웹 불법 개인정보 유통 모니터링 시스템

KIBOA 1기

2025.09.07

한우언

Contents

개요

01 프로젝트 개요

웹 크롤링 기반 키워드 탐지 시스템 개발

프로젝트 배경

- 불법 개인정보 유통의 심각성
- 기존 수동 감시의 한계점
- 자동화 솔루션의 필요성

핵심 목표

- 효율적인 웹 크롤링 시스템 구현
- 수집 데이터의 체계적인 DB 구조화
- 키워드 매칭 및 필터링 알고리즘 구현
- 필터링 및 정확도 향상 기능 구현

4주 개발 계획

1주	2주	3주	4주
기술 스택 선정	웹 크롤링 개발	DB 연동 및 알고리즘 구현	필터링 기능 및 테스트

진행 과정

02 진행 과정

1주차: 기술 스택 선정

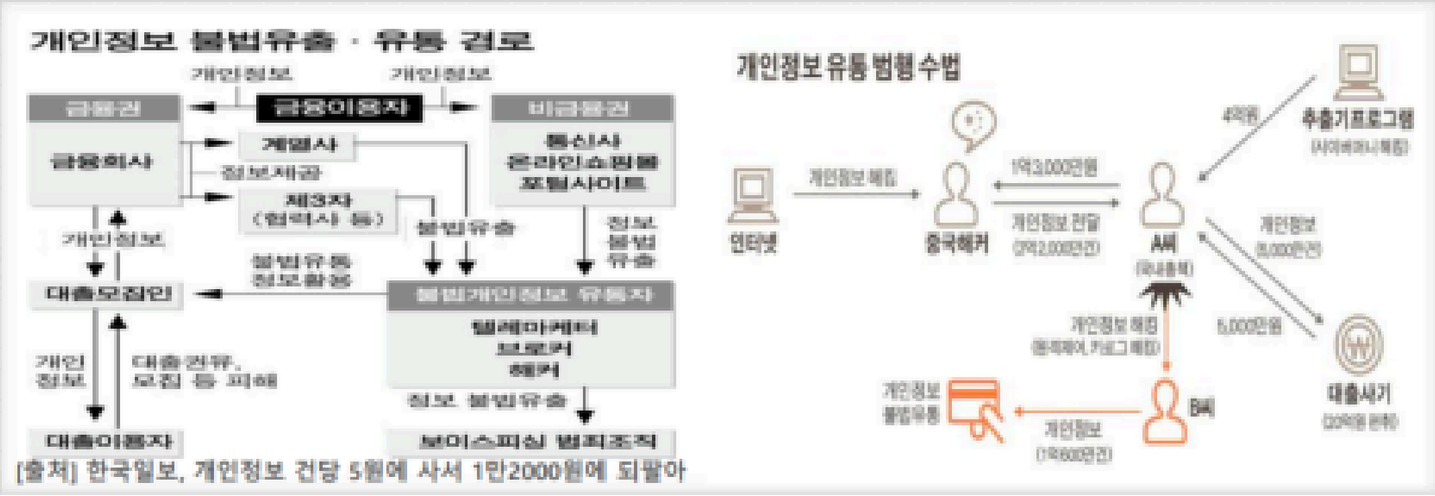
Python

Playwright

PostgreSQL

02 진행 과정

1주차: 키워드 선정



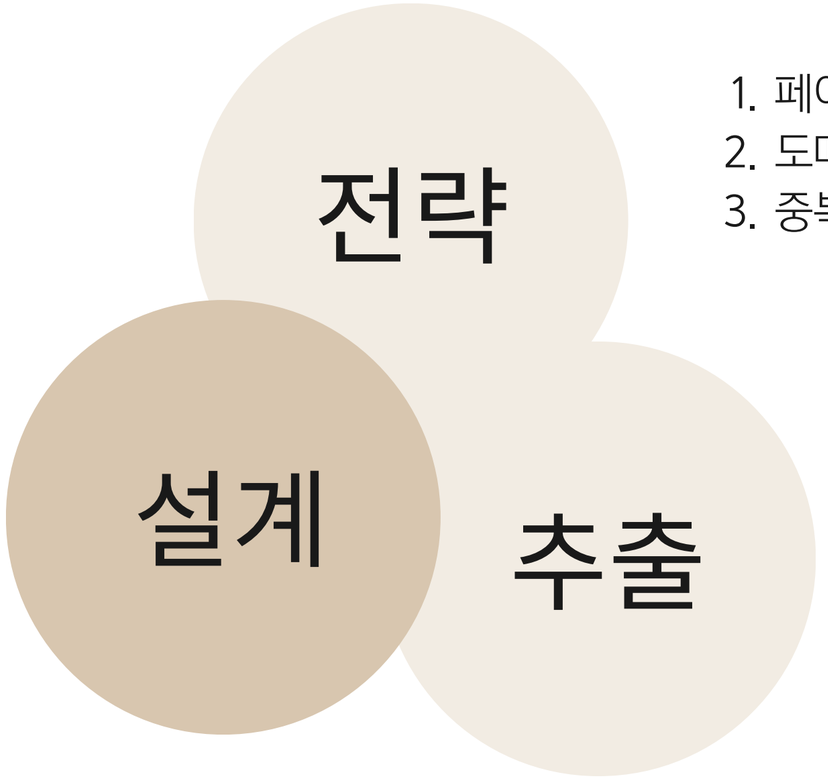
[출처] 한국일보, 개인정보 건당 5원에 사서 1만2000원에 되팔아

구 분	개인정보 DB 불법 브로커가 사용하는 은어 예시
부결DB	신용이 낮아 대출 신청이 거절된 정보
완콜DB	한 번 전화를 통해 확인된 정보. 대출희망금액 포함
전날DB	전날 혹은 최근 대출신청자 정보
실시간DB	당일 혹은 최신 대출신청자 정보
문자DB	이름과 전화번호 등만 남긴 문자발신용 정보
내구재DB	신용이 낮아 휴대전화 등을 담보로 대출한 정보
토토DB	스포츠 복권 사이트에서 얻어낸 정보
주식DB	주식 사이트에서 얻어낸 정보
막DB	여러 경로를 통해서 얻은 정보를 취합한 정보

구 분	아이디(계정) 불법 브로커가 사용하는 은어 예시
최적화ID	유명한 네이버 카페(중고나라, 맘카페 등)등급 중 3등급일 경우 ① 게시글을 올리고 15분 이후에 네이버 검색 결과에 상위 노출을 확인 ② 0등급 : 네이버 검색 결과에 안 나올 경우 ③ 1등급 : 네이버 검색 결과 제목+내용 1페이지에 나올 경우 ④ 2등급 : 네이버 검색 결과 제목만 1페이지에 나올 경우 ⑤ 3등급(최적화) : 모바일 네이버 검색 결과에 2페이지 안에 나올 경우
네이버실명	실명 확인 방법을 통해 실명 전환된 이용자(네이버 전체 서비스 이용)
네이버비실명	실명 전환이 않은 이용자(일부 서비스에 제한) ① 비실명 아이디로 사용 가능한 서비스 • 메일, 블로그, 쪽지 등 개인 영역의 서비스 ② 비실명 아이디로 사용 불가한 서비스 • 뮤직, 영화, 웹툰 다운로드 및 결제가 포함된 서비스

02 진행 과정

2주차: 웹 크롤링 개발



1. 초기화
2. 게시판 순회
3. 링크 수집
4. 본문 파싱
5. 정제/검증
6. 저장 포맷
7. 리소스 정리

1. 페이지네이션 인식
2. 도메인 고정
3. 중복 제거

1. 제목 추출
2. 본문 추출
3. 메타데이터 추출
4. 텍스트 정규화

02 진행 과정

3주차: DB 연동 및 중복 확인

```
# 키워드 테이블 생성
cursor.execute('''
    CREATE TABLE IF NOT EXISTS keywords (
        id SERIAL PRIMARY KEY,
        category VARCHAR(100) NOT NULL,
        keyword TEXT NOT NULL,
        UNIQUE(category, keyword)
    )
''')

# 크롤링 결과 테이블 생성
cursor.execute('''
    CREATE TABLE IF NOT EXISTS crawl_results (
        id SERIAL PRIMARY KEY,
        url TEXT NOT NULL UNIQUE,
        title TEXT,
        content TEXT,
        content_hash TEXT NOT NULL,
        detected_keywords JSONB,
        is_processed BOOLEAN DEFAULT FALSE
    )
''')
```

- **크롤링 결과와 키워드를 관계형 DB에 통합 관리**
- **정규화된 키워드 관리 + JSON 기반 탐지 결과 관리**
- **해시 기반 내용 중복 체크**

데이터 무결성

URL/키워드 중복 방지

확장성

키워드 추가 시 DB 변경 최소화

후처리 용이성

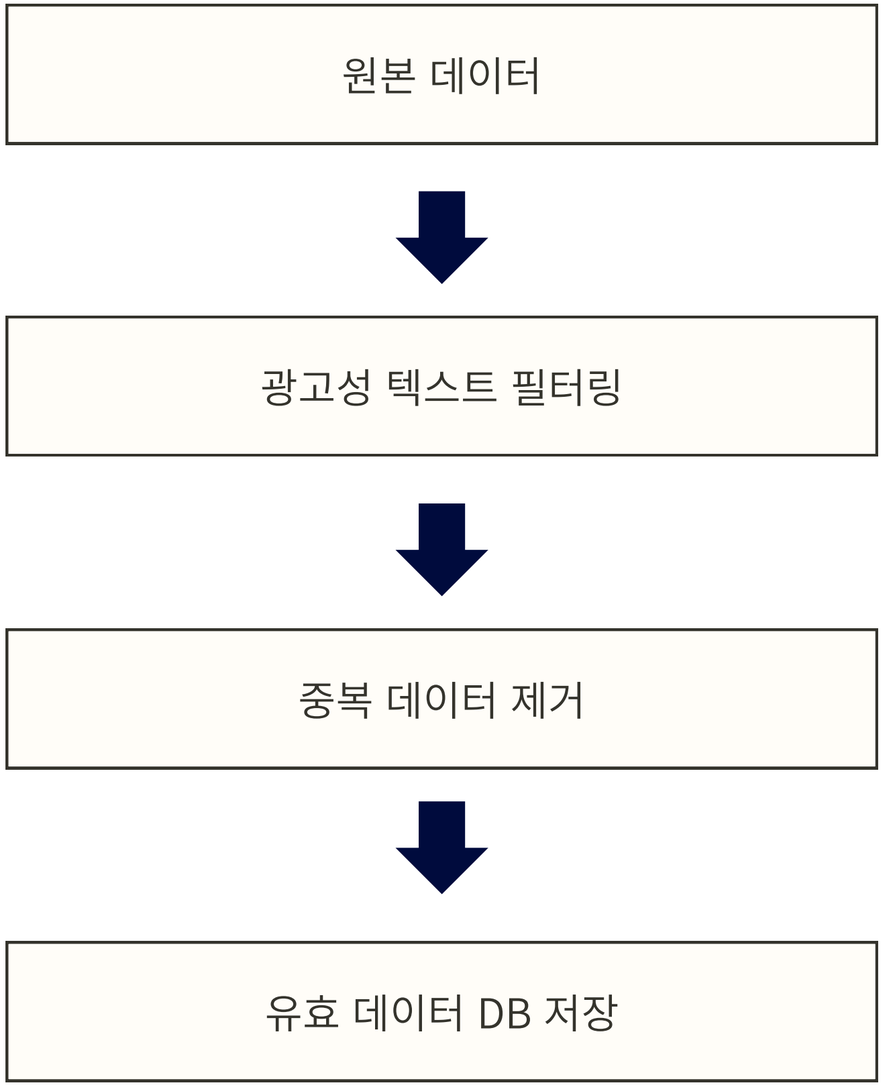
탐지 결과를 활용해 통계·시각화·분석 자동화

02 진행 과정

3주차: 스팸 필터링

- 1단계: 불법 광고 차단
  - 불법 키워드 패턴을 미리 정의
  - 텍스트에서 이런 키워드가 발견되면 즉시 차단
- 2단계: 스팸 특수문자 감지
  - 광고에서 자주 사용하는 특수문자 패턴 분석
- 3단계: 업체 정보 필터링
  - 전화번호 패턴 (010-1234-5678)
  - 운영시간 패턴 (09:00~18:00)
- 4단계: 동적 패턴 학습
  - 같은 단어가 40% 이상 반복되는 메뉴성 텍스트 자동 감지

	id integer	category character varying	keyword text
1	1301	personal_db	부결DB
2	1302	personal_db	부결디비
3	1303	personal_db	완콜DB
4	1304	personal_db	완콜디비
5	1305	personal_db	전날DB
6	1306	personal_db	전날디비



02 진행 과정

3주차: 키워드 탐지 알고리즘

- 1차 필터: 빠른 스크리닝
  - 전체 페이지를 분석하기 전에 제목, 메타 설명, 주요 헤더만 먼저 확인
  - "JSON", "데이터베이스" 같은 핵심 키워드가 있는지 빠르게 판단
  - 정규식 패턴 매칭

```
DB_PATTERNS = [
    r'부결\s*(?:db|디비|d\.?b)',
    r'완콜\s*(?:db|디비|d\.?b)',
    r'전날\s*(?:db|디비|d\.?b)',
    r'실시간\s*(?:db|디비|d\.?b)'
    r'문자\s*(?:db|디비|d\.?b)',
    r'내구재\s*(?:db|디비|d\.?b)'
    r'토토\s*(?:db|디비|d\.?b)',
    r'주식\s*(?:db|디비|d\.?b)',
    r'막\s*(?:db|디비|d\.?b)',
    r'대출\s*(?:db|디비|d\.?b)'
]
```

02 진행 과정

3주차: 키워드 탐지 알고리즘

```
# 해당 카테고리의 필수 지표 확인
category_required = REQUIRED_INDICATORS.get(category, [])
has_required_in_context = any(indicator in context_lower for indicator in category_required)

if keyword_has_trade or context_has_trade:
    has_required = True
else:
    has_required = has_required_in_context

has_negative = any(indicator in context_lower for indicator in NEGATIVE_INDICATORS)

return has_required and not has_negative
```

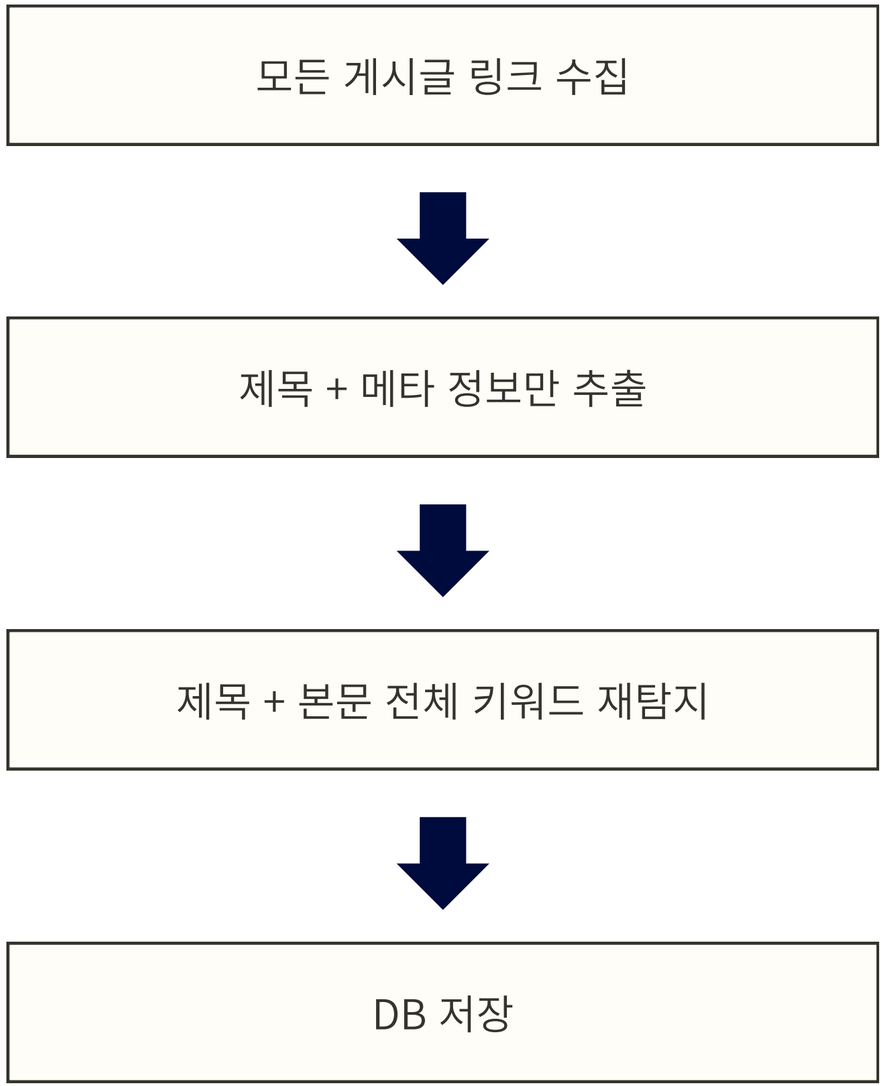
```
REQUIRED_INDICATORS = {
    "personal_db": [
        '판매', '구매', '삽니다', '팝니다', '매입',
        '가격', '문의', '연락', '저렴', '급매', '급구'
    ],
    "account_trade": [
        '판매', '구매', '삽니다', '팝니다', '가격',
        '문의', '연락', '거래', '저렴'
    ],
    "bank_account": [
        '판매', '구매', '삽니다', '팝니다', '매입',
        '가격', '문의', '연락', '거래'
    ],
    "hacking": [
        '의뢰', '대행', '문의', '연락', '가격',
        '서비스', '업체'
    ]
}

# 거래 의도를 나타내는 키워드
TRADE_INDICATORS = ['판매', '구매', '삽니다', '팝니다', '매입']
```

- 2차 필터: 정밀 분석
  - 1차를 통과한 페이지만 전체 본문 상세 분석
  - 키워드 조합, 문맥, 위치까지 종합적으로 고려
  - 단순 키워드 매칭이 아닌 의미 기반 정확한 탐지

02 진행 과정

3주차: 키워드 탐지 알고리즘



프로젝트 성과

03 프로젝트 성과

핵심 시스템 구현

웹 크롤링 시스템 구축

Playwright 기반 동적 크롤링

게시판 자동 순회

도메인 검증 시스템

안정적인 브라우저 관리

데이터 관리 / 텍스트 처리

효율적인 스키마 설계

완벽한 중복 방지

2단계 키워드 탐지

동적 노이즈 제거

03 프로젝트 성과

기술적 성과

성능 최적화

조기 종료 시스템

메모리 효율성

브라우저 리소스 관리

확장 가능한 아키텍처

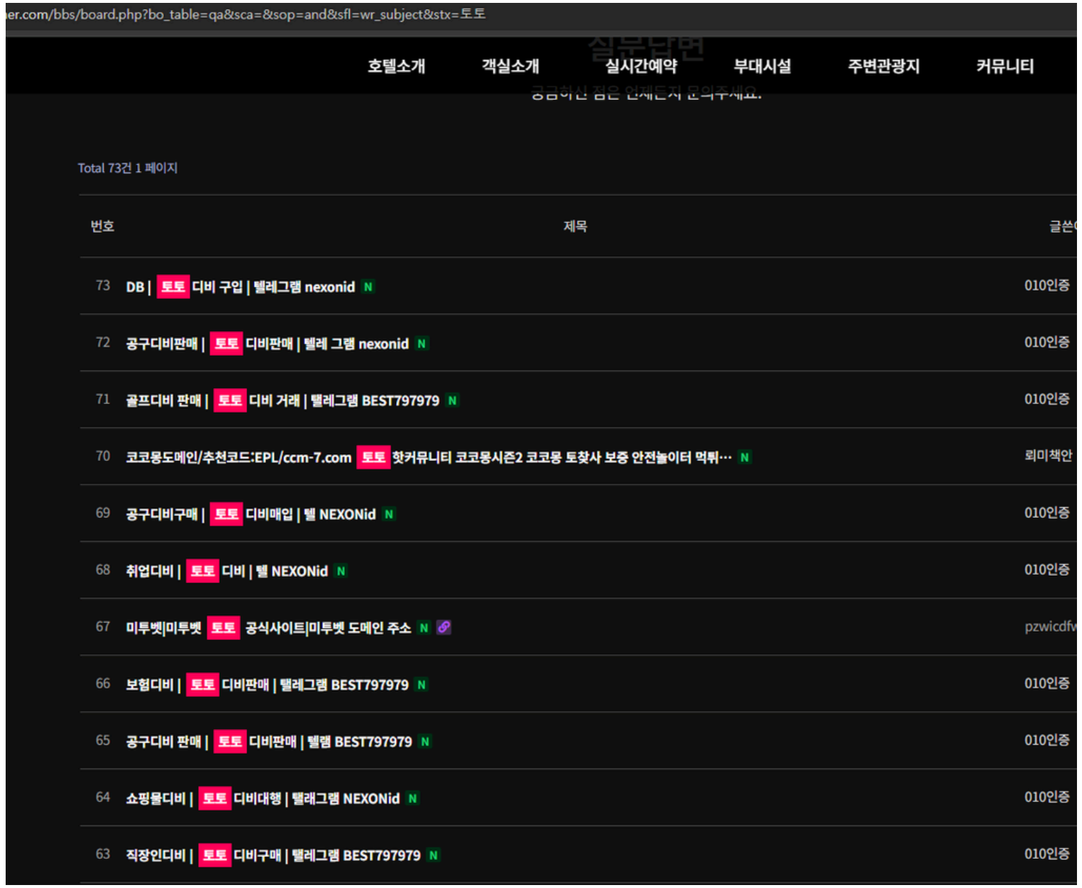
모듈화 설계

설정 기반 제어

플러그인 구조

최종 산출물

04 최종 산출물




```
분석할 게시판 URL을 입력하세요: http://www.hoteldemer.com/bbs/board.php?bo_table=
r_subject&stx=%ED%86%A0%ED%86%A0
최대 분석 페이지 수: 1
총 17개 게시글 크롤링 시작...
```

```
중복 제거 결과:
   총 키워드 탐지 게시글: 8개
   고유 내용 게시글: 4개
   중복 제거된 게시글: 4개
   중복률: 50.0%
```

04 최종 산출물

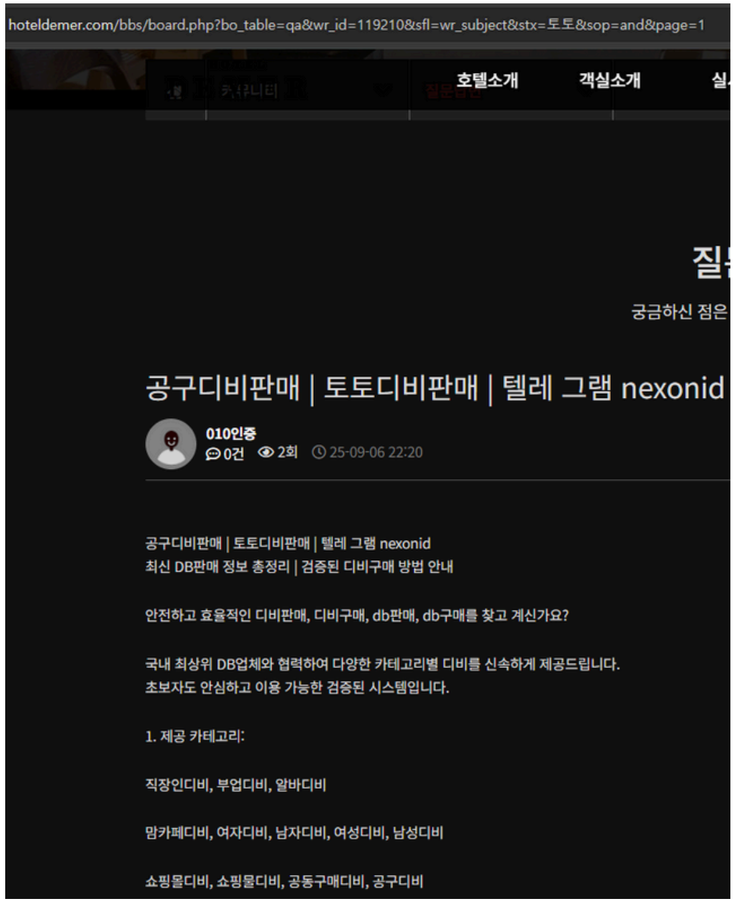
```
[2] 키워드 탐지: http://www.hoteldemer.com/bbs/board.php?bo_table=qa&wr_id=119210&sfl=wr_subject&stx=%ED%86%
   제목: 공구디비판매...
   → 새로운 내용으로 저장 진행
진짜 엑셀 파일(.xlsx)에 저장: results/crawl_results.xlsx (클릭 가능한 하이퍼링크 포함)
새로운 크롤링 결과 저장 완료 (ID: 5) - Excel 파일에 저장됨
URL: http://www.hoteldemer.com/bbs/board.php?bo_table=qa&wr_id=119210&sfl=wr_subject&stx=%ED%86%A0%ED%86%A0&
제목: 공구디비판매
본문: 최신 DB판매 정보 총정리 | 검증된 디비구매 방법 안내
안전하고 효율적인 디비판매, 디비구매, db판매, db구매를 찾고 계신가요?
국내 최상위 DB업체와 협력하여 다양한 카테고리별 디비를 신속하게 제공드립니다.
초보자도 안심하고 이용 가능한 검증된 시스템입니다.
직장인디비, 부업디비, 알바디비
맘카페디비, 여자디비, 남자디비, 여성디비, 남성디비
쇼핑몰디비, 쇼핑몰디비, 공동구매디비, 공구디비
재테크디비, 코인디비, 로또디비, 보험디비
의사디비, 병원디비, 연령대디비
학원디비, 취업디비, 취직디비
유흥디비, 오피디비, 골프디비
2. 이런 분들께 추천드립니다:
타겟 마케팅용 디비가 필요하신 분
소자본으로 재테크와 부업을 시작하실 분
검증된 데이터로 효율을 극대화하고 싶은 분
100% 실시간 업데이트, 검수 완료, 불량율 0% 목표로 빠르고 정확한 데이터를 제공합니다.

[3] 키워드 탐지: http://www.hoteldemer.com/bbs/board.php?bo_table=qa&wr_id=119090&sfl=wr_subject&stx=%ED%86%
   제목: 골프디비 판매...
   → 중복 내용으로 스킵

[4] 키워드 탐지: http://www.hoteldemer.com/bbs/board.php?bo_table=qa&wr_id=118667&sfl=wr_subject&stx=%ED%86%
   제목: 공구디비구매...
   → 중복 내용으로 스킵
```

추출 게시글

중복 게시글



04 최종 산출물

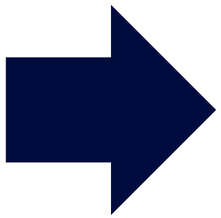
	A	B	C	D	E	F	G
1	ID	Timestamp	URL	Title	Content	Keywords	Hash
2	1	2025-09-06 22:36:57	https://leaderslight.com/article/%EC%83%81%ED%92%88-qa/6/94886/	DB 판매 | 해킹DB구매 | 텔레그램 BEST797979 상품	› 판매 | 해킹DB구매 | 텔레그램 BEST797979 DB 판매 | 해킹DB구매 | 텔레그램 BEST797979 최신 DB 매 정보 총정리 | 검증된 디비구매 방법 안내 안전하고 효율적인 디비판매, 디비구매, db판매, db구매 찾고 계신가요? 국내 최상위 DB업체와 협력하여 다양한 카테고리별 디비를 신속하게 제공드립니다. 보자도 안심하고 이용 가능한 검증된 시스템입니다. 1. 제공 카테고리: 직장인디비, 부업디비, 알바디비 카페디비, 여자디비, 남자디비, 여성디비, 남성디비 쇼핑몰디비, 쇼핑몰디비, 공동구매디비, 공구디비 테크디비, 코인디비, 로또디비, 보험디비 의사디비, 병원디비, 연령대디비 학원디비, 취업디비, 취직디 유흥디비, 오피디비, 골프디비 2. 이런 분들께 추천드립니다: 타겟 마케팅용 디비가 필요하신 분 소자 으로 재테크와 부업을 시작하실 분 검증된 데이터로 효율을 극대화하고 싶은 분 100% 실시간 업데이 검수 완료, 불량율 0% 목표로 빠르고 정확한 데이터를 제공합니다.	[personal_db]: DB구매, DB판매, 해킹DB	56d6cbd41bac
3	2	2025-09-06 22:40:46	http://sjcon.com/bbs/board.php?bo_table=52_2&wr_id=9336&sst=wr_hit&sod=desc&sop=and&page=453	코인DB판매	25-08-21 20:28:05 코인DB판매 | 텔레 그램 _ BEST797979	[personal_db]: DB판매	8dcca251ff60
4	3	2025-09-06 22:43:11	https://example.com/test	DB판매 테스트	것은 DB판매 테스트 내용입니다.	[personal_db]: DB판매	6548b30610a8
5	4	2025-09-06 23:32:32	http://www.hoteldemer.com/bbs/board.php?bo_table=qa&wr_id=119235&sfl=wr_subject&stx=%ED%86%A0%ED%86%A0&sop=and&page=1	- 슈페리어 뷰(산 전망)	신 DB판매 및 디비구매 전문 안내 | 믿을 수 있는 DB업체와 안전한 거래 현재, 다양한 분야에서 맞춤형 비판매와 디비구매에 대한 수요가 급증하고 있습니다. 효과적인 마케팅과 사업 확장을 위해 신뢰도 높 DB 자료를 찾는 분들을 위해 전문 업체와 협력하여 다양한 카테고리의 검증된 디비를 제공합니다. 공 디비, 공동구매디비, 알바디비, 부업디비, 맘카페디비 여자디비, 남자디비, 여성디비, 남성디비, 직장인 비 재테크디비, 코인디비, 로또디비, 보험디비, 취업디비, 취직디비 검디비, 검찰디비, 유흥디비, 오피디 골프디비 의사디비, 병원디비, 학원디비, 연령대디비, 쇼핑몰디비, 쇼핑몰디비 각 분야별 맞춤 필터링 실시간 업데이트를 통해 최신 데이터를 제공하며, 100% 검수 완료된 안전한 데이터만을 취급합니다. 비판매, 디비구매, db판매, db구매를 원하시는 분들께 빠르고 정확한 상담을 약속드립니다. 합법적인 적에 맞는 안전한 거래 환경을 제공합니다. 공구디비판매 | 토토디비판매 | 텔레 그램 nexonid 등록된 글이 없습니다.	[personal_db]: DB판매	fe473f21aa1e
6	5	2025-09-06 23:32:36	http://www.hoteldemer.com/bbs/board.php?bo_table=qa&wr_id=119210&sfl=wr_subject&stx=%ED%86%A0%ED%86%A0&sop=and&page=1	공구디비판매	신 DB판매 정보 총정리 | 검증된 디비구매 방법 안내 안전하고 효율적인 디비판매, 디비구매, db판매, 구매를 찾고 계신가요? 국내 최상위 DB업체와 협력하여 다양한 카테고리별 디비를 신속하게 제공드 니다. 초보자도 안심하고 이용 가능한 검증된 시스템입니다. 직장인디비, 부업디비, 알바디비 맘카페디 여자디비, 남자디비, 여성디비, 남성디비 쇼핑몰디비, 쇼핑몰디비, 공동구매디비, 공구디비 재테크디 코인디비, 로또디비, 보험디비 의사디비, 병원디비, 연령대디비 학원디비, 취업디비, 취직디비 유흥디 오피디비, 골프디비 2. 이런 분들께 추천드립니다: 타겟 마케팅용 디비가 필요하신 분 소자본으로 재 크와 부업을 시작하실 분 검증된 데이터로 효율을 극대화하고 싶은 분 100% 실시간 업데이트, 검수 완 불량율 0% 목표로 빠르고 정확한 데이터를 제공합니다.	[personal_db]: DB구매, DB판매	119763f55ac9

한계점

05 한계점

검증 불가능성 문제

고정 키워드의 한계



오탐

확장 가능성

06 프로젝트 확장 가능성

기술적 성과

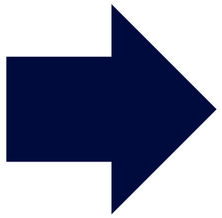
기존 모니터링

URL 수동 추가

실시간 대응 지연

범위 제한

사후 대응



지능형 모니터링

URL 자동 발견

실시간 대응

범위 제한

선제적 대응