

TEAM: 000
MALWARE DETECT BY AI
AI SEC

2025.07.10

Mission.1

**선수조사
자료수집**

Mission.2

**선수조사
자료분석**

Mission1

선수조사
자료수집

자료를 어떻게 수집해야하는가?

자료수집의 MAIN ROUTE

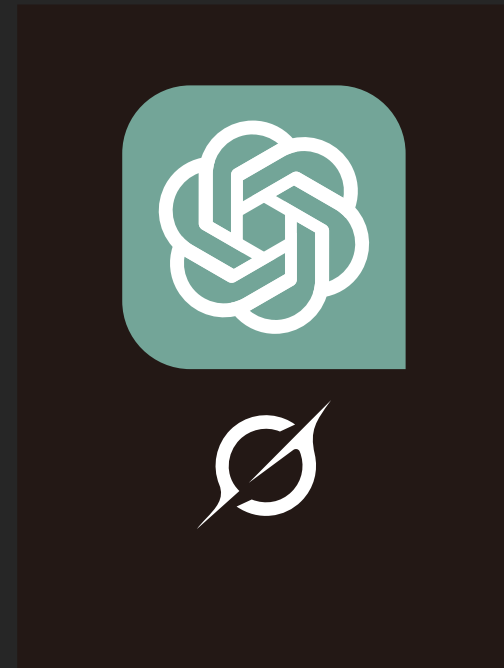
MAJOR DATA



[우리의 가이드] 코드가 존재하는 곳
깃허브



[우리의 여전한 친구] 구글
구글



[새로운 기술] ChatGPT, Grok 등등
AI

Why ?

Github

소스코드

코드가 있는 것과 없는 것은 매우 다르다.
참고할 소스가 많을 수록 합치면 된다.

논문

논문은 설명은 좋지만 소스코드가 없을 수 있다.
키워드 기반 검색 Machine Learning, Malware, Detect

GPT

ChatGPT

요청을 키워드 기반으로 Github중심 검색도 가능
답리서치로 여러 소 주제들을 미리 알고 깊게 생각도 가능

Other AI

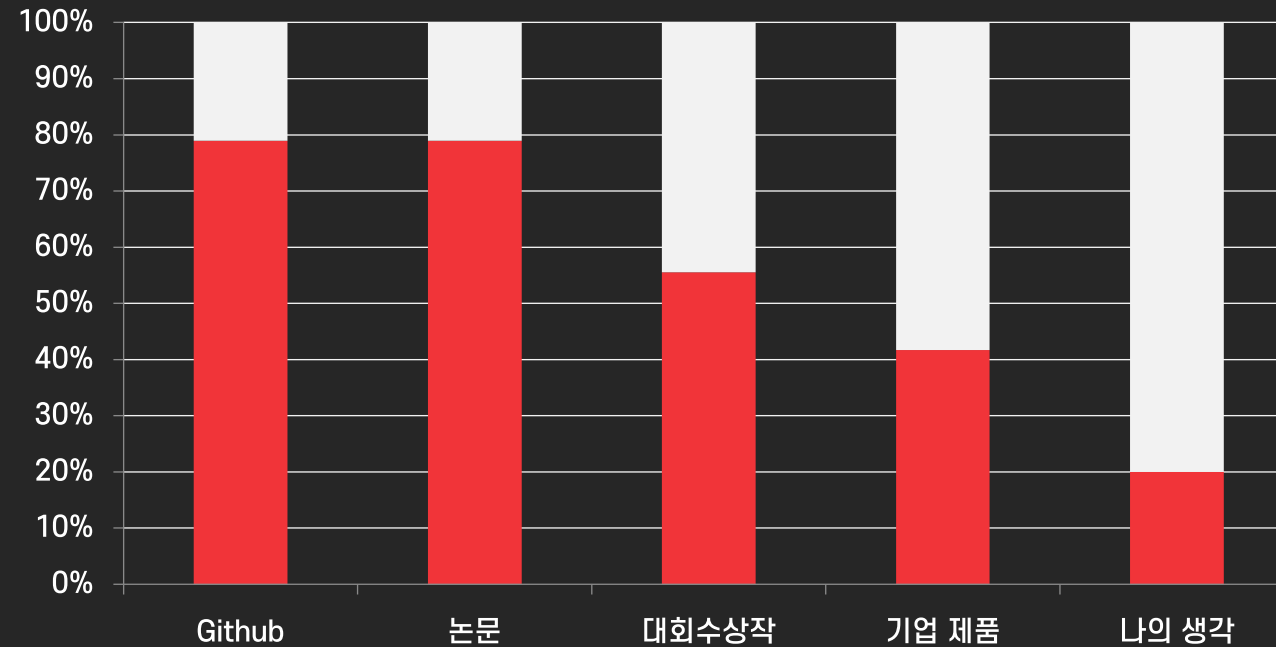
ChatGPT외 다른 AI인 Gemini, Grok등도 사용하길 권장
여러곳에 일시키고 보고 판단만 인간이 하면 된다.

좋은 레퍼런스 (자료)

구분	설명	양식
Github	<ul style="list-style-type: none"> ▪ F1 스코어, 문서, 코드의 주석이 잘 적혀있는가? ▪ 쓴 저자가 색다른 방법을 했는가? 	<ul style="list-style-type: none"> ▪ F1 스코어, 그림 ▪ 코드의 주석
논문	<ul style="list-style-type: none"> ▪ 가능하면 국외 영어 논문이 좋을 수 있습니다.(ex IEEE) ▪ 논문 같은 경우 데이터, 소스 코드에 대한 설명이 부실할 수 있다. 그러므로 어떤 특성을 주로 사용했는지 보는 것 또한 좋은 관점이 될 수 있다. ▪ 주의해야 할 점 절대 학사 논문은 이미지나 결과값이 좋은지를 꼭 체크할 것 	<ul style="list-style-type: none"> ▪ 전처리 방법론 ▪ 어떤 특성(Feature)를 추출하는가?
대회 수상작	<ul style="list-style-type: none"> ▪ 대회를 수상했다면 그만한 이유가 있다. ▪ 소개서라도 어떻게 들고오면 잘 분석하면 된다. 	<ul style="list-style-type: none"> ▪ 우승한 이유 분석 ▪ 평가 기준. 해당 대회 분석
기업 제품	<ul style="list-style-type: none"> ▪ 기업을 팔기 위해 스토리텔링을 하는 전문 사기꾼이다. ▪ 해당 제품이 어떤 상품인지 분석하는 것도 좋을 수 있다. 	<ul style="list-style-type: none"> ▪ 제품 분석
나의 생각	<ul style="list-style-type: none"> ▪ 어찌면 세상에 나오지 않는 새로운 아이디어 일 수 있다. ▪ 하지만 위험하다. 	<ul style="list-style-type: none"> ▪ 안하는 걸 권장한다.

데이터 중요도 차트

데이터 중요도



- 이 차트는 팀장의 주관적 의견을 밝힙니다.
- 이의있을 시 팀원의 의견이 맞습니다.

그래서 어떻게 하세요?

step
1

[How?]

AI를 이용해 찾는법
구글을 이용해 찾는법

step
2

[What?]

키워드 기반 검색방법
AI에 맞게 검색하기

step
3

[Where]

찾은 데이터를 어디에 정리해
야하나



Step 1

어떤 도구로 검색을 해야하나?(HOW?)



무엇을 리서치하고 싶으신가요?

내가 "AI를 이용해서 악성코드를 탐지하는 모델을 만들고 싶은데 선수자료가 필요해
정확도가 좋고 Github, 논문등 실질적으로 의미있고 소스코드 및 데이터가 잘되어있
는 자료를 원해!

+ 오픈 | 🔍 리서치 ✕



답 리서치

ChatGPT에 검색능력을 보시면 됩니다.

키워드 기반 검색

AI, 악성코드, 멀웨어, ML, 머신러닝, Github, 논문등

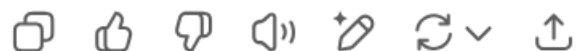
정확한 지시

- 선수자료가 필요해
목적은 먼저 AI에게 설명
- 실질적으로 의미있는 소스코드 및 데이터
내가 어떤걸 원하는지 설명
- 정확도가 좋은
내가 바라는 데이터에 포함되어야할 요소를 설명

AI 기반 악성코드 탐지 모델 관련 자료를 찾는 데 도움이 될 수 있도록 몇 가지를 더 여쭙보고 싶어요:

1. 어떤 종류의 악성코드를 중점적으로 다루고 싶으신가요? (예: 바이너리, 안드로이드 앱, 웹 기반, 랜섬웨어 등)
2. 선호하는 AI/ML 접근 방식이 있으신가요? (예: 딥러닝, 트리 기반 모델, 트랜스포머 등)
3. 데이터셋이 반드시 공개되어 있어야 하나요, 아니면 논문이나 코드만 있어도 괜찮을까요?
4. 목표는 실제 프로젝트 구현인가요, 아니면 논문 연구 용도인가요?

이 정보들을 알려주시면 보다 정밀하고 유용한 자료를 찾아드릴 수 있어요!



1. 윈도우 악성코드
2. 너가 보기에 좋은 정확도로 결과나오는 걸로
3. 데이터셋이 공개되어있으면 좋고 논문이나 코드가 있어도 괜찮긴한데 공개되었으면 좋겠어
4. 실제 프로젝트 구현 및 발표

+ 오피스 | 리서치 x





1. 윈도우 악성코드
 2. 너가 보기에 좋은 정확도로 결과나오는 걸로
 3. 데이터셋이 공개되어있으면 좋고 논문이나 코드가 있어도 괜찮긴한데 공개되었으면 좋겠어
 4. 실제 프로젝트 구현 및 발표
- 내가 "AI를 이용해서 악성코드를 탐지하는 모델을 만들고 싶은데 선수자료가 필요해 정확도가 좋고 Github, 논문등 실질적으로 의미있고 소스코드 및 데이터가 잘되어있는 자료를 원해!

윈도우 악성코드 탐지를 위한 AI 모델 개발에 필요한 정확도가 높고, 데이터셋 및 소스코드가 공개된 자료를 아래에 정리했습니다. 실제 프로젝트 구현 및 발표에 활용할 수 있도록 논문, GitHub 리포지토리, 데이터셋 중심으로 추천하며, 최신성과 실용성을 고려했습니다.

1. 추천 데이터셋

아래 데이터셋은 윈도우 악성코드 탐지에 널리 사용되며, 공개적으로 접근 가능하고 품질이 높아 프로젝트에 적합합니다.



Step 1

어떤 도구로 검색을 해야하나?(고전[구글])



대학생 신분의 활용

대학교 도서관을 이용하여 자료 검색또한 방법

키워드 기반 검색

AI, 악성코드, 멀웨어, ML, 머신러닝, Github, 논문등

구글 검색 TIP

- 이 키워드가 꼭 포함하게 해줘 큰따옴표("")
"Keyword"로 검색시 해당 키워드를 반드시 포함한 결과를 검색
- type:pdf 나는 pdf파일을 보고 싶어
type:확장자명으로 해당 확장자 파일 전반적으로 검색
- site:웹주소 해당 사이트를 기준으로 검색
내가 원하는 웹주소를 기반으로 검색

Google 학술검색

malware detection by machine learning



☒ 모든 언어 ☐ 한국어 웹

거인의 어깨에 올라서서 더 넓은 세상을 바라보라 - 아이작 뉴턴

<https://scholar.google.com/>



Mission2

선수조사
자료분석

자료를 어떻게 정리해야하는가?

자료수집의 MAIN ROUTE

MAJOR DATA



Google Sheets

[우리의 여전한 친구] 구글
구글 스프레드시트

참고 그림 (구글 스프레드시트)

선수조사 자료 ☆ 📁 ☁

파일 수정 보기 삽입 서식 데이터 도구 확장 프로그램 도움말

메뉴 100% 123 기본값 ... 10 B I A

A1 데이터셋

	A	B
1	데이터셋	어떤 특성 추출
2	1. www.kaggle.com/datasets/dasarijayanth/pe-header-data 2. www.kaggle.com/competitions/malware-classification/data	데이터 세트1의 경우: 데이터 세트에는 이미 해당 기능에 대한 값이 있습니다. .exe, .dll 파일 유형과 같은 주어진 PE 파일에서 해당 기능과 헤더 정보를 추출하는 스크립트를 만 해 Extra-Tries 분류기를 대해 별도의 디렉토리를 prefixes/keywords/regist 해 이미지 픽셀을 특징으로 용하고 각 기능 세트에 대

내 드라이브 > AI_SEC > 자료(래퍼런스) ▾

1개 선택됨

이름 ↑

선수조사 자료 정렬, 내림차순

100 Kbytes

200 Image Pixels.

여러분은 권한이 없습니다. 접근 권한을 요청하세요!

개인 설문

B *I* U  

설문지 설명

이름

단답형 텍스트

전화번호

단답형 텍스트

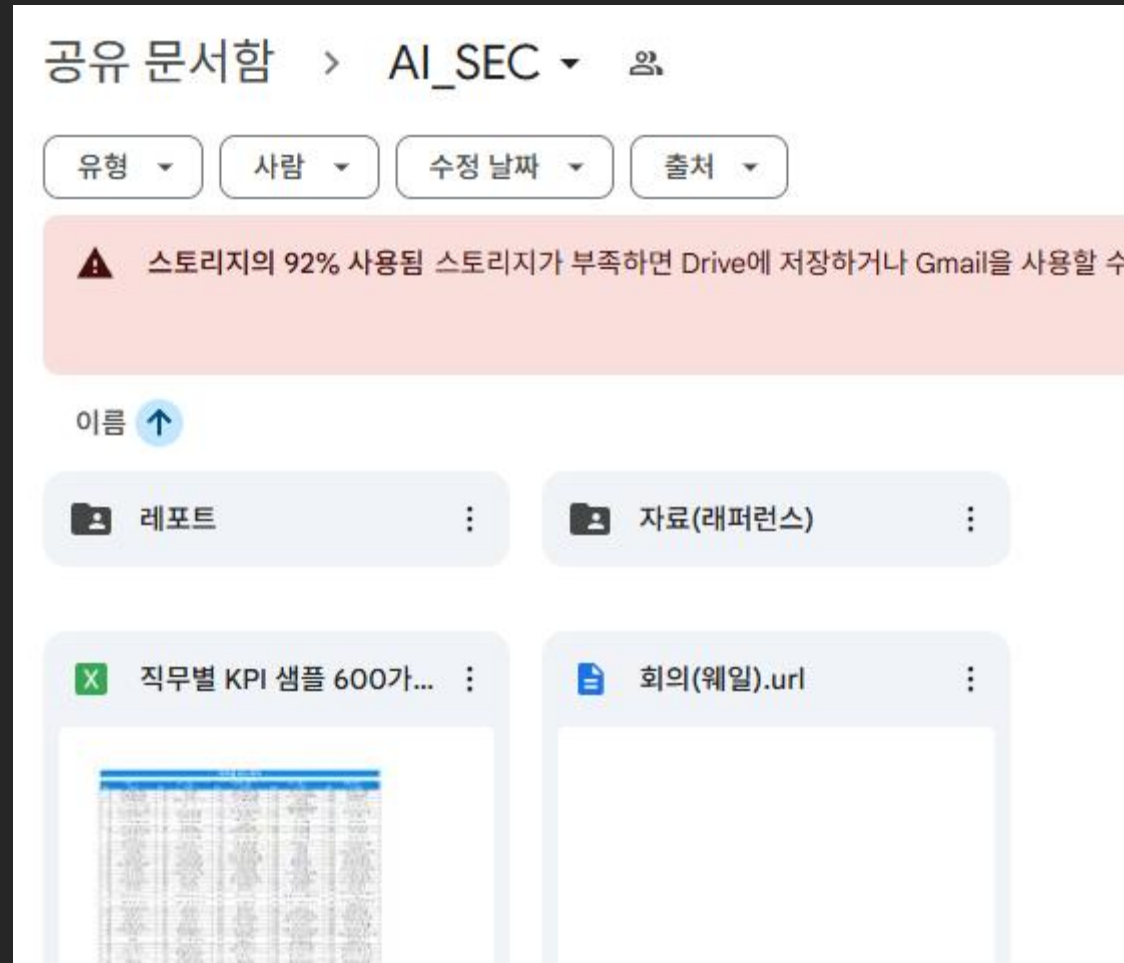
이메일

단답형 텍스트

여러분은 권한이 없습니다. 접근 권한을 요청하세요!

https://docs.google.com/forms/d/e/1FAIpQLSf_M2uu5v61SHWccCSxDp0wrUrYcl8cBP85U4AcxsiBrRA5jw/viewform?usp=header

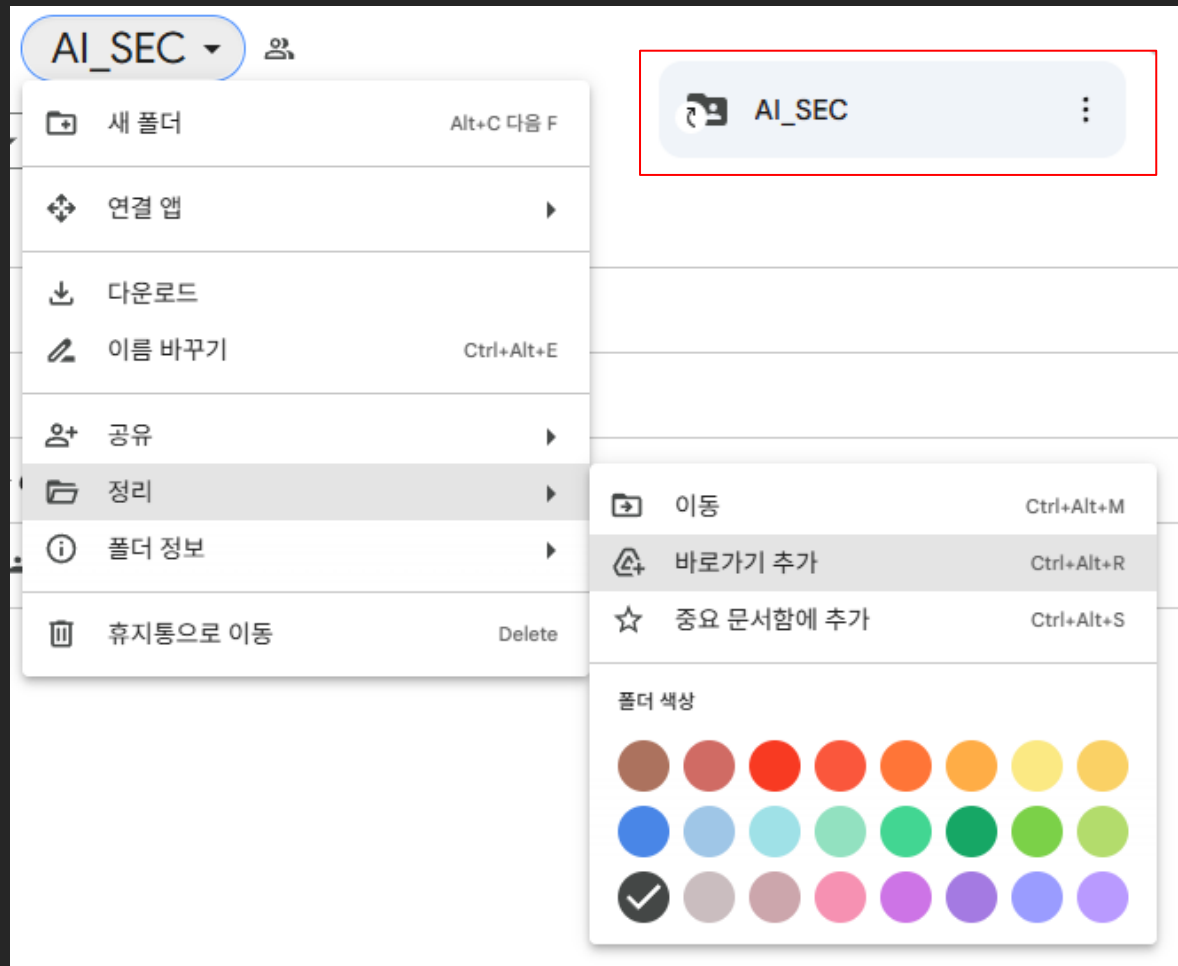
팀 프로젝트 공유폴더



권한을 얻으면 위에처럼 들어가집니다.

<https://drive.google.com/drive/u/7/folders/1fjBwp8dPf4wPQFCKrLPsQUkH6MwmRIAF>

구글 드라이브 바로가기



바로가기 만들면 자기드라이브에서 편하게 접근할 수 있습니다.

<https://drive.google.com/drive/u/7/folders/1fjBwp8dPf4wPQFCKrLPsQUkH6MwmRIAF>

스프레드시트

	A	B	C	D	E	F
1	데이터셋	어떤 특성 추출	어떤 모델 학습	결과(F1 Score)	비고	url
2						
3						



Github(ML) ▾

Github(Deep Learning) ▾

논문(국외) ▾

논문(국내) ▾

대회(수상작) ▾

학생프로젝트 ▾

아이디어(나의 생각) ▾

같이 봤으면 좋을 것들 ▾

Github(ML)

Github에서 찾은 소스중 Machine Learning

Github(Deep Learning)

Github에서 찾은 소스중 Deep Learning

논문 TIP

■ 논문(국내)

논문(국내)을 적으시면 됩니다. 대학생이시면 대학교 도서관 사이트중 RISS로 검색해서 찾는 걸 추천

■ 논문(국외)

국외 논문도 대학교 도서관 사이트 검색으로 Science나 IEEE등 공신력있는데 찾으시면 됩니다,

■ 대회 수상작, 학생프로젝트

대회수상작, 학생프로젝트는 Github이나 블로그에서 찾으셔도 이쪽에 분류 부탁드립니다.

■ 아이디어

좋은 아이디어 프로젝트용으로 아쉬운데 아이디어 괜찮은건 여기에 적어주시면 감사하겠습니다.

■ 같이 봤으면 좋을 것들

같이 봤어면 좋을 것들 적으시면 됩니다.

스프레드시트

	A	B	C	D	E	F
1	데이터셋	어떤 특성 추출	어떤 모델 학습	결과(F1 Score)	비고	url
2						
3						

www.kaggle.com/datasets/dasarijayanth/pe-header-data www.kaggle.com/competitions/malware-classification/data	데이터 세트1의 경우: 데이터 세트에는 이미 해당 기능에 대한 값이 있습니다. .exe, .dll 파일 유형과 같은 주어진 PE 파일에서 해당 기능과 헤더 정보를 추출하는 스크립트를 만들었습니다. 모든 사용 가능한 정보에서 중요한 기능 세트를 선택하여 기능 선택을 위해 Extra-Tries 분류기를 사용했습니다. 데이터 세트 2의 경우: 데이터셋에는 원시 바이트 파일과 asm 파일이 있습니다. 각 파일 유형에 대해 별도의 디렉토리를 생성하고 각 파일의 파일 크기를 특성으로 추출했습니다. 각 파일의 특징으로 바이트(byte-n-grams, 여기서 n=1,2)와 asm 파일(prefixes/keywords/registers/opcode-n-grams, 여기서 n=1,2,3,4)에서 추출된 N-grams입니다. asm 파일을 이미지로 변환하고 해당 이미지에서 성능이 가장 좋은 200개 이미지 픽셀을 특징으로 추출했습니다. 위의 모든 기능 중에서 중요한 기능을 선택하기 위해 랜덤 포레스트를 사용하고 각 기능 세트에 대해 별도로 병합했습니다. 최종 데이터 세트에는 다음과 같은 특징이 포함됩니다. 14 PE header features Byte file features(unigram+byte file size) + 200 Byte Bi-gram features Asm file features(opcode uni-gram+ Asm file size + keywords, registers, prefixes) 100 Opcode Bi-gram features 100 Opcode Tri-gram features 200 Image Pixels.	다양한 ML 모델 Random Forest 0.995699 knn 0.992096 Decision Tree 0.988022 Support vector Machine 0.561446 Gradient Boost Classification 0.990698		https://github.com/DasariJ
--	---	--	--	---

데이터셋

데이터셋을 kaggle이나 다른사이트에서 있으면 url로 표기해주시면 감사하겠습니다.
가능하면 자세하게 샘플수, 용량, 제작 년도등 정보도 같이 표기해주시면 좋습니다.

특성

데이터에서 어떤 정보를 추출했는지 적으시면 됩니다.
어떤 내용인지 모르겠어도 최대한 적어주시면 제가 조금씩 고치겠습니다.

모델

Random Forest, KNN등 모델 사용한걸 적어주시면 감사하겠습니다.

비고

팀원들에게 이프로젝트에 대한 알려주고 싶은 것 및 국내등 특이한 사항있으면 기록해주시면 감사하겠습니다.

url

해당 프로젝트나 논문등의 url을 적어주시면 됩니다.

「

」

「

」

감사합니다