불법 · 유해 사이트 차단 프로그램

인터넷 위의 작은 영웅,
당신의 안전한 길잡이 HitAnt

중부대학교 정보보안S/W융합전공 | 정진호 | 박우경 | 최수민 | 홍준희

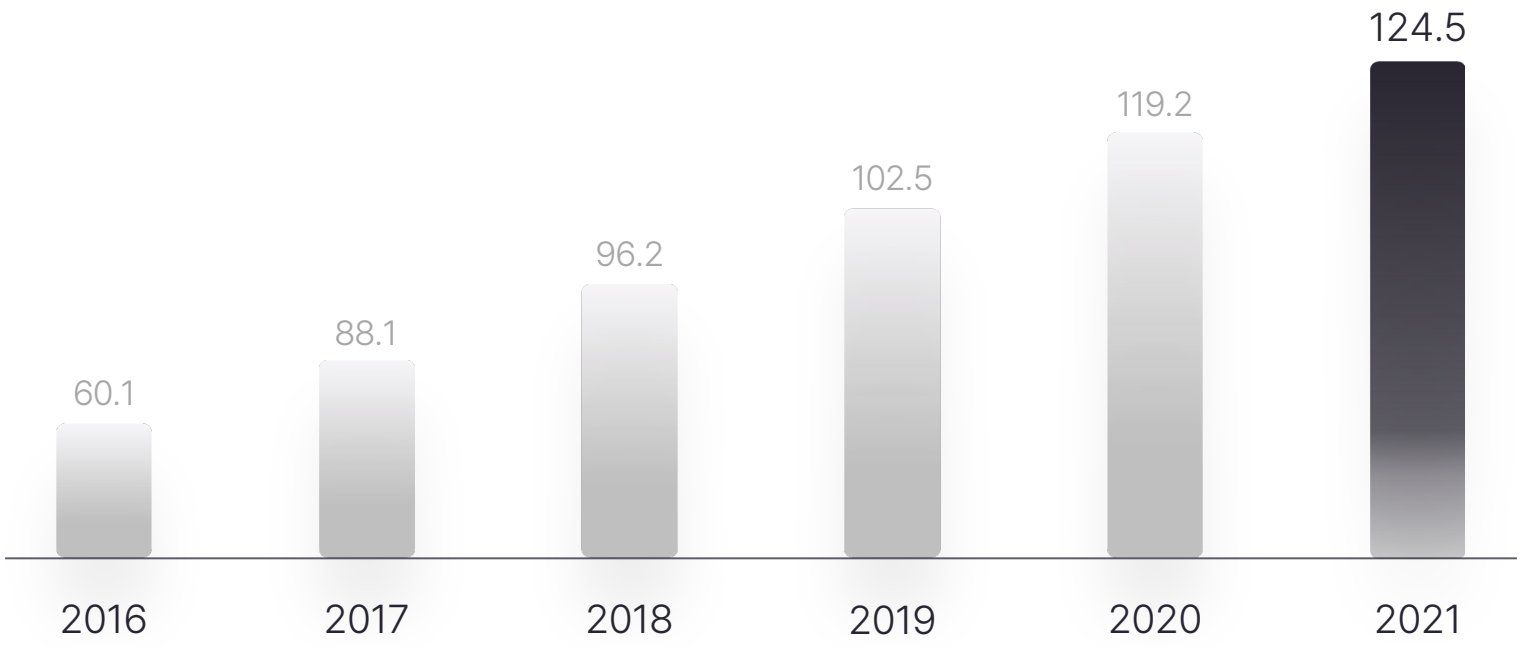
JOONGBU UNIVERSITY

1

프로젝트 선정 이유 및 목표

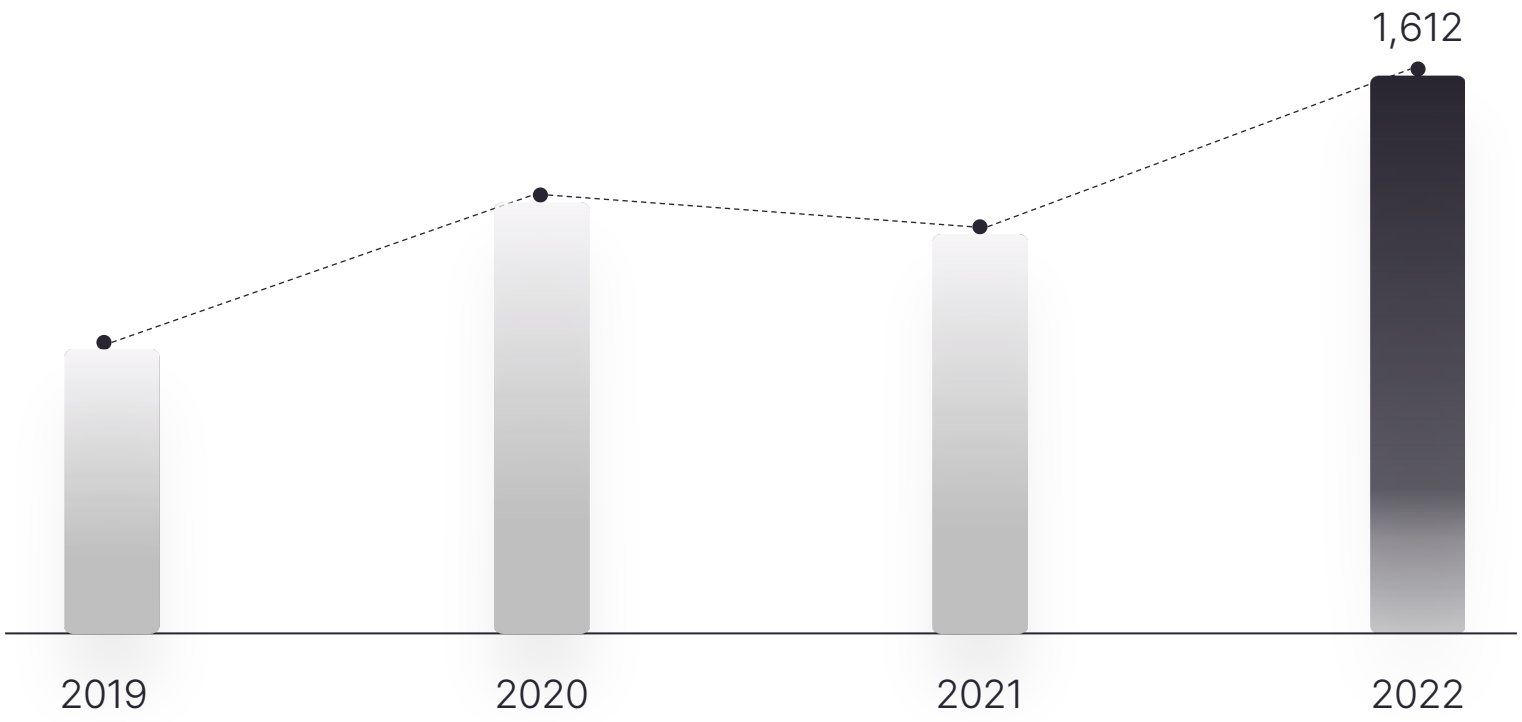
2021년 기준 K-콘텐츠 산업조사

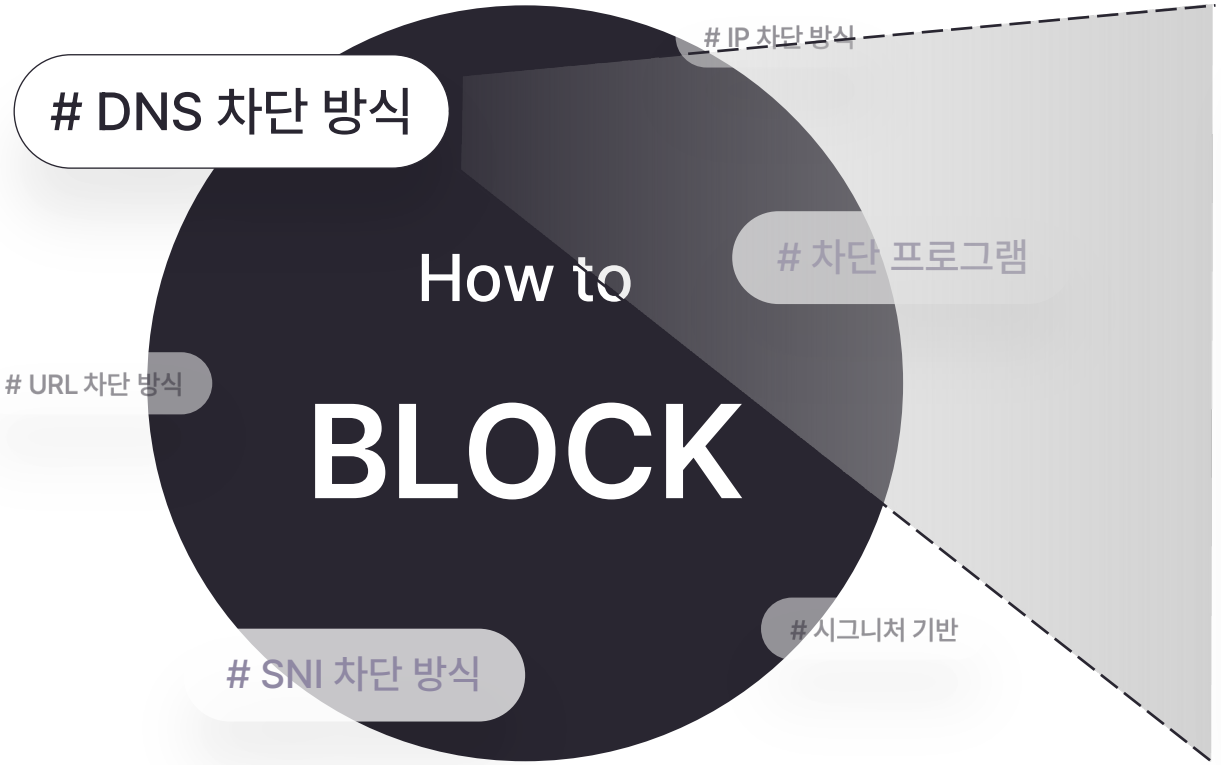
수출액 (단위: 억 달러)

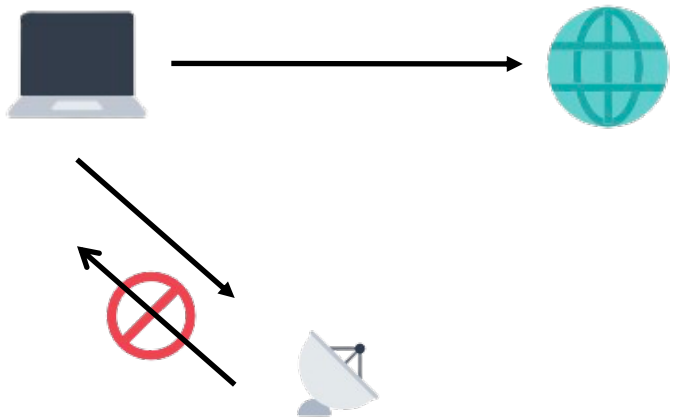


2023년도 불법 복제물 이용량

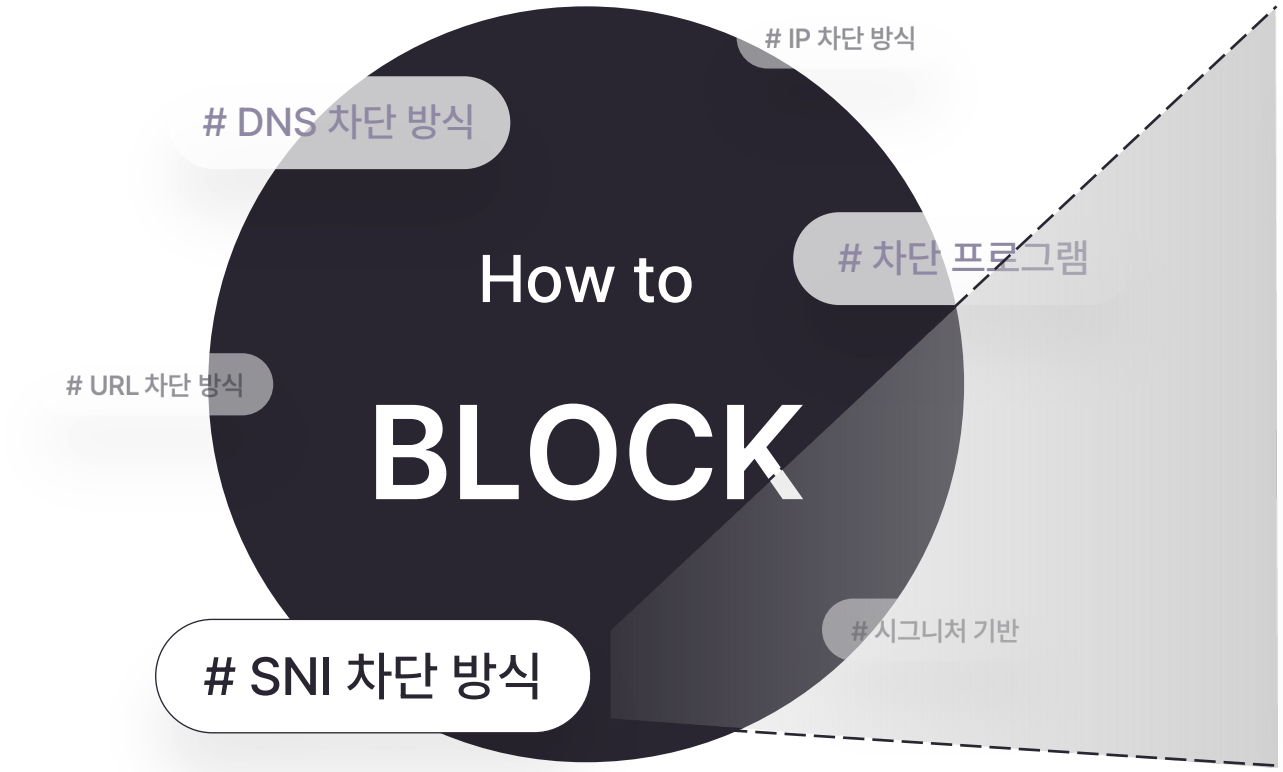
(단위: 천 개)

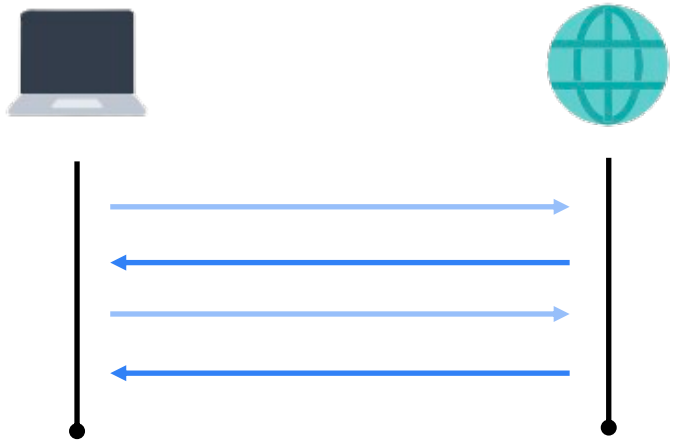






클라이언트와 DNS 서버 간 질의 응답 과정에서 특정 블랙 리스트 호스트 네임인 경우 차단 사이트로 이동





암호화 통신 이전(HTTPS) 평문 통신 단계에서 접속
하고자 하는 호스트 정보를 파악해 차단 사이트로 이동


IP 차단 방식
DNS 차단 방식
How to
차단 프로그램
BLOCK
URL 차단 방
시그니처 기반
SNI 차단 방식
PC보안케어
K9 Web Protection
from BlueCoat
맘아이momi
엑스키퍼 LITE
안심
i-NOON 아이눈
사이버가디언

자살·도박 정보 판치는데...
불법사이트 차단 규정 부족
2023. 10. 03.

청소년 27% 이상 불법 유해
사이트 접속

설날연휴특선영화 '서울의 봄'
누누티비로 볼 수 있나
2024. 02. 09.

K-웹툰 절반 침해... 해외 불법
사이트와 질긴 숨바꼭질
2024. 05. 01.

불법사이트 홍보하려 경복궁
낙서 테러...
2024. 05. 31.

누누티비 방지법 국회 본회의
통과... "접속차단 의무화"
2023. 12. 21.

스위트 홈 시청이 무료?
누누티비 시즌 2 버젓이 활개
2023. 12. 12.

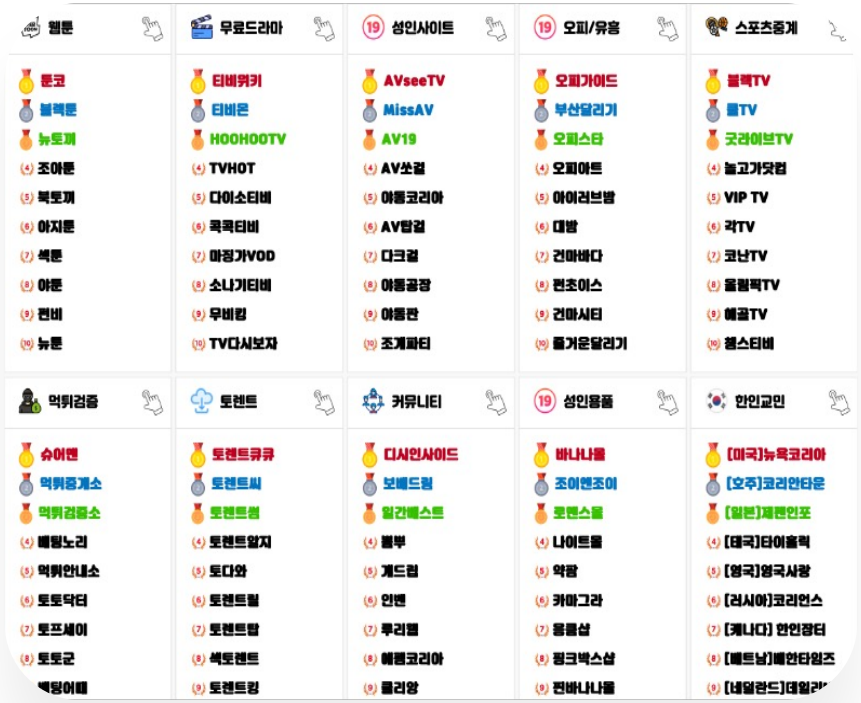
규제의 필요성
차단에 대한 규정 부족 문제

접속차단 의무화
그러나, 현실적인 문제 해결 X

K-콘텐츠의 위협
불법 · 유해 사이트를 통한 유포

1

링크 모음 사이트 및 연관 링크 사이트



2

HTML의 a Tag 속 링크 크롤링

3

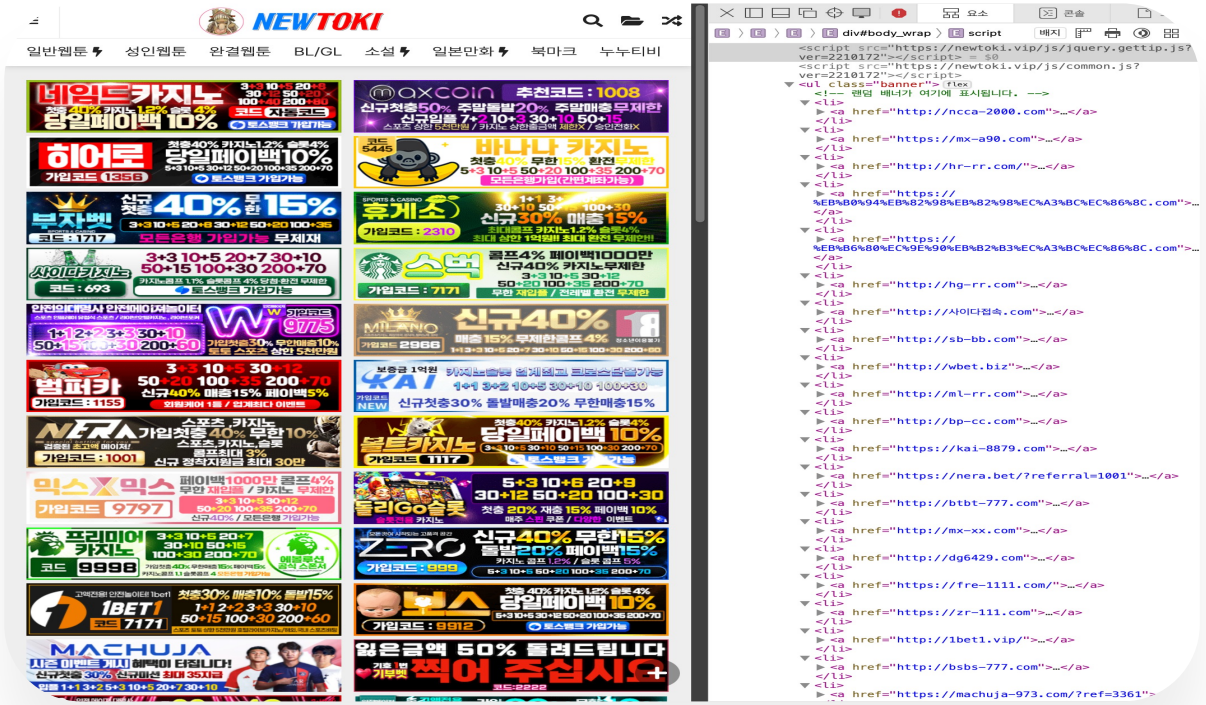
수집한 데이터 셋 정보

1

링크 모음 사이트 및 연관 링크 사이트

2

HTML의 a Tag 속 링크 크롤링



3

수집한 데이터 셋 정보

1 링크 모음 사이트 및 연관 링크 사이트

2 HTML의 a Tag 속 링크 크롤링

3 수집한 데이터 셋 정보

비유해 사이트

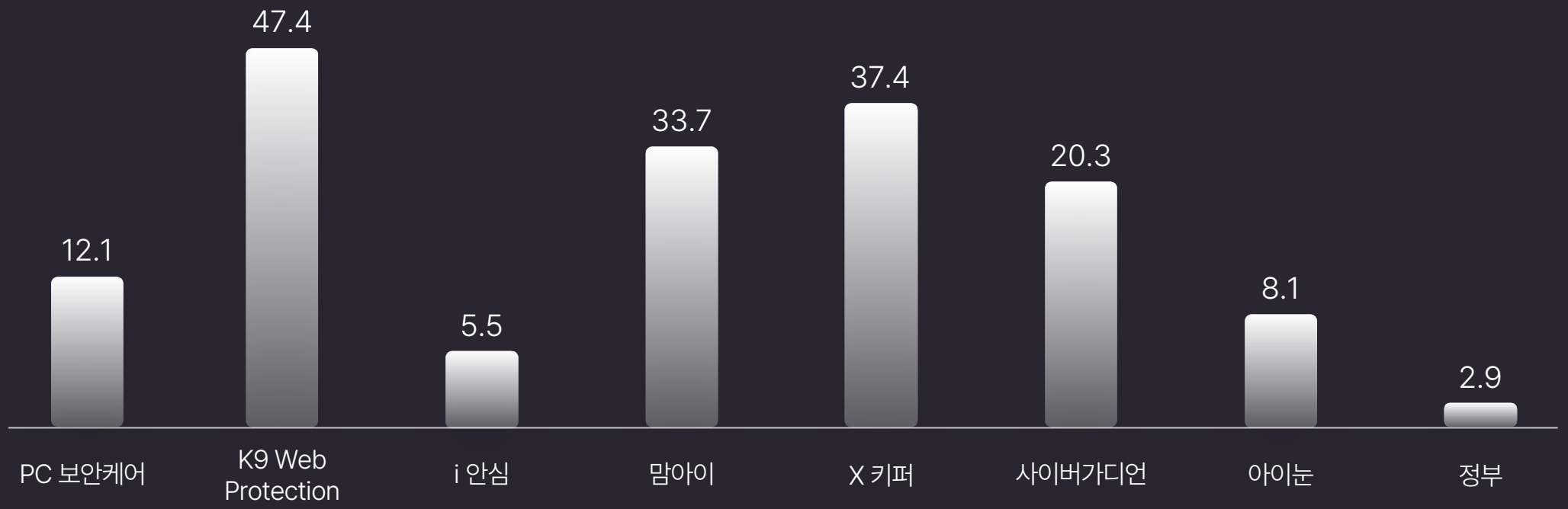
	URL
2	https://www.naver.com
3	https://www.daum.net
4	https://www.google.com
5	https://comic.naver.com/index
6	https://webtoon.kakao.com
7	https://www.netflix.com/kr/
8	https://page.kakao.com
9	https://www.lezhin.com/ko
10	https://www.toomics.com
11	https://www.mrblue.com
12	https://bufftoon.plaync.com
13	https://toptoon.com
14	https://www.myktoon.com/
15	https://www.tving.com/
16	https://www.coupangplay.com/
17	https://www.wavve.com
18	https://www.disneyplus.com/ko-kr
19	https://watcha.com
20	https://ko-kr.facebook.com
21	https://www.instagram.com
22	https://www.threads.net/
23	https://www.youtube.com/
24	https://www.tiktok.com/ko-KR/
25	https://section.blog.naver.com/
26	https://www.tistory.com
27	https://www.notion.so/
28	https://www.pinterest.co.kr
29	https://twitter.com/
30	https://www.afreecatv.com
31	https://chzzk.naver.com
32	https://news.naver.com
33	https://www.coupang.com
34	https://www.joongbu.ac.kr
35	https://everytime.kr
36	https://wrtn.ai
37	https://chat.openai.com
38	https://www.hancomdocs.com/
39	https://top.cafe.daum.net
40	https://cafe.naver.com/
41	https://finance.naver.com
42	https://search.naver.com/
43	https://papago.naver.com/
44	https://edu.joongbu.ac.kr
45	https://laftel.net
46	https://www.mrblue.com
47	https://page.kakao.com/
	https://www.yna.co.kr

유해 사이트

	URL
2	http://1p-888.com/
3	http://블링주소.com/
4	http://이프로벳.com/
5	http://인디뱃주소03.com/
6	http://b-time511.com/
7	http://bbox1212.com
8	http://bn2222.com/#/
9	http://bp-cc.com
10	http://bwd07.com/
11	http://bz-0101.com/
12	http://cg-mvp10.com/
13	http://dgdg-777.com/front
14	http://fst-234.com/
15	http://fsw-333.com
16	http://fsw-555.com/
17	http://fw-0001.com/
18	http://gcity-222.com/
19	http://gqb-333.com/
20	http://gn1020.com
21	http://gws42.com/
22	http://h-two41.com/
23	http://jg-5555.com/login.asp
24	http://ka-01.com/login
25	http://lg-1010.com
26	http://lg0011.com
27	http://mmcc234.com/
28	http://mz1-one.com/
29	http://pgss1122.com
30	http://rk9284.com/
31	http://site-99.com/
32	http://spo-one.com/
33	http://spst-1111.com
34	http://sun-9909.com/?regcode=4557
35	http://we-324fg.com/
36	http://woori-333.com/main
37	http://www.fu2024.com/#/
38	http://www.gbqx93.com/#/
39	http://www.mxr-36.com/#/
40	http://www.pnt357.com/front
41	http://www.snxz91.com/#/
42	http://www.solslotgm111.com/#/
43	http://www.xvs-49.com/#/
44	https://17b-00.com/
	https://365kor.bet/
	tps://888-sm.com/?ref=4545

현 차단 프로그램의 차단율

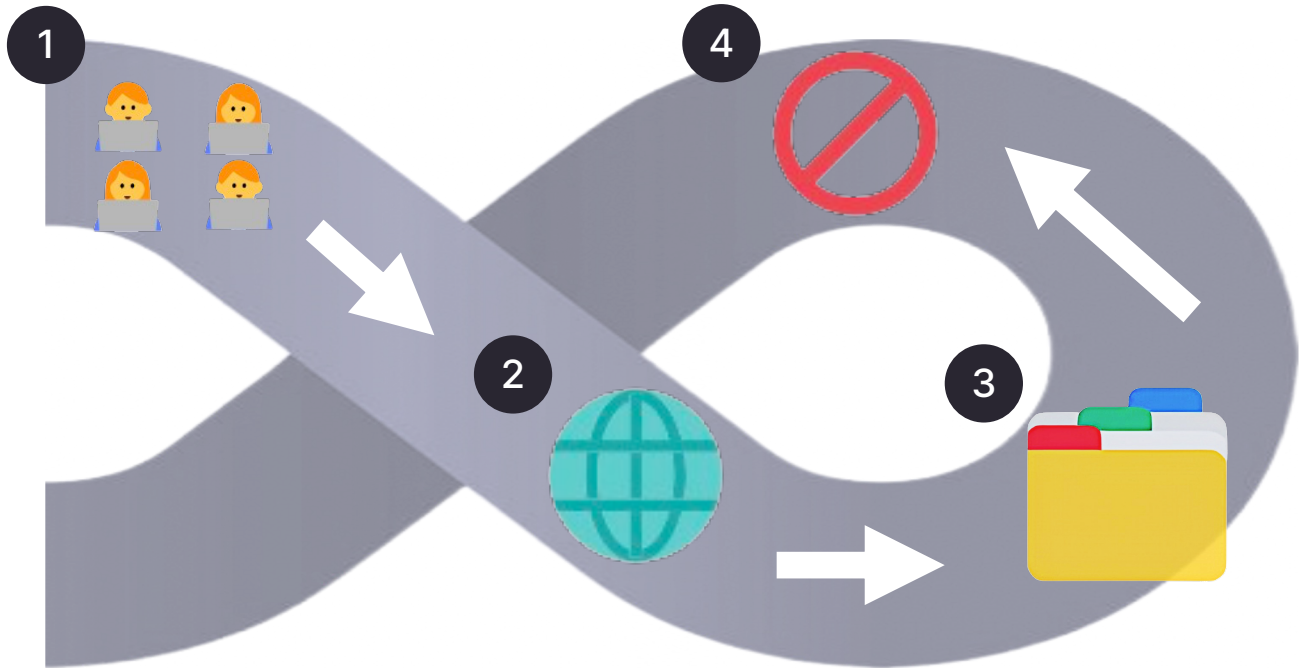
(단위: %)



데이터 셋 크기: 19,479개

시그니처 방식

수동적 차단 방식



1 사용자

2 인터넷 접속

3 불법 · 유해 데이터베이스

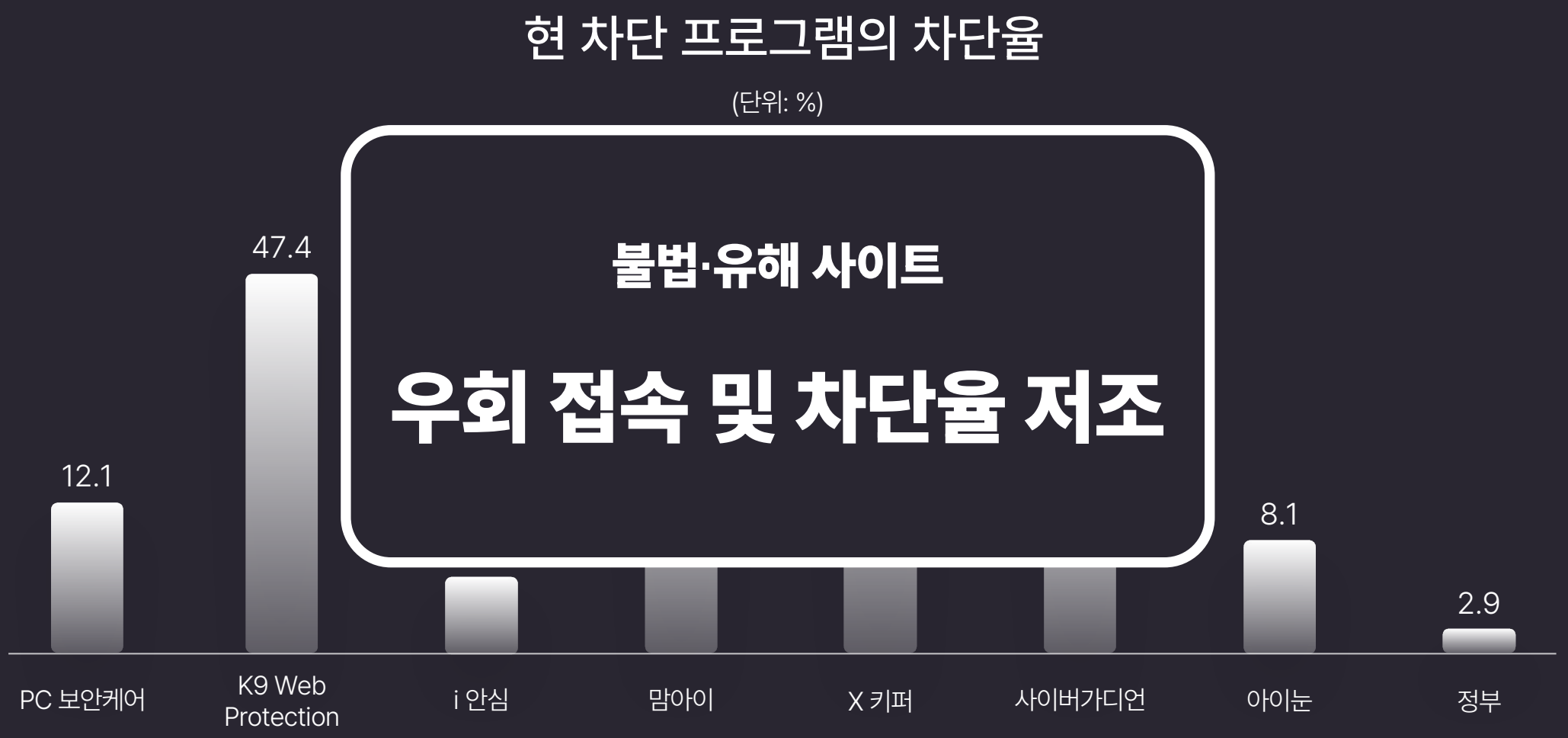
4 차단 유무 판정

주기적인

데이터 베이스 업데이트 필수

불법 · 유해 사이트

도메인 주소 변경 시 차단 회피



데이터 셋 크기: 19,479개

프로젝트 목표

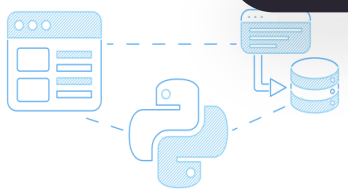
과정

분류	크롤링	머신러닝	최종
성인, 도박, 웹툰, 스트리밍	Requests 방식 Selenium 방식	능동적인 차단 신속한 대처	차단율 97% 이상

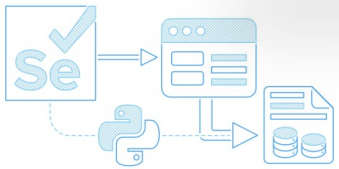
2

프로젝트 소개 및 진행 과정

requests



selenium



>

HTML

HTML 문서의 **전체 범위**

<!DOCTYPE html> HTML5 문서

<html lang="en"> ▶ lang : 어떤 언어로 작성되었는지 나타냄 (영어 en / 한국어 ko)

<head>

<meta charset="UTF-8"> ▶ 문자 인코딩 정보

<meta http-equiv="X-UA-Compatible" content="IE=edge"> ▶ 인터넷 익스플로러그 호환성 관련 태그

<meta name="viewport" content="width=device-width, initial-scale=1.0"> ▶ 뷰포트 정보

<title>Document</title> ▶ 문서의 제목 (브라우저 탭에 나타남)

</head>

문서의 **정보**를 나타내는 범위

<body>

문서의 **구조**를 나타내는 범위

</body>

</html>

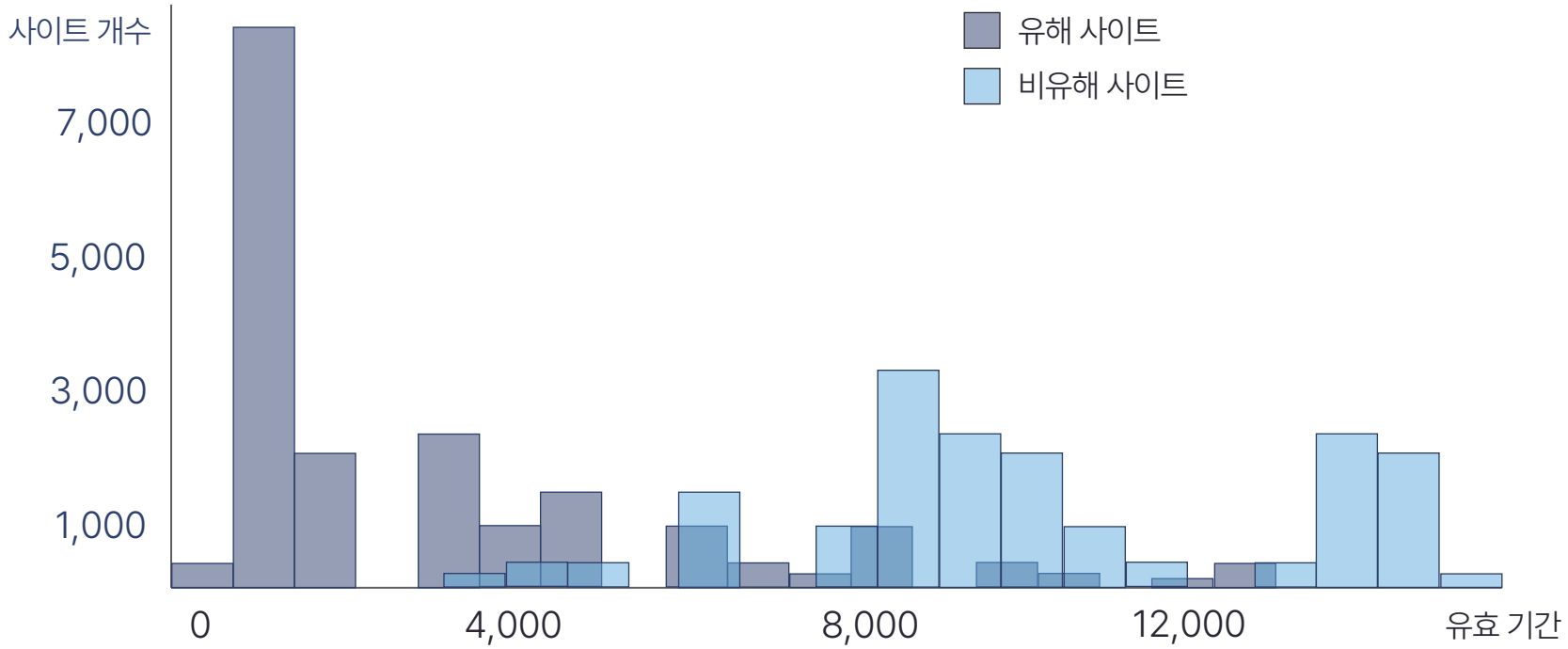
피쳐 선정

24가지

연결 상태	로딩 시간	주소 길이	생성 기간	최종 주소	리다이렉트
HTTPS	Meta	A	IP	Title	Div
Popup Count	Js_Len	Js_density	Link	Country	Label
City	Img	Img_density	Tld	SSL	Registrar

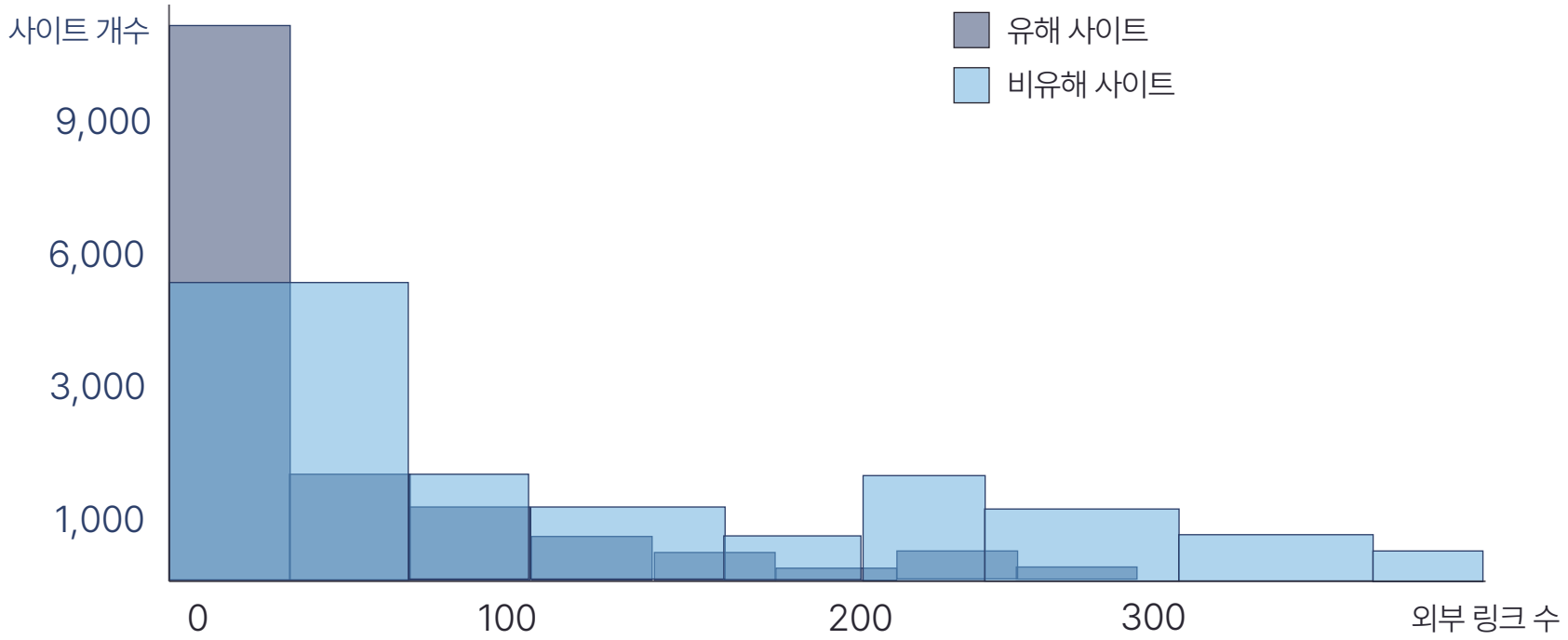
유효 기간 분포 비교

(단위: 개)



외부 링크 수 분포 비교

(단위: 개)



IP	Js_Len	외부 링크 수	도메인 생성일	도메인 마감일	Div 수
213.166.69.26	7391	0	2022-02-13 15:58:13	2025-02-13 15:58:13	25
104.21.17.249	2074	83	2023-08-30 10:09:23	2024-08-30 10:09:23	14
167.114.35.174	9789	5	2021-07-12 16:05:39	2024-07-12 16:05:39	118
104.21.59.9	5108	57	2023-08-21 00:10:21	2024-08-21 00:10:21	27
104.21.61.222	22634	18	2005-02-23 17:07:11	2025-02-23 05:00:00	617
172.67.172.38	79591	11	2002-04-19 22:42:04	2027-04-19 22:42:04	1135
172.67.165.12	76516	11	2020-10-30 09:36:57	2025-10-30 09:36:57	1138
172.67.163.235	5135	57	2023-08-21 00:10:08	2024-08-21 00:10:08	27
104.21.65.109	5010	59	2023-08-21 00:10:34	2024-08-21 00:10:34	30
104.21.19.16	1162	2	2024-01-19 04:24:53	2025-01-19 04:24:53	333

팝업 수	이미지 수	로딩 시간	리다이렉트	Js_density	Img_density
0	7	0.00297904	0	0.103187346	0.28
0	4	0.002000093	0	0.036336878	0.285714286
1	21	0.003004789	0	0.101218049	0.177966102
0	2	0.002022743	0	0.088093267	0.074074074
0	25	0.001998186	0	0.157867939	0.040518639
0	229	0.002015591	0	0.099017181	0.201762115
0	248	0.001999855	0	0.094648587	0.217926186
0	2	0.001999617	0	0.088368411	0.074074074
0	2	0.00199914	0	0.068043841	0.066666667
1	47	0.002000332	0	0.017107859	0.141141141

Numeric

전처리

Int

Float

∨

별도의 전처리 없이
머신러닝 학습 데이터로

바로 이용 가능

연결 상태	Meta	A	Title	Tld	Registrar
False	[ ]	[ ]		ca	0
True	['Ibp8V0h90drtN...']	['#site-navigation'...]	Wildblaster Casino	com	NAMECHEAP INC
True	['X_uHNqglElAWnl...']	['https://adappgeo'...]	M11HKB : Raja...	net	NAMECHEAP INC
True	['width=device-wid...']	['#content', '#'...]	A Beginner's Guide t...	com	Dynadot Inc
True	['3p0Q0q3gFDaG'...]	['https://basahter'...]	PAPASLOT : Link...	org	NameCheap, Inc.
True	['width=device-wid...']	['#fl-main-content'...]	Home – Canadian...	ca	GoDaddy Domains Canada, Inc
True	['text/html; charset...']	['https://casinome'...]	Casinos Online...	com	Key-Systems GmbH
True	['width=device-wid...']	['https://casinome'...]	Casinos Online...	com	Key-Systems GmbH
True	['text/html; charset...']	['https://casinosc'...]	Online Casino...	com	NAMECHEAP INC
True	['Ps9Z0Mp9ayNFm...']	['https://basahter...']	AGEN123 : Link Raja...	org	NAMECHEAP INC

TF-IDF / One-Hot 인코딩

String

Konolpy | Kss

∨

문자열 속 문장 중 단어만 추출,
단어를 숫자형으로 변환


머신러닝
NumPy
XGBoost
TensorFlow
Random Forest
matplotlib
PyTorch
learn
Keras
LightGBM
pandas

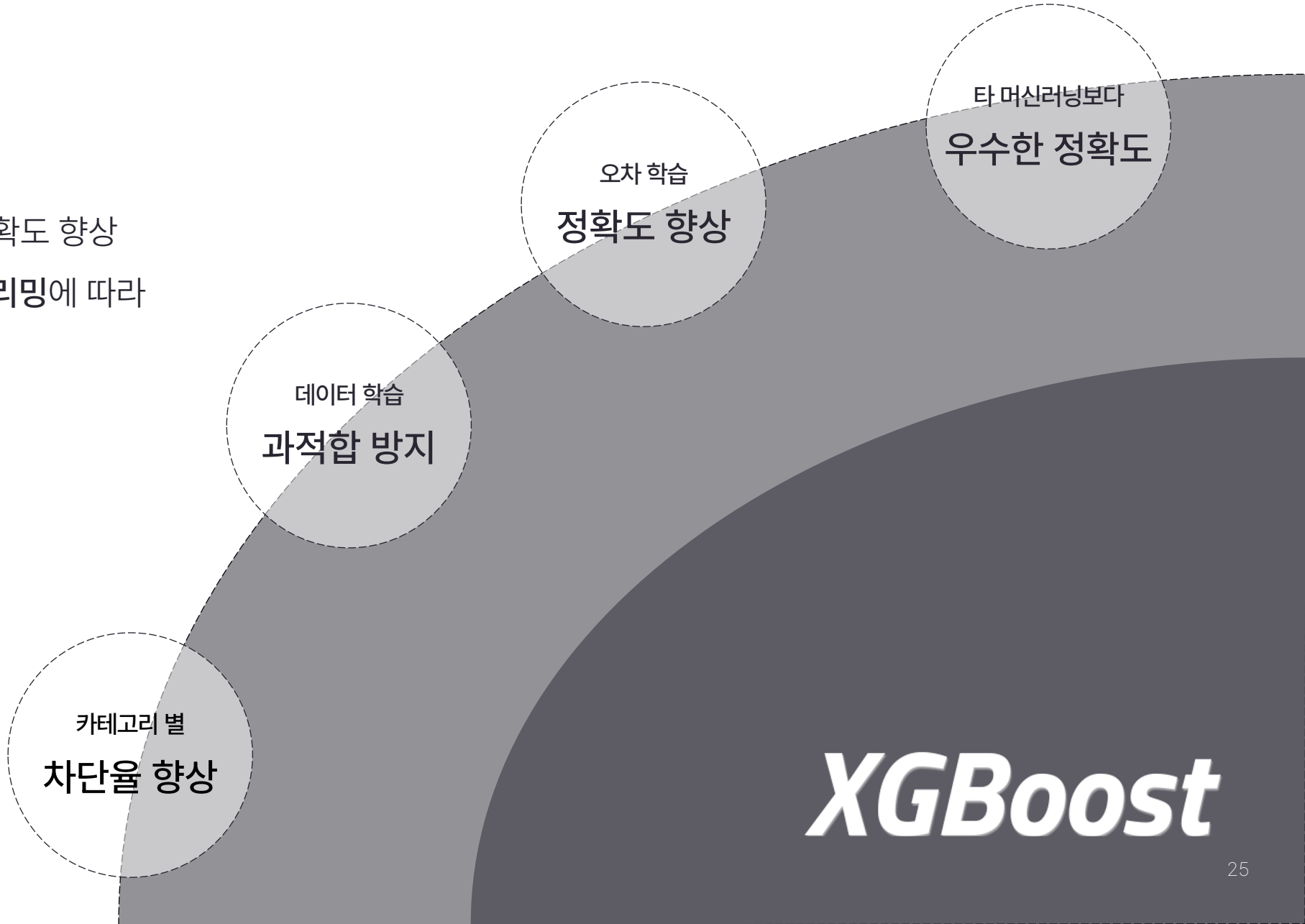
머신러닝, XGBoost

데이터 학습 과정에서 과적합 방지

오차 데이터만 반복 학습을 통해 정확도 향상

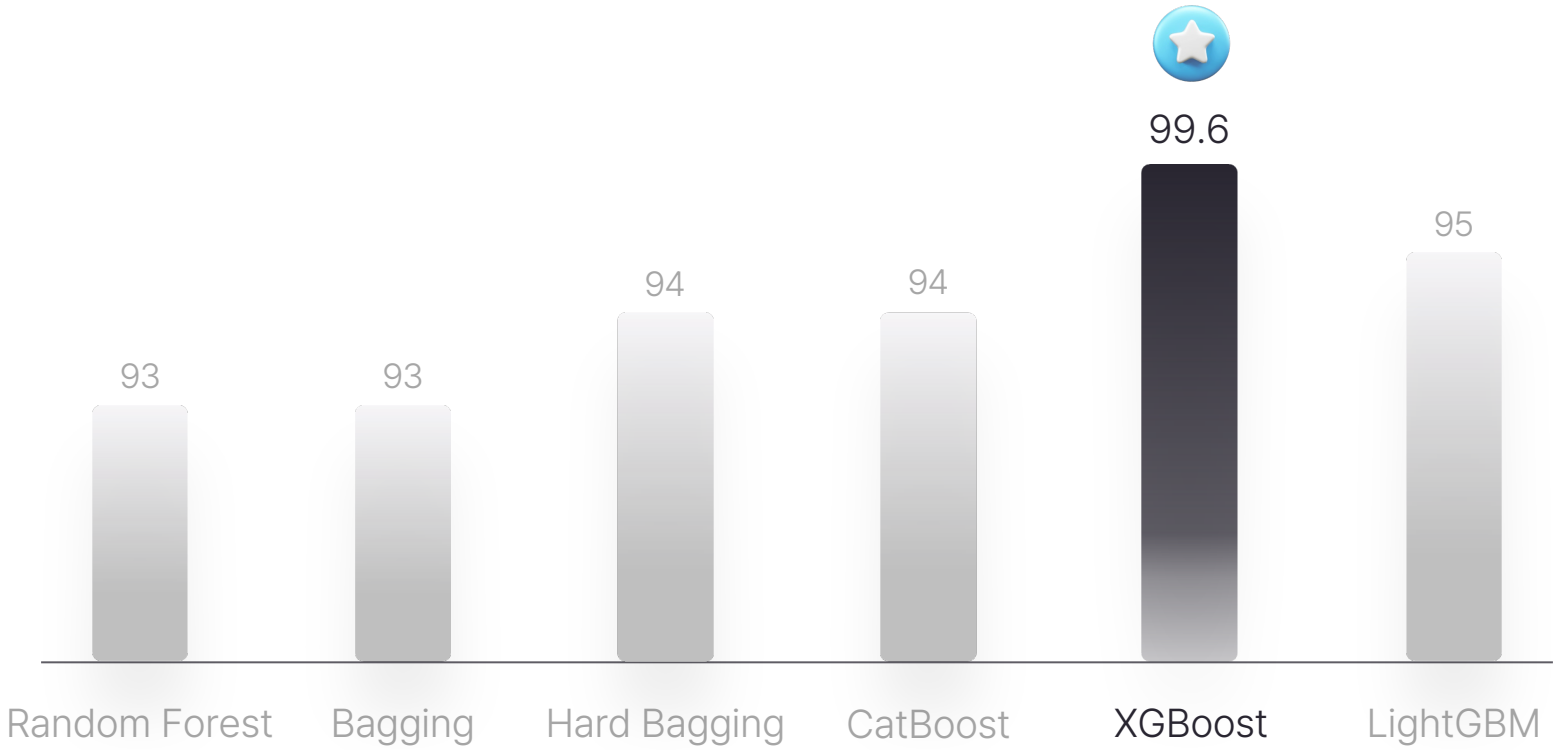
각 카테고리 **성인, 도박, 웹툰, 스트리밍**에 따라

우수한 정확도 제공

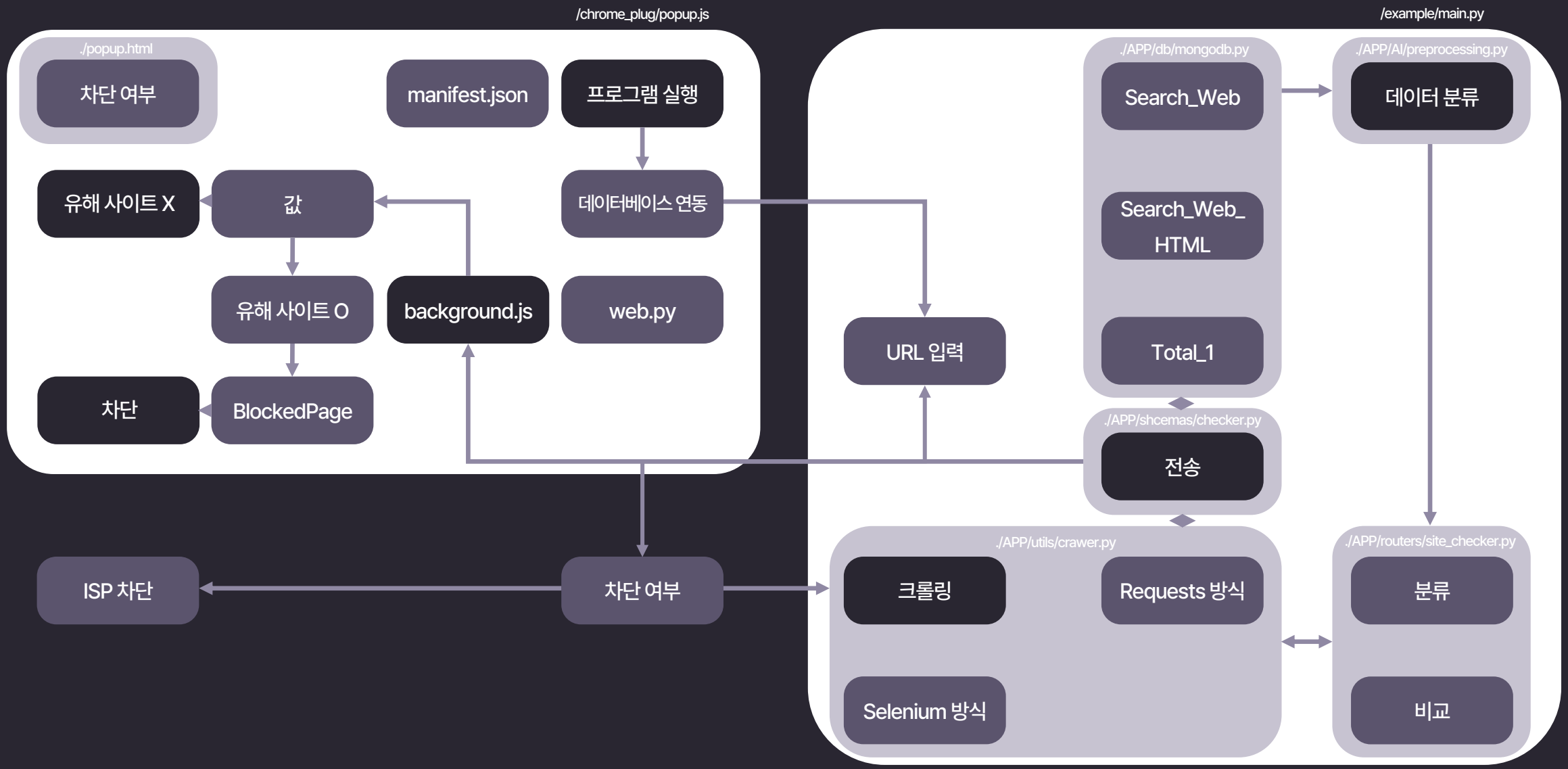


머신러닝 학습 정확도

(단위: %)

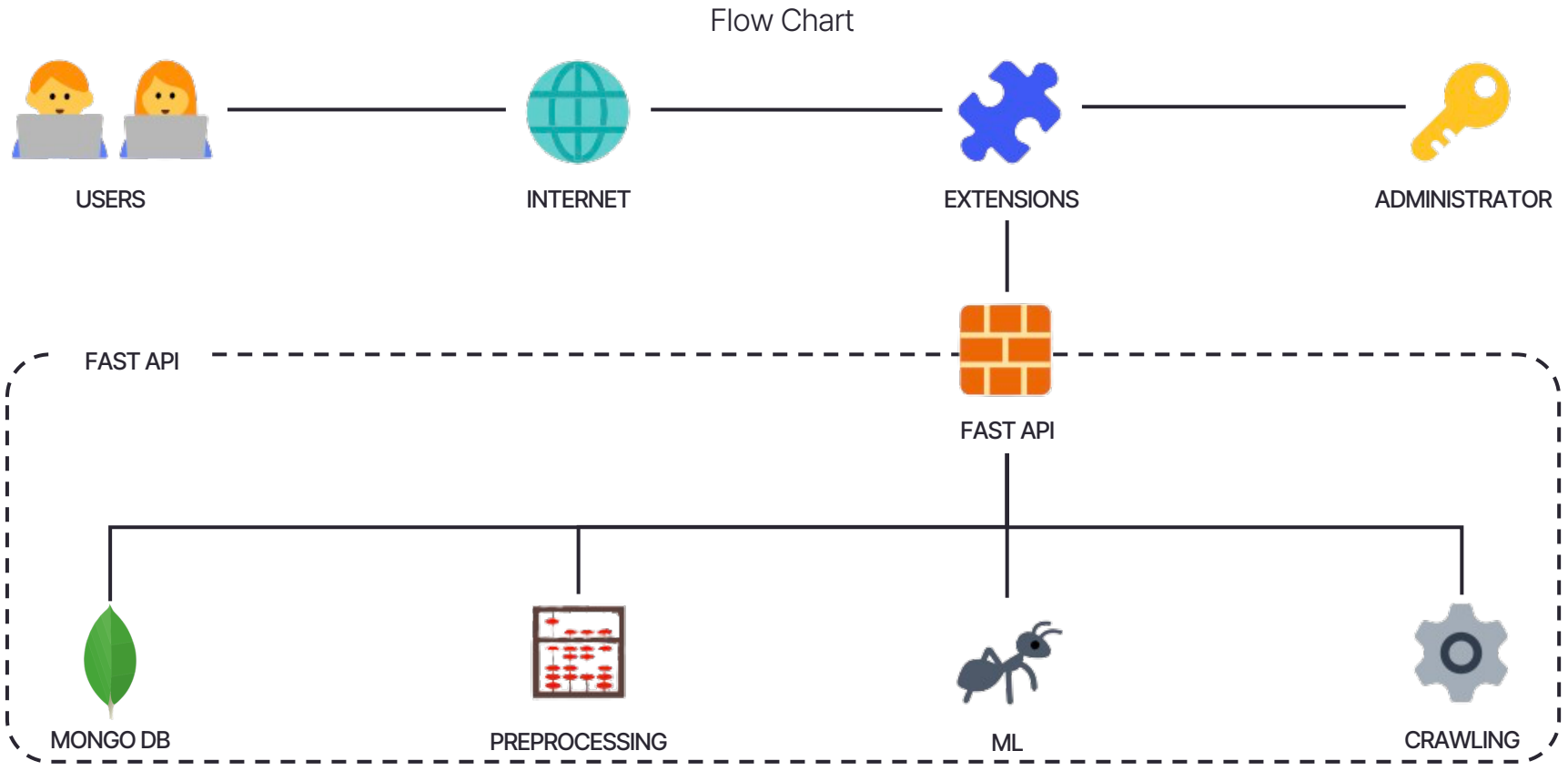


데이터 셋 크기: 19,479개


/chrome_plug/popup.js
./popup.html
차단 여부
manifest.json
프로그램 실행
유해 사이트 X
값
데이터베이스 연동
유해 사이트 O
background.js
web.py
차단
BlockedPage
ISP 차단
차단 여부
URL 입력
/example/main.py
./APP/db/mongodb.py
Search_Web
Search_Web_HTML
Total_1
./APP/AI/preprocessing.py
데이터 분류
./APP/shcemas/checker.py
전송
./APP/utils/crawer.py
크롤링
Requests 방식
Selenium 방식
./APP/routers/site_checker.py
분류
비교

불법 · 유해 사이트 차단 프로그램

Flow Chart



화이트 / 블랙 리스트 구축

TLD 검증

.ORG .COM .INFO .CO .NET .CO.UK

>

HitAnt 머신러닝 검증

>

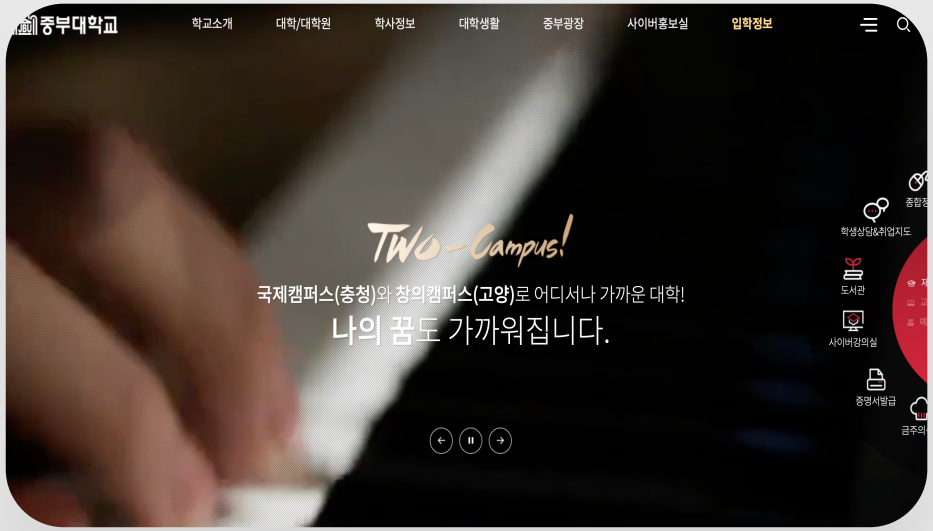
차단 유무 판정

불법 · 유해 사이트 판정

차단 페이지로 리다이렉트

비유해 사이트 판정

정상적으로 이용 가능



3

프로그램 검증 및 배포

현 차단 프로그램의 차단율

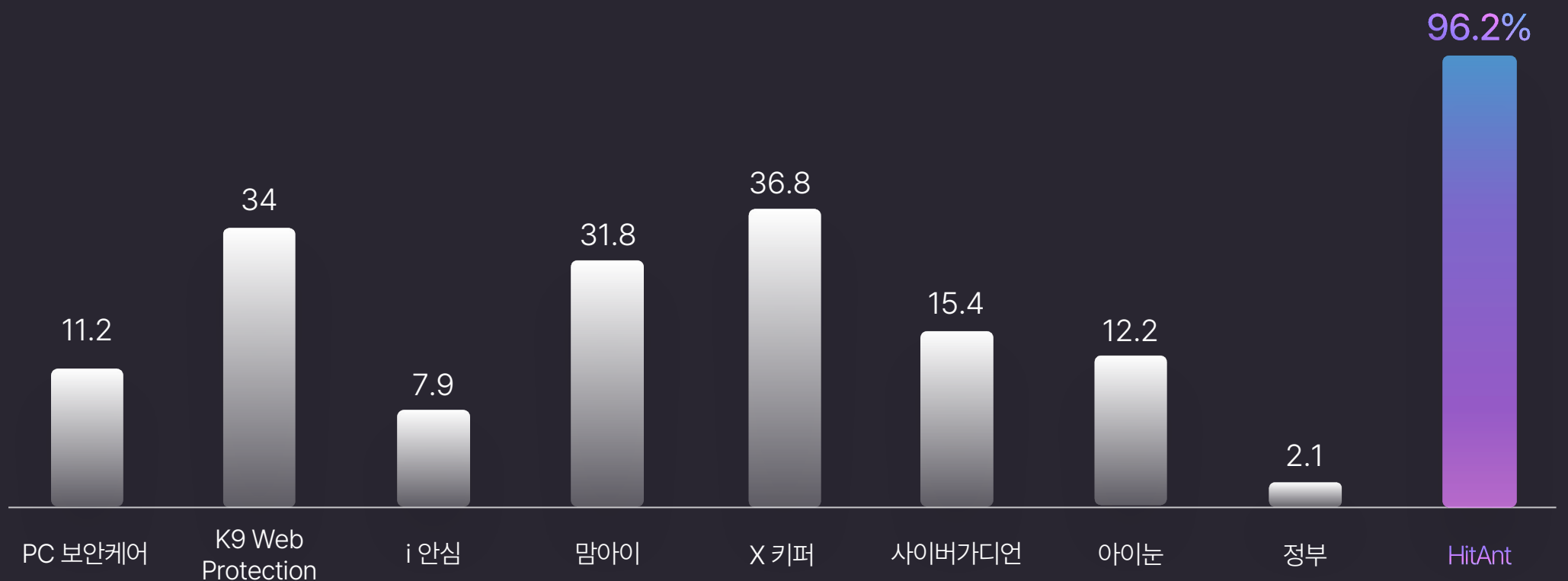
(단위: %)

데이터 셋 크기: 19,479개

JOONGBU UNIVERSITY

현 차단 프로그램의 차단율

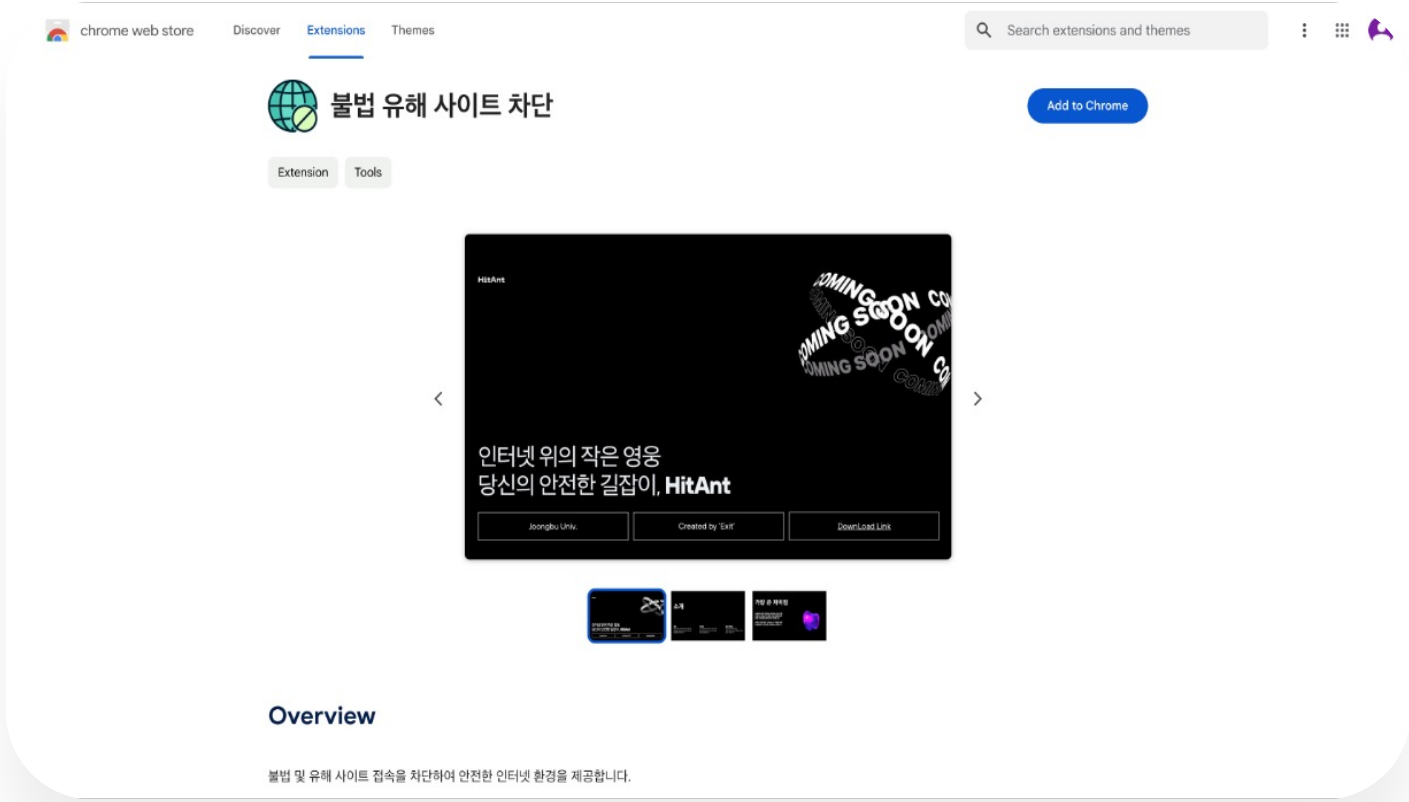
KISA 데이터 자료 (단위: %)

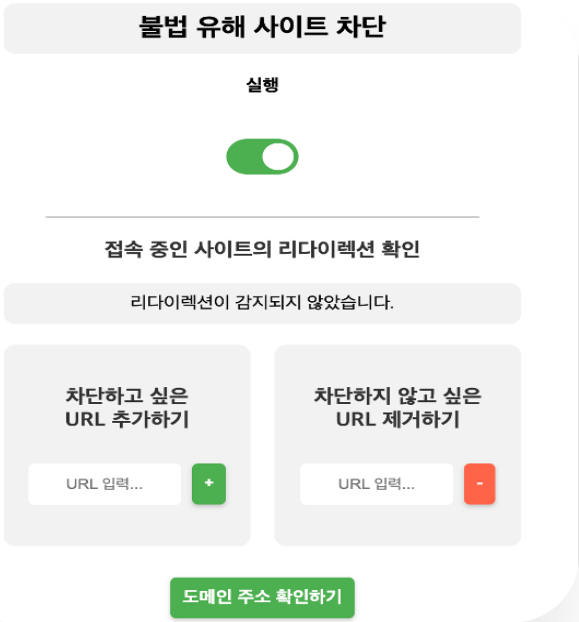


데이터 셋 크기: 1,276개

불법 · 유해 사이트 차단 프로그램 배포

크롬 확장자 플러러그인 프로그램

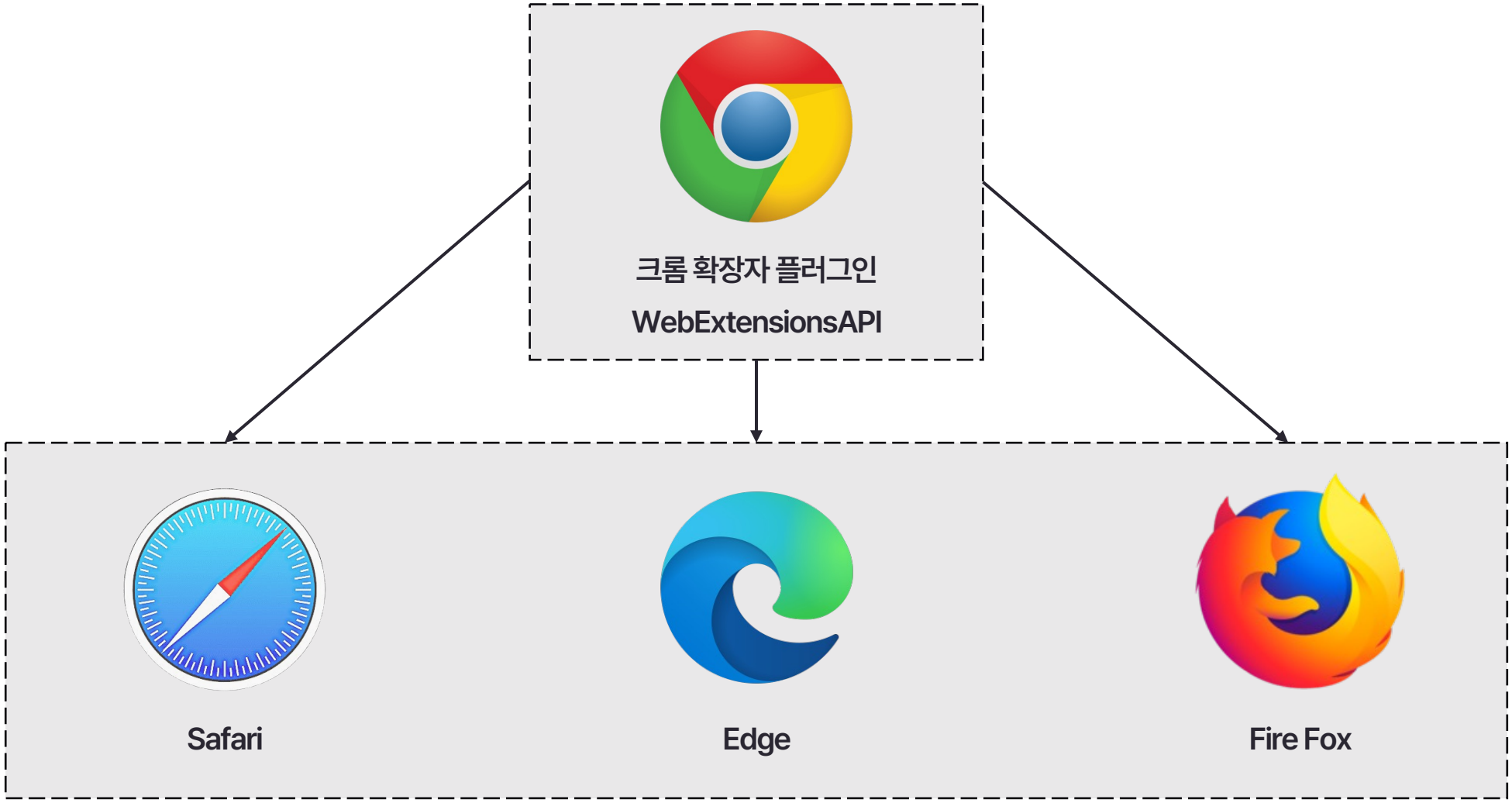


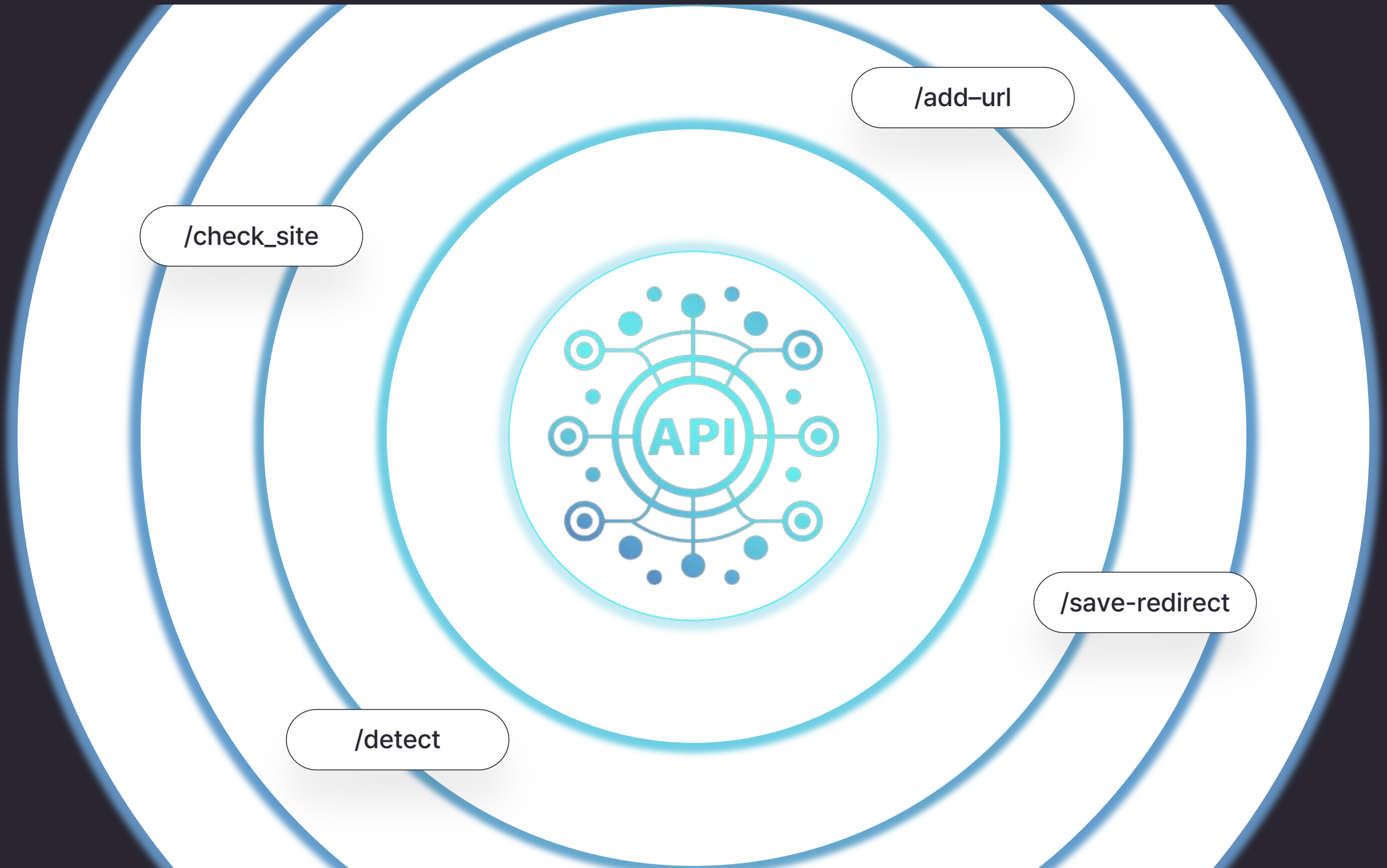


🔗 크롬 확장자 스토어 링크

4

마무리


크롬 확장자 플러그인
WebExtensionsAPI
Safari
Edge
Fire Fox


/add-url
/check_site
API
/save-redirect
/detect

/add-url

데이터베이스 추가

차단하고 싶은 URL 추가하기

URL 입력...

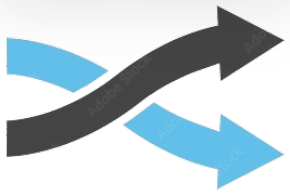
차단하지 않고 싶은 URL 제거하기

URL 입력...

유해

비유해

/save_redirect



REDIRECT

리다이렉션 확인 [전]

접속 중인 사이트의 리다이렉션 확인

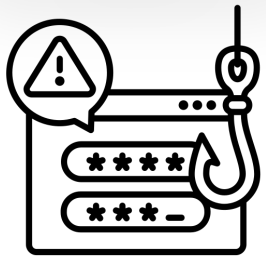
리다이렉션이 감지되지 않았습니다.

리다이렉션 확인 [후]

접속 중인 사이트의 리다이렉션 확인

URL: https://blacktoon.me/ -> https://blacktoon310.com/ (상태 코드: 301)

/detect



피싱 사이트 탐지 [전]

피싱 사이트 탐지

현재 안전한 사이트입니다!

피싱 사이트 탐지 [후]

피싱 사이트 탐지

현재 피싱 사이트에 접속하셨습니다!

구글 API 활용한 피싱 탐지

Google Cloud Platform

정규 표현식을 이용한 피싱 URL 탐지

/Reg[ex]+/

머신러닝 비교 전 실행

/check_site

SAFE

DANGEROUS

도메인 정규 표현식 검증

/Reg[ex]+/

화이트 / 블랙 리스트 검증

TLD 검증

.ORG .COM .INFO

.CO .NET .CO.UK

HitAnt 머신러닝 검증

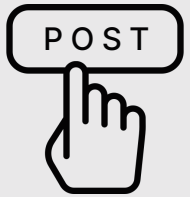
검증 결과 형 변환

TRUE → 0

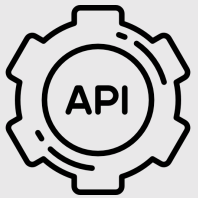
FALSE → 1

1

외부 검증 요청



Check_site API 유해・비유해 판별



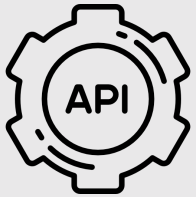
차단 유무 결과 전송 및
해당 결과의 가중치 부여

차단 시스템 정확도 향상에 도움

2

검증이 필요한 데이터 자료 제공

Check_site API 유해・비유해 판별



판별 데이터 전송 및
데이터베이스 업데이트

실시간 모니터링

감사합니다.

JOONGBU UNIVERSITY