코드하우스 코딩아카데미 기업연계 셀 프로젝트 계획서

소프트웨어 명	크롤링을 통한 데이터 수집 프로그램
소프트웨어 목적	다양한 웹 사이트의 데이터를 자동으로 수집하고 정제하여 내부 데이터 분석 및 시스템 연동에 활용할 수 있는 크롤링 프로그램 개발
협업 내용 (개발 범위, 추후 상세 협의 예정)	1. URL 및 CSS 선택자 기반 데이터 수집 2. 다중 페이지 게시판 자동 순회 크롤링 3. 요청 간 지연 시간 설정 기능 4. robots.txt 준수 로직 5. User-Agent 커스터마이징 기능 6. 수집 데이터 DB 저장 및 테이블 구조 설계
협업 기능 정의 (추후 기능정의서 검토 후 협업 기능 확정)	1. URL 및 선택자 입력 기능 2. 다중 페이지 순회 크롤링 3. 요청 간 지연 시간 설정 4. robots.txt 준수 여부 설정 5. User-Agent 설정 기능 6. DB 연결 및 데이터 저장 7. 크롤링 로그 기록 및 저장 8. 사용자에게 편리한 UI/UX
기대 효과	1. 수작업 수집업무 자동화 2. 수집 품질 및 정확도 향상 3. 추후 공공기관 제안 시 PoC 형태로 기능 포함 가능 4. 타 데이터 수집 업무에도 확장 활용 가능
협업시 필요 사항	1. 사용자 매뉴얼 2. 개발자용 기술문서 (DB구조, 크롤링 흐름) 3. 설치/배포 가이드 4. 크롤링 결과 리포트 샘플 5. 요구사항 정의서 6. 테스트 시나리오

※ 프로젝트 진행을 통해 개발하고자 하는 기업 소프트웨어 관련 자료 별도첨부 가능 (PDF or PPT)

* 프로젝트 요구사항 정의

업무영역	● 업무명칭 : 웹 크롤러 기반 데이터 수집 자동화 ● 업무개요 : 특정 외부 웹 사이트로부터 필요한 텍스트·링크 데이터를 자동 수집하여 내부 DB에 저장하는 프로그램 개발
요구사항	1. URL 및 선택자 입력 기능 2. 다중 페이지 순회 크롤링 3. 요청 간 지연 시간 설정 4. robots.txt 준수 여부 설정 5. User-Agent 설정 기능 6. DB 연결 및 데이터 저장 7. 크롤링 로그 기록 및 저장 8. 사용자에게 편리한 UI/UX
요구사항 설명	1. URL 및 선택자 입력 기능 : 사용자가 크롤링할 대상 URL과 CSS 선택자를 입력하면 해당 위치의 데이터를 수집 2. 다중 페이지 순회 크롤링 : 게시판처럼 pagination이 있는 경우, 다음 페이지로 자동 이동하며 반복 수집 3. 요청 간 지연 시간 설정 : 수집대상서버의 부하를 방지하기 위해 요청 사이 간격(ms/초 단위)을 설정 가능 4. robots.txt 준수 여부 설정 : 크롤링 대상의 robots.txt 파일을 확인, 접근 가능 여부 판단 후 진행 5. User-Agent 설정 기능 : 크롤러의 User-Agent 값을 설정할 수 있도록 하여 서버 차단 회피 가능 6. DB 연결 및 데이터 저장 : 수집된 데이터를 지정된 데이터베이스 테이블에 저장 가능 (MySQL, MariaDB 등) 7. 크롤링 로그 저장 : 수집 시도 시간, 결과, 성공/실패 여부 등 로그로 저장
세부내용 및 요건	● 웹사이트 게시판, 목록형 콘텐츠에서 지정된 영역의 텍스트, 이미지, 날짜, 링크 등 정보 수집 ● 다중 페이지 자동 순회(Pagination 처리) ● 특정 키워드 포함 여부 필터링 기능 ● 데이터 정제 및 구조화 ● 크롤링 대상 URL 및 선택자(Selector) 사용자 입력 가능 ● 요청 간 간격 및 수집 속도 조절 기능 포함 ● 정기 수집 스케줄링 (예: 매일 9시 자동 수집) ● 수집 이력 관리 및 중복 수집 방지 기능
제약사항 및 전제조건	● robots.txt 및 사이트 이용약관 준수 필요 ● 수집대상 페이지의 비동기 로드 여부 (예: ajax, fetch 등) 감안 후 개발 필요