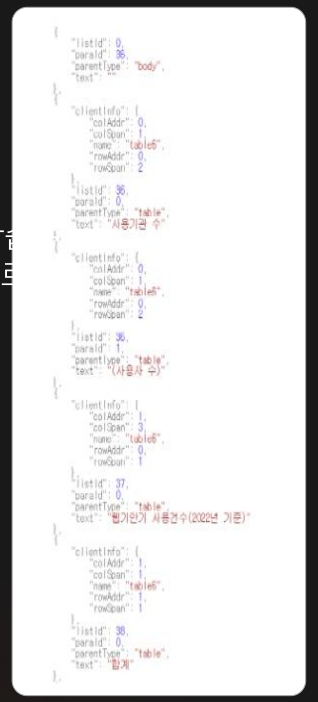
PDFminer

2

AI 학
데이터

추출된 데이터

MD JSON XML



개요

HANCOM

오픈데이터로더 PDF

출시일: 2025년 9월 17일

버전: v1.1.3 (2025.10 기준)

라이선스: Mozilla Public License 2.0 (MPL-2.0)

https://opendataloader.org

공동 개발

HANCOM × dual lab

지원 언어 및 플랫폼

- Java
- Python
- Node.js

핵심 기능

1. PDF 변환
   - JSON
   - HTML
   - Markdown
2. 레이아웃 인식
   - 리스트 인식
   - 헤딩 인식
   - 표인식

활용 분야

- AI 학습 데이터 전처리
- 비정형 문서 정형화
- RAG(Vector Search) 등과 연동
- 보안·프라이버시 중요 환경

Github: https://github.com/opendataloader-project/opendataloader-pdf

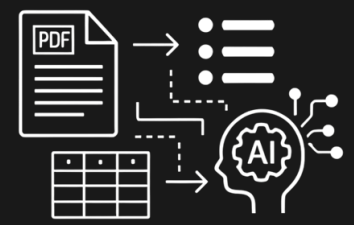
주요 특징

HANCOM

Safe, Open, High-Performance

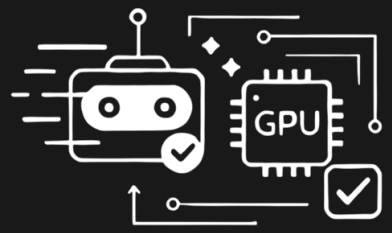
OpenDataLoader PDF for AI

고품질 AI 데이터 추출



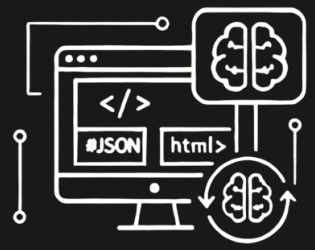
제목, 문단, 목록, 표 등의 구조를 이해하여 고품질 AI 학습용 데이터를 추출

빠르고 효율적인 성능



휴리스틱 추출 방식과 AI 방식을 결합하여 GPU 의존성을 최소화

다양한 AI 포맷 지원



단순 텍스트 추출을 넘어 AI 학습에 필요한 의미와 구조 정보를 제공

로컬 기반 & AI-Safety 지원

로컬 기반 동작, 데이터 보호 AI-Safety 지원을 통한 데이터 무결성 지원

기술적 차별화 포인트

HANCOM

우수한 성능으로 Tagged PDF, AI-Safety를 지원하는 유일한 글로벌 오픈소스 솔루션

차별화 포인트 ①

하이브리드 엔진

휴리스틱 접근 방식 + AI 기술
접목을 통해 성능과 효율 동시 확보

성능과 효율성

차별화 포인트 ②

Tagged PDF

PDF에 태그 정보 추가로
AI 학습의 정확도와 속도 향상

확장성

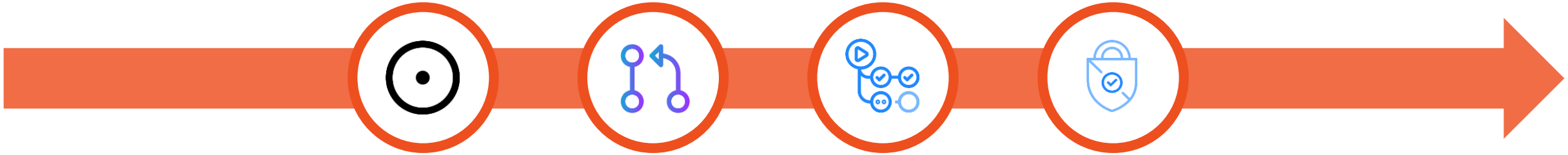
차별화 포인트 ③

AI-Safety

데이터 침해 위험이 낮고, 잠재적인
악성 콘텐츠를 사전에 식별·무력화

안정성

GitHub 생태계



GitHub Issue / Project	GitHub Pull Request	GitHub Action	GitHub Security
JIRA	Merge Request Code Review	Jenkins Webhook Cron	보안 취약점 의존성 취약점

DevOps - GitHub Actions

- 멀티 플랫폼
- 버전 관리
- 빌드
- 단위/통합 테스트
- 보안 취약점 정적분석
- 의존성 취약점 정적분석
- 서명
- 배포

Windows
Linux
MacOS
iOS
Android
Emscription

x86/x64
arm/arm64
WASM

Docker
Java/Kotlin
C/C++
Python
Jupyter
Javascript
Typescript

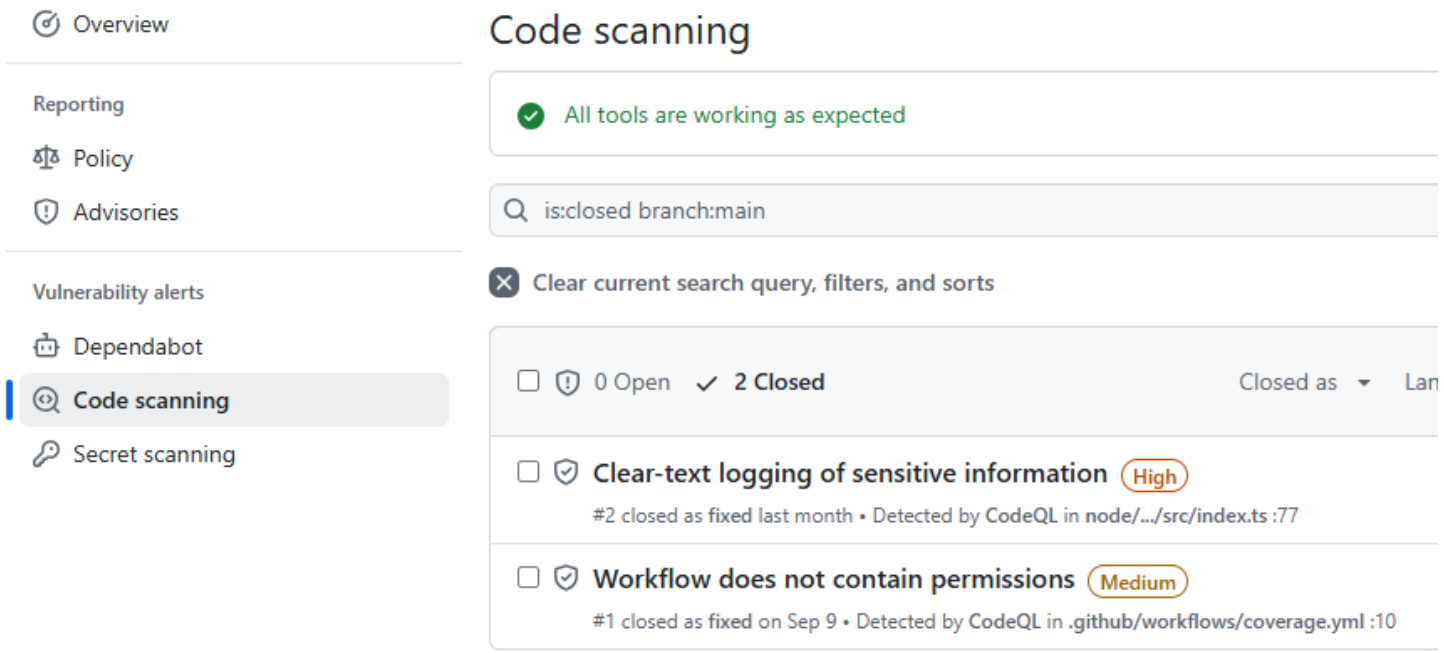
CodeQL
Dependabot
Webhook
Cron
Apps

GitHub Action 멀티 플랫폼 예제

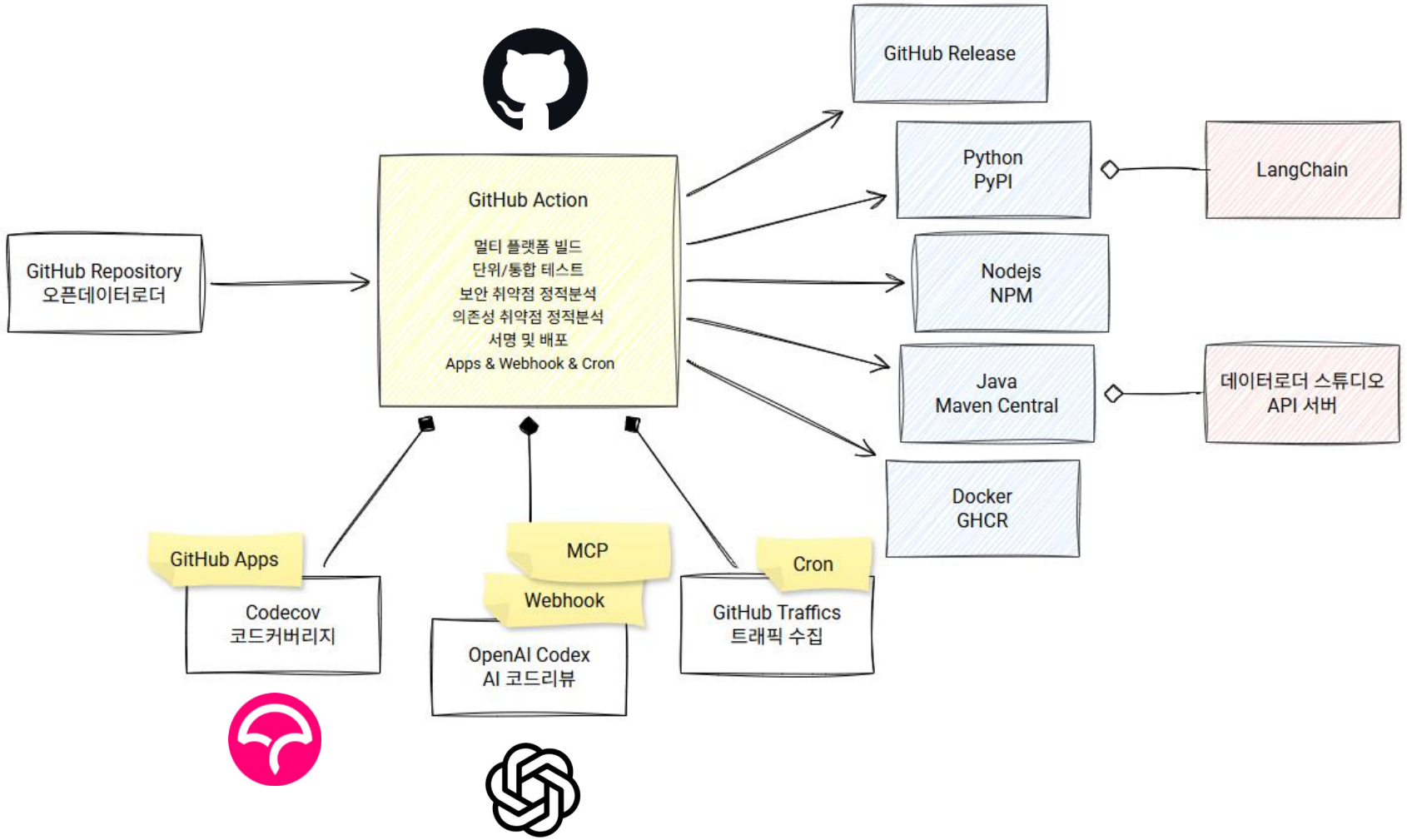
```yaml
jobs:
  build-linux:
    runs-on: ubuntu-latest
    steps:
      - run: echo "✅ Linux”



  build-windows:
    runs-on: windows-latest
    steps:
      - run: echo "✅ Windows”



  build-macos:
    runs-on: macos-latest
    steps:
        run: echo "✅ macOS"
```

```yaml
jobs:
  build-wasm:
    runs-on: ubuntu-latest
    steps:
      - uses: mymindstorm/setup-emsdk@v14
        with:
          version: 3.1.58
      - run: echo "✅ WebAssembly"
```

GitHub Security



DevOps – https://github.com/opendataloader-project/opendataloader-pdf



AI 코드리뷰
GitHub Action + MCP + OpenAI Codex

PR (Pull Request)

GitHub Action
Webhook

OPENAI CODEX

MCP

hnc-leebd commented 2 days ago

@codex review

chatgpt-codex-connector bot reviewed 2 days ago — View reviewed changes

chatgpt-codex-connector bot left a comment

Codex Review

Here are some automated review suggestions for this pull request.

▶ About Codex in GitHub

`java/opendataloader-pdf-core/src/main/java/org/opendataloader/pdf/pdf/PDFWriter.java` Outdated

Comment on lines 139 to 142

```
139 +        BoundingBox box = DocumentProcessor.getPageBoundingBox(boundingBox.getPageNumber());
140 +        if (box != null) {
141 +            boundingBox.move(box.getLeftX(), box.getBottomY());
142 +        }
```

chatgpt-codex-connector bot 2 days ago

P1 **Avoid mutating bounding boxes while drawing annotations**

The new translation in `draw` moves the passed `BoundingBox` directly. These bounding boxes live in the `contents` collection, which is reused after PDF generation to emit JSON/HTML/text in `DocumentProcessor.processFile`. As a result, calling `updatePDF` now permanently shifts every bounding box and the coordinates serialized later depend on whether PDF generation ran beforehand. The filtering code in the same commit avoids this by translating a copy. Consider shifting a temporary copy for annotation placement instead of mutating the shared object.

Useful? React with 👍 / 👎.

Reply...

Resolve conversation

GitHub Traffics

GitHub

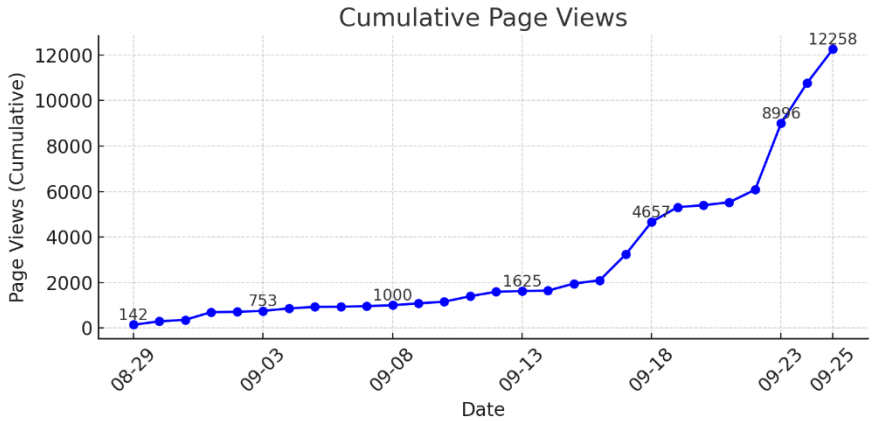
GitHub Action
Cron

GitHub API

opendataloader-pdf-traffics / metrics / traffic.csv

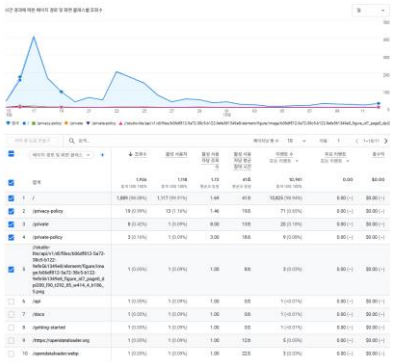
Preview | Code | Blame

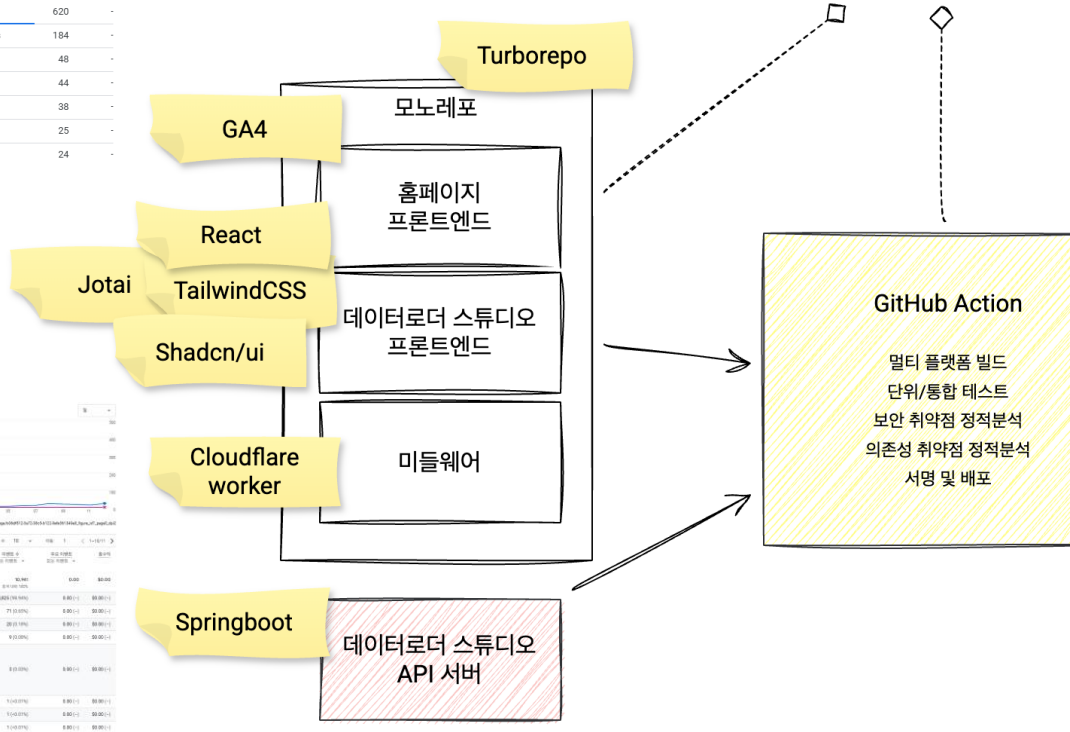
66	2025-09-22	views	471	97
67	2025-09-23	clones	29	22
68	2025-09-23	views	2858	1984
69	2025-09-24	clones	24	14
70	2025-09-24	views	1527	784
71	2025-09-25	clones	35	15
72	2025-09-25	views	1433	662
73	2025-09-26	clones	137	31
74	2025-09-26	views	1214	379



- 방문자 수 (PV, UV)
- Stars/Watch/Fork
- 자주 방문한 페이지
- 유입 경로

DevOps – https://opendataloader.org





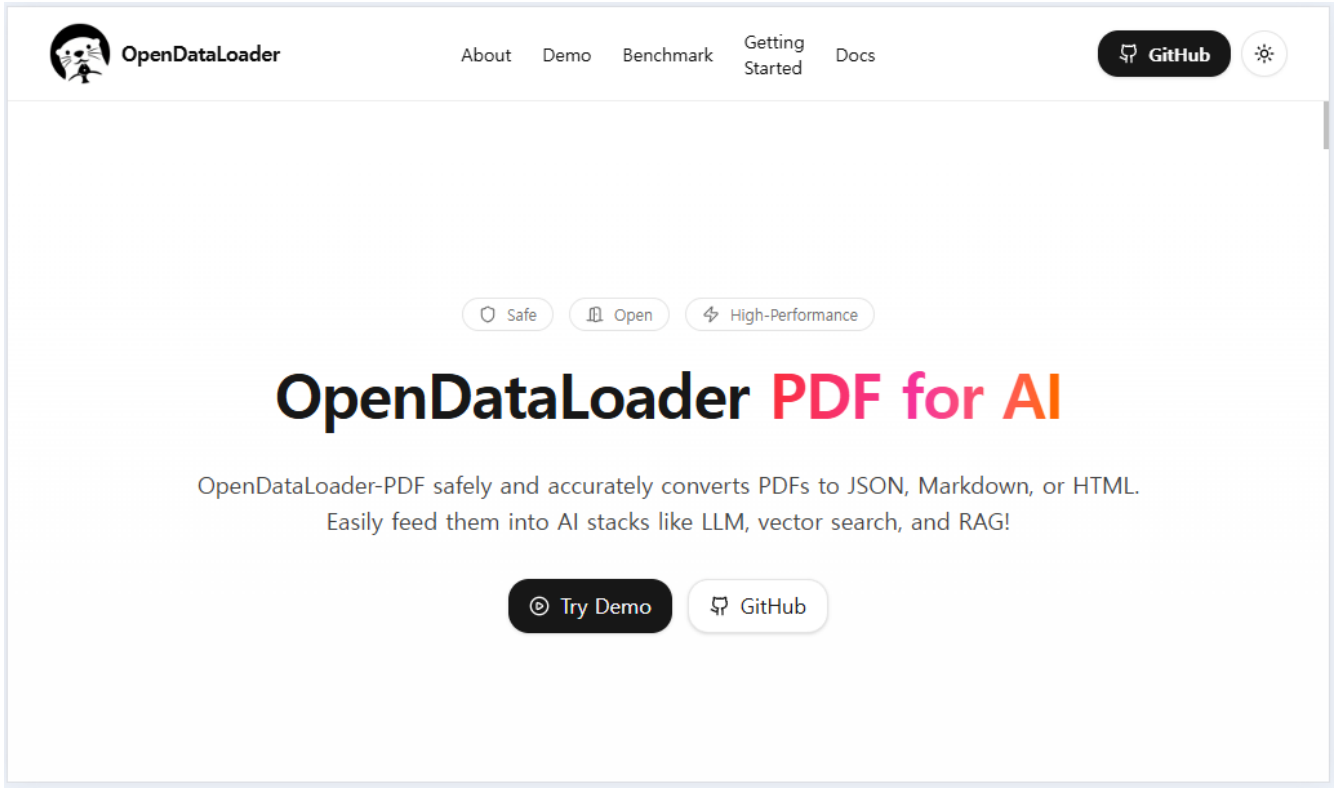
모노레포 소개 및 도입기

2025년 09월 15일 이분도, 박서연, 공소나

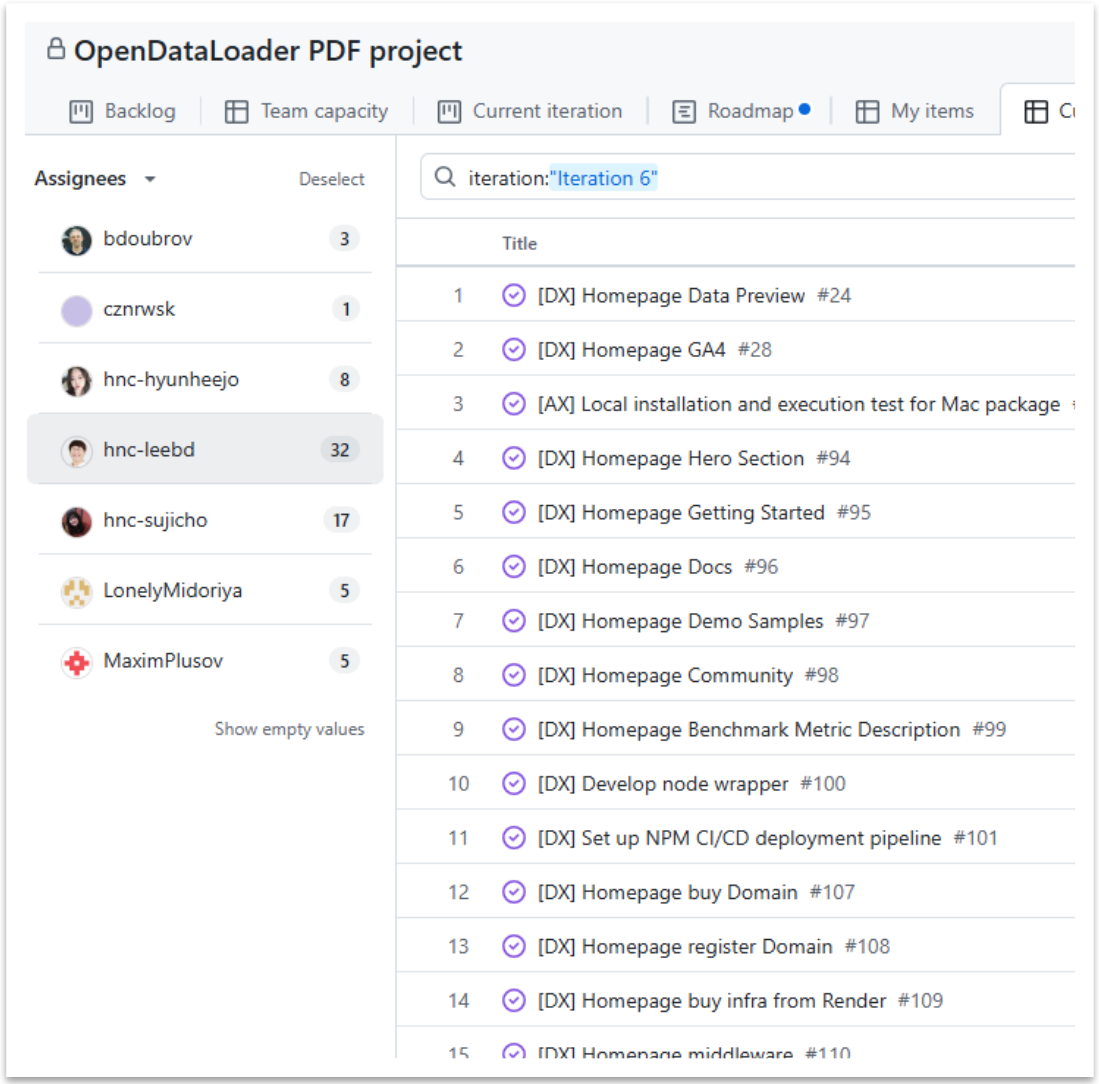
이 글은 모노레포와 멀티레포의 구조적 차이와 선택 기준을 다룹니다. pnpm workspace, Turborepo, 특징을 비교하며, 실제 프로젝트 도입...

AI Agent

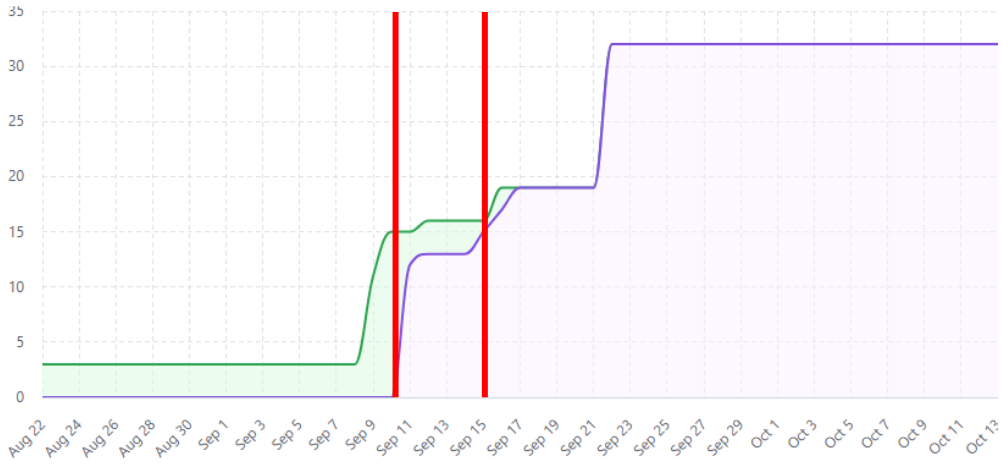
- ✓ **OpenAI Codex**
  - Tailwind CSS 기반 모던 웹 디자인 잘함
- ✓ **Google Gemini Pro**
  - Code, GitHub Action 작성 잘함
- ✓ **Google NanoBanana**
  - 최고의 이미지 AI
- ✓ **GitHub spec-kit**
  - 컨텍스트 엔지니어링 도구 (CLAUDE.md, GEMINI.md)
  - 3단계 설계
    - Spec : 요구사항 정리
    - Plan : 기능 및 단위 테스트 설계
    - Task : TDD



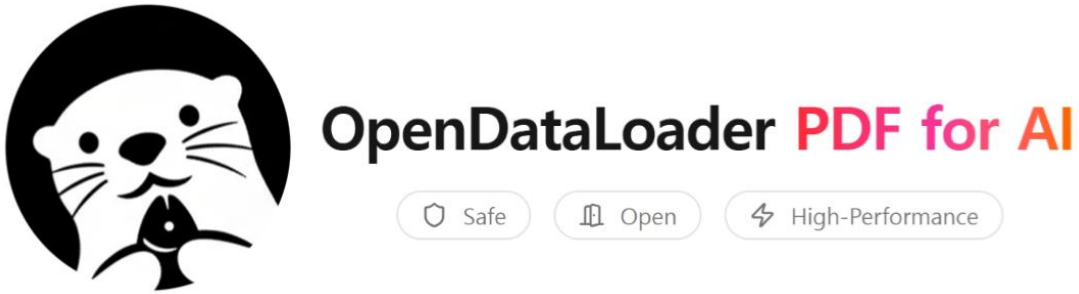
GitHub Project (Issue Manager)



DevOps = 2일
개발 = 2일



회고



10X

40일 → 4일

- ✓ AI x GitHub 생태계로 생산성을 크게 개선한 경험이었습니다.
- ✓ 실험적 사이드 프로젝트의 진입 장벽이 많이 낮아졌다고 느끼고 있습니다.
- ✓ 기술적 인사이트와 경험을 공유하는 문화가 점점 발전하고 있는데 도움이 되었으면 합니다.