

AI 기반 웹 크롤러 기획안

다양한 구조의 웹사이트에서 데이터를 자동 수집하는
지능형 크롤러 설계안 발표

AI/LLM

Web Crawler

Data Pipeline

2025 Web Crawler Architecture & Design

목차

01  개요

02  아키텍처

03  주요 기능

04  UI 사용 흐름

05  AI 모델 활용

06 </> 기술 스택

07  결론

프로젝트 배경

- ⚠ 사이트마다 다른 복잡한 DOM 구조
- ⚠ 동적으로 생성되는 콘텐츠 처리의 어려움
- ⚠ 비정형 데이터 추출의 한계
- ⚠ 수동 설정이 필요한 크롤러 개발
- ⚠ 페이지 구조 변경 시 유지보수 부담



프로젝트 목표

 AI와 LLM을 활용한 지능적 크롤링

 다양한 사이트 구조 자동 적응

 자동 필드 추출 및 제안 기능

 사람처럼 유연한 데이터 추출

 손쉬운 데이터 수집 및 구조화



핵심 가치 제안



지능적 데이터 추출

AI/LLM 기반으로 웹 구조를 자동 이해하고 핵심 정보 식별



뛰어난 범용성

뉴스, 쇼핑몰, 커뮤니티 등 다양한 사이트 구조 자동 대응



높은 효율성

수동 설정 최소화로 크롤러 개발 시간 90% 단축



완전한 자동화

사람의 판단처럼 유연하고 지능적인 자동 크롤링



전체 시스템 아키텍처



사용자 인터페이스



작업 설정 패널

Task Configuration Panel

- 크롤링 대상 URL/도메인 입력 및 관리
- 크롤링 범위 및 깊이 설정
- 동적 콘텐츠 로딩 옵션 선택
- 스케줄링 및 반복 작업 설정



필드 구성 에디터

Field Configuration Editor

- AI가 제안한 필드 확인 및 검증
- 필드명 수정 및 추가/삭제
- 페이지 요소 하이라이트 및 직접 선택
- 추출 규칙 실시간 미리보기



실시간 모니터링

Real-time Monitoring Dashboard

- 진행 상황 및 처리 속도 실시간 표시
- 수집된 페이지 수 및 데이터 항목 통계
- 오류 및 경고 로그 실시간 출력
- 작업 일시정지/재개/중지 제어



결과 뷰어

Result Viewer

- 추출된 데이터 표 형태 미리보기
- 원본 페이지 링크 및 검증 기능
- CSV/JSON/DB 등 다양한 형식 출력
- 크롤링 설정 저장 및 재사용

크롤링 엔진



병렬/비동기 처리

다중 스레드 및 비동기 방식으로 여러 페이지를 동시에 수집하여 크롤링 속도를 극대화합니다.

- Python Asyncio 기반 비동기 처리
- 다중 워커를 통한 병렬 크롤링
- URL 큐 관리 및 우선순위 제어
- 요청 속도 제한 및 리소스 최적화



Headless 브라우저

자바스크립트로 생성되는 동적 콘텐츠를 실제 브라우저 렌더링을 통해 완벽하게 수집합니다.

- Playwright를 통한 브라우저 제어
- 무한 스크롤 및 클릭 이벤트 처리
- AJAX 동적 로딩 콘텐츠 대응
- 스크린샷 및 디버깅 기능 지원



robots.txt 준수

웹사이트의 크롤링 정책을 존중하여 윤리적이고 안전한 데이터 수집을 보장합니다.

- robots.txt 자동 파싱 및 규칙 적용
- Crawl-delay 설정 자동 반영
- 사용자 정의 User-Agent 설정
- 크롤링 허용 경로 사전 검증



Anti-bot 우회

다양한 봇 탐지 메커니즘을 우회하여 안정적인 크롤링 작업을 수행합니다.

- 프록시 서버 자동 교체 시스템
- 랜덤 User-Agent 및 헤더 변경
- 요청 간 지연 시간 자동 조절
- CAPTCHA 감지 시 알림 및 대응

AI 분석 모듈



AI Analysis Module

LLM 기반 지능형 분석



구조 자동 인식

웹 페이지의 DOM 구조를 자동으로 분석하고 핵심 콘텐츠 영역을 식별합니다

- HTML 구조 패턴 분석
- 콘텐츠 밀도 알고리즘
- Boilerplate 제거



필드 자동 추출

LLM을 활용해 제목, 본문, 날짜 등 주요 필드를 자동으로 추출합니다

- 의미 단위 정보 추출
- 비정형 데이터 구조화
- 문맥 기반 이해



페이지 유형 분류

뉴스, 제품, 게시글 등 페이지 타입을 자동으로 분류하고 맞춤 전략을 적용합니다

- 뉴스 기사 페이지
- 쇼핑몰 상품 페이지
- 커뮤니티 게시글



지능형 추출 전략

규칙 기반 파싱과 AI 추론을 결합하여 최적의 추출 전략을 수립합니다

- 룰 기반 + AI 하이브리드
- 선택자 자동 생성
- 적응형 크롤링 로직

데이터 처리 및 저장



유효성 검사

스키마 기반으로 데이터 정확성을 검증하고 오류를 사전에 탐지합니다

- 필수 필드 누락 검사
- 데이터 타입 및 형식 검증
- AI 기반 문맥 적합성 판정
- 실시간 오류 로그 및 경고



데이터 정규화

일관된 형식으로 데이터를 변환하여 품질을 향상시킵니다

- HTML 태그 및 특수문자 제거
- 날짜/금액 형식 표준화
- 언어 감지 및 인코딩 통일
- 공백 정리 및 텍스트 정제



품질 관리

AI 기반 품질 검증으로 데이터 신뢰성을 확보합니다

- 이상치 및 중복 데이터 탐지
- 분류 모델 기반 품질 평가
- 구조 변경 자동 감지
- 데이터 분포 분석 및 리포팅



다양한 출력 형식

목적에 맞는 다양한 형태로 데이터를 제공합니다

- 파일 형식: CSV, JSON, Excel
- 데이터베이스: MySQL, PostgreSQL, MongoDB
- API 연동: REST API, GraphQL
- 스트리밍: Kafka, RabbitMQ

다양한 사이트 구조 지원

뉴스, 포럼, 쇼핑몰 등 다양한 웹사이트 구조를 자동으로 인식하고 최적화된 데이터 추출 전략을 적용합니다



뉴스 사이트

News Portal

- 기사 제목 및 부제
- 본문 콘텐츠
- 작성자 및 기자명
- 게시일 및 수정일
- 카테고리 및 태그
- 이미지 및 멀티미디어

자동 감지



커뮤니티 포럼

Forum & Community

- 게시글 제목
- 작성자 및 프로필
- 게시 내용
- 댓글 및 답글
- 조회수 및 추천수
- 첨부파일 링크

자동 감지



쇼핑몰

E-commerce

- 상품명 및 브랜드
- 가격 및 할인정보
- 상품 설명 및 스펙
- 재고 및 배송정보
- 고객 리뷰 및 평점
- 옵션 및 선택사항

자동 감지



블로그

Blog Platform

- 포스트 제목
- 저자 정보
- 본문 및 이미지
- 작성일 및 카테고리
- 태그 및 키워드

자동 감지



리뷰 사이트

Review Platform

- 제품/서비스명
- 평점 및 별점
- 리뷰 내용
- 작성자 및 날짜
- 추천 및 반응

자동 감지



기타 사이트

Other Structures

- 위키 및 문서
- 채용 공고
- 부동산 매물
- 이벤트 정보
- 맞춤형 구조

자동 감지

DOM 기반 크롤링 및 동적 콘텐츠 처리

⚡ 정적 콘텐츠 크롤링

Static Content Crawling

핵심 기술 스택

requests BeautifulSoup lxml XPath/CSS

- 빠른 HTTP 요청으로 HTML 소스 직접 수집
- 서버에서 렌더링된 완성된 HTML 파싱
- XPath 및 CSS 선택자로 정확한 요소 추출
- 빠른 처리 속도와 낮은 리소스 소비
- 대량 페이지 병렬 수집에 효율적

</> 동적 콘텐츠 처리

Dynamic Content Handling

핵심 기술 스택

Playwright Selenium Headless Chrome

- JavaScript 실행 후 렌더링된 DOM 수집
- 무한 스크롤, AJAX 로딩 콘텐츠 처리
- 버튼 클릭, 폼 입력 등 사용자 동작 자동화
- SPA(Single Page Application) 대응
- 복잡한 인터랙션이 필요한 사이트 지원

👉 추가 최적화 전략

공식 API 우선 활용

사이트에서 제공하는 공식 API가 있을 경우 우선적으로 활용하여 효율성 극대화

RSS 피드 연동

뉴스 사이트 등에서 RSS 피드 제공 시 구조화된 데이터 직접 수집

하이브리드 접근

정적 파싱 우선, 실패 시 단계적으로 동적 렌더링 방식 적용

UI 자동 필드 제안

✍️ 하이브리드 필드 인식 프로세스



✍️ 하이라이트 기능

- 페이지 미리보기에서 추출 영역 시각적 표시
- 각 필드별 색상으로 구분하여 표시
- 클릭으로 해당 요소 직접 선택 가능

📌 커스터마이징 기능

- 필드 이름 및 타입 수정
- 추출 규칙(셀렉터) 직접 편집
- 실시간 추출 결과 미리보기

☰ 자동 식별 항목

- 제목, 본문, 작성자, 날짜 등 주요 필드
- 이미지, 링크, 메타데이터 자동 감지
- 사이트 유형별 맞춤 필드 제안

⌚ 실시간 검증

- 추출된 예시 값 즉시 표시
- 필드별 데이터 품질 확인
- 오류 발견 시 자동 경고 알림

구조 인식 및 추출 로직

룰 기반 추출

Rule-Based Extraction

- HTML 태그 기반 식별 (main, article, h1 등)
- CSS 선택자 및 XPath 활용
- 콘텐츠 밀도 알고리즘 적용
- 정형화된 패턴 빠른 처리
- 높은 정확도와 안정성



AI 추론 추출

AI-Powered Extraction

- LLM 기반 문맥 이해
- 비정형 데이터 유연 처리
- 의미 단위 정보 추출
- 다양한 표현 방식 대응
- 사람처럼 판단하는 추출



하이브리드 접근 방식

정형 패턴은 규칙으로 정확하게, 비정형 표현은 AI로 유연하게 처리하여 최적의 추출 정확도와 범용성 확보

 높은 정확도

 광범위한 적용

 효율적 처리

 자동 최적화

리스트 페이지 탐색



페이지네이션

번호로 구분된 페이지 자동 탐색

- 다음/이전 버튼 자동 인식
- URL 패턴 분석 및 자동 생성
- 페이지 번호 범위 자동 추출
- 마지막 페이지 자동 감지



무한 스크롤

동적 콘텐츠 로딩 자동 처리

- 스크롤 이벤트 자동 트리거
- AJAX 요청 모니터링
- 콘텐츠 로딩 완료 감지
- 전체 항목 수집 완료 판단



링크 패턴 추출

규칙적 URL 구조 자동 분석

- 정규식 기반 패턴 매칭
- 카테고리/섹션 링크 식별
- 불필요한 링크 자동 필터링
- 유사 URL 구조 자동 탐색

💡 자동 탐색 최적화

목록/상세 페이지를 자동 구분하고, 개별 항목의 링크와 메타정보를 추출하여 체계적인 크롤링 경로를 구축합니다. 외부 링크나 광고는 자동으로 제외됩니다.

중복 방지 및 유효성 검사



URL 중복 방지

해시 기반 URL 관리로 중복 수집 방지

- 모든 방문 URL을 해시로 기록하여 관리
- 중복 URL 자동 스kip 처리
- 동일 콘텐츠의 다른 경로 감지
- 효율적인 메모리 사용 최적화



スキ마 기반 검증

필수 필드 및 형식 유효성 자동 검사

- 필수 필드 누락 여부 확인
- 데이터 타입 형식 검증
- 날짜/숫자/URL 형식 체크
- 오류 발생 시 로그 기록 및 경고



콘텐츠 중복 탐지

동일 내용의 중복 저장 차단

- 콘텐츠 해시값 비교
- 유사도 기반 중복 판정
- 여러 카테고리 중복 게시글 처리
- 한 번만 저장하는 스마트 수집



AI 기반 품질 검증

지능형 이상값 탐지 및 품질 관리

- 추출 값의 문맥 적합성 판별
- 분류 모델로 필드 내용 검증
- 이상치 자동 탐지 및 표시
- 품질 점수 기반 필터링



데이터 품질 관리 프로세스

STEP 1

URL 중복 체크



STEP 2

데이터 수집



STEP 3

スキ마 검증



STEP 4

AI 품질 검증



STEP 5

최종 저장

다국어 처리 및 정규화



언어 감지 및 처리

- 자동 언어 감지로 다양한 언어의 콘텐츠 식별
- UTF-8 기반 유니코드 인코딩 지원
- 언어별 로케일에 맞는 데이터 해석
- 다국어 LLM 모델을 활용한 추출 정확도 향상



콘텐츠 정규화

- HTML 태그 및 스크립트 코드 제거
- 불필요한 공백, 줄바꿈 정리
- 광고 문구 및 노이즈 콘텐츠 필터링
- 특수문자 및 이모지 처리

형식 통일 및 표준화

날짜 형식

2024-01-15



ISO8601: 2024-01-15T00:00:00Z

통화 금액

₩1,234,500원



1234500 (KRW)

전화번호

02-1234-5678



+82-2-1234-5678

주소 형식

서울 강남구



서울특별시 강남구

숫자 단위

1.5k, 2M



1500, 2000000

인코딩 통일

EUC-KR, CP949



UTF-8 표준화

오류 처리 및 로깅



예외 처리

- HTTP 오류 처리 (404, 500, 타임아웃)
- 파싱 예외 및 데이터 타입 오류
- 네트워크 연결 실패 격리
- 실패 URL 별도 관리



재시도 메커니즘

- 자동 재시도 로직 (exponential backoff)
- 최대 재시도 횟수 설정
- 부분 실패 시 작업 계속 진행
- 실패 항목 별도 리포트



실시간 로깅

- 모든 이벤트 상세 기록
- UI 대시보드 실시간 표시
- 로그 레벨별 필터링
- 에러 추적 및 디버깅 지원



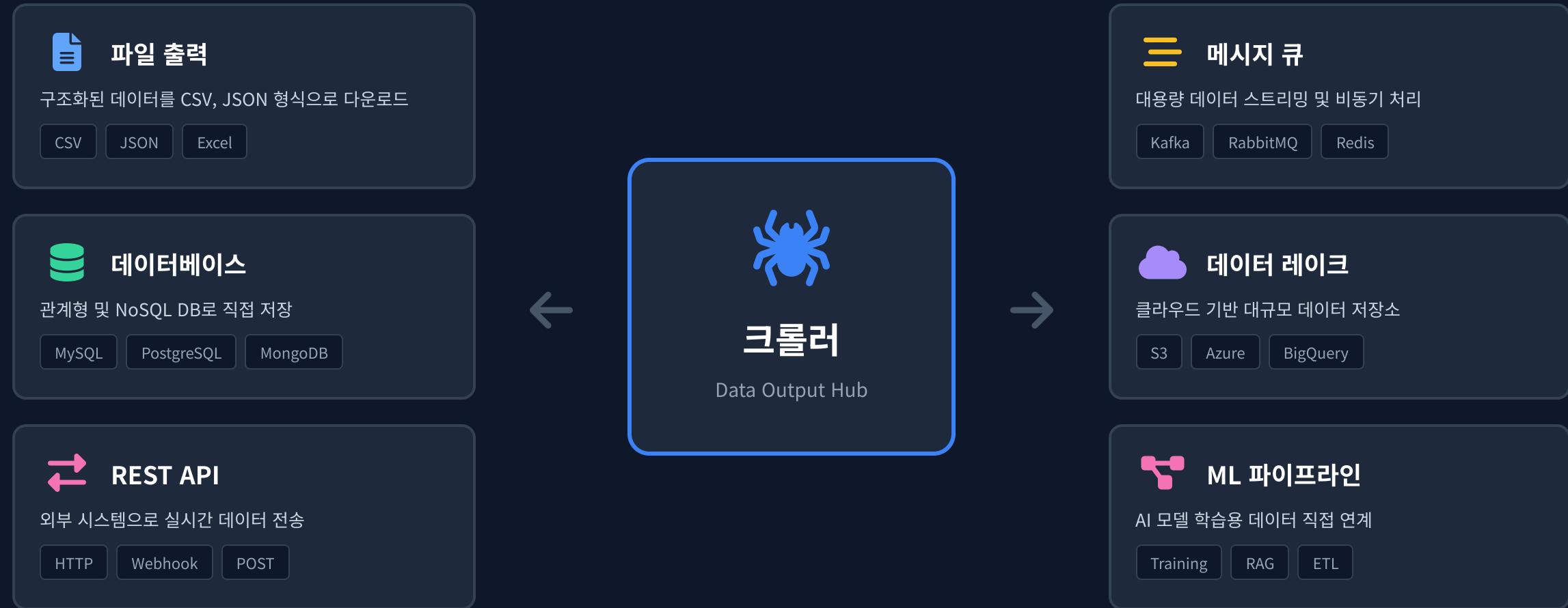
반복 차단 대응

- IP 프록시 자동 교체
- 랜덤 User-Agent 사용
- 요청 지연 시간 조절
- CAPTCHA 감지 시 사용자 알림

>_ 실시간 크롤링 로그 예시

```
14:32:15 [INFO] 크롤링 작업 시작 - 대상: example.com
14:32:18 [SUCCESS] 페이지 1 수집 완료 - 10개 항목 추출
14:32:22 [WARN] 페이지 5 타임아웃 - 재시도 중 (1/3)
14:32:28 [SUCCESS] 페이지 5 재시도 성공
14:32:35 [ERROR] 페이지 12 파싱 실패 - 스키마 불일치
14:32:40 [INFO] 진행률: 85% (170/200 페이지)
```

데이터 파이프라인 연계



UI 사용 흐름: 1단계 작업 생성

+ STEP 1

크롤링 작업 생성

사용자가 웹 크롤러 UI에 접속하여 새로운 크롤링 작업을 생성합니다. 대상 URL 또는 도메인을 입력하고, 수집하고자 하는 데이터의 목적을 설명합니다.

- 대상 URL 또는 도메인 입력
- 크롤링 목적 설명 (예: 뉴스 기사 수집, 상품 리뷰 수집)
- 원하는 데이터 필드 목록 입력 (선택사항)
- LLM 분석을 위한 힌트로 활용

+ 새 크롤링 작업 생성

대상 URL / 도메인

`https://example.com/news`

크롤링 목적

예: 최근 뉴스 기사 수집

수집할 데이터 필드 (선택사항)

예: 제목, 본문, 작성자, 날짜, 이미지

→ 다음 단계: 구조 분석

UI 사용 흐름: 2단계 페이지 구조 분석

입력된 페이지 분석

사용자가 입력한 URL 또는 도메인의 페이지를 불러와 구조를 분석합니다. 목록 페이지인 경우 대표 상세 페이지도 함께 분석하여 전체 구조를 파악합니다.

AI + 규칙 기반 구조 해석

LLM과 규칙 기반 파서를 결합하여 페이지의 핵심 콘텐츠 요소를 식별합니다.
HTML 구조, 콘텐츠 밀도, 시맨틱 태그 등을 종합적으로 분석합니다.

 LLM 추론

 규칙 기반

자동 제안된 필드

제목 (Title)

"AI 기반 웹 크롤링의 미래와 전망"

셀렉터: article > h1.post-title

본문 (Content)

"최근 대규모 언어모델의 발전으로..."

셀렉터: article > div.content-body

작성일 (Date)

2025-01-15

셀렉터: time.publish-date

작성자 (Author)

김개발

셀렉터: span.author-name

 페이지 미리보기에서 각 필드에 해당하는 요소가 하이라이트되어 표시됩니다. 사용자는 이를 확인하고 수정할 수 있습니다.

UI 사용 흐름: 3단계 사용자 검증

STEP 3

필드 검증 및 커스터마이징

제목 (Title)

"AI 기반 웹 크롤러..."



본문 (Content)

"다양한 구조의 웹사이트..."



작성일 (Date)

"2025-10-20"



💡 원본 페이지에서 해당 요소들이 하이라이트되어 시각적으로 확인 가능



필드 수정

필드 이름 변경, 추출 영역 조정, 데이터 타입 설정



직접 선택

원하는 요소를 직접 클릭하여 추출 대상 지정



고급 옵션

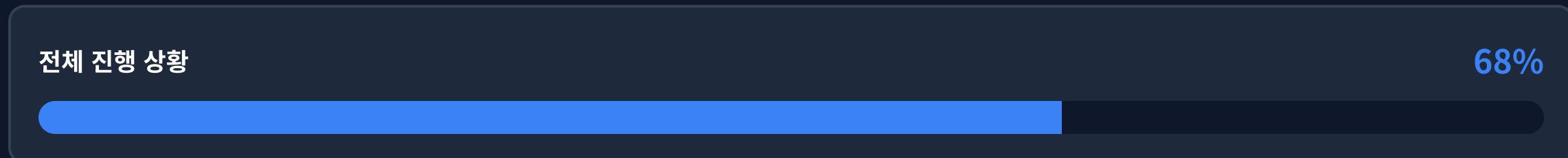
크롤링 범위, 탐색 깊이, 동적 콘텐츠 로딩 설정



실시간 미리보기

변경된 규칙으로 예시 추출 결과를 즉시 재계산

UI 사용 흐름: 4단계 - 크롤링 실행 및 모니터링



>_ 실시간 크롤링 로그	
14:32:45	✓ Page 342 loaded successfully - 15 items extracted
14:32:43	✓ Processing: https://example.com/news/article-1234
14:32:40	⚠ Retrying page 341 (timeout) - attempt 2/3
14:32:38	✓ Page 340 loaded - 12 items extracted

UI 사용 흐름: 5단계 결과 확인 및 저장

트 크롤링 결과 요약

총 소요 시간	2분 34초
크롤링한 페이지 수	156개
추출된 데이터 레코드	1,248건
에러 발생 건수	3건
데이터 품질 점수	98.5%

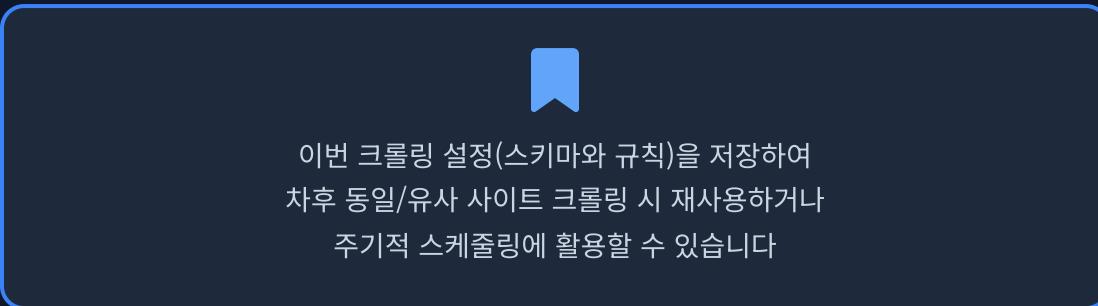
田 데이터 미리보기

제목	날짜	작성자	조회수
AI 기술의 최신 동향	2025-10-20	김철수	1,234
웹 크롤링 자동화 방법	2025-10-19	이영희	856
데이터 분석 실전 가이드	2025-10-18	박민수	2,103

⬇ 데이터 출력 옵션



▣ 크롤러 설정 저장



AI 모델 활용 전략

 LLM 기반 페이지 구조 이해 및 분석

 자동 필드 추출 및 데이터 변환

 크롤러 설정 및 코드 자동 생성

 지능형 크롤링 경로 탐색 및 제어

 AI 기반 품질 검증 및 예외 처리



LLM 기반 구조 이해 및 추출



페이지 구조 이해

HTML과 텍스트 내용을 분석하여 페이지의 구조와 핵심 요소를 자동으로 파악

- LLM이 HTML DOM을 읽고 의미 단위로 구조 요약
- 제목, 본문, 메타정보 등 핵심 요소 자동 식별
- 페이지 유형 분류 및 적합한 추출 전략 선택

프롬프트 예시

"이 페이지에서 뉴스 기사 제목과 본문이 어디에 있는지 찾아줘"



데이터 필드 추출

비정형 텍스트에서 구조화된 데이터를 지능적으로 추출하고 변환

- 복잡한 문장에서 의미 단위로 정보 분리
- 사용자 생성 콘텐츠의 핵심 데이터 추출
- 추출된 데이터를 JSON 형식으로 자동 변환

활용 사례

리뷰 텍스트에서 제품 장단점, 구매자 정보, 평점 등을 자동으로 구조화



크롤러 코드 생성

사용자 요구사항을 바탕으로 크롤러 스크립트와 설정을 자동 생성

- 자연어 지시로부터 크롤링 로직 자동 구성
- XPath/CSS 선택자 및 Spider 코드 제안
- 비개발자도 쉽게 크롤러 설정 완료 가능

입력 예시

"이 블로그에서 모든 게시글의 제목과 날짜를 수집하는 크롤러"

품질 검증 및 최적화

✓ AI 기반 품질 검증

- **스키마 기반 유효성 검사**
필수 필드 누락, 형식 불일치 자동 감지
- **LLM 문맥 검증**
추출된 값의 의미적 적합성 판단
- **이상치 탐지**
다른 데이터와 패턴이 다른 항목 식별
- **자동 피드백 루프**
문제 발견 시 크롤링 규칙 자동 조정

● 성능 최적화 전략

- **샘플링 전략**
초기 페이지만 LLM 분석 후 패턴 재사용
- **캐싱 메커니즘**
유사 구조 페이지의 분석 결과 저장
- **규칙 우선 처리**
정형 데이터는 규칙, 비정형만 AI 활용
- **비용 모니터링**
API 호출 추적 및 예산 관리

⚡ AI/규칙 병합 최적화



효율성 우선

규칙 기반으로 빠른 처리, 복잡한 경우만 AI 활용



안정성 확보

AI 응답 불일치 대비 백업 로직 항시 준비



지속적 개선

품질 메트릭 수집 및 모델 파인튜닝

프론트엔드/백엔드 기술



프론트엔드

> React.js / Vue.js

싱글 페이지 애플리케이션(SPA) 기반 UI 구현

> WebSocket

실시간 크롤링 진행 상황 및 로그 업데이트

> Bootstrap / Tailwind CSS

직관적인 대시보드 및 반응형 디자인



백엔드

> Python

풍부한 크롤링 생태계 및 AI 라이브러리 지원

> FastAPI

고성능 API 엔드포인트 및 WebSocket 통신

> Asyncio

비동기 처리로 대량 URL 병렬 크롤링



크롤링 프레임워크

> Scrapy

강력한 오픈소스 크롤러 프레임워크

> Playwright / Selenium

동적 콘텐츠 렌더링을 위한 Headless 브라우저

> BeautifulSoup / lxml

HTML DOM 파싱 및 데이터 추출



지원 도구

> Celery / Redis

비동기 작업 큐 및 분산 처리

> Docker / Kubernetes

컨테이너화 및 확장 가능한 배포

> Proxy API

IP 로테이션 및 Anti-bot 우회

AI/ML 및 데이터 처리 기술



AI/ML 프레임워크

OpenAI GPT-4	핵심	LLM 기반 구조 분석
LangChain	핵심	LLM 애플리케이션 개발
Transformers	핵심	오픈소스 LLM 활용
spaCy	선택	자연어 처리
fastText	선택	텍스트 분류



데이터 처리

Pandas	핵심	데이터 분석 및 변환
SQLAlchemy	핵심	관계형 DB ORM
MongoDB	선택	NoSQL 데이터 저장
CSV/JSON	핵심	표준 데이터 포맷
Apache Kafka	선택	메시지 큐 연동



인프라 및 배포

Docker	핵심	컨테이너 기반 배포
Kubernetes	선택	오케스트레이션
AWS EC2	선택	클라우드 호스팅
Redis	선택	캐싱 및 세션 관리
Nginx	핵심	리버스 프록시



확장 도구

Luminati/Zyte	선택	프록시 API
Sentry	선택	오류 모니터링
Celery	핵심	비동기 작업 큐
WebSocket	핵심	실시간 통신
dateparser	핵심	날짜 파싱

결론 및 기대효과

AI 기반 웹 크롤러가 만드는 데이터 수집의 미래



비즈니스 혁신

시장 분석, 경쟁사 모니터링, 고객 리뷰 분석 등
실시간 데이터 기반 의사결정



연구 효율화

대규모 웹 데이터 수집 자동화로 학술 연구 및
데이터 과학 프로젝트 가속화



AI 학습 데이터

다양한 웹 데이터를 구조화하여 LLM 파인튜닝
및 RAG 시스템 구축에 활용

지능형 웹 크롤러로 데이터 수집의 새로운 시대를 열어갑니다

자동화 · 지능화 · 범용성