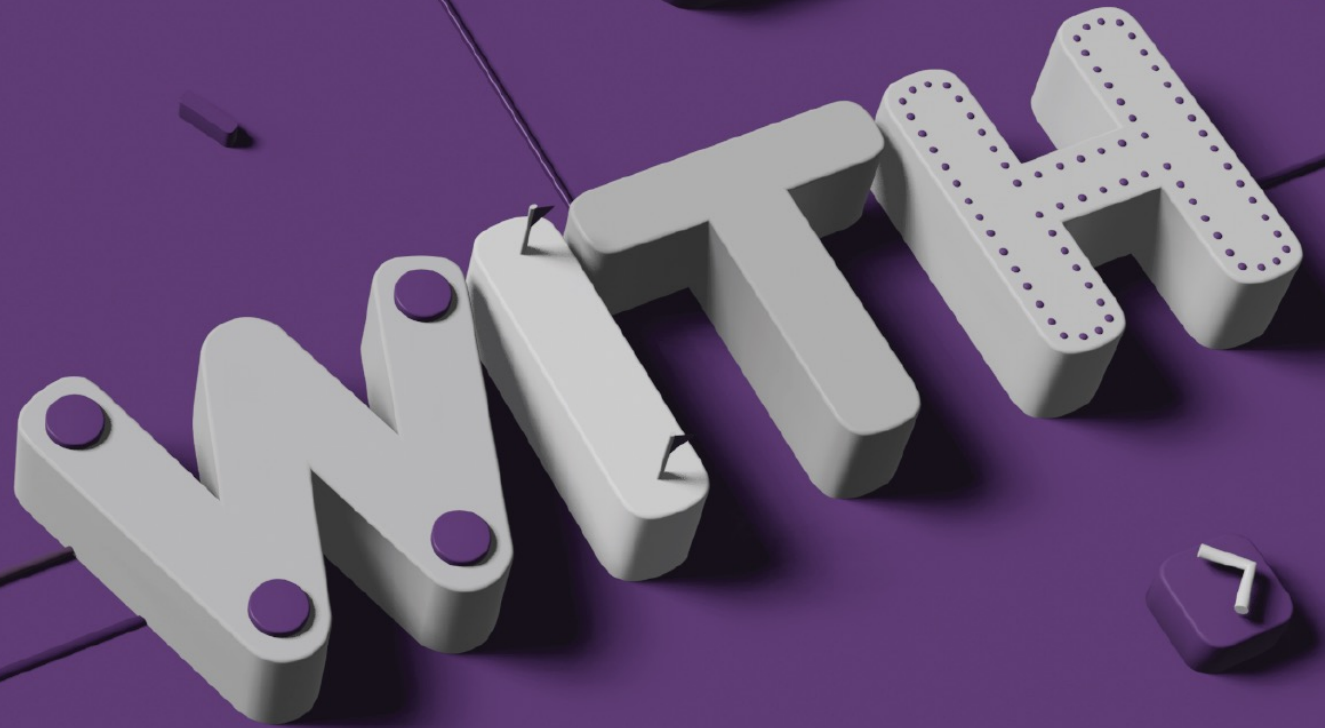


S2W WEBINAR

with < DarkBERT >



제3회 S2W WITH 웨비나

2023년 6월 15일 (목) LIVE
오후 2:00 - 3:30

For June

AI 기술과 언어모델(DarkBERT)

Agenda

2:00 - 2:20	키노트: AI와 사이버 보안
2:20 - 2:50	NLP와 위협 콘텐츠 자동 분류 및 사례
2:50 - 3:20	익명 채널 및 보안에 특화된 '다크웹 언어모델(DarkBERT)'
3:20 - 3:30	QnA



제3회 S2W WITH 웨비나

Session 1. Keynote

박근태, CTO

AI와 사이버 보안



Introduction

- **(현) S2W CTO ('22.8~)**
- **SKT Big data & AI 팀장 ('10.7~'22.8)**
 - 에이닷 서비스 개발 (쿠팡드)
 - AI 맞춤형 고객 상품 추천 서비스 개발
 - 센서 데이터 기반 예지 정비 AI 개발
 - 이미지 기반 불량 검출 AI 개발
 - Tmap 실시간 운전 습관 시스템 개발
 - MPP(Massively parallel processing) DB 기술 개발
- **TmaxCore Kernel 개발팀('09.1~'10.7)**
- **ETRI 저장시스템 개발팀('07.3~'08.12)**

Contents

- **AI를 쓰는 이유**
- **LLM(거대 언어 모델) 이전과 이후의 AI**
- **S2W의 사이버 보안 AI**

AI를 쓰는 이유



AI를 쓰는 이유



AI를 쓰는 이유



AI를 쓰는 이유



전문가 대비 시간과 공간의 제약없이 일정 수준의 성능을 제공

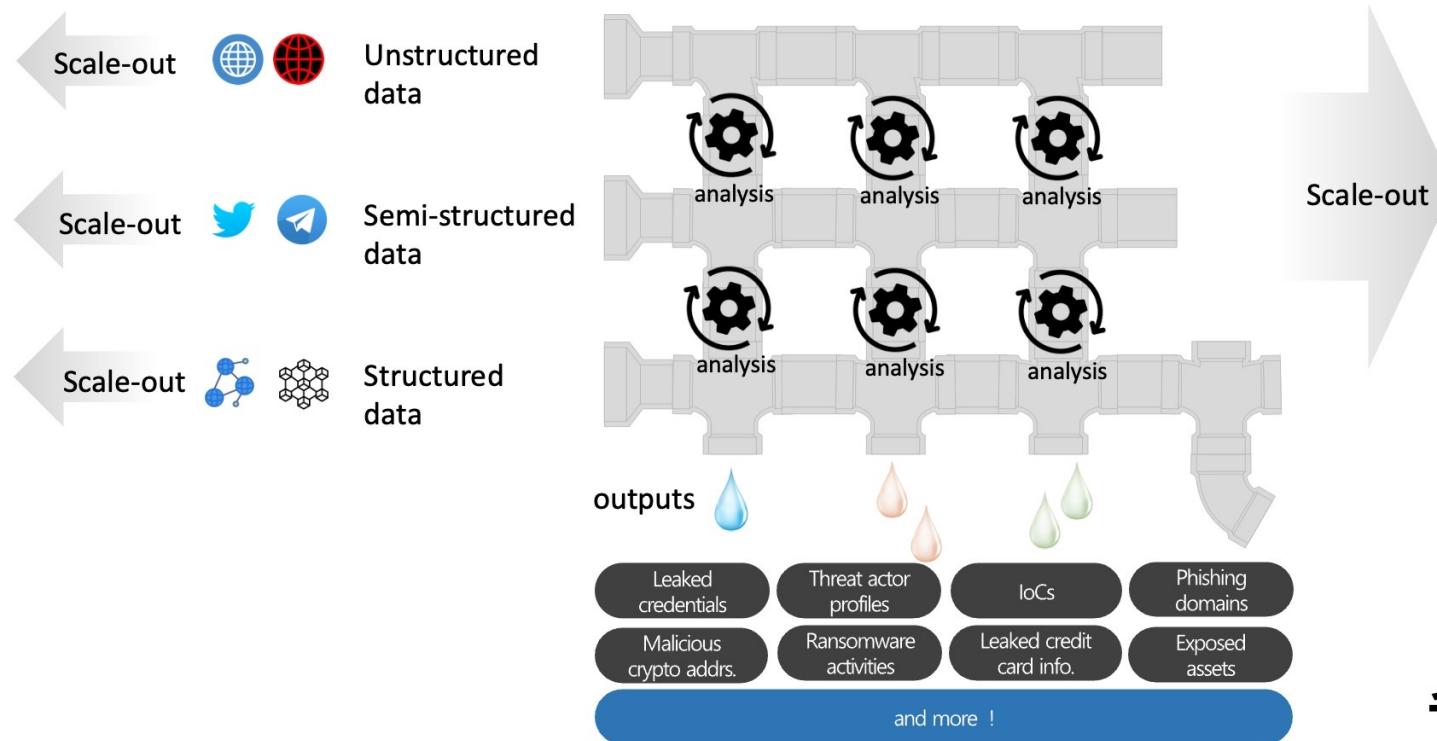


AI를 쓰는 이유

시간과 공간의 제약 없는 일정 수준 성능?

폭발적인 분석 대상 데이터의 증가 (Big data handling)

- 인력으로 더 이상 전체 확인이 불가능



최종 분석
결과물이
수억 건 이상

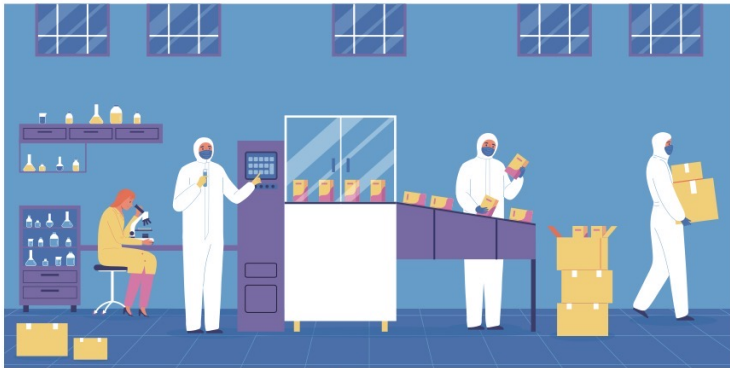
AI를 쓰는 이유

시간과 공간의 제약 없는 일정 수준 성능?

동시 다발적인 수요에 대응 (Scalability)

- 특히 Global business의 경우 인력 기반으로서는 즉각 대응이 힘들

인력 기반 제품 검수



이미지: Freepik.com 작가: macrovector

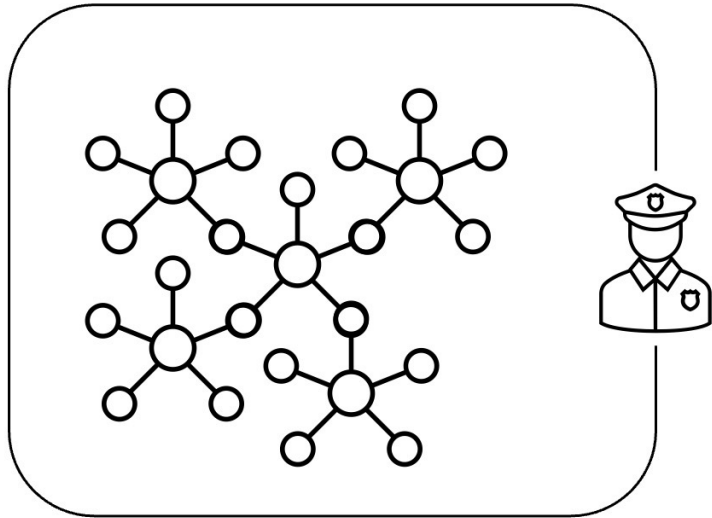


AI를 쓰는 이유

시간과 공간의 제약 없는 일정 수준 성능?

전문가의 상태에 의존하지 않는 일정한 성능

- 365일 24시간, 식사/수면/피로/의도? 등의 영향 배제

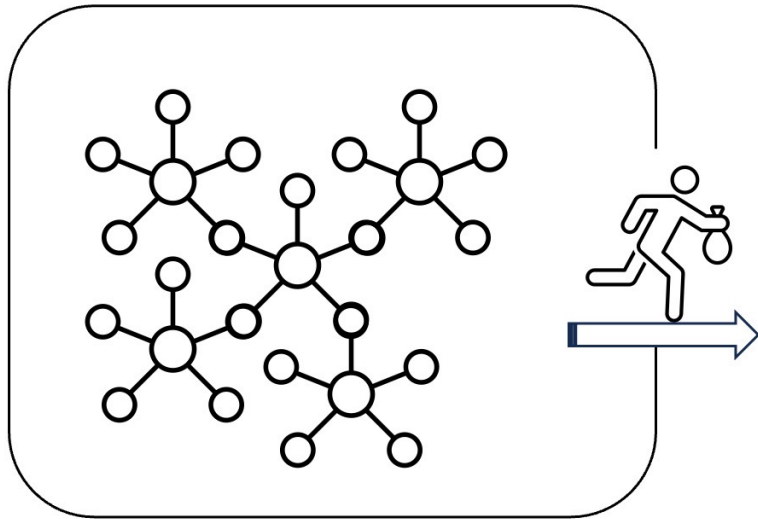


AI를 쓰는 이유

시간과 공간의 제약 없는 일정 수준 성능?

전문가의 상태에 의존하지 않는 일정한 성능

- 365일 24시간, 식사/수면/피로/의도? 등의 영향 배제



해외 네트워크 보안 업체 case

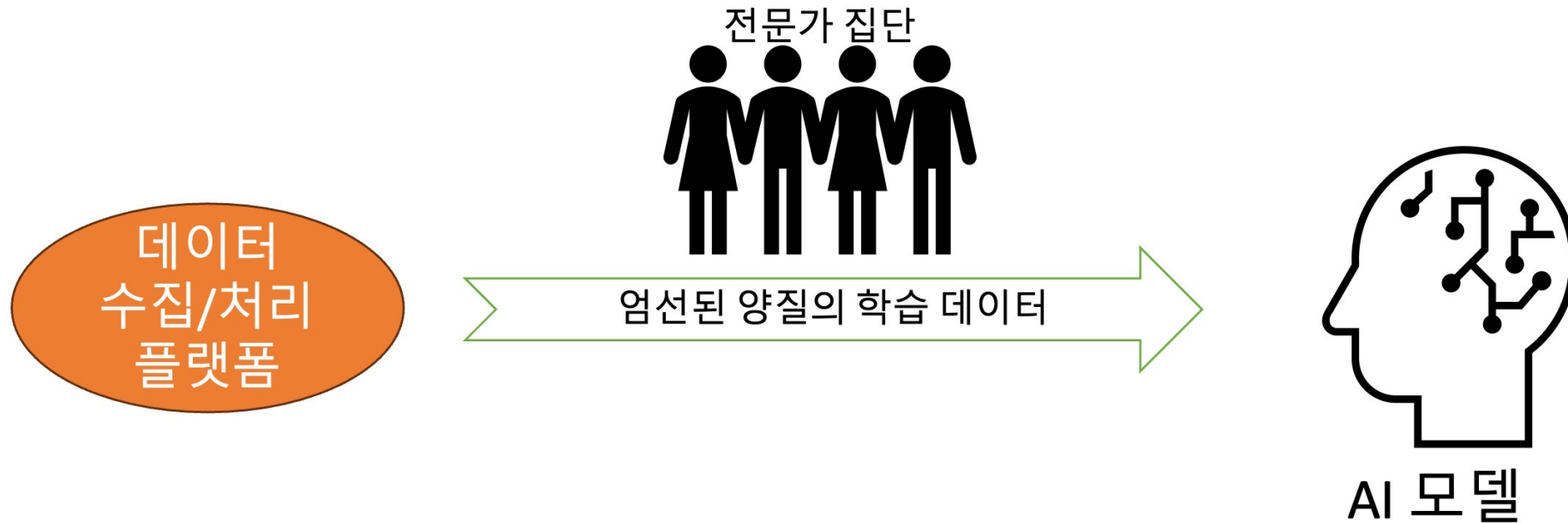
- 자체 AI 테스트 중 내부 데이터 유출 정황 포착
- 데이터 유출 탐지 rule 관리자가 backdoor 통해 유출

AI를 쓰는 이유

현실적인 목표치

전문가보다 더 깊이 있는 지식을 주는 것은 어려움

- 얼마나 양질의 데이터를 보유하고 있는지가 중요함
- 이 경우에도 목표는 전문가에 버금가는 수준 도달

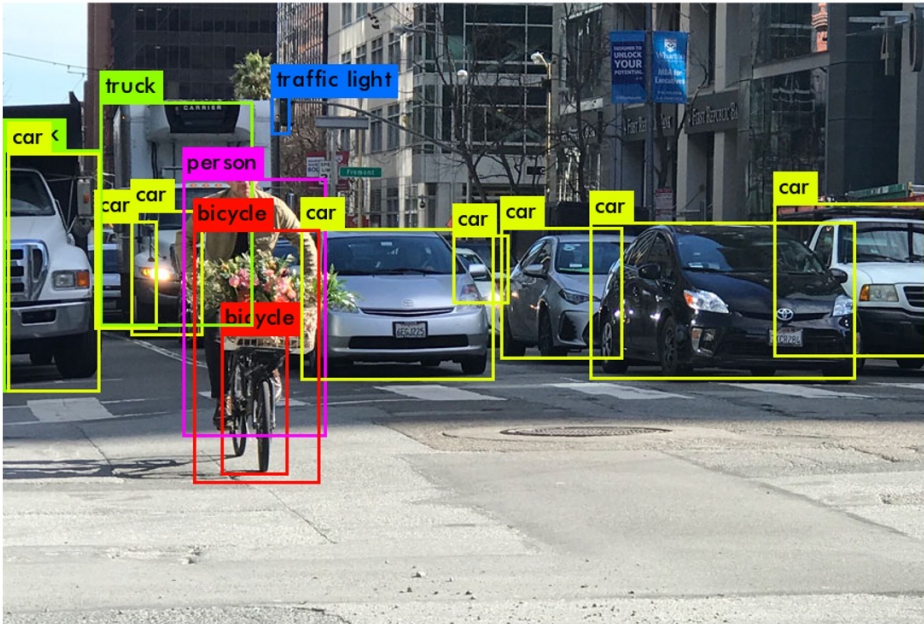


LLM 기술 이전과 이후의 AI

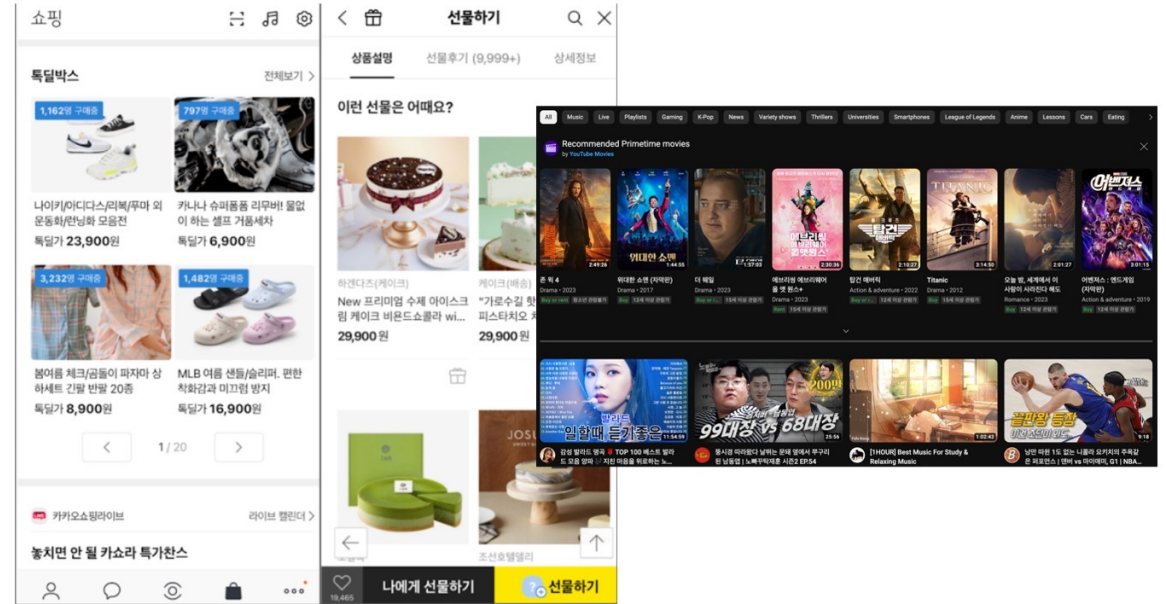
1) LLM 기술 이전

특화된 분야의 AI 기술들은 이미 존재하고 활용되고 있었음

- Supervised Learning: 학습을 위한 정답이 존재하는 경우



물체 인식

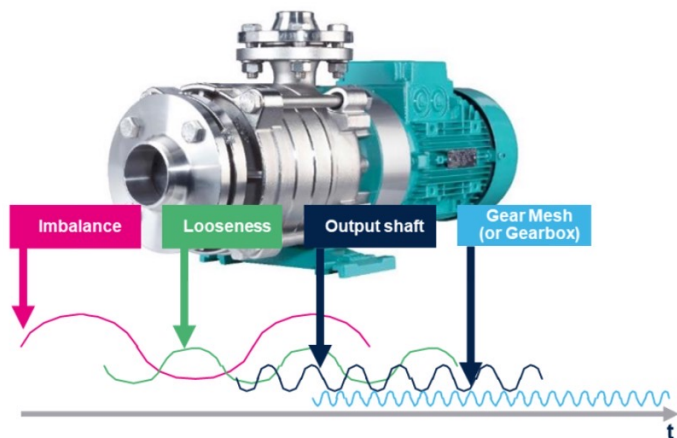


추천 시스템 (쇼핑/OTT 등)

LLM 기술 이전과 이후의 AI

1) LLM 기술 이전

- **Unsupervised Learning: 정답을 특정하기 힘든 경우**
ex. 정상적이지 않은 모든 경우를 다 잡아내기 (Anomaly detection)



장애 예측



출처: <https://www.xenonstack.com/>

사기 탐지

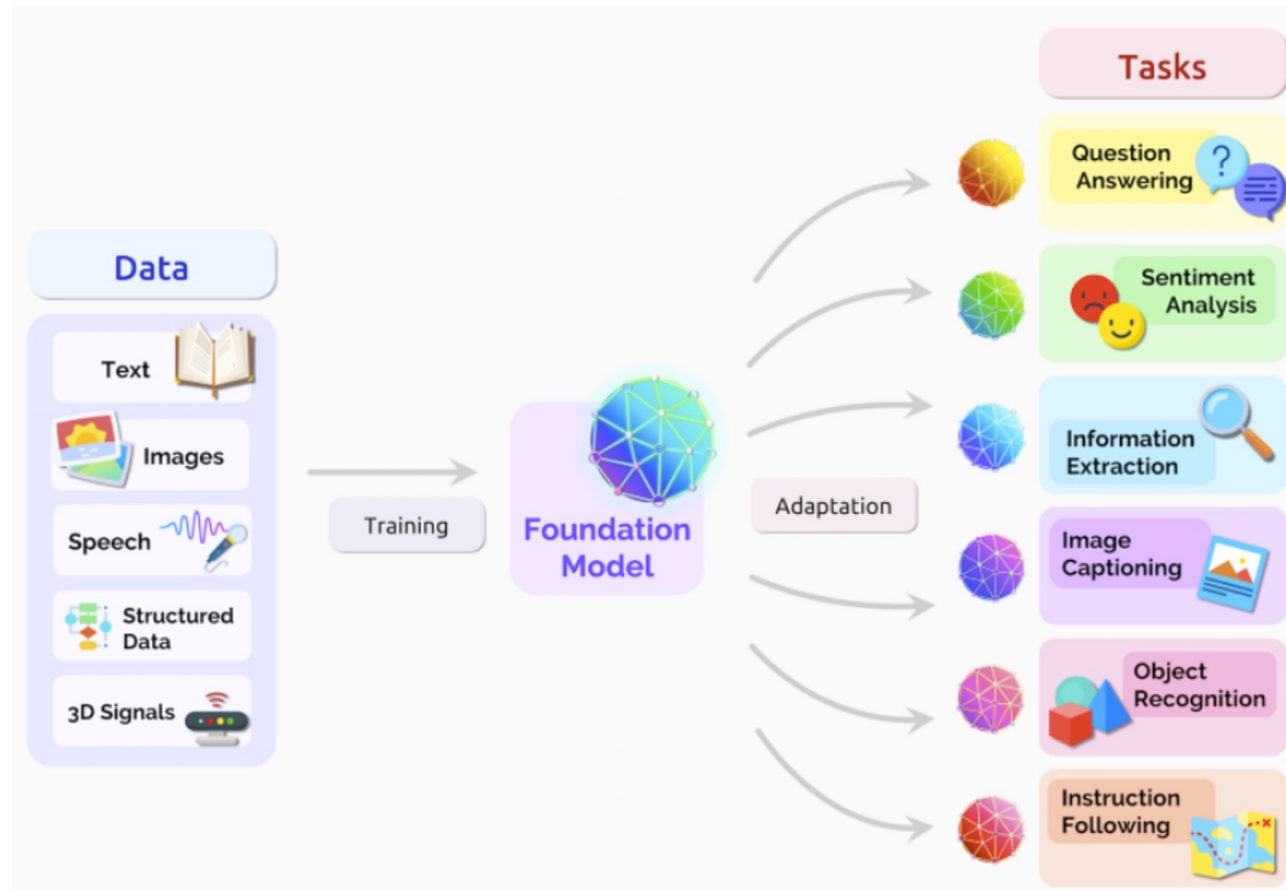
각 분야에 특화된 AI가 제한적으로 활용되고 있었음

LLM 기술 이전과 이후의 AI

2) LLM 기술의 등장

LLM(Large Language Model: 거대 언어 모델)

- 데이터 기반의 다양한 문제에 공통적으로 적용할 수 있는 기반이 되는 모델






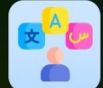







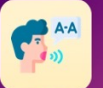















LLM 기술 이전과 이후의 AI

2) LLM 기술의 등장

- 일상의 다양한 문제들에 실제 적용되어 활용

ACHIEVEMENTS UNLOCKED BY LLMs

EMERGENT ABILITIES OF LARGE LANGUAGE MODELS (APR/2023)

S	GPT-3 13B, PaLM 8B	 Mod.Arithmetic*	 Debugging*	 Comprehension*			
M	GPT-3 175B, LaMDA 137B, PaLM 64B, Chinchilla 7B	 LinguisticsPuzzles*	 EmojiMovie*	 GRE-Comprehension*	 MetaphorUnderstanding*	 PhysicalIntuition*	 LogicalDeduction*
L	PaLM 540B, Chinchilla 70B	 GeometricShapes*	 Proverbs*	 PhoneticAlphabet*	 ElementaryMath*	 CausalJudgment*	 CodeLineDescription*
XL	GPT-4, Gemini (est.)	 College-LevelExams	 Self-Critique/Reflection	 AppBuilding	 SpatialReasoning	 AdvancedCreativity	 EmbodimentOptions
Next...		 Grounding	 Long-HorizonPlanning	 Persuasion	 AdvancedEmbodiment	 Awareness	 More...

Selected highlights only. Sources: Original papers and Dr. Jason Wei's summary: <https://www.lifearchitect.ai/blog/emergence-big-bench/>. * Using images from Flaticon.com. Alan D. Thompson, April 2023. <https://lifearchitect.ai>

[LifeArchitect.ai/models](https://lifearchitect.ai/models)

LLM 기술 이전과 이후의 AI

3) LLM 등장 의의

가장 큰 차이점

- 과거 일상적인 삶을 사는 사람들에게 AI는 남는 시간이나 관심을 가질 만한 유희의 아이템 중 하나였음
- 이제는, AI가 본인 업무나 삶에 큰 도움을 줄 수 있겠다는 공감대가 형성되어 실제 다양한 분야에 능동적으로 적용하기 시작

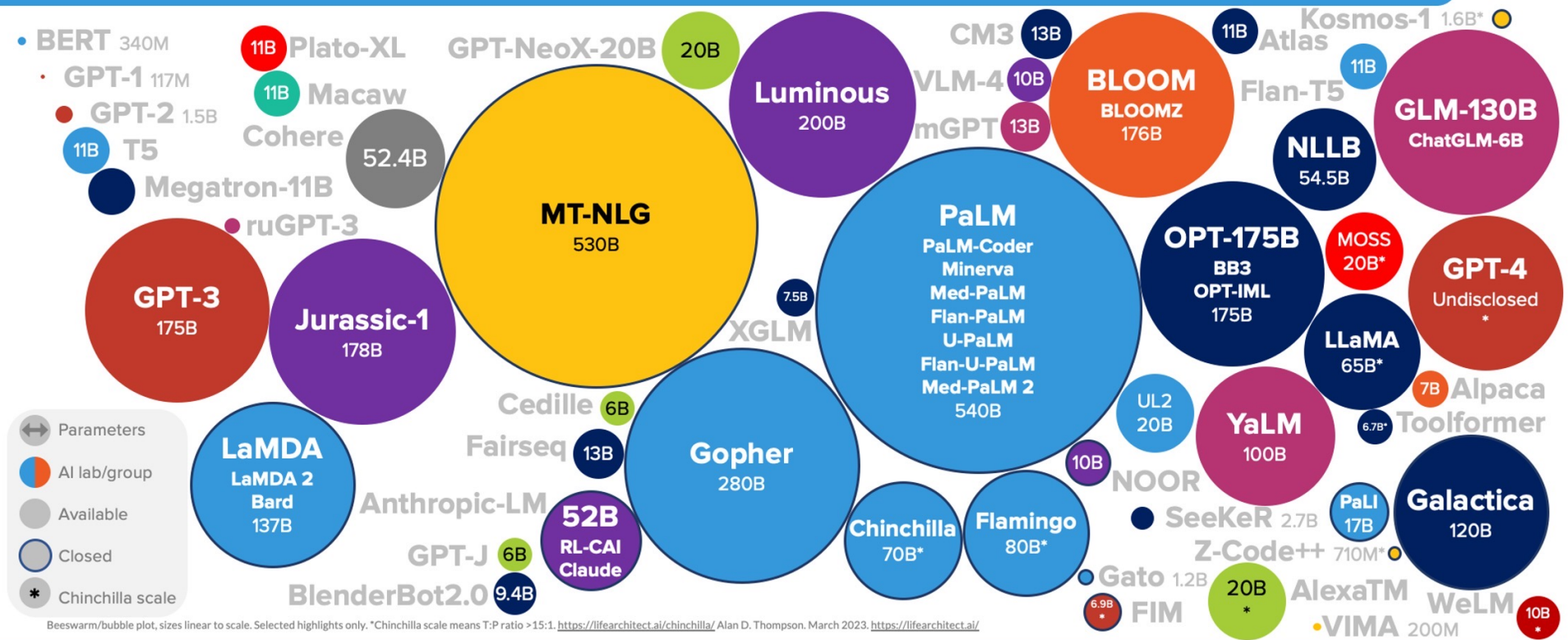
업무에 도움이 된다는 것 == 큰 돈이 유입 된다는 것

- 선순환 구조에 따라서 LLM의 발전 속도가 더욱 가속되고 있음

LLM 기술 이전과 이후의 AI

3) LLM 등장 의의

LANGUAGE MODEL SIZES TO MAR/2023



LifeArchitect.ai/models

LLM 기술 이전과 이후의 AI

3) LLM 등장 의의

- 결과물의 임팩트도 훨씬 커진다는 의미
- 환각(Hallucination)

Bard (Experiment)

recent south korean data breach incidents

View other drafts

Sure, here are some recent South Korean data breach incidents:

- In March 2023, a data breach at a South Korean hospital exposed the personal information of over 1 million patients, including their names, addresses, phone numbers, and medical records.
- In April 2023, a data breach at a South Korean government agency exposed the personal information of over 100,000 employees, including their names, Social Security numbers, and home addresses.
- In May 2023, a data breach at a South Korean telecommunications company exposed the personal information of over 50,000 customers, including their names, phone numbers, and email addresses.

These are just a few examples of the many data breaches that have occurred in South Korea in recent years. The country has a growing problem with data breaches, and it is important for businesses and individuals to take steps to protect their personal information.

Here are some tips for protecting your personal information from data breaches:

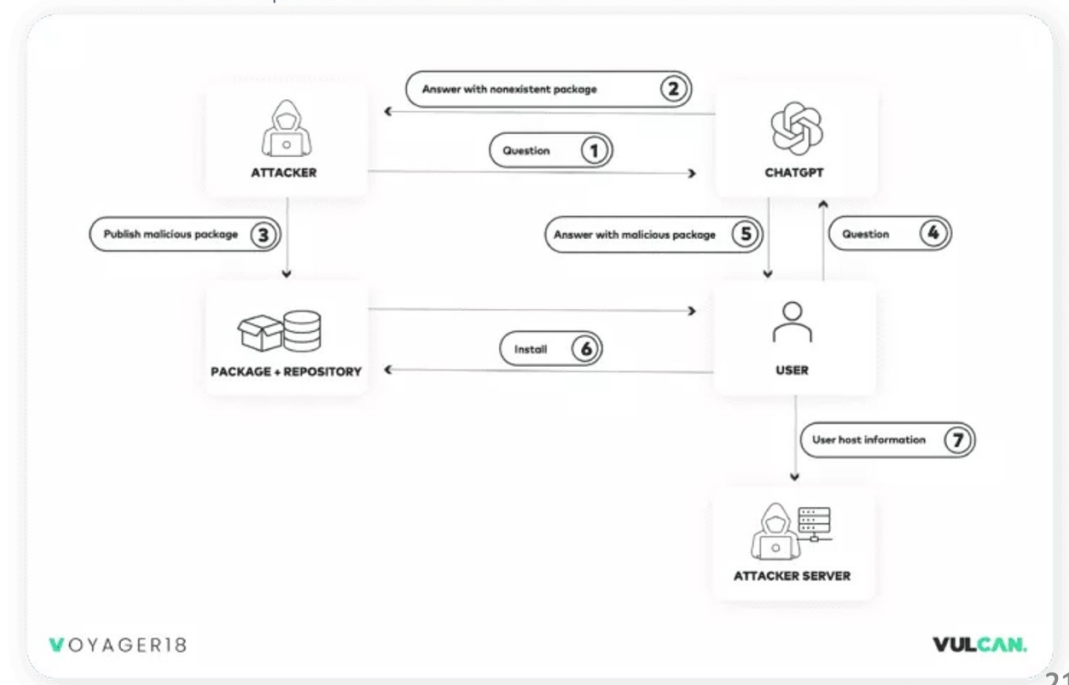
- Use strong passwords and change them regularly.
- Be careful about what information you share online.
- Use a VPN when connecting to public Wi-Fi networks.
- Keep your software up to date.
- Be aware of the latest scams and phishing attacks.

By following these tips, you can help to protect your personal information from data breaches.

Application Security | 5 MIN READ | NEWS

ChatGPT Hallucinations Open Developers to Supply Chain Malware Attacks

Attackers could exploit a common AI experience – false recommendations – to spread malicious code via developers that use ChatGPT to create software.



LLM 기술 이전과 이후의 AI

3) LLM 등장 의의

- 악의적인 비전문가에 위험 정보가 전달 - 해킹 등 범죄에 활용 가능



tom's HARDWARE

US Edition

Reviews Best Picks Raspberry Pi CPUs GPUs Coupons

TRENDING Computex 2023 Memorial Day Deals Try Our AI Chatbot

When you purchase through links on our site, we may earn an affiliate commission. [Here's how it works.](#)

Home > News

ChatGPT Can Generate Mutating Malware That Evades Modern Security Techniques

By Ash Hill published 1 day ago

ChatGPT complicates modern IT security with the ability to create mutating malware that's hard to detect with endpoint detection and response (EDR).

S2W의 AI

1) 보안에서 AI 기술의 활용은 이제는 필수

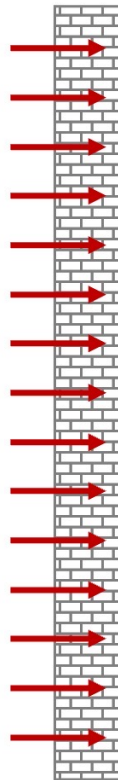
LLM의 보편화로 보안에서 공격과 방어의 비대칭성이 더욱 심화

- 인력으로 대응할 수 있는 범위를 점점 벗어나고 있음

Dark Web

Deep Web

LLM



00은행의 내부망 및 정보보호체계

FDS	방화벽	NAC	IPS	매체제어	네트워크 포렌식	DB암호화	
PC 보안관제	DDoS	1 롤기반의 대응 			WIPS	문서반출입	
ARS 사전승인	PC개인정보검출				망연계	백신	
암호초기화	보안USB 관리				DB접근제어	웹취약점	
문서보안	시스템취약점				출력물관리	APT	
서버개인 정보 검출	각종협의회				외주직원 관리	불법프로세스통제	
업무용PC보안점검	대외보고서작성				소스코드취약점	망분리	
컴플라이언스 대응	서무 및 감사				각종 보고자료	EAM	
개인정보보호	신용정보보호				개인정보 상시 모니터링	패치관리	
고객정보 관리	협력사고고객정보 공유				취약점이력관리	스팸차단	EDR
					외부메일	통합인증 로그인	

2



재택근무 보편화
업무 외 목적 사용으로 악성코드 감염사례 증가

3



클라우드 전환 증가
직원/협력사 관리 실수 등으로 내부 정보 유출

4



오픈소스, 협업툴 사용 증가
외부 페이지에 소스코드 등 중요정보 노출 위험도 증가

S2W의 AI

2) AI 기술의 활용 중심에는 Language Model이 있음

방어하는 입장에서 AI에 대한 손쉬운 접근이 필요하고,
이에 있어서 역시나 Language Model이 중요한 역할을 수행

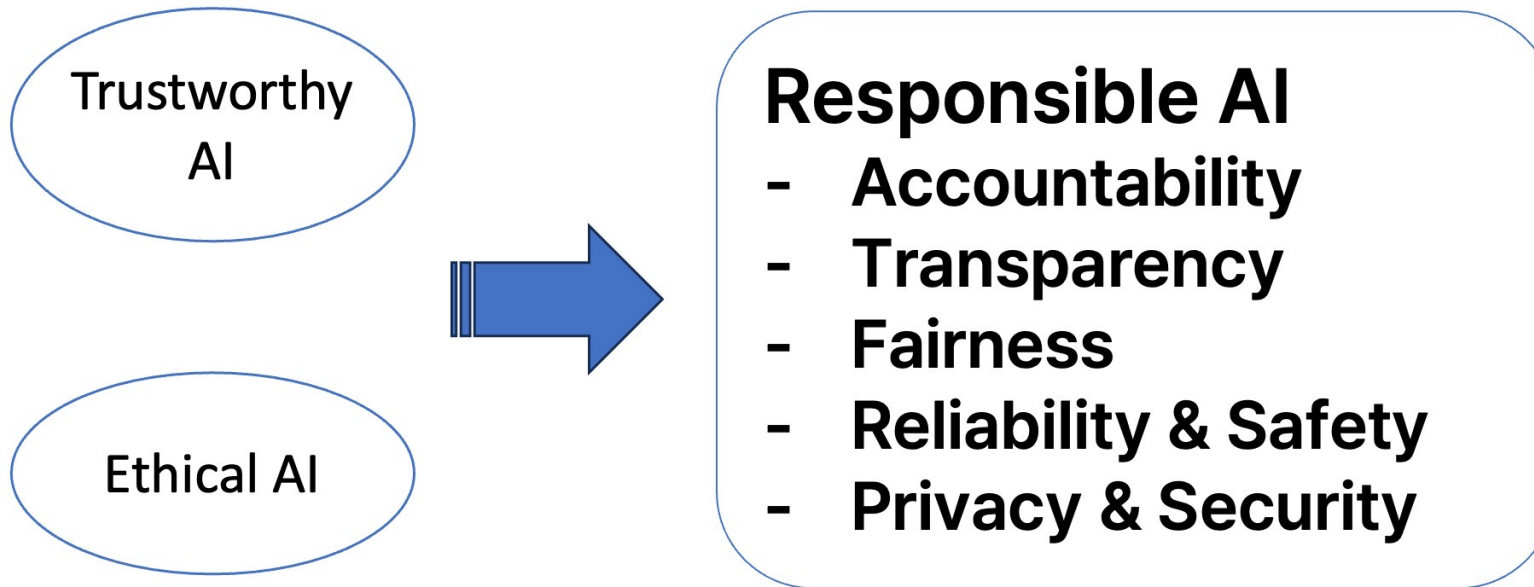
- Named entity recognition: 주요 보안 관련 요소에 대한 인지
- Relation extraction: 요소들 간의 인과 관계 도출
- Security knowledge base: 모든 데이터 간 관계를 구조화

전문가가 아닌 사용자도 AI를 통해서 답을 얻을 수 있도록 함

- 현 상황의 이해
- 사건 발생의 원인 탐색
- 향후 진행 방향에 대한 예측 및 대응책

S2W의 AI

3) 특히 고려해야할 점





장우진

AI팀 연구원

NLP 연구 및 AI 데이터 인사이트 분석

The information contained in this document is proprietary and confidential.
If you are not the intended recipient, please note that any use or circulation of this document may be cause for legal action.

제3회 S2W WITH 웨비나

Session 2. Data Insight

장우진, Researcher of AI Team

NLP와 위협 콘텐츠 자동 분류 및 사례



장우진 Eugene Jang

- KAIST

B.S. Computer Science (2019)

M.S. Computer Science (2021)

- S2W

Senior Researcher, AI Team (2021~)

Shedding New Light on the Language of the Dark Web

Youngjin Jin¹ Eugene Jang² Yongjae Lee² Seungwon Shin¹ Jin-Woo Chung^{2*}

¹KAIST, Daejeon, South Korea

²S2W Inc., Seongnam, South Korea

¹{ijjinjin, claude}@kaist.ac.kr

²{genesith, lee, jwchung}@s2w.inc

Abstract

The hidden nature and the limited accessibility of the Dark Web, combined with the lack of public datasets in this domain, make it difficult to study its inherent characteristics such as linguistic properties. Previous works on text classification of Dark Web domain have sug-

Ghosh et al., 2017; He et al., 2019; Choshen et al., 2019) and content analysis (Biryukov et al., 2014; Avarikioti et al., 2018). However, not much work has been done on the linguistic analysis of the Dark Web (Choshen et al., 2019).

In addition, the Dark Web has been studied and analyzed in the security research community to

NAACL 2022

<https://aclanthology.org/2022.naacl-main.412.pdf>

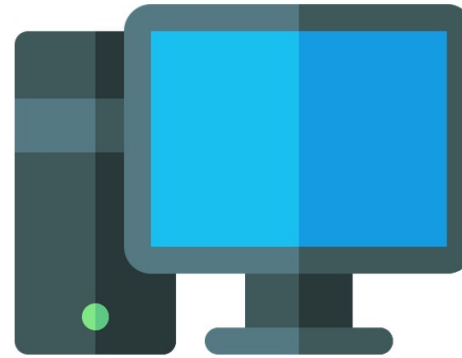
Natural Language Processing (자연어 처리)

Natural Language Processing 자연어 처리



Natural Language Processing (자연어 처리)

Natural Language Processing 자연어 처리



자연어로 수집되는 대용량 데이터

Home Page of Ransomr Leaks site



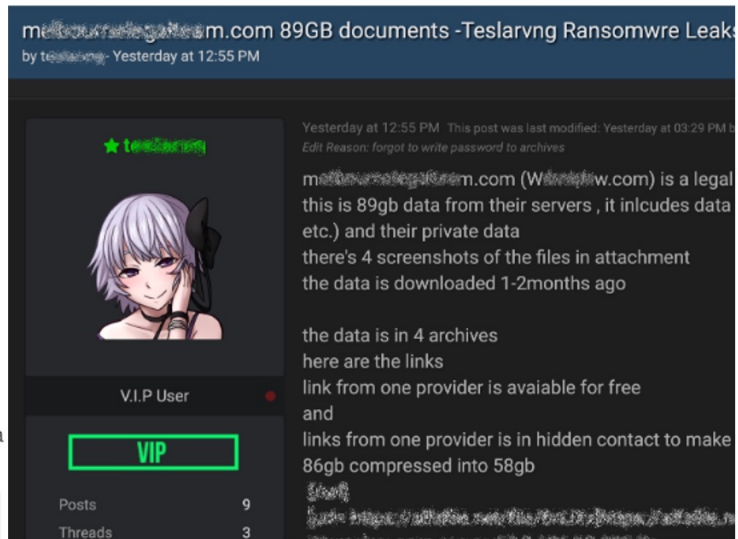
WALL OF SHAME

Here will be permanent list of companies who would like to keep in secret the info leakage, exposing themselves and their customers, partners to even greater risk than a bug-hunting reward!

Financial Holdings, Inc. - Leaked

views: 30311 | Published: 10/06/2021 16:43:51

다크웹 웹페이지



해킹 포럼 게시물



텔레그램

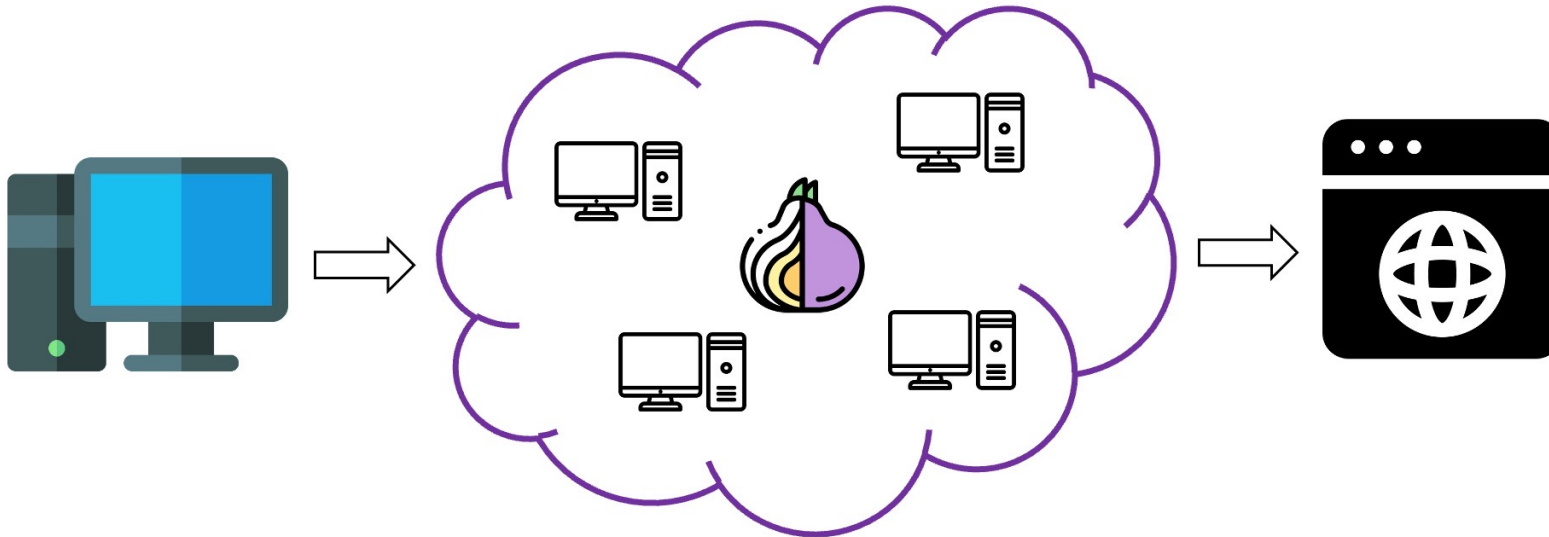
다크웹이란?

인덱싱이 되지 않고, 특수 브라우저로만 접근 가능한 웹



다크웹이란?

인덱싱이 되지 않고, 특수 브라우저로만 접근 가능한 웹



다크웹 웹페이지 유형

Home Page of R...r Leaks site



WALL OF SHAME

Here will be permanent list of companies who would like to keep their data secret, exposing themselves and their customers, partners to a public bug-hunting reward!

Financial Holdings, Inc. - Leaked

Internal UID	Balance	Account type	Card	Country	Our Price	Add to cart
[REDACTED]	927 USD	Personal	Yes (confirmed)	United States	\$ 51	Buy this!
[REDACTED]	705 USD	Personal	Yes (confirmed)	United States	\$ 40	Buy this!
[REDACTED]	693 USD	Personal	Yes (confirmed)	United States	\$ 39	LOCKED
[REDACTED]	1821 USD	Premier	Yes (confirmed)	United States	\$ 96	LOCKED
[REDACTED]	747 EUR	Personal	Yes (confirmed)	France	\$ 50	Buy this!
[REDACTED]	1892 USD	Premier	Yes (confirmed)	United States	\$ 100	Buy this!
[REDACTED]	1777 USD	Premier	Yes (confirmed)	United States	\$ 99	Buy this!
[REDACTED]	1992 USD	Personal	Yes (confirmed)	United States	\$ 104	Buy this!
[REDACTED]	681 USD	Premier	Yes (confirmed)	United States	\$ 39	LOCKED
[REDACTED]	1564					

ECSTASY

wearamsterdam

1gr dutch mdma 84% usa...
84% CHAMPAGNE MDMA CRYSTALS. Nice champagne mdma. Instant high of pure dutch quality... mdma. You will love the supreme high
[Netherlands > Worldwide](#)

★★★★★

4517 49 22

9.00 USD [View](#)

joyinc

1g dutch mdma champagn...
YOU will get natural, purest MDMA with no side effects, no acid reflux, no jaw clenching, only HIGH MOOD... and LOVE We can 100% guarantee
[United States > REG](#)

★★★★★

3614 36 17

50.00 USD [View](#)

wvanoranjie

260 mg- xtc pills
Substance: 280 MG XTC PILLS
Psychic effects: Perception and sounds become more intense. War... oen. In love, with a strong need for
[Netherlands > Worldwide](#)

★★★★★

4205 36 17

0.59 USD [View](#)

vanillasurf

xtc-pills - meetoo
Well Pressed ecstasy sparkling with different closer inspection, will fill with thousands of green
[United Kingdom](#)

★★★★★

3624 33

18.71 USD [View](#)

our price

\$550=0.0213 BTC

our price

\$950=0.0368 BTC

our price

\$650=0.0251 BTC

1-2-3-4-5-6-7-8-9

**Savage Mark II TRR-SR 22 Ir TB
22L-1 Thunderbeast**

Features:
100% titanium construction

1-2-3-4-5-6-7-8

**TSS Custom AK 47 AKMS
Underfolder 24k GOLD!**

Firearm has been finished with a dual coating of nickel under a top coat of 24-KARAT GOLD, and completed

1-2-3-4-5

**S&W 686 Competitor 357
Weighted Barrel**

S&W 686 Competitor .357 Magnum with Weighted Barrel and all Accessories including Aluminum

다크웹 웹페이지

다크웹 웹페이지 유형

Home Page of RLeaks site



WALL OF SHAME

Here will be permanent list of companies who would like to keep their data secret, exposing themselves and their customers, partners to a public bug-hunting reward!

Financial Holdings, Inc. - Leaked

Internal UID	Balance	Account type	Card	Country	Our Price	Add to cart
[REDACTED]	927 USD	Personal	Yes (confirmed)	United States		
[REDACTED]	705 USD	Personal	Yes (confirmed)	United States		
[REDACTED]	693 USD	Personal	Yes (confirmed)	United States		
[REDACTED]	1821 USD	Premier	Yes (confirmed)	United States		
[REDACTED]	747 EUR	Personal	Yes (confirmed)	France		
[REDACTED]	1892 USD	Premier	Yes (confirmed)	United States		
[REDACTED]	1777 USD	Premier	Yes (confirmed)	United States		
[REDACTED]	1992 USD	Personal	Yes (confirmed)	United States		
[REDACTED]	681 USD	Premier	Yes (confirmed)	United States		
[REDACTED]	1564					

sethforprivacy.com

menu

[About Me](#) [Guides](#) [Podcast](#) [Posts](#) [Recommended Privacy Tools](#) [RSS Feed](#)

Donating to FOSS Projects and Pro-Privacy Organizations

2021-04-13 [Updated: 2021-05-02] :: Seth

#Monero #donations #open source #FOSS

Table of Contents

- [Introduction](#)
- [Monero](#)
- [Privacy-preserving services](#)
- [Mobile OSs and Applications](#)
- [Pro-privacy Organizations](#)
- [Others](#)
- [Something Missing?](#)

ECSTASY

wearamsterdam

1gr dutch mdma 84% usa...
84% CHAMPAGNE MDMA CRYSTALS. Nice champagne mdma. Instant high of pure dutch quality... mdma. You will love the supreme high
[Netherlands > Worldwide](#)

★★★★★

4517 49 22

9.00 USD [View](#)

joyinc

1g dutch mdma champagn...
YOU will get natural, purest MDMA with no side effects, no acid reflux, no jaw clenching, only HIGH MOOD... and LOVE We can 100% guarantee
[United States > REG](#)

★★★★★

3614 36 17

50.00 USD [View](#)

wvanbranje

260 mg- xtc pills
Substance: 280 MG XTC PILLS
Psychic effects: Perception and sounds become more intense. War... oen. In love, with a strong need for
[Netherlands > Worldwide](#)

★★★★★

4205 36 17

0.59 USD [View](#)

vanillasurf

xtc-pills - meetoo
Well Pressed ecstasy sparkling with different closer inspection, will fill with thousands of area
[United Kingdom](#)

★★★★★

3624 33

18.71 USD [View](#)

our price
\$550=0.0213 BTC

Savage Mark II TRR-SR 22 Ir TB 22L-1 Thunderbeast

Features:
100% titanium construction

our price
\$950=0.0368

TSS Custom AK 47 A Underfolder 24k GOLD.

Firearm has been finished with a dual coating of nickel under a top coat of 24-KARAT GOLD, and completed

S&W 686 Competitor .357 Magnum with Weighted Barrel and all Accessories including Aluminum

다크웹 웹페이지

다크웹 웹페이지 텍스트

```
dont let your rival kill you ....hire a hitman  
(ID_EMAIL) are you looking for hit man? hire a  
hit man now  
hire a killer | rent a killer | hire a hitman  
| rent a hitman | hire an assassin | rent an a  
ssassin | hire a gunman | rent a gunman | hire  
a murderer | rent a murderer | order the kill  
one of the few legitimate men on the darknet t  
hat you will meet your demand in america europ  
ean europe asia africa middle east australia i  
f you are interested send me a message we can  
talk about the problem you need to solve. cont  
act: email: ID_EMAIL hire a killer | rent a ki  
ller | hire a hitman | rent a hitman | hire an  
assassin | rent an assassin | hire a gunman |  
rent a gunman | hire a murderer | rent a murd  
erer | order the kill  
reply quote ID_NUMBER
```

Category: *Violence*

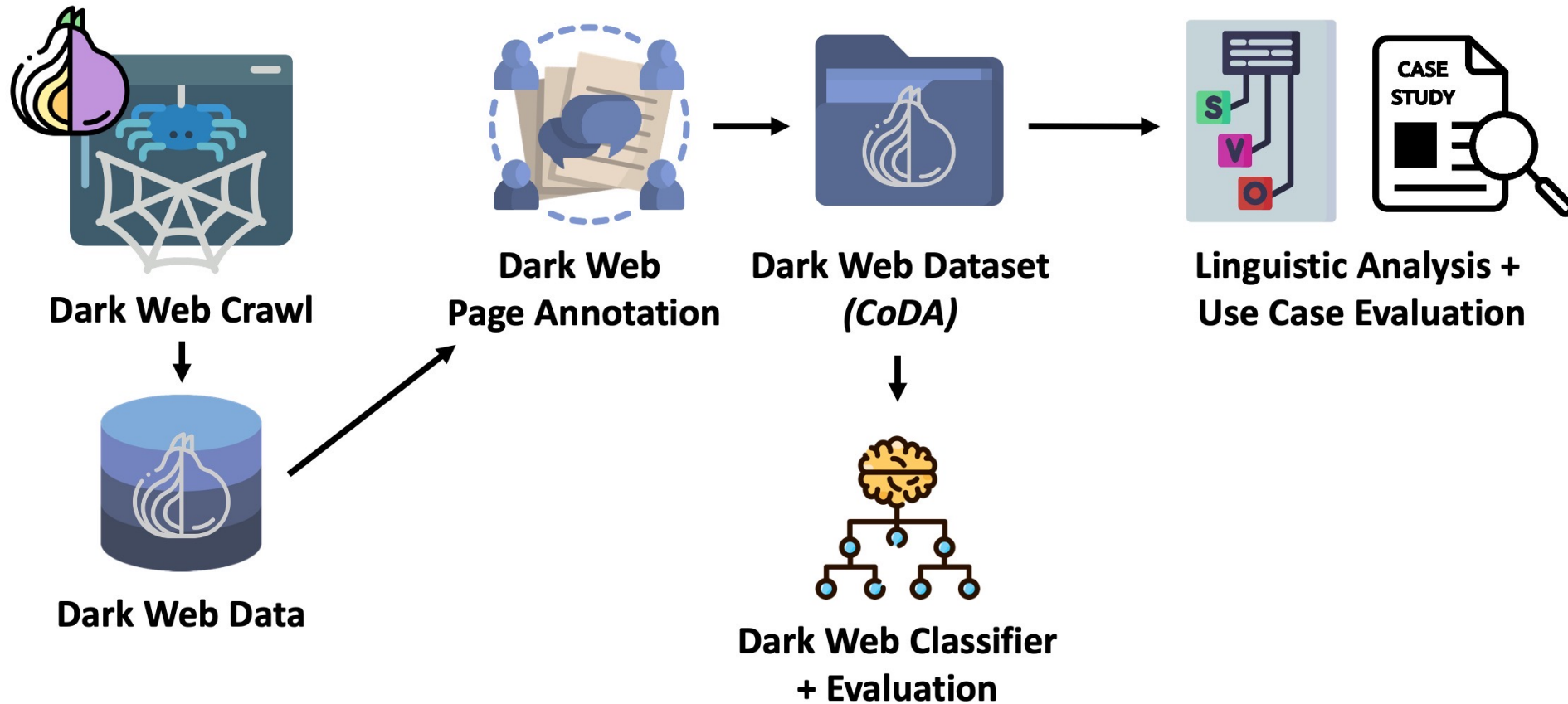
```
visa cards  
attacks  
take ad down  
take site down  
exploits  
doc exploit  
pdf exploit  
snapchat exploit  
instagram media  
telegram hacking  
windows  
websites  
whatsapp hacking  
wifi hacking  
spam  
emails  
php mailers  
servers
```

Category: *Hacking*

```
ID_NUMBER samsung  
ID_NUMBER ipad  
ID_NUMBER macbook  
ID_NUMBER playstation  
ID_CRYPTOMONEY miner  
ID_NUMBER and more...  
christmas packs  
pack - iphone ID_FILESIZE (space gray) + iphon  
e ID_FILESIZE (black) category:pack price: ID_  
CRYPTOMONEY (ID_GENERAL_MONEY) / shipping is  
free  
pack - iphone ID_FILESIZE (silver) + iphone 4s  
ID_FILESIZE (black) category:pack price: ID_  
CRYPTOMONEY (ID_GENERAL_MONEY) / shipping is f  
ree  
pack - iphone ID_NUMBER+ ID_FILESIZE (space gr  
ay) + iphone ID_FILESIZE (space grey) category  
:pack price: ID_CRYPTOMONEY (ID_GENERAL_MONEY  
) / shipping is free
```

Category: *Electronics*

다크웹 웹페이지 분석 Overview



웹페이지 카테고리 정리

Category

Pornography

Drugs

Financial

Gambling

Cryptocurrency

Hacking

Arms / Weapons

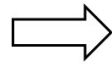
Violence

Electronics

Others

웹페이지 카테고리 정리

```
dont let your rival kill you ...hire a hitman  
(ID_EMAIL) are you looking for hit man? hire a  
hit man now  
hire a killer | rent a killer | hire a hitman  
| rent a hitman | hire an assassin | rent an a  
ssassin | hire a gunman | rent a gunman | hire  
a murderer | rent a murderer | order the kill  
one of the few legitimate men on the darknet t  
hat you will meet your demand in america europ  
ean europe asia africa middle east australia i  
f you are interested send me a message we can  
talk about the problem you need to solve. cont  
act: email: ID_EMAIL hire a killer | rent a ki  
ller | hire a hitman | rent a hitman | hire an  
assassin | rent an assassin | hire a gunman |  
rent a gunman | hire a murderer | rent a murd  
erer | order the kill  
reply quote ID_NUMBER
```



Category

Pornography

Drugs

Financial

Gambling

Cryptocurrency

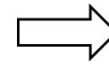
Hacking

Arms / Weapons

Violence

Electronics

Others

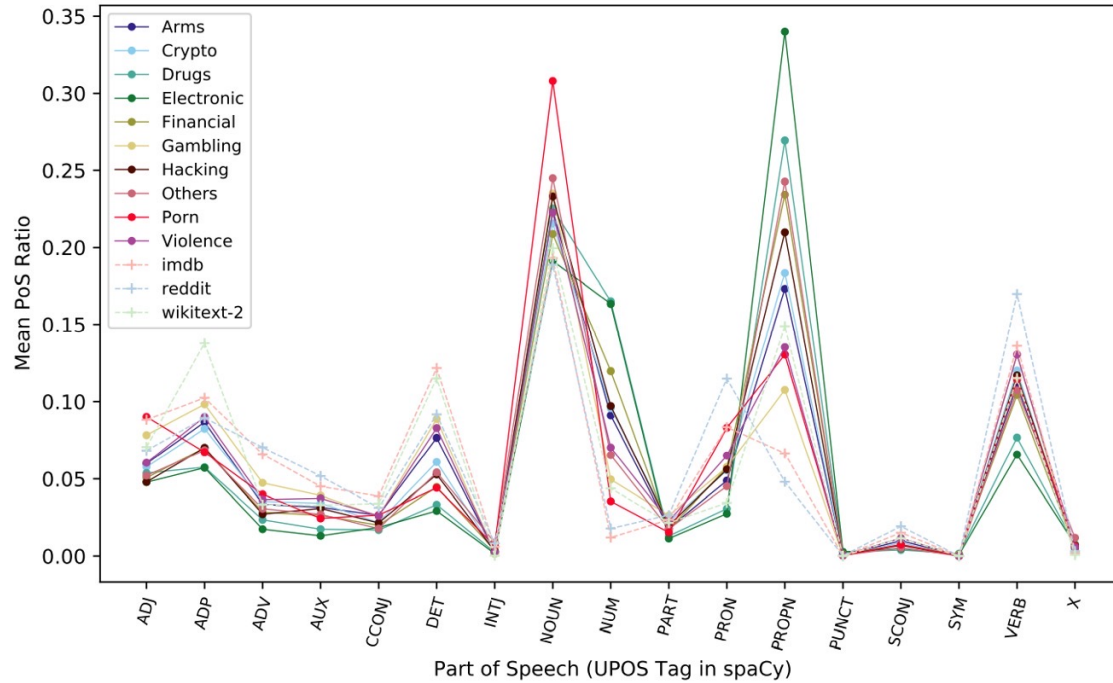


Category: *Violence*

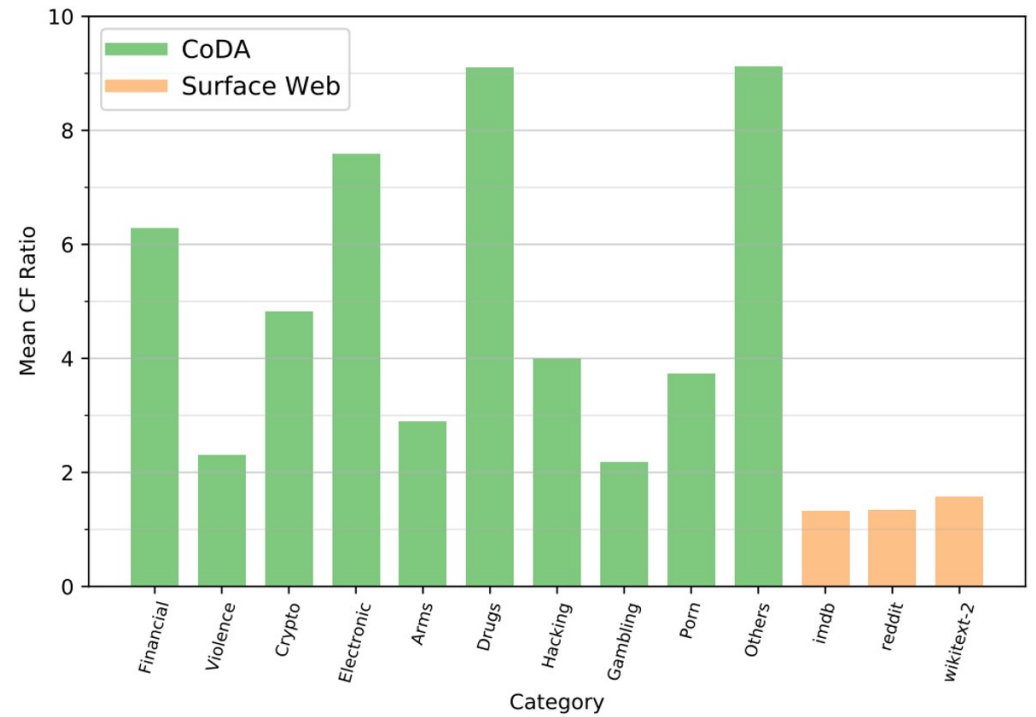
CoDA (Comprehensive Darkweb Annotations)

Category	Document Count	Ratio	Short guideline description
<i>Pornography</i>	1195	12.0%	general / child pornography and other explicit content
<i>Drugs</i>	1172	11.7%	various types of legal / illegal drugs such as medications, steroids, pain killers, viagra, cannabis, hashish, meth, benzos, ecstasies, opioids, and psychedelics
<i>Financial</i>	1003	10.0%	counterfeit / cloned / stolen money or identifications (e.g., bills, credit cards, certificates, passports), money transfers (e.g., PayPal), fiat money, ATM skimmers, magnetic card readers, etc.
<i>Gambling</i>	787	7.87%	any type of gambling, betting, casinos, lotteries, etc.
<i>Cryptocurrency</i>	763	7.63%	cryptocurrency-specific services or technologies such as wallets, generators, mining, laundering, mixing, multiplying, doubling, scamming, and escrow
<i>Hacking</i>	649	6.49%	hacking tools, hacking guides, hacking groups, hacking services, ransomware, malware, exploits, DDoS attacks, cracking, botnet, etc.
<i>Arms / Weapons</i>	599	5.99%	any type of non-lethal / lethal weapons such as guns, ammunition, explosives, knives, missiles, and chemical weapons
<i>Violence</i>	485	4.85%	human trafficking, hitman, kidnapping, poisoning, torture, extortion, sextortion, sex slavery, blackmail, etc.
<i>Electronics</i>	426	4.26%	sale of or information on (stolen / hacked) mobile phones, laptops, tablet computers, etc.
<i>Others</i>	2921	29.2%	all other content that does not fit the above categories, including log-in pages, error messages, etc.
Total	10000	100.0%	

다크웹 언어 분석 (PoS tags)



$$\bar{r}_{cf}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{d \in \mathcal{C}} \frac{N_c(d)}{N_f(d)}$$



다크웹 언어 분석 (TF-IDF)

Rank	Crypto		Drugs		Electronics		Financial	
	Term	TF-IDF	Term	TF-IDF	Term	TF-IDF	Term	TF-IDF
1	ID_NUMBER	0.684	ID_NUMBER	0.864	ID_NUMBER	0.799	ID_NUMBER	0.906
2	ID_CRYPTOMONEY	0.418	ID_WEIGHT	0.255	ID_GENERAL_MONEY	0.291	ID_GENERAL_MONEY	0.220
3	bitcoin	0.354	ID_GENERAL_MONEY	0.249	iphone	0.279	card	0.215
4	ID_TIME	0.233	buy	0.112	ID_FILESIZE	0.184	ID_TIME	0.130
5	ID_BTC_ADDRESS	0.227	weed	0.101	apple	0.142	ID_EMAIL	0.067
6	btc	0.110	pot	0.089	pro	0.111	buy	0.065
7	use	0.071	pill	0.084	ipad	0.110	account	0.051
8	wallet	0.069	online	0.080	macbook	0.092	credit	0.048
9	address	0.068	cocaine	0.073	imac	0.089	paypal	0.048
10	buy	0.068	cannabis	0.068	airpod	0.082	transfer	0.044
11	transaction	0.061	lsd	0.057	ID_TIME	0.079	order	0.039
12	get	0.060	mdma	0.051	buy	0.073	get	0.038
13	blockchain	0.057	drug	0.050	card	0.071	dump	0.038
14	invest	0.049	adderrall	0.050	watch	0.058	cc	0.037
15	ID_GENERAL_MONEY	0.048	viagra	0.049	ipod	0.055	good	0.036
16	coin	0.046	xanax	0.049	case	0.052	money	0.036
17	service	0.046	ID_PERCENTAGE	0.046	gopro	0.051	new	0.036
18	make	0.045	product	0.045	product	0.051	use	0.035
19	ID_PERCENTAGE	0.042	order	0.044	xs	0.051	cvv	0.034
20	double	0.041	quality	0.042	order	0.047	shop	0.034
Rank	Gambling		Hacking		Pornography		Violence	
	Term	TF-IDF	Term	TF-IDF	Term	TF-IDF	Term	TF-IDF
1	ID_NUMBER	0.593	ID_NUMBER	0.792	porno	0.605	ID_NUMBER	0.915
2	casino	0.396	hack	0.364	porn	0.557	ID_TIME	0.173
3	game	0.256	facebook	0.262	video	0.272	kill	0.074
4	br	0.190	ID_TIME	0.167	free	0.220	anonymous	0.066
5	online	0.177	account	0.162	sex	0.154	hitman	0.066
6	slot	0.173	hacker	0.092	girl	0.125	ID_FILENAME	0.065
7	play	0.161	password	0.087	teen	0.123	murder	0.064
8	ID_TIME	0.103	use	0.070	ID_NUMBER	0.105	people	0.063
9	bet	0.098	ID_GENERAL_MONEY	0.068	boy	0.093	like	0.056
10	win	0.096	service	0.059	fuck	0.091	one	0.054
11	poker	0.094	email	0.058	child	0.083	get	0.052
12	page	0.094	software	0.052	cock	0.077	file	0.052
13	get	0.093	ransomware	0.050	cp	0.070	post	0.047
14	free	0.089	download	0.045	young	0.068	site	0.043
15	money	0.087	get	0.044	pussy	0.053	comment	0.043
16	time	0.087	free	0.040	pedo	0.053	ID_NORMAL_URL	0.042
17	card	0.083	instagram	0.039	say	0.051	hire	0.042
18	player	0.081	attack	0.039	mom	0.050	make	0.041
19	good	0.077	hacking	0.039	gay	0.050	say	0.040
20	roulette	0.075	online	0.038	get	0.049	use	0.040

Common Abbreviations

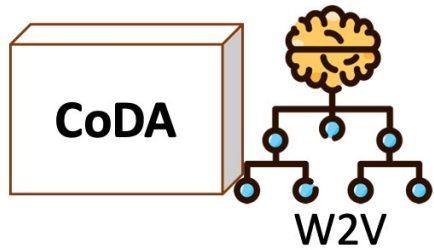
CoDA (Dark Web)

Surface Web

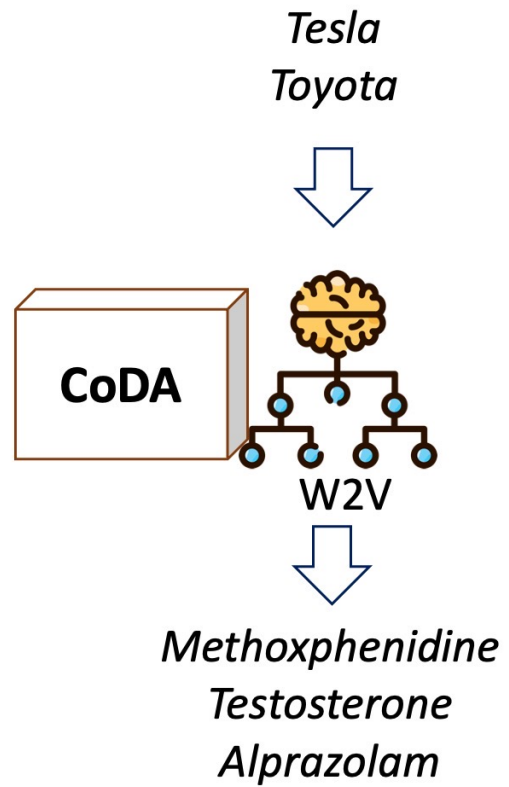
btc	cp	cvc	btw	dvd	idk
cvv	hd	irc	imo	km	lmao
lsd	mg	pthc	mph	pc	st
ssn	vpn	xxx	tv	vs	wtf

Frequently appearing abbreviations for each web domain

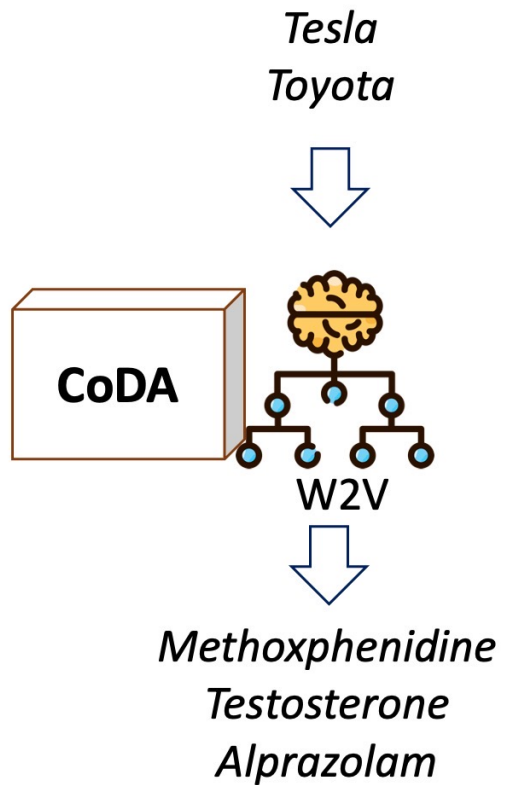
다크웹 내 유사 단어 추출 (Word2Vec)




다크웹 내 유사 단어 추출 (Word2Vec)



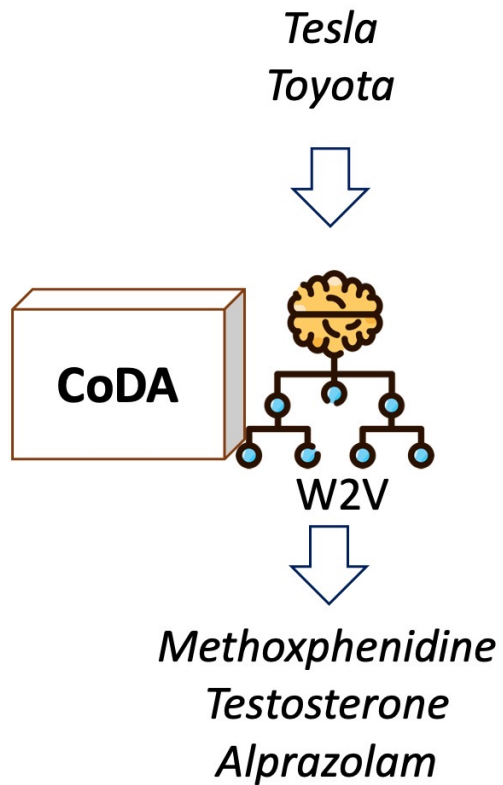
다크웹 내 유사 단어 추출 (Word2Vec)



Viewing desktop version. [Switch to mobile](#)

Toyota	Sold as: MDMA	ID: 3809
	<p>ID: 3809</p> <p>Name: Toyota</p> <p>Other Names:</p> <p>GC/MS: ❗</p> <ul style="list-style-type: none">• MDMA : 162.40 mg ✅	<p>Test Date: Sep 01, 2015</p> <p>Pub. Date: Sep 08, 2015</p> <p>Src Location: Zurich, Switzerland</p> <p>Submitter: Zurich, --</p> <p>Loc: Switzerland</p> <p>Color: Red / Pink / Speckled</p>

다크웹 내 유사 단어 추출 (Word2Vec)



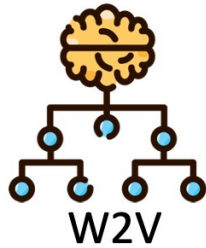
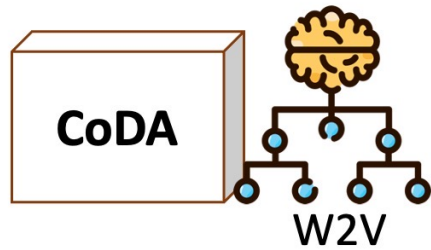
Orange 'Tesla' drug warning after children hospitalised

6 February 2018

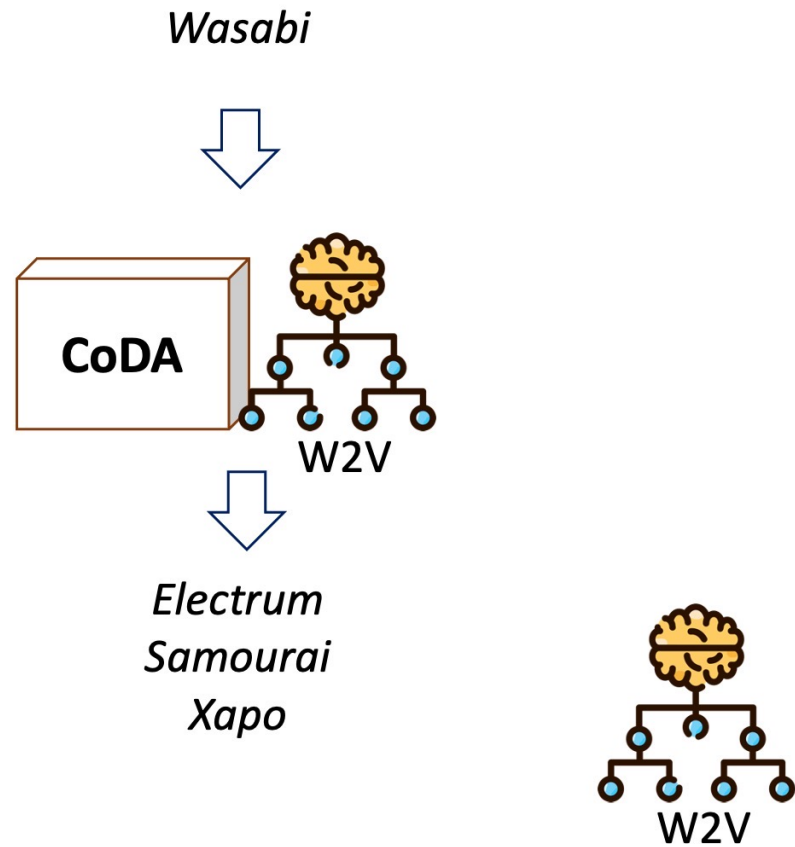


ID: 3809	Test Date: Sep 01, 2015
Name: Toyota	Pub. Date: Sep 08, 2015
Other Names:	Src Location: Zurich, Switzerland
GC/MS: !	Submitter: Zurich, --
• MDMA : 162.40 mg !	Loc: Switzerland
	Color: Red / Pink / Speckled

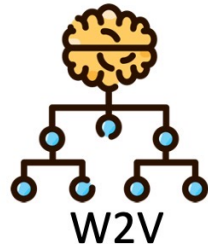
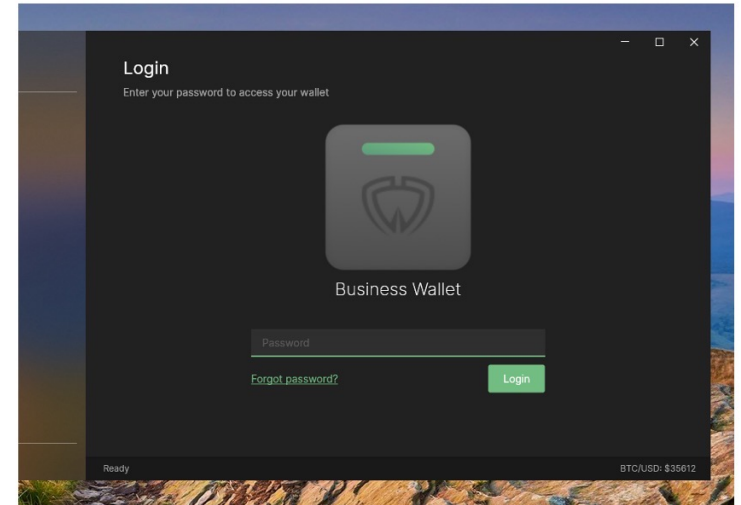
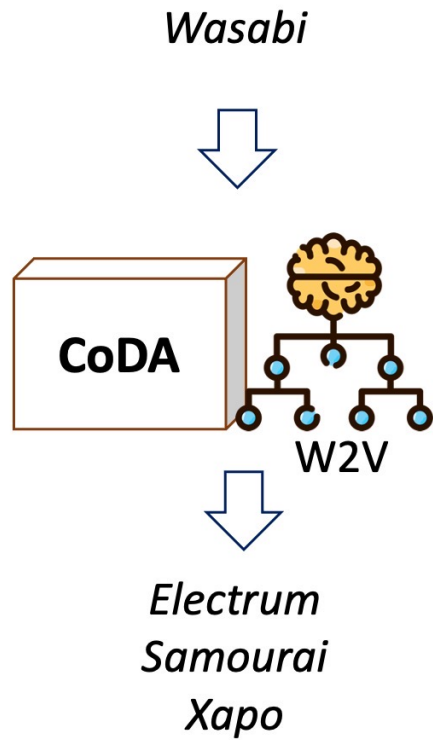
다크웹 내 유사 단어 추출 (Word2Vec)



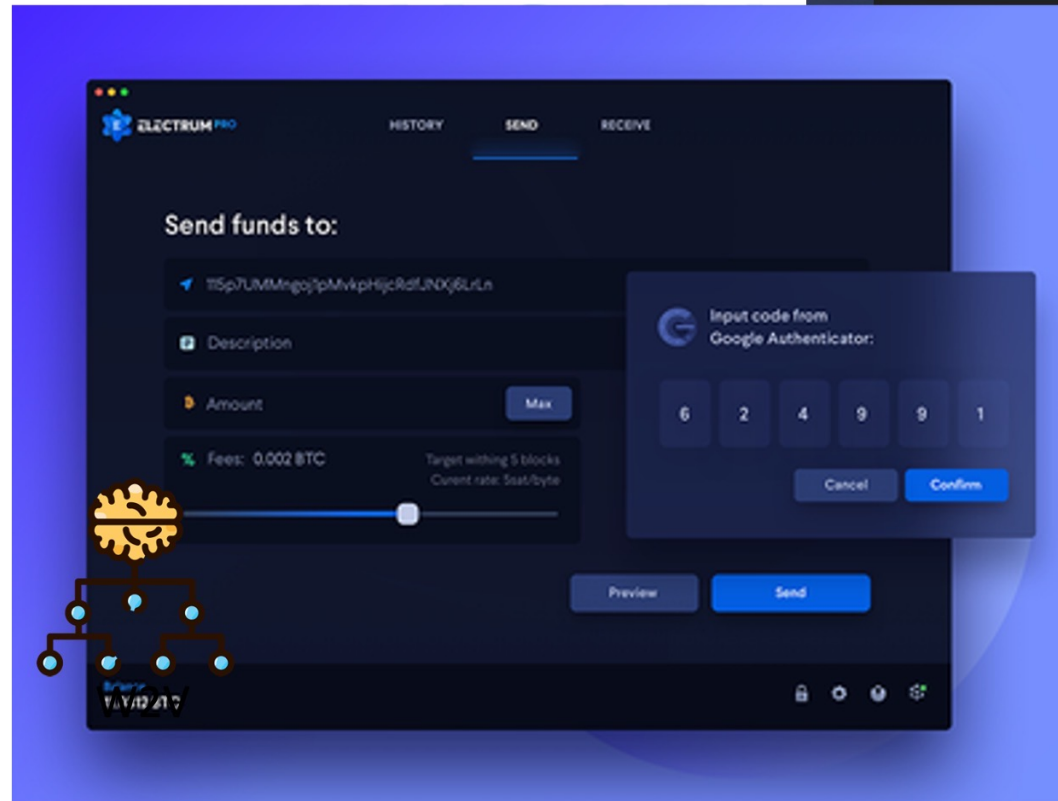
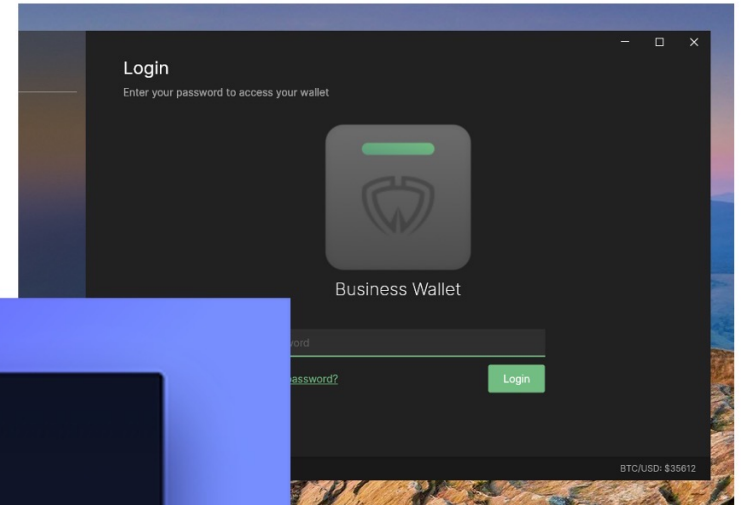
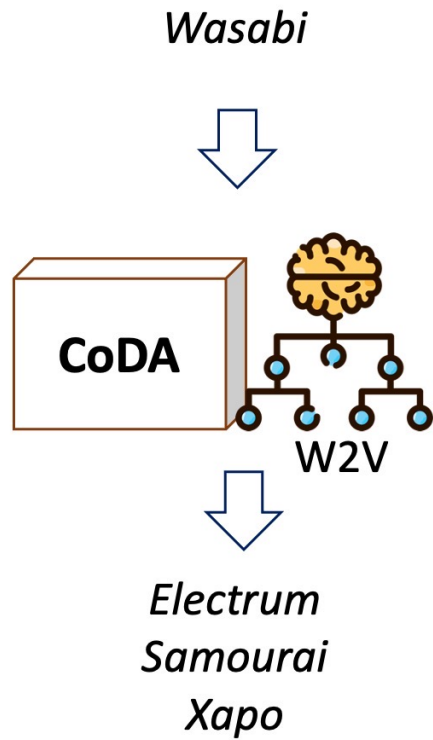
다크웹 내 유사 단어 추출 (Word2Vec)



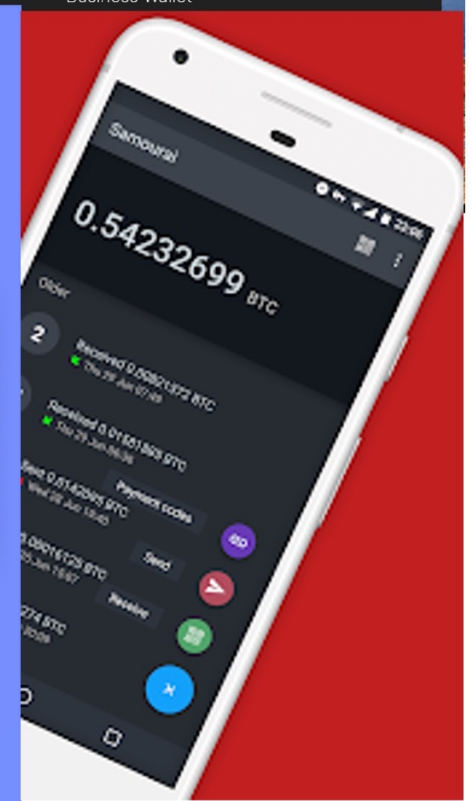
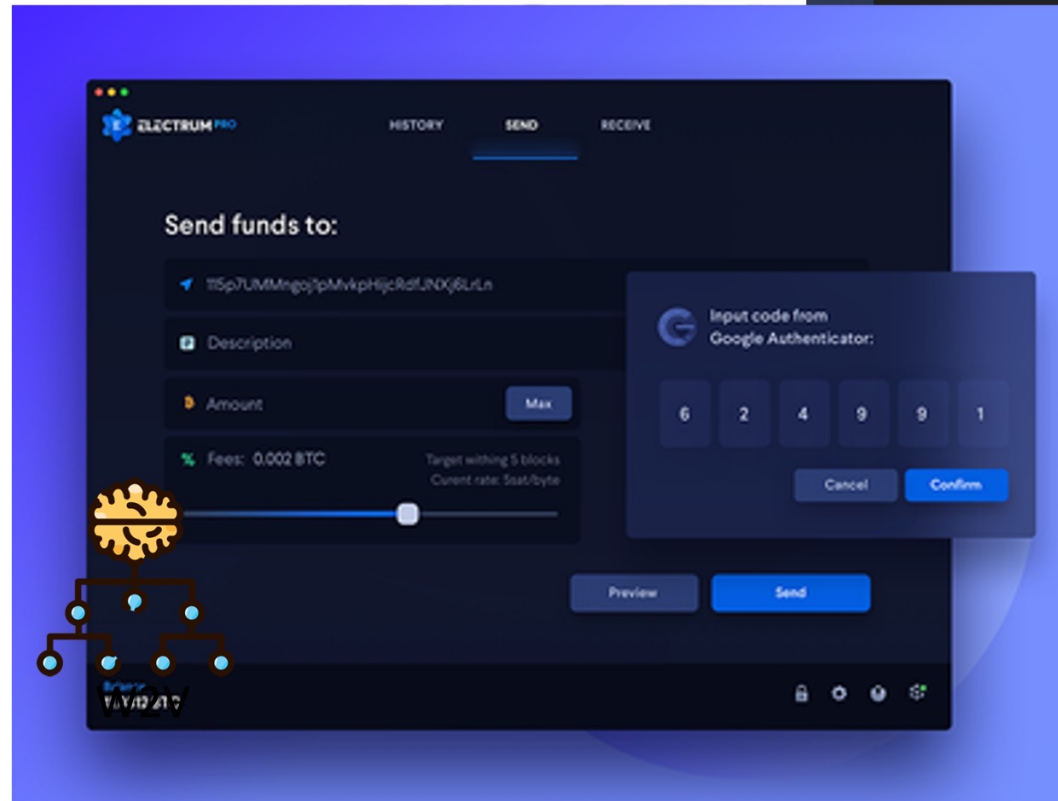
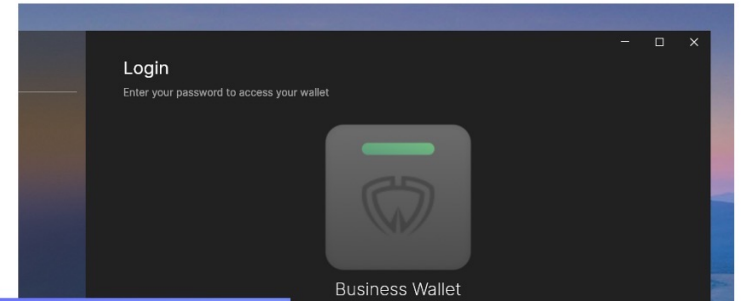
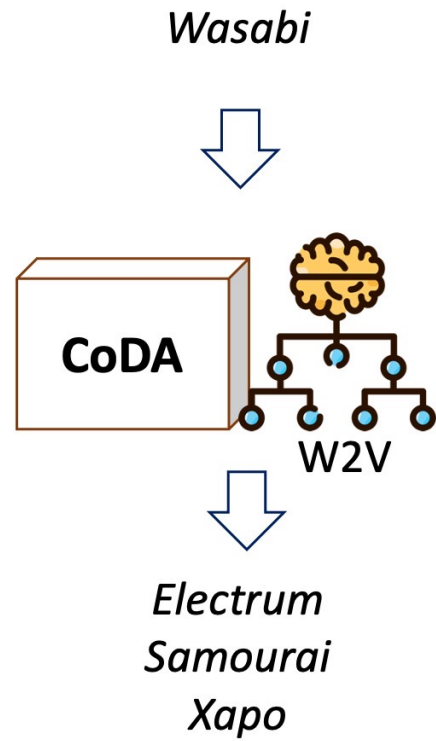
다크웹 내 유사 단어 추출 (Word2Vec)



다크웹 내 유사 단어 추출 (Word2Vec)



다크웹 내 유사 단어 추출 (Word2Vec)

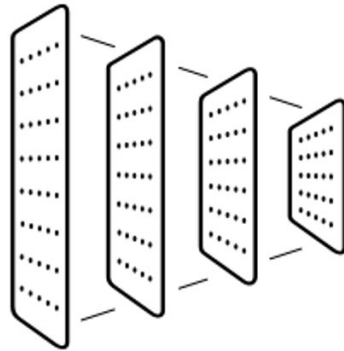


Traditional ML vs. CNN vs. Transformer-based LM



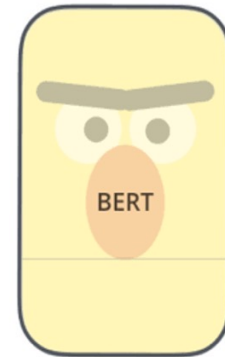
Multi-class SVM + Bag of words

- 단어 분포를 통한 비교
- 키워드 위주



CNN + GloVe embedding

- GloVe: 단어당 의미를 벡터화
- CNN: 벡터끼리의 위치를 고려하는 모델
- 단어의 의미를 뽑아 문맥을 이해



BERT

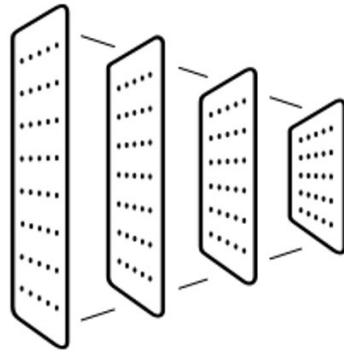
- 문맥을 통째로 벡터화하는 모델
- 문맥을 고려해 단어 의미 이해

Traditional ML vs. CNN vs. Transformer-based LM



Multi-class SVM + Bag of words

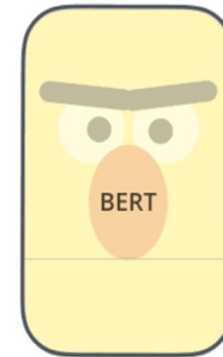
- 단어 분포를 통한 비교
- 키워드 위주



CNN + GloVe embedding

- GloVe: 단어당 의미를 벡터화
- CNN: 벡터끼리의 위치를 고려하는 모델
- 단어의 의미를 뽑아 문맥을 이해

	Precision	Recall	F1-score
SVM	91.59	91.17	91.19
CNN	88.08	87.30	87.23
BERT	92.51	92.50	92.49



BERT

- 문맥을 통째로 벡터화하는 모델
- 문맥을 고려해 단어 의미 이해

Traditional ML vs. CNN vs. Transformer-based LM



Multi-class SVM + Bag of words

- 단어 분포를 통한 비교
- 키워드 위주

	Precision	Recall	F1-score
→ SVM	91.59	91.17	91.19
CNN	88.08	87.30	87.23
BERT	92.51	92.50	92.49

카테고리별 대표하는 키워드가 많이 있다

Traditional ML vs. CNN vs. Transformer-based LM



Multi-class SVM + Bag of words

- 단어 분포를 통한 비교
- 키워드 위주

	Precision	Recall	F1-score
→ SVM	91.59	91.17	91.19
CNN	88.08	87.30	87.23
BERT	92.51	92.50	92.49

카테고리별 대표하는 키워드가 많이 있다

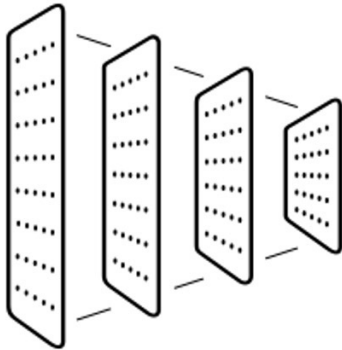


Need **PayPal Accounts?**

Your one stop shop for quickly buying hacked paypal accounts.

Bitcoin Only. Fully Anonymous.

Traditional ML vs. CNN vs. Transformer-based LM



CNN + GloVe embedding

- GloVe: 단어당 의미를 벡터화
- CNN: 벡터끼리의 위치를 고려하는 모델
- 단어의 의미를 뽑아 문맥을 이해

	Precision	Recall	F1-score
SVM	91.59	91.17	91.19
→ CNN	88.08	87.30	87.23
BERT	92.51	92.50	92.49

단어의 의미를 뽑는 것이 한계적 (은어 등)

Common Abbreviations					
CoDA (Dark Web)			Surface Web		
btc	cp	cvc	btw	dvd	idk
cvv	hd	irc	imo	km	lmao
lsd	mg	pthe	mph	pc	st
ssn	vpn	xxx	tv	vs	wtf

Traditional ML vs. CNN vs. Transformer-based LM



BERT

- 문맥을 통째로 벡터화하는 모델
- 문맥을 고려해 단어 의미 이해

	Precision	Recall	F1-score
SVM	91.59	91.17	91.19
CNN	88.08	87.30	87.23
→ BERT	92.51	92.50	92.49

문맥을 고려하기 때문에 키워드 구분이 가능
BERT는 다크웹 데이터를 학습한 적이 없지만,
어느 정도 유추 능력이 있음

Traditional ML vs. CNN vs. Transformer-based LM



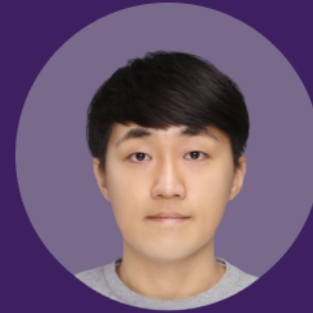
BERT

- 문맥을 통째로 벡터화하는 모델
- 문맥을 고려해 단어 의미 이해

	Precision	Recall	F1-score
SVM	91.59	91.17	91.19
CNN	88.08	87.30	87.23
→ BERT	92.51	92.50	92.49

문맥을 고려하기 때문에 키워드 구분이 가능
BERT는 다크웹 데이터를 학습한 적이 없지만,
어느 정도 유추 능력이 있음

-> BERT가 다크웹 데이터를 학습한다면...?



정진우

AI팀 총괄 팀장

AI 언어모델 'DarkBERT' 개발 리드

The information contained in this document is proprietary and confidential.
If you are not the intended recipient, please note that any use or circulation of this document may be cause for legal action.

제3회 S2W WITH 웨비나

Session 3. Technology

정진우, Head of AI Team

익명 채널 및
위협 콘텐츠에 특화된
다크웹 언어모델 'DarkBERT'



- **KAIST 전산학부 박사 (자연어처리 전공) (2018) & 박사후 연구원 (2019)**
 - 뉴스 기사로부터 사건의 시공간 속성 정보 추출
 - 생물학 문헌에서 미생물-서식지 관계 정보 자동 추출
- **삼성전자 종합기술원 머신러닝랩 전문연구원 (2019~2020)**
 - 모바일 환경에서 구동되는 자동 동시통역 알고리즘 개발
 - 화학 문헌에서 물질 합성 경로 및 물성 정보 자동 추출 연구
- **S2W AI 팀장 (2020~현재)**
 - 딥/다크웹 내 위협 콘텐츠(해킹, 음란물, 마약 등) 자동 분류 및 추출 기술
 - 온라인 텍스트로부터 사이버 보안 정보 및 위협 인텔리전스(악성코드, 취약점, 유출) 자동 추출 기술

DarkBERT: 다크웹을 학습한 AI 언어모델

<https://arxiv.org/abs/2305.08596>

DarkBERT: A Language Model for the Dark Side of the Internet

Youngjin Jin¹ Eugene Jang² Jian Cui² Jin-Woo Chung² Yongjae Lee² Seungwon Shin¹

¹KAIST, Daejeon, South Korea

²S2W Inc., Seongnam, South Korea

¹{*ijinjin, claude*}@kaist.ac.kr

²{*genesith, geeoon19, jwchung, lee*}@s2w.inc

Abstract

Recent research has suggested that there are clear differences in the language used in the Dark Web compared to that of the Surface Web. As studies on the Dark Web commonly require textual analysis of the domain, language models specific to the Dark Web may provide valuable insights to researchers. In this work, we introduce DarkBERT, a language model pretrained on Dark Web data. We describe the steps taken to filter and compile the text data used to train DarkBERT to combat the extreme lexical and structural diversity of the Dark Web that may be detrimental to build-

To handle the ever-changing landscape of modern cyber threats, cybersecurity experts and researchers have started to employ natural language processing (NLP) methods. Gaining evidence-based knowledge such as indicators of compromise (IOC) to mitigate emerging threats is an integral part of modern cybersecurity known as *cyber threat intelligence* (CTI) (Liao et al., 2016; Bromiley, 2016), and modern NLP tools have become an indispensable part of CTI research. As such, the use of NLP techniques has also been extended to the Dark Web (Jin et al., 2022; Yoon et al., 2019; Choshen et al., 2019; Al Nabki et al., 2017; Al-

- S2W x KAIST
- ACL 2023 정규 논문 채택
- 자연어처리 분야 top-tier 학술대회
 - NAACL에 이어 2년 연속 채택
- 2023년 7월 토론토 개최 예정

Dark Web ChatGPT Unleashed: Meet DarkBERT

By Francisco Pires published 29 days ago

A language model trained on the fringes of the dark web... for science.



The ONLY Model Trained on the DARK WEB (Dark Bert)

조회수 6.5천회 · 1일 전

Matthew Berman

In this video, we look at DarkBERT, the titled ...

새 동영상 4K

DarkBERT: A Language Model for the Dark Side of the Internet
Youngjin Jin¹ Eugene Jang² Jian Cui² Jin-Woo Chung² Yongjae Lee² Seungwon Shin¹

¹KAIST, Daejeon, South Korea

²S2W Inc., Seongnam, South Korea

¹{ijinjin, claude}@kaist.ac.kr

²{genesith, geeoon19, jwchung, lee}@s2w.inc

Abstract

Recent research has suggested that there are clear differences in the language used in the Dark Web compared to that of the Surface Web. As studies on the Dark Web commonly require textual analysis of the domain, language models specific to the Dark Web may provide valuable insights to researchers. In this work,

To handle the ever-changing landscape of modern cyber threats, cybersecurity experts have started to employ natural language processing (NLP) methods. Gained based knowledge such as indicators of compromise (IOC) to mitigate emerging threats is a great part of modern cybersecurity.



Hackread.com

@HackRead

The Twitter account of the most reliable cybersecurity news platform brings exclusive dark web, tech, hacking news, and much more. Contact admin@hackread.com.

미디어 및 뉴스 업체 United Kingdom

hackread.com 2011년 11월에 가입함

203 팔로잉 117,094 팔로워

트윗 답글 미디어 마음에 들어요

Hackread.com @HackRead · 1시

AI for cybersecurity on the Dark Web - #DarkBERT is based on the RoBERTa architecture and leverages the power of AI to navigate the #DarkWeb.

Read: hackread.com/darkbert-ai-da..

#Security #CyberSecurity #DarkNet #Ransomware



@Moongie_Phowngi · 11h

Already done abusing Chat-gpt Now its time to abuse #DarkBERT

Andrey Lukyanenko @AndLukyane · May 18

#DarkBERT, a new language model tackling the Dark Web's unique linguistic challenges! My review dives deep into its creation and potential. #AI #NLP #Cybersecurity

arxiv.org/abs/2305.08596
andlukyane.com/blog/paper-rev...

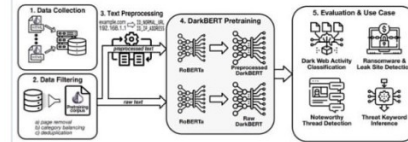


Figure 1: Illustration of the DarkBERT pretraining process and the various use case scenarios for evaluation.

Table 9: Dark Web data page classification statistics. We use the categorization method following Jin et al. (2022).

Category	Page Count	Total Size (GB)	Average Size per Page (in KB)
Personality	2,267,020	9.16 GB	4.04 KB
Financial	507,403	1.74 GB	3.43 KB
Political	407,017	2.16 GB	5.30 KB
Geography	414	1.34 GB	3.24 KB
Hacking	862,162	1.74 GB	2.02 KB
AI/ML/Response	48,616	0.14 GB	2.76 KB
Violence	323,124	1.23 GB	3.74 KB
Education	481,126	0.89 GB	1.85 KB
Total	3,447,038	20.34 GB	

Table 10: Statistics for the pretraining corpus used for DarkBERT.

Category	Page Count	Total Size	Disproportionate Size	Total Relative Size
Personality	2,267,020	9.16 GB	12.0%	40.8%
Financial	507,403	1.74 GB	1.8%	8.5%
Political	407,017	2.16 GB	2.2%	10.6%
Geography	414	1.34 GB	0.0%	6.5%
Hacking	862,162	1.74 GB	1.8%	8.5%
AI/ML/Response	48,616	0.14 GB	0.1%	0.7%
Violence	323,124	1.23 GB	1.3%	6.0%
Education	481,126	0.89 GB	0.9%	4.3%
Total	3,447,038	20.34 GB	13.6%	100%

Table 3: Dark Web activity classification evaluation results. Boldface indicates best performance.

Dataset	Model	Precision	Recall	F1 score
DarkBERT	DarkBERT	0.88	0.88	0.88
DarkBERT	RoBERTa	0.78	0.78	0.78
DarkBERT	GPT-4	0.78	0.78	0.78

Table 5: Noteworthy and non-noteworthy threads in the dataset.

Language Model	Summarizability
DarkBERT	fills, ignores, lation, open, change, change
RoBERTa	fills, ignores, lation, open, change, change
GPT-4	fills, ignores, lation, open, change, change

Table 6: Fill-mask task results. Precise words related to drugs and general words.

Language Model	Summarizability
DarkBERT	fills, ignores, lation, open, change, change
RoBERTa	fills, ignores, lation, open, change, change
GPT-4	fills, ignores, lation, open, change, change

Table 7: Quantitative performance of word inference. Precision at k (P@k) varying at increments of 10.

Language Model	Summarizability
DarkBERT	fills, ignores, lation, open, change, change
RoBERTa	fills, ignores, lation, open, change, change
GPT-4	fills, ignores, lation, open, change, change



Security News Bot

Follow

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

Security News Bot

오늘의 주제

언어모델이란 무엇인가?

다크웹을 위한 언어모델을 왜 만들었는가?

어디에 활용할 수 있는가?

언어모델(Language Model)이란?

사람의 언어를 이해하고 있는 AI 프로그램

≡ 문장 내 특정 자리에 어떤 단어가 등장할지 확률적으로 알고 있는 프로그램

- Causal Language Model (예: GPT)
 - 현재 문장에서 마지막에 이어서 등장할 단어를 잘 맞추도록 학습
 - Paris is the capital of ... → **France**
- Masked Language Model (예: BERT)
 - 현재 문장에서 중간에 등장할 단어를 잘 맞추도록 학습
 - Paris is the [MASK] of France. → **capital**

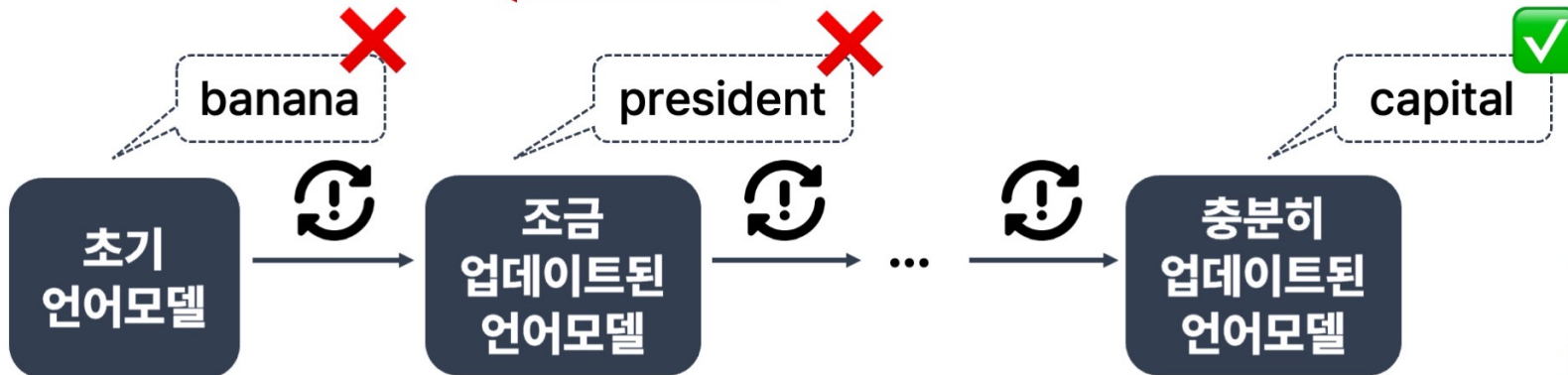
언어모델(Language Model)이란?

Masked Language Model (MLM)

Paris is the capital of France.

↓ 전체 단어 중 랜덤으로 15%를 마스킹

Paris is the **[MASK]** of France.



주변 문맥을 이용하여 마스킹된 단어를 스스로 예측하는 훈련을 반복하여 사람의 언어와 지식을 이해



Google BERT: 영문 위키피디아 25억 & 도서 8억 단어로 구성된 텍스트로 4일 동안 훈련

언어모델(Language Model)이란?

Mask token: <mask> 📌

Paris is the <mask> of France.

BERT

```
▼ [
  ▼ 0 : {
    | "capital" : 0.9969369173049927
    | }
  ▼ 1 : {
    | "heart" : 0.0005914904759265482
    | }
  ▼ 2 : {
    | "center" : 0.0004378768790047616
    | }
  ▼ 3 : {
    | "centre" : 0.00033783656544983387
    | }
  ▼ 4 : {
    | "city" : 0.00026996011729352176
    | }
]
```

Mask token: <mask> 📌

The most important thing in life is <mask>.

BERT

```
▼ [ 📌
  ▼ 0 : {
    | "money" : 0.2775939404964447
    | }
  ▼ 1 : { 📌
    | "love" : 0.14565502107143402 📌
    | }
  ▼ 2 : {
    | "family" : 0.047264765948057175
    | }
  ▼ 3 : {
    | "you" : 0.03244999423623085
    | }
  ▼ 4 : {
    | "safety" : 0.024119790643453598
    | }
]
```

언어모델의 활용

언어모델은 사람의 언어를 이해하고 방대한 지식을 사전 학습(pre-training)하여 여러 자연어처리 문제 해결 능력이 뛰어남



도메인 특화 언어모델

언어모델을 이용하여 여러 문제를 잘 푸는 건 좋은데
특정 도메인의 여러 문제를 더 잘 풀기 위한 언어모델을 만들 수 있을까?

Mask token: <mask> 🚩

Paris, Tokyo, and Washington are types of <mask>.

BERT

```
[
  0: {
    "cities": 0.2712933421134949
  },
  1: {
    "capitals": 0.05180267617106438
  },
  2: {
    "capital": 0.03752797096967697
  },
  3: {
    "airports": 0.029108094051480293
  },
  4: {
    "paris": 0.02577260695397854
  }
]
```

Mask token: <mask> 🚩

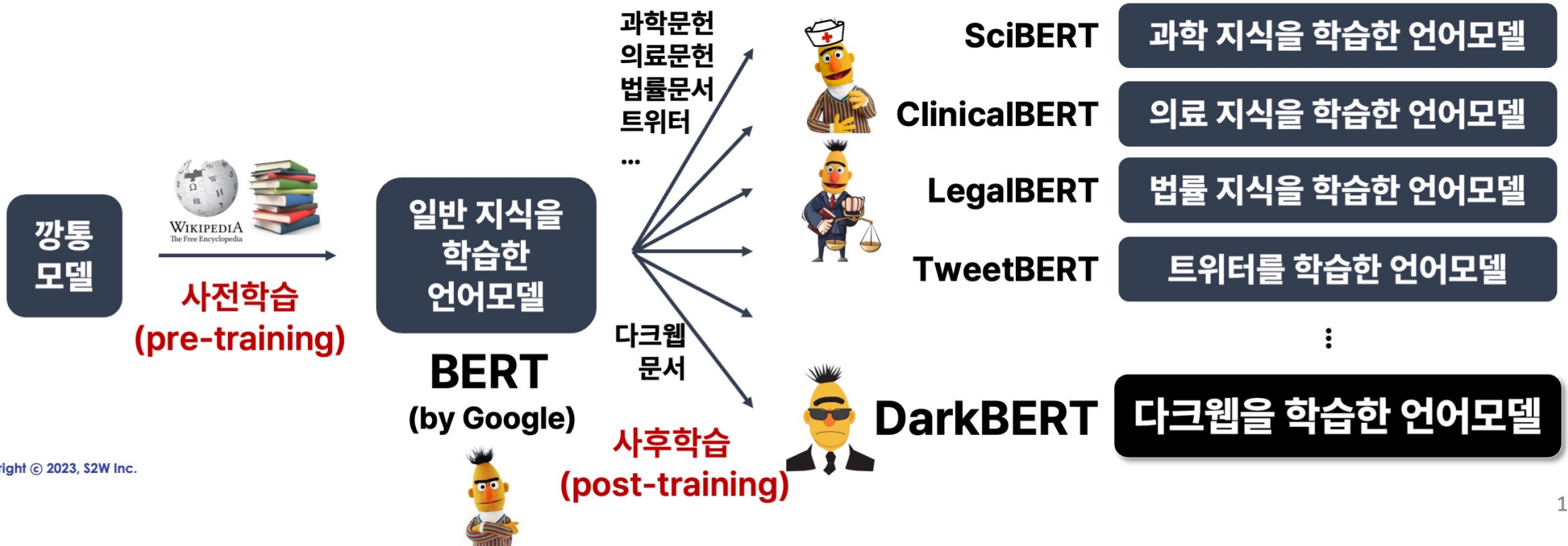
LockBit, RagnarLocker, and REvil are types of <mask>.

BERT

```
[
  0: {
    "game": 0.02942637726664543
  },
  1: {
    "traps": 0.01350460760295391
  },
  2: {
    "cards": 0.012321463786065578
  },
  3: {
    "games": 0.010584800504148006
  },
  4: {
    "dice": 0.010569826699793339
  }
]
```

도메인 특화 언어모델

언어모델을 이용하여 여러 문제를 잘 푸는 건 좋는데
특정 도메인의 여러 문제를 더 잘 풀기 위한 언어모델을 만들 수 있을까?
기존 언어모델을 가져와서 **특정 도메인 데이터**를 추가로 학습 ("**Post-training**")



다크웹을 학습한 언어모델 DarkBERT

마스킹된 단어 예측하기

LockBit, RagnarLocker, and REvil are types of **<mask>**.

BERT

```
▼ [
  ▼ 0 : {
    "game" : 0.02942637726664543
  }
  ▼ 1 : {
    "traps" : 0.01350460760295391
  }
  ▼ 2 : {
    "cards" : 0.012321463786065578
  }
  ▼ 3 : {
    "games" : 0.010584800504148006
  }
  ▼ 4 : {
    "dice" : 0.010569826699793339
  }
]
```

DarkBERT-Raw

```
▼ [
  ▼ 0 : {
    "ransomware" : 0.8941881656646729
  }
  ▼ 1 : {
    "malware" : 0.011506946757435799
  }
  ▼ 2 : {
    "attacks" : 0.003964886534959078
  }
  ▼ 3 : {
    "encryption" : 0.0032116961665451527
  }
  ▼ 4 : {
    "crypto" : 0.0027745813131332397
  }
]
```

다크웹을 학습한 언어모델 DarkBERT

마스킹된 단어 예측하기

Selling high quality <mask>. DM for more info.

BERT

```
▼ [
  ▼ 0 : {
    | "products" : 0.18069440126419067
    | }
  ▼ 1 : {
    | "software" : 0.062272243201732635
    | }
  ▼ 2 : {
    | "books" : 0.034454114735126495
    | }
  ▼ 3 : {
    | "product" : 0.028478382155299187
    | }
  ▼ 4 : {
    | "materials" : 0.02108207531273365
    | }
]
```

DarkBERT-Raw

```
▼ [
  ▼ 0 : {
    | "pills" : 0.09229528158903122
    | }
  ▼ 1 : {
    | "hash" : 0.04963463917374611
    | }
  ▼ 2 : {
    | "drugs" : 0.04293271154165268
    | }
  ▼ 3 : {
    | "product" : 0.03319595754146576
    | }
  ▼ 4 : {
    | "MDMA" : 0.0313962884247303
    | }
]
```

다크웹을 학습한 언어모델 DarkBERT

마스킹된 단어 예측하기

Want to buy high quality <mask>.

BERT

```
▼ [  
  0 : {  
    "products" : 0.04338027536869049  
  }  
  1 : {  
    "clothes" : 0.03938606008887291  
  }  
  2 : {  
    "shoes" : 0.03168926388025284  
  }  
  3 : {  
    "food" : 0.030154826119542122  
  }  
  4 : {  
    "books" : 0.028179103508591652  
  }  
]
```

DarkBERT-Raw

```
▼ [  
  0 : {  
    "ammo" : 0.036419112235307693  
  }  
  1 : {  
    "etc" : 0.02162759006023407  
  }  
  2 : {  
    "gun" : 0.0207974836230278  
  }  
  3 : {  
    "firearms" : 0.018258778378367424  
  }  
  4 : {  
    "guns" : 0.01799386367201805  
  }  
]
```

다크웹을 학습한 언어모델 DarkBERT

마스킹된 단어 예측하기

Donald Trump is a type of <mask>.

BERT

```
▼ [
  ▼ 0 : {
    | "president" : 0.22427412867546082
    | }
  ▼ 1 : {
    | "trump" : 0.140208438038826
    | }
  ▼ 2 : {
    | "politician" : 0.05716140940785408
    | }
  ▼ 3 : {
    | "personality" : 0.026182906702160835
    | }
  ▼ 4 : {
    | "man" : 0.01624278537929058
    | }
  ]
```

DarkBERT-Raw

```
▼ [
  ▼ 0 : {
    | "racist" : 0.18189860880374908
    | }
  ▼ 1 : {
    | "nationalist" : 0.03924461826682091
    | }
  ▼ 2 : {
    | "Trump" : 0.030204061418771744
    | }
  ▼ 3 : {
    | "evil" : 0.027612563222646713
    | }
  ▼ 4 : {
    | "person" : 0.02696448378264904
    | }
  ]
```

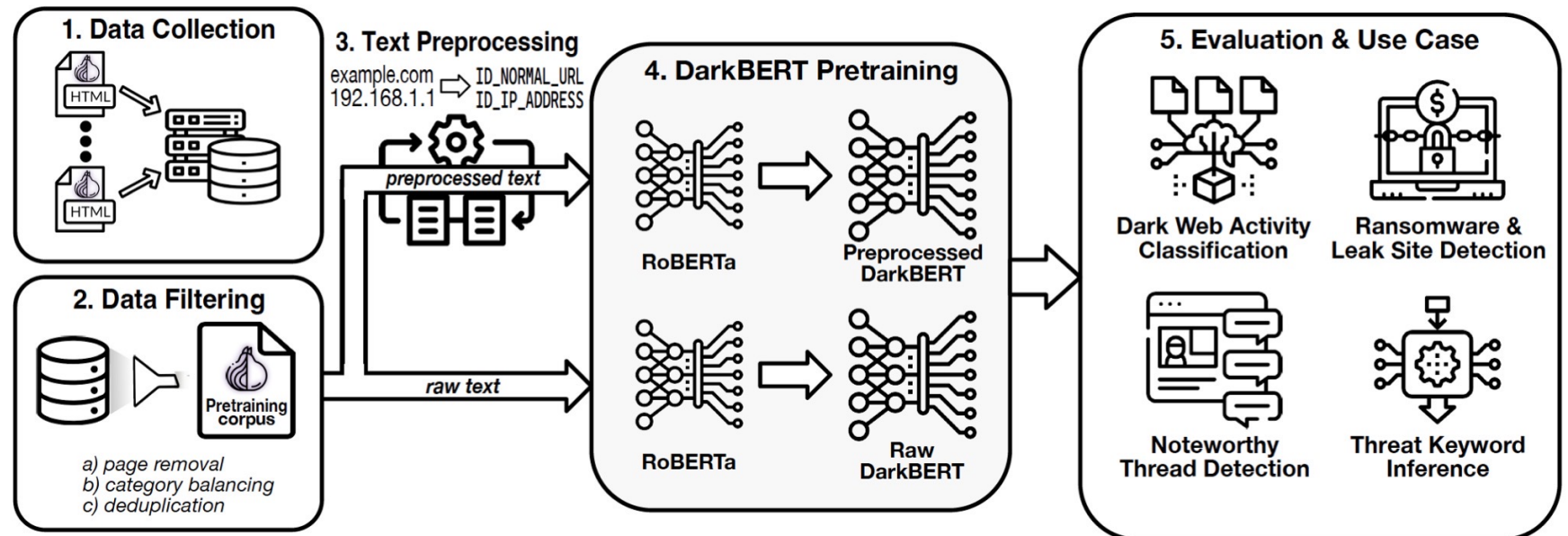
다크웹을 학습한 언어모델 DarkBERT

- 학습 방법

- 실험

1. 다크웹 페이지 주제 분류
2. 랜섬웨어 유출 사이트 탐지
3. 해킹 포럼 중요 게시물 분류
4. 위협 키워드 추론

- 활용 사례 및 의의



다크웹 학습 과정

BERT 대신 RoBERTa를 초기 모델로 사용하여 다크웹 학습 시작

- **RoBERTa**: BERT의 개량 모델 (by Meta)
 - 더 많은 데이터에, 더 오래 학습된 모델
 - 문장 경계 및 선후 관계를 구분하지 않은 모델

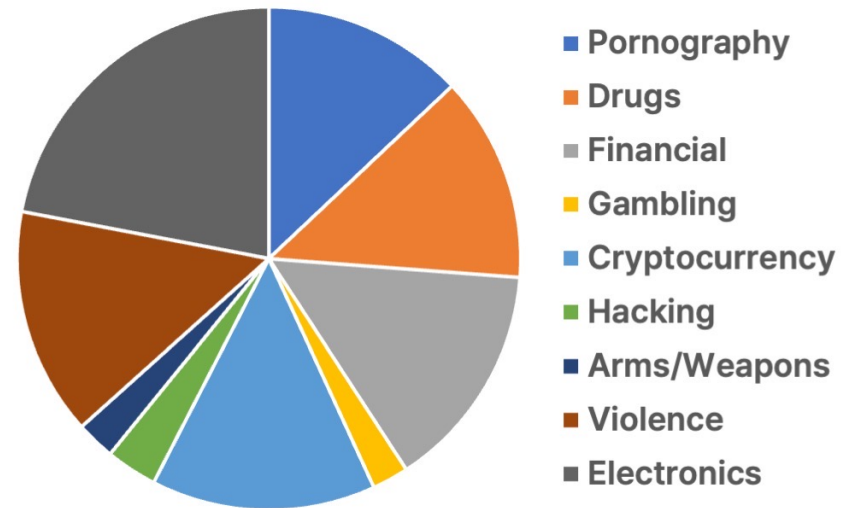
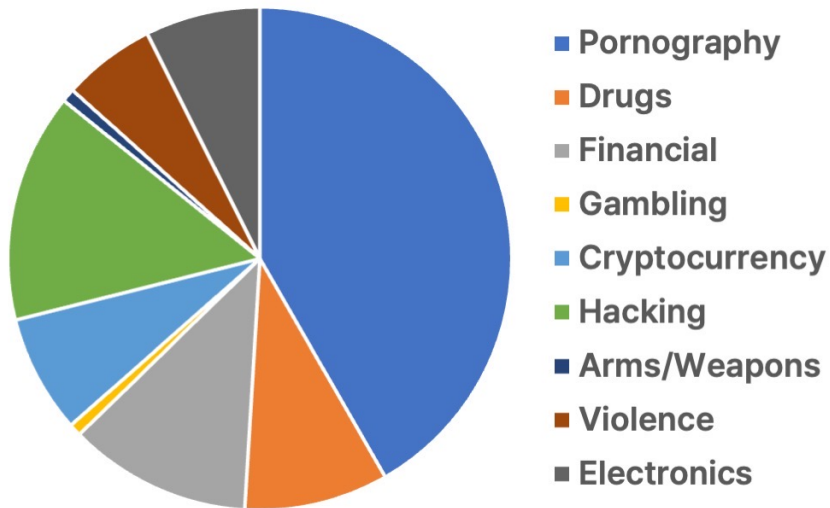
학습 데이터

- 다크웹에서 610만 페이지 수집 (영어로 자동 언어탐지된 것만)
- 전처리: 짧은 문서 & 중복 문서 필터링, 카테고리 밸런싱, 식별자 마스킹
- 학습 시간: 15일 (Nvidia A100 64GB * 8개)

다크웹 학습 과정

카테고리 밸런싱

- 다크웹 페이지의 대략 절반은 음란물 → PornBERT ?? 🤯🤯
- 카테고리 분류기를 활용하여 다량의 음란물 페이지 제거
- 중복 페이지가 많은 해킹 카테고리 데이터 상당수 제거



다크웹 학습 과정

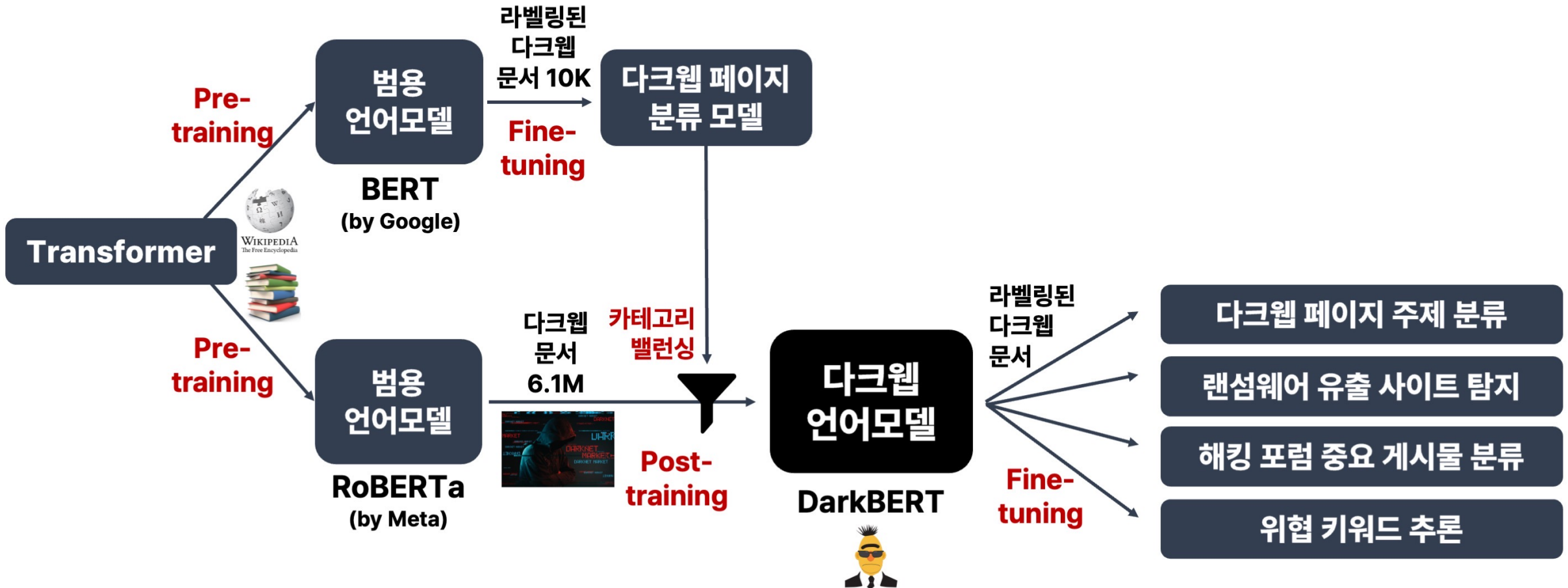
식별자 마스킹: 이메일, URL, IP주소, 가상화폐 주소 등
- 구체적인 문자열 형태보다는, 식별자의 "종류"가 중요하다고 가정

두 가지 버전: DarkBERT_{Raw} (마스킹X) & DarkBERT_{Preproc} (마스킹O)

Table 10: The types of identifier masks and the list of preprocessed texts.

Identifier Type	Example Text or Description	Preprocess Action Type	Identifier Mask Token
Email Addresses	example@email.com	Replace with token	ID_EMAIL
URLs (non-onion domain)	www.example.com https://www.example.com/home	Replace with token	ID_NORMAL_URL
URLs (onion domain)	facebookwkhpilnemxj7asaniu7vnjjbiltxjqhye3mhshg7kx5tfyd.onion	Replace with token	ID_ONION_URL
IP Addresses (IPv4 & IPv6)	192.168.1.1 fe80::1ff:fe23:4567:890a%eth2	Replace with token	ID_IP_ADDRESS
Cryptocurrency Addresses	BTC, ETH, LTC addresses	Replace with token	ID_BTC_ADDRESS ID_ETH_ADDRESS ID_LTC_ADDRESS
Lengthy "Words"	Any group of non-whitespace characters that are 38 or more letters long	Replace with token	ID_LONGWORD
Uncommon Characters	Any characters out of Unicode range from U+0000 to U+00FF	Remove from text	-
Whitespaces	Newline characters, tabs, spaces, etc.	Truncate to a single space	-

다크웹 학습 과정



실험 결과

실험1: 다크웹 페이지 주제 분류

- 문제: 주어진 다크웹 페이지를 10개의 주제 중 하나로 분류하기
 - 음란물, 마약, 해킹, 금융, 암호화폐, 무기, 도박, 폭력, 전자제품, 기타
- 데이터: 10개의 주제가 라벨링된 10,000개의 페이지 사용 (**CoDA dataset**)
 - 학습데이터:평가데이터 = 90%:10%

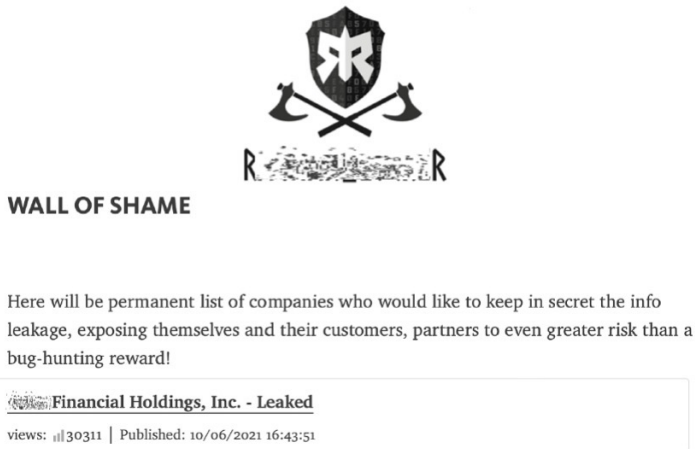
Dataset	Model	Precision	Recall	F1 score
CoDA _{cased}	BERT _{cased}	92.12	92.16	92.13
	BERT _{uncased}	92.83	92.67	92.75
	RoBERTa	93.36	93.27	93.31
	DarkBERT _{raw}	94.15	94.35	94.25
	DarkBERT _{preproc}	94.26	94.33	94.29
CoDA _{uncased}	BERT _{cased}	92.86	92.85	92.85
	BERT _{uncased}	92.83	92.67	92.75
	RoBERTa	93.30	93.40	93.34
	DarkBERT _{raw}	94.46	94.45	94.46
	DarkBERT _{preproc}	94.31	94.53	94.42

실험 결과

실험2: 랜섬웨어 유출 사이트 탐지

- 문제: 주어진 다크웹 사이트가 랜섬웨어 유출 사이트인지 판단하기 (이진 분류)
- 데이터: 랜섬웨어 유출 사이트와 일반 해킹 카테고리 사이트
 - 유출(positive): 54개의 랜섬웨어 그룹 사이트 내 총 105개 페이지
 - 일반(negative): 랜섬웨어 그룹 사이트가 아닌, 해킹 카테고리 679개 페이지
 - 학습데이터:평가데이터 = 80%:20%

Home Page of Ransomr Leaks site



(a) A ransomware leak site sample

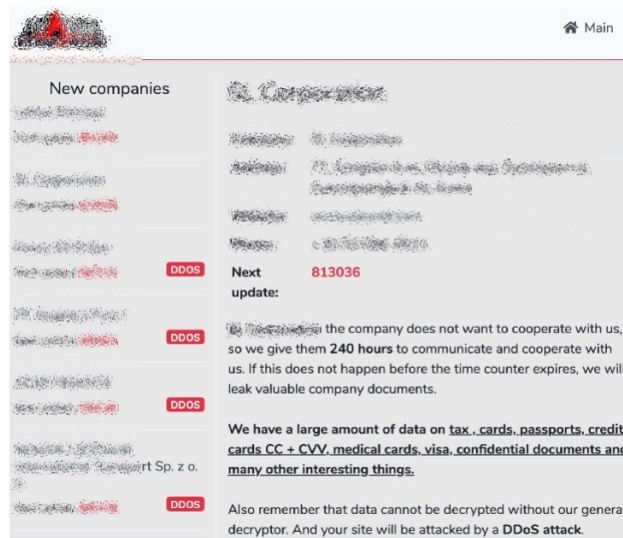


Figure 6: A leak site page sample in the dataset.

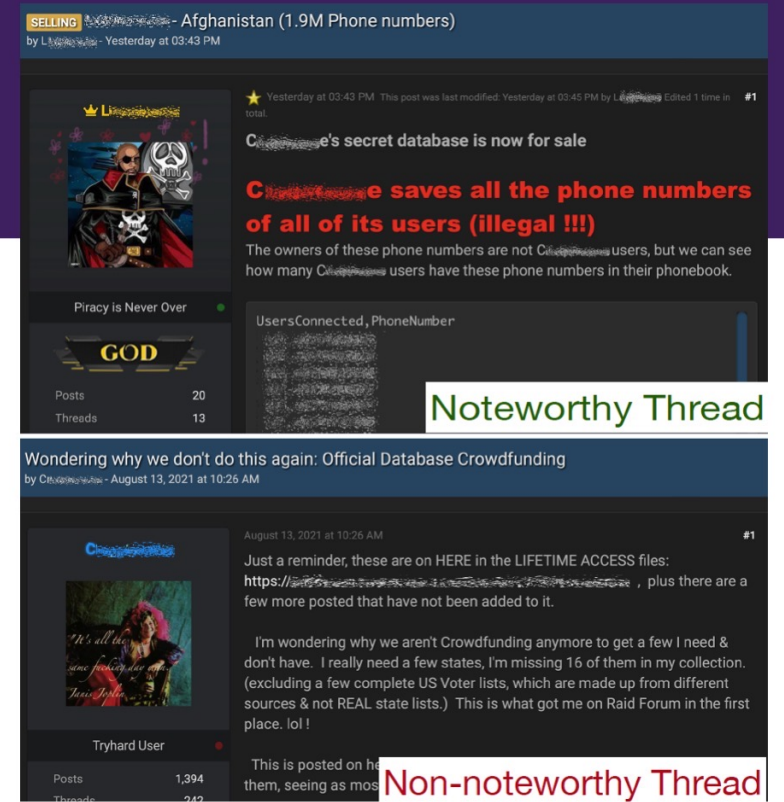
Table 4: Ransomware leak site detection performance. Boldface indicates the best performance.

Input	Model	Precision	Recall	F1 score
Raw	BERT _{cased}	75.83	69.52	71.01
	BERT _{uncased}	77.18	73.90	72.77
	RoBERTa	39.83	36.00	36.27
	DarkBERT _{raw}	78.81	83.62	79.98
Preprocessed	BERT _{cased}	76.81	68.19	70.13
	BERT _{uncased}	71.97	71.62	70.77
	RoBERTa	48.36	45.14	44.31
	DarkBERT _{preproc}	85.16	84.57	84.11

실험 결과

실험3: 해킹 포럼 중요 게시물 분류

- 문제: 해킹 포럼 내 주어진 게시물이 중요 게시물인지 여부 판단하기 (이진 분류)
- 데이터: Raidforums, Breached 포럼 게시물 1,873개 (21년 6월~22년 3월)
 - 249개 중요 게시물 vs 1624개 일반 게시물
 - 학습데이터:평가데이터 = 80%:20%
 - 중요 게시물 예시
 - 기업의 기밀 자산 공유/판매: 관리자 권한, 임직원/고객 정보, 거래내역, 소스코드, 계약서 등
 - 민감한 개인 정보 공유/판매: 신용정보, 의료정보, SSN, 여권, 신분증, 시민권 등
 - 유명 소프트웨어 및 기업을 대상으로 한 악성코드, 취약점 공유/판매 등



Noteworthy Thread

Non-noteworthy Thread

Table 5: Noteworthy thread detection performance. Boldface indicates best performance.

Input	Model	Precision	Recall	F1 score
Raw	BERT _{cased}	55.09	19.91	26.90
	BERT _{uncased}	52.34	23.49	28.51
	RoBERTa	28.97	17.89	21.38
	DarkBERT _{raw}	75.93	43.08	52.85
Preprocessed	BERT _{cased}	61.43	20.48	28.81
	BERT _{uncased}	45.46	21.52	26.16
	RoBERTa	29.04	15.27	18.71
	DarkBERT _{preproc}	72.44	45.13	54.17

실험 결과

실험4: 위협 키워드 추론

- 문제: 마약 이름이 언급된 문장에서 이를 마스킹한 후, 언어모델이 예측한 단어가 실제로 마약 관련 단어인지 확인
- 데이터: 약물 관련 Reddit 내 텍스트 및 마약 은어 데이터 (Zhu et al., 2021)

25 X XTC 230 MG DUTCH MDMA PHILIPP PLEIN



25 X XTC 230 MG DUTCH [MASK] PHILIPP PLEIN

Language Model	Semantically Related Words
DarkBERT	-pills, import, md, dot, translation, speed, up, oxy, script, champagne
BERT _{Reddit}	##man, champion, singer, rider, driver, sculptor, producer, manufacturer, ##er, citizen

Table 7: Quantitative performance metric of threat keyword inference. Precision at k ($P@k$) is measured with varying k in increments of 10.

	Top-10	Top-20	Top-30	Top-40	Top-50
DarkBERT _{CoDA}	0.60	0.60	0.50	0.42	0.42
BERT _{CoDA}	0.40	0.40	0.50	0.50	0.40
BERT _{Reddit}	0.40	0.45	0.60	0.57	0.52

25 X XTC 230 MG DUTCH MDMA PHILIPP PLEIN



Figure 4: An MDMA sales page excerpted from the Dark Web

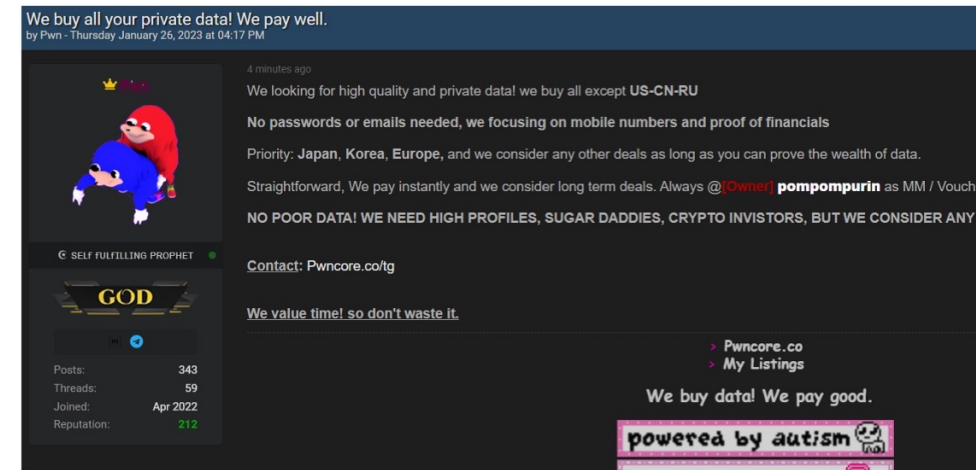
패션 브랜드
PHILIPP PLEIN 공식 로고



적용 사례

자체 수집망을 통해 수집된 위협 채널 게시물에 대해 DarkBERT로 개체명 추출 및 위협 스코어링 수행

Curated ThreatInfo								
Country: Korea			Country: Japan			Country: China		
Top 300 (out of 557 in total)								
NO	SITE	DATE	BOARD	COUNTRY	INDUSTRY	CONTENTS	USER	AUTO-REPORT
1	EXPD	5/19	Databases Exposed - Never been exposed! Leaks	China	Retail	Forum JD.COM [3GB-] @ Browser > JD.COM LEAKED DATABASE SIZE: [3GB-] Enjoy.	koko	Export
2	EXPD	5/16	Buyers Place Exposed - Never been exposed! Marketplace	China	Govt./Military/Edu.	Forum Chinese Government Database @ Browser > If you have Chinese GOV data, and you are not a reseller. Contact XMPP:AGUNcc0ND@ubuntu-jabber.net	jackjones	Export
3	EXPD	5/26	Exposed - Never been exposed! Leaks Market Marketplace Sellers Place	Japan	Govt./Military/Edu.	Forum Japan Government Ministry Database @ Browser > Hello, Today i want to sell a Japan Government ministry Database. This database is fresh, (i hacked it in a few of minutes) This database contain the following: - Japan Government Emails - Name of workers, etc.... read more	g9d	Export
4	XSS	5/19	ДОСТУПЫ: сети, rdp, шеллы, ftp, sq-l-inj, DB's Торговая площадка	China	Govt./Military/Edu.	Forum Selling - China Internet Network Information Center (Chinese ... @ Browser > Selling - China Internet Network Information Center (Chinese government cyberspace control) Price: 600\$ Telegram: Vars_Secc Contact me promptly.	Vars_Secc	Export
5	EXPT	5/26	Buying/Selling Commerce [Payment systems] - exchange, sale, identification, distribution	China	Finance	Forum нужен ищущий китайский банк акк search china bank acc @ Browser > Для вывода денег писать можно на телегу https://t.me/nomadshop to withdraw money contact telegram or jabber nomadshop@xabber.org	nomadshop2	Export



단순 정보성 및 구매글은 낮은 점수를 부여하고 판매/공유글은 (특히 구체적인 데이터를 첨부한 경우) 높은 점수 부여

연구 의의

다크웹을 학습한 최초의 언어모델 연구 성과

**언어모델이 다크웹의 언어와 지식을 학습할 수 있고,
학습된 모델을 유용한 작업에 활용할 수 있음을 학술적으로 보임**

**다크웹 내 발생하는 다양한 범죄 활동을 탐지/분류하고,
핵심 위협 정보를 추출하는 데 이용 가능할 것으로 기대**

Q&A



S2W와 솔루션에 대해 더 알고 싶으신가요?

아래의 메일 주소로 문의주세요.

info@s2w.inc

The information contained in this document is proprietary and confidential.
If you are not the intended recipient, please note that any use or circulation of this document may be cause for legal action.