Inferring Discussion Topics about Exploitation of Vulnerabilities from Underground Hacking Forums

Felipe A. Moreno-Vera

PPGI PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA
Universidade Federal do Rio de Janeiro
UFRJ

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

CAPES

FUNDAÇÃO COPPETEC

Content

- Introduction
- Dataset
- Methodology
- Experimental Results
- Conclusion

Introduction

Motivation

A necessity of understand what is the content shared in underground forums.

Exploitation of vulnerabilities in the wild are a threat to internet ecosystem and software communities.

Context

We identify the necessity of known what are being discussed in underground hacking forums:

- Relevant Topics
- Word Frequency
- Language used

Our Key Contributions

- Methodology to analyze, model and identify discussion topics, and significant information within underground hacking forums.

- **Topic Modeling:** What are the discussion topic on each forum?
    - Topic modeling for identify relevant topics and words.
    - Interpretation of topics based on the content.

Dataset

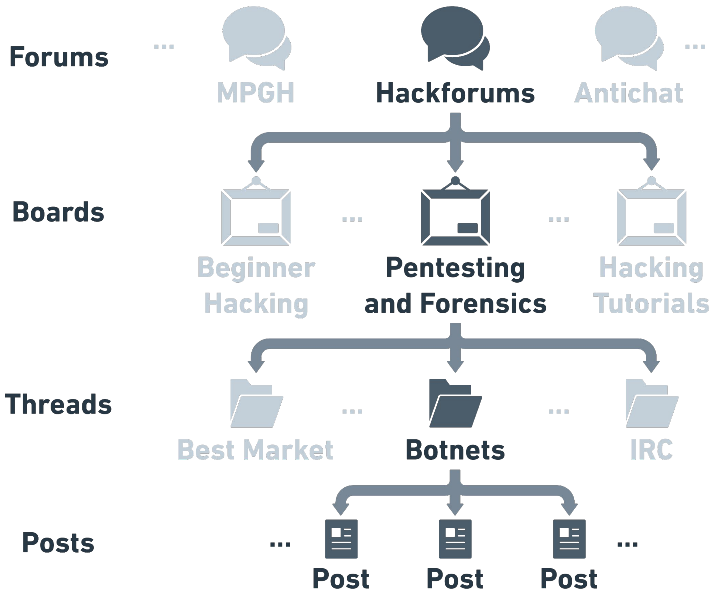
Dataset

Made available by Cambridge Cybercrime Centre

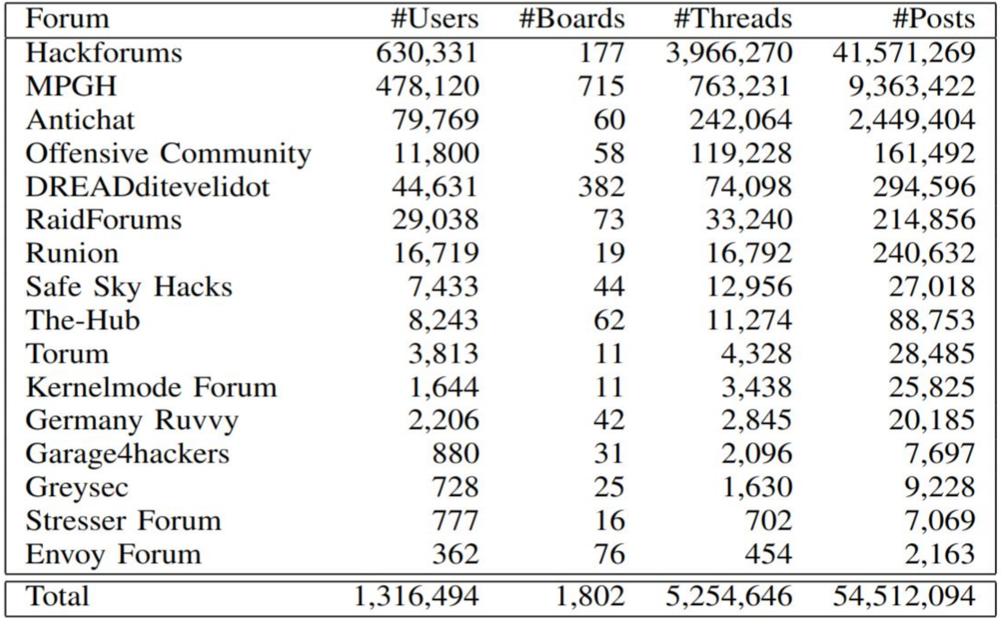
Contains data scraped from multiple underground forums (16 studied)

Organized in forums, boards, threads and posts

Provide about 54,512,094 lines of textual information.



CrimeBB



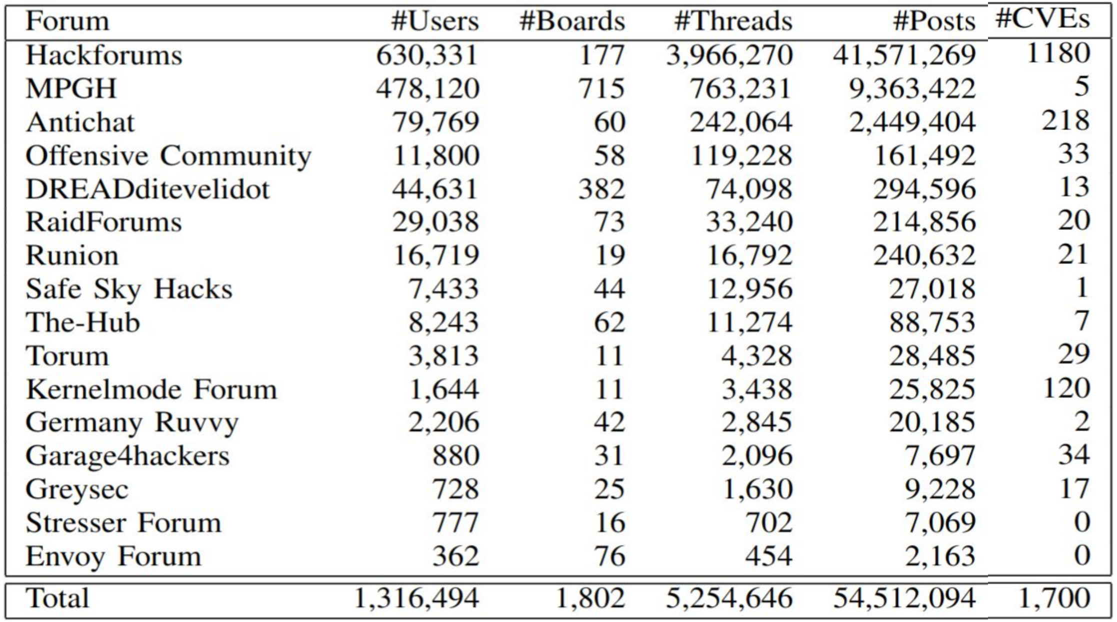
Forum	#Users	#Boards	#Threads	#Posts
Hackforums	630,331	177	3,966,270	41,571,269
MPGH	478,120	715	763,231	9,363,422
Antichat	79,769	60	242,064	2,449,404
Offensive Community	11,800	58	119,228	161,492
DREADditevelidot	44,631	382	74,098	294,596
RaidForums	29,038	73	33,240	214,856
Runion	16,719	19	16,792	240,632
Safe Sky Hacks	7,433	44	12,956	27,018
The-Hub	8,243	62	11,274	88,753
Torum	3,813	11	4,328	28,485
Kernelmode Forum	1,644	11	3,438	25,825
Germany Ruvvy	2,206	42	2,845	20,185
Garage4hackers	880	31	2,096	7,697
Greysec	728	25	1,630	9,228
Stresser Forum	777	16	702	7,069
Envoy Forum	362	76	454	2,163
Total	1,316,494	1,802	5,254,646	54,512,094

Natural Vulnerability Database (NVD)

The NVD is the U.S. government repository of standards based vulnerability. This data enables automation of vulnerability management, and is fully synchronized with the **CVE** list and **CVSS** scores.

- **Common Vulnerabilities and Exposures (CVE)** is a list of publicly disclosed vulnerabilities and exposures.
- **Common Vulnerability Scoring System (CVSS)** provides a numerical (0-10) representation of the severity of a vulnerability.

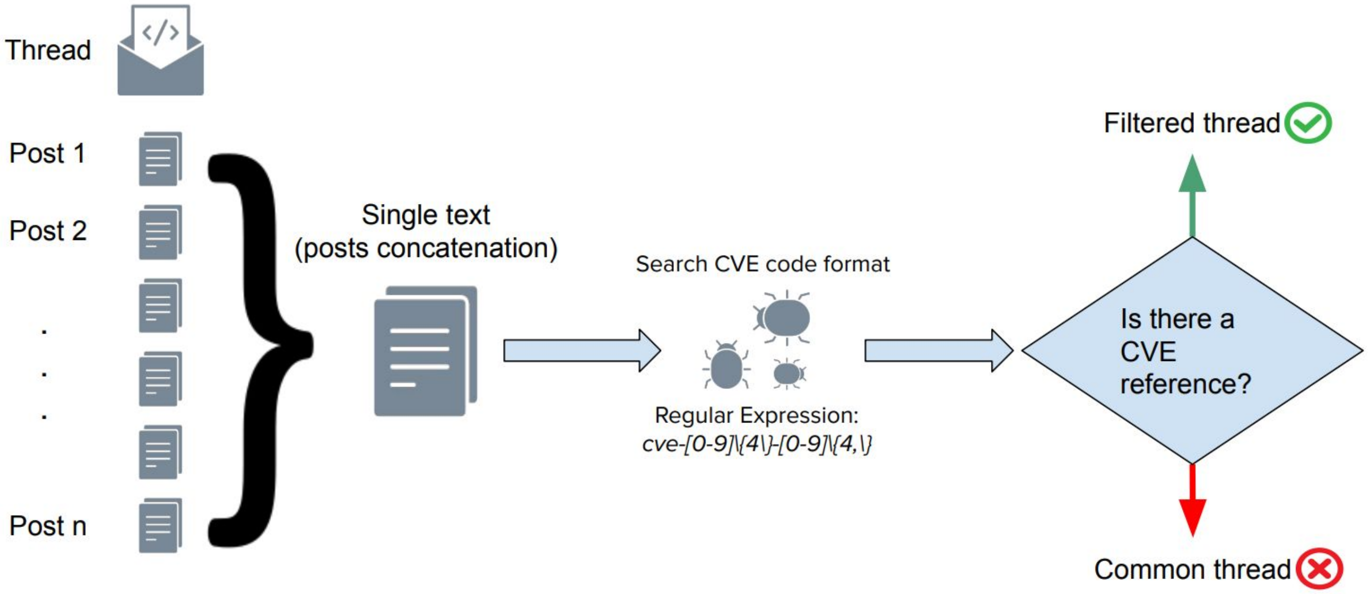
CrimeBB + NVD



Forum	#Users	#Boards	#Threads	#Posts	#CVEs
Hackforums	630,331	177	3,966,270	41,571,269	1180
MPGH	478,120	715	763,231	9,363,422	5
Antichat	79,769	60	242,064	2,449,404	218
Offensive Community	11,800	58	119,228	161,492	33
DREADditevelidot	44,631	382	74,098	294,596	13
RaidForums	29,038	73	33,240	214,856	20
Runion	16,719	19	16,792	240,632	21
Safe Sky Hacks	7,433	44	12,956	27,018	1
The-Hub	8,243	62	11,274	88,753	7
Torum	3,813	11	4,328	28,485	29
Kernelmode Forum	1,644	11	3,438	25,825	120
Germany Ruvvy	2,206	42	2,845	20,185	2
Garage4hackers	880	31	2,096	7,697	34
Greysec	728	25	1,630	9,228	17
Stresser Forum	777	16	702	7,069	0
Envoy Forum	362	76	454	2,163	0
Total	1,316,494	1,802	5,254,646	54,512,094	1,700

Methodology

Data Preparation - Filtering threads



Thread

Post 1

Post 2

.
.
.

Post n

Single text
(posts concatenation)

Search CVE code format

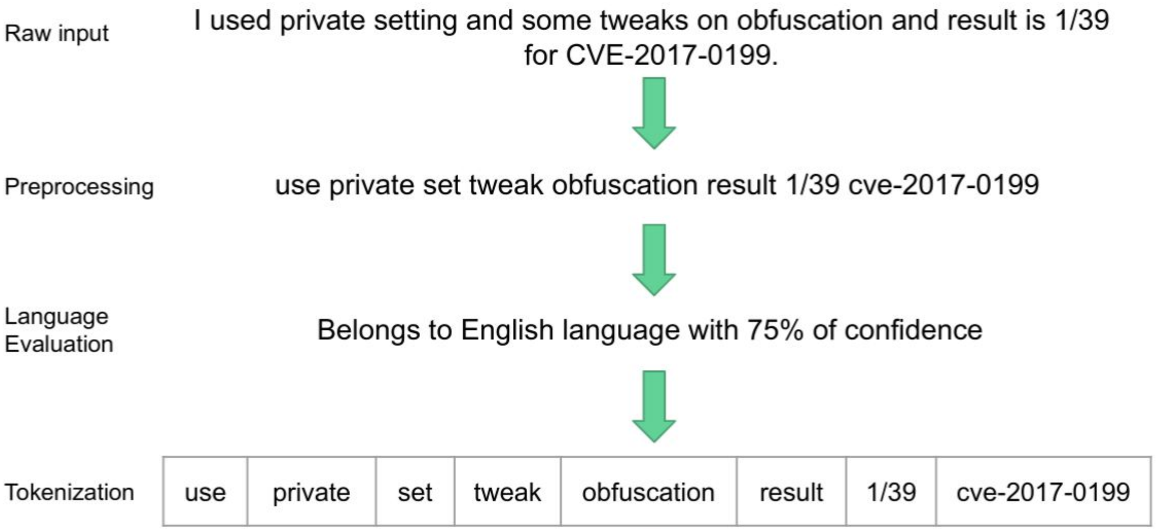
Regular Expression:
cve-[0-9]\{4\}-[0-9]\{4,\}

Is there a CVE reference?

Filtered thread ✓

Common thread ☒

Data Preparation - Text Processing



Raw input: I used private setting and some tweaks on obfuscation and result is 1/39 for CVE-2017-0199.

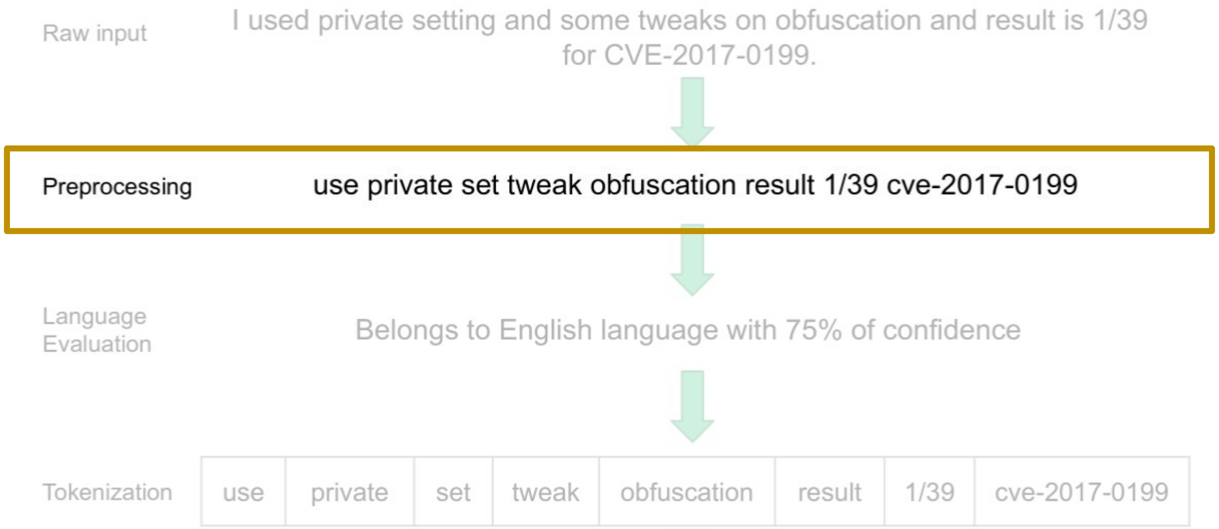
Preprocessing: use private set tweak obfuscation result 1/39 cve-2017-0199

Language Evaluation: Belongs to English language with 75% of confidence

Tokenization:

use	private	set	tweak	obfuscation	result	1/39	cve-2017-0199

Data Preparation - Text Processing



Raw input — I used private setting and some tweaks on obfuscation and result is 1/39 for CVE-2017-0199.

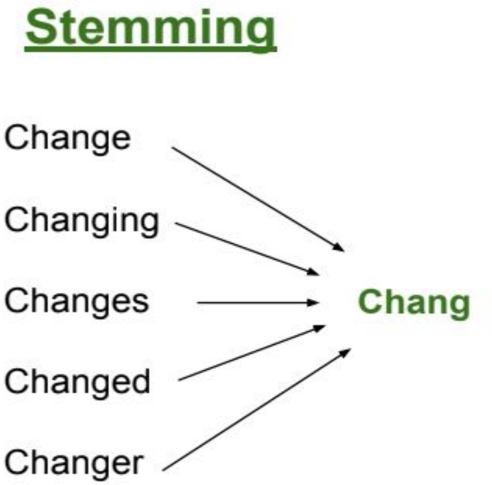
Preprocessing — use private set tweak obfuscation result 1/39 cve-2017-0199

Language Evaluation — Belongs to English language with 75% of confidence

Tokenization

use	private	set	tweak	obfuscation	result	1/39	cve-2017-0199

Text Processing - Text Normalization



Stemming

Change

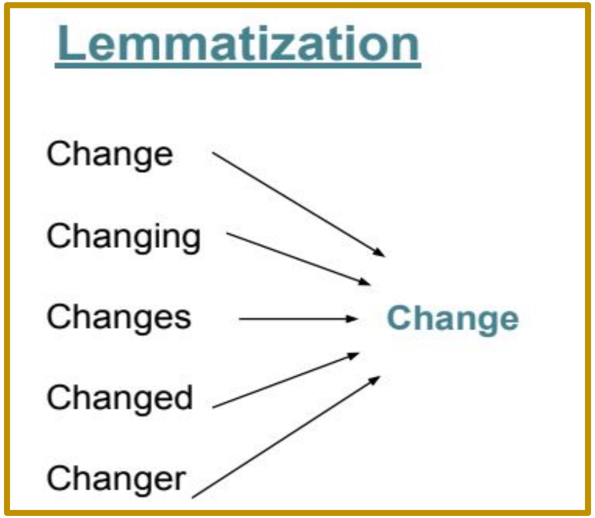
Changing

Changes

Changed

Changer

Chang



Lemmatization

Change

Changing

Changes

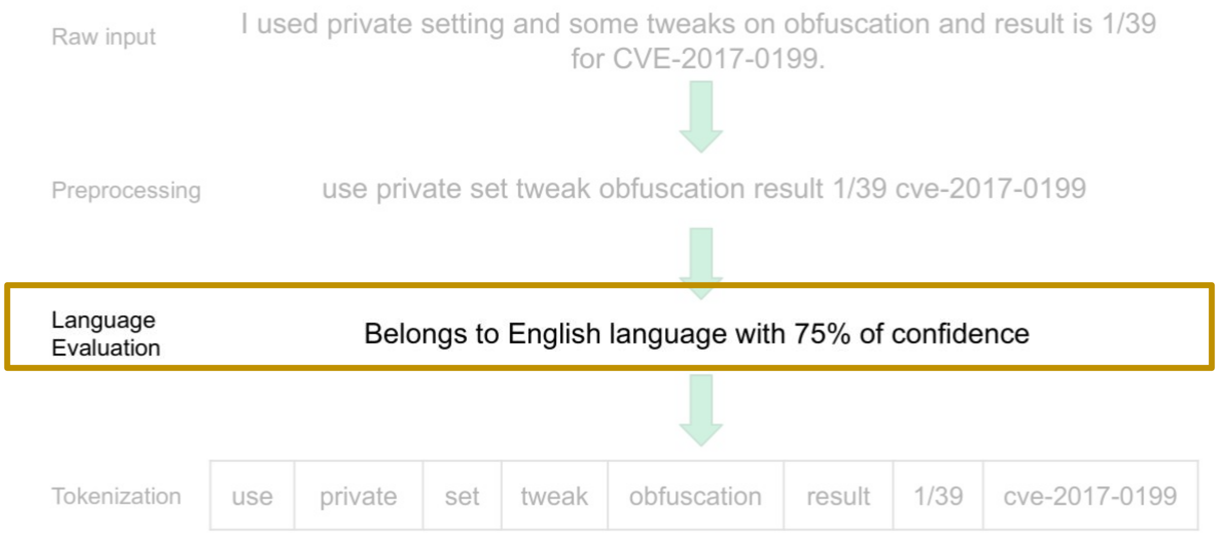
Changed

Changer

Change

* **Stemming**: Keeps the roots (stem) base form of the word to do content analysis.
* **Lemmatization**: Keep the meaningful base form of the word to do morphological analysis (context).
* **None**: Keep all words.

Data Preparation - Text Processing



Raw input: I used private setting and some tweaks on obfuscation and result is 1/39 for CVE-2017-0199.

Preprocessing: use private set tweak obfuscation result 1/39 cve-2017-0199

Language Evaluation: Belongs to English language with 75% of confidence

Tokenization	use	private	set	tweak	obfuscation	result	1/39	cve-2017-0199

Text Processing - Language Evaluation

We define an Indicator Language function (ilf) or $\mathbb{1}_{ilf}$ as:

$$\mathbb{1}_{ilf}(word, language) = \begin{cases} 1, & \text{if word belongs to language} \\ 0, & \text{otherwise} \end{cases}$$

We define a Language Ratio Function (lrf) as:

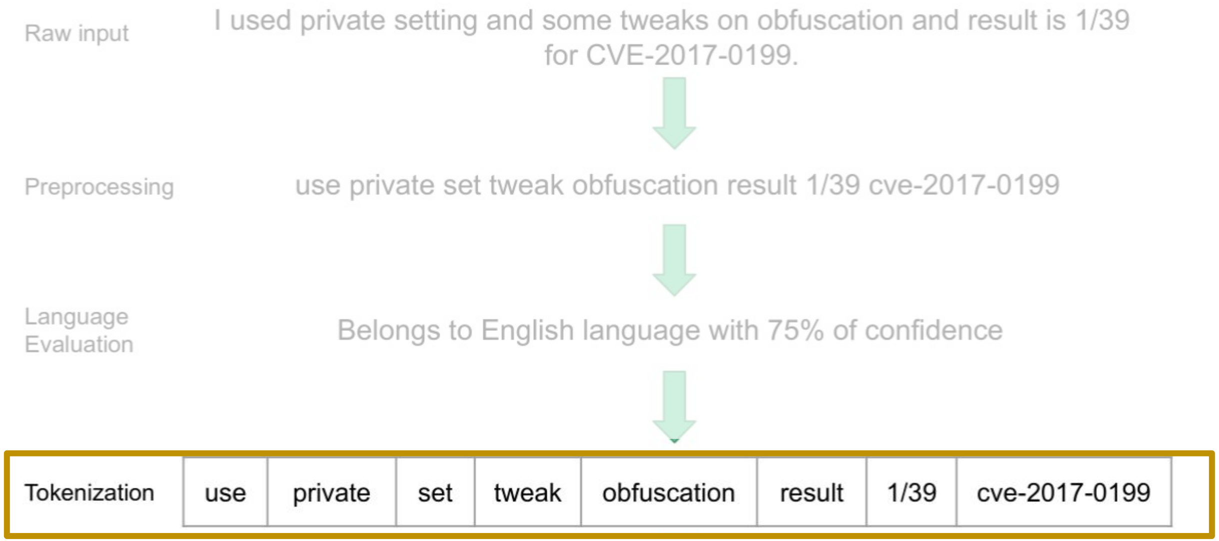
$$Ratio_{lrf}(text, language_j) = \frac{1}{\text{Total words in text}} \sum_{\substack{i=1 \\ \text{word}_i \in \text{ text}}}^{n} \mathbb{1}_{ilf}(word_i, language_j)$$

We determine which language is the most probable to be after evaluate a text as:

$$language(text) = \max_{\forall lang \in \text{languages list}} Ratio_{lrf}(text, lang)$$

* **NLTK**: Language punctuations & stopwords
* **Spacy**: Language Lemmatization
* **Enchant**: Language vocabulary

Data Preparation - Text Processing



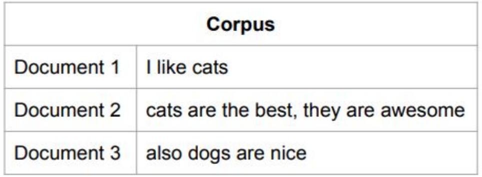
Raw input: I used private setting and some tweaks on obfuscation and result is 1/39 for CVE-2017-0199.

Preprocessing: use private set tweak obfuscation result 1/39 cve-2017-0199

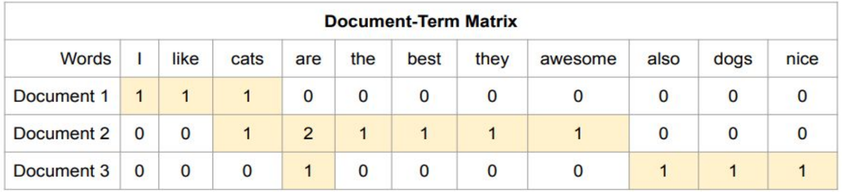
Language Evaluation: Belongs to English language with 75% of confidence

Tokenization	use	private	set	tweak	obfuscation	result	1/39	cve-2017-0199

Data Preparation - Feature Extraction

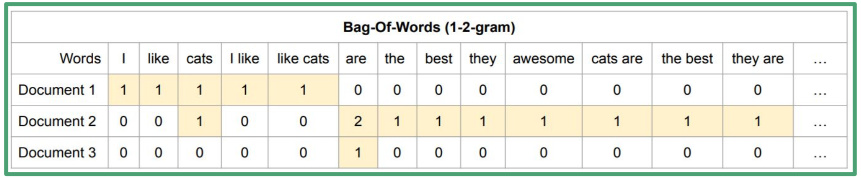


Corpus	
Document 1	I like cats
Document 2	cats are the best, they are awesome
Document 3	also dogs are nice



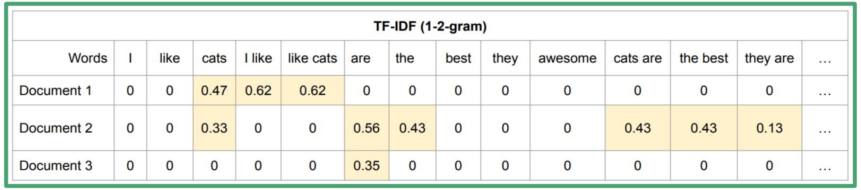
Document-Term Matrix

Words	I	like	cats	are	the	best	they	awesome	also	dogs	nice
Document 1	1	1	1	0	0	0	0	0	0	0	0
Document 2	0	0	1	2	1	1	1	1	0	0	0
Document 3	0	0	0	1	0	0	0	0	1	1	1



Bag-Of-Words (1-2-gram)

Words	I	like	cats	I like	like cats	are	the	best	they	awesome	cats are	the best	they are	…
Document 1	1	1	1	1	1	0	0	0	0	0	0	0	0	…
Document 2	0	0	1	0	0	2	1	1	1	1	1	1	1	…
Document 3	0	0	0	0	0	1	0	0	0	0	0	0	0	…



TF-IDF (1-2-gram)

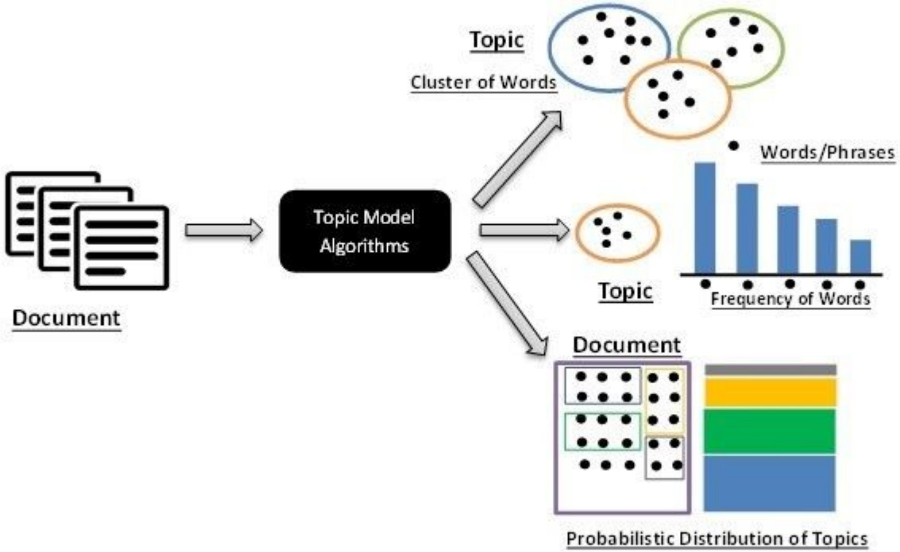
Words	I	like	cats	I like	like cats	are	the	best	they	awesome	cats are	the best	they are	…
Document 1	0	0	0.47	0.62	0.62	0	0	0	0	0	0	0	0	…
Document 2	0	0	0.33	0	0	0.56	0.43	0	0	0	0.43	0.43	0.13	…
Document 3	0	0	0	0	0	0.35	0	0	0	0	0	0	0	…

We also apply standard NLP pre-processing techniques, e.g., filtering stopwords and punctuation

Topic Modeling - Latent Dirichlet Allocation (LDA)

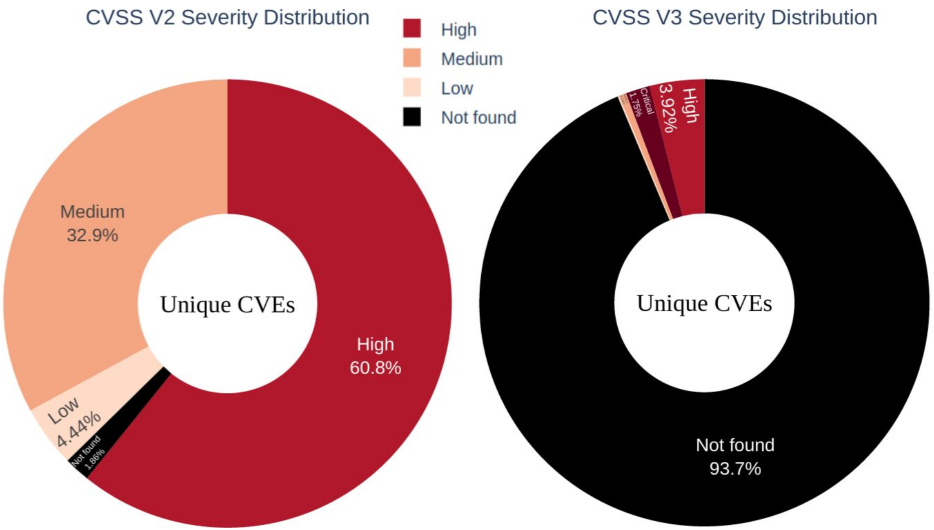
In this work, we define 4 main topics of interest: **Proof Of Concept (PoC)**, **Weaponization, Exploitation**, and **Others**.

We select this 4 topics based on previous manual analysis of the content. We find that threads content discuss about CVE codes and vulnerabilities.



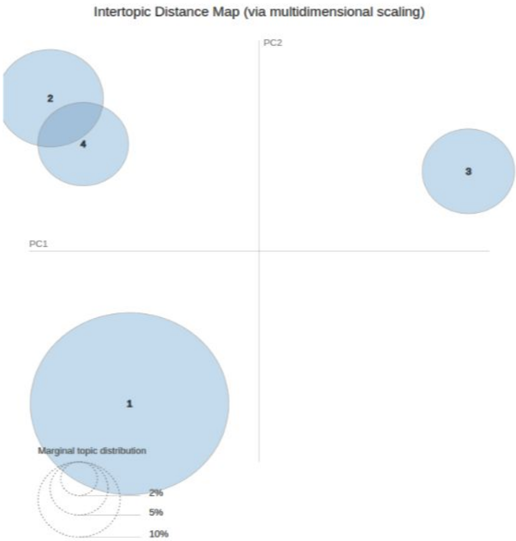
Experimental results

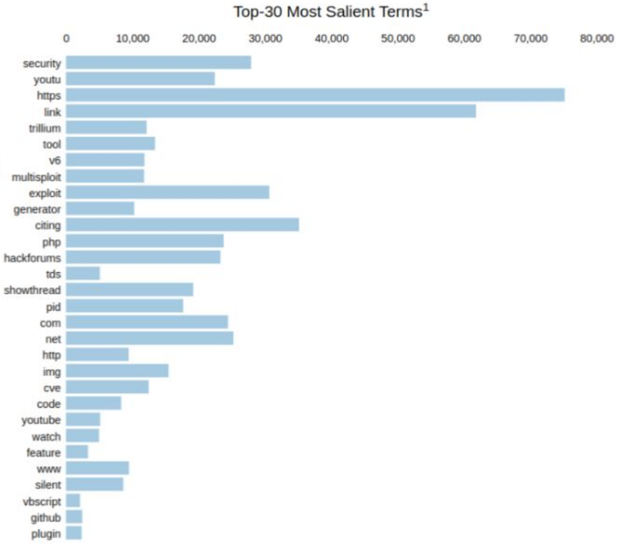
CVSS



- Notably, more than **60%** of the cited **CVEs** are designated with a **high severity** level in version 2. However, in version 3.1, we encountered difficulty in determining the present severity level of our CVE codes.

Findings

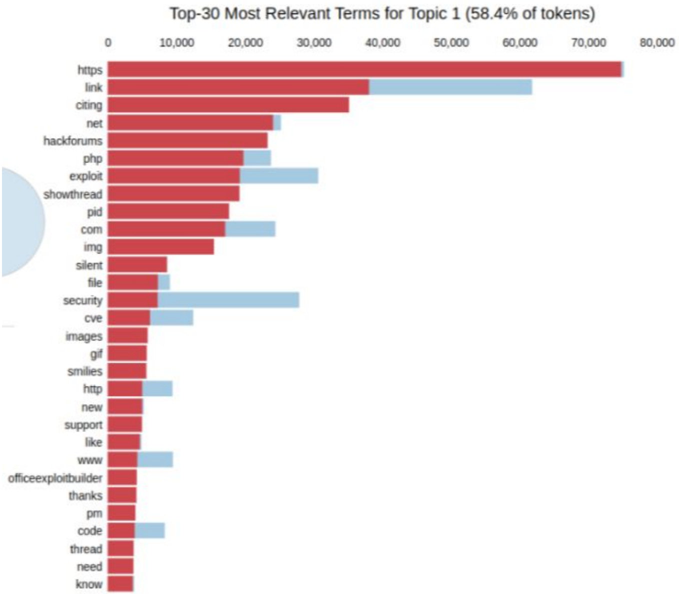




- We identify 4 topics: 1 (PoC), 2 (Weaponization), 3 (Exploitation), and 4 (Others)
- We show the 30 top words in all topics, we note that words such as "https", "link", "citing", and "exploit" are the most relevant.

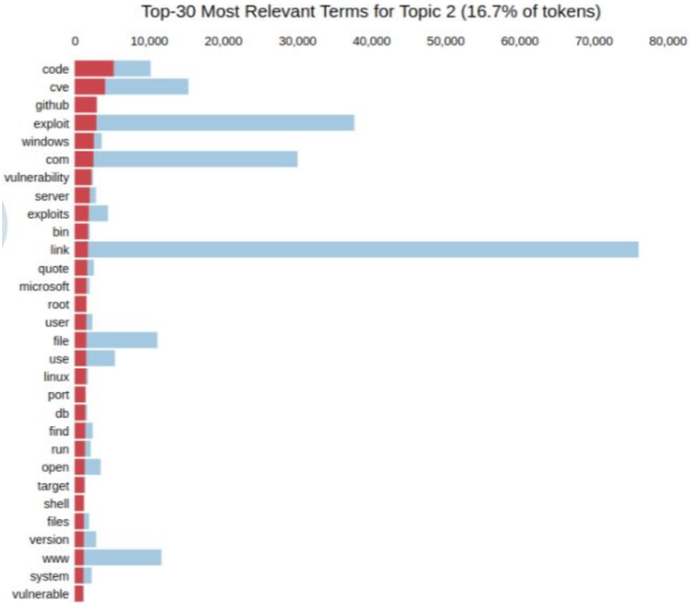
Topic 1

PoC Topic: Using only the **58.4%** of tokens, the most relevant words are "https", "link", "php", etc. We note that in general, those words are relevant for all topics except for "code" and "security".


Top-30 Most Relevant Terms for Topic 1 (58.4% of tokens)

Topic 2

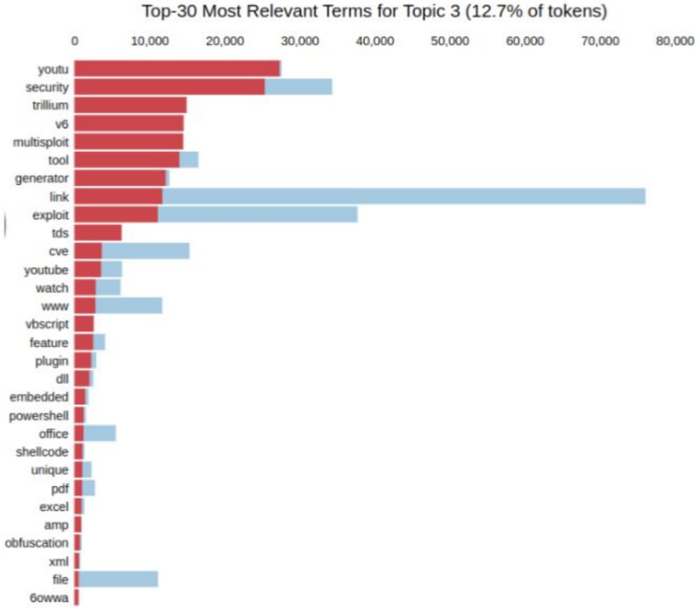
Weaponization Topic: Using only the **16.7%** of tokens, the most relevant words are "code", "cve", "github", etc.



Topic 3

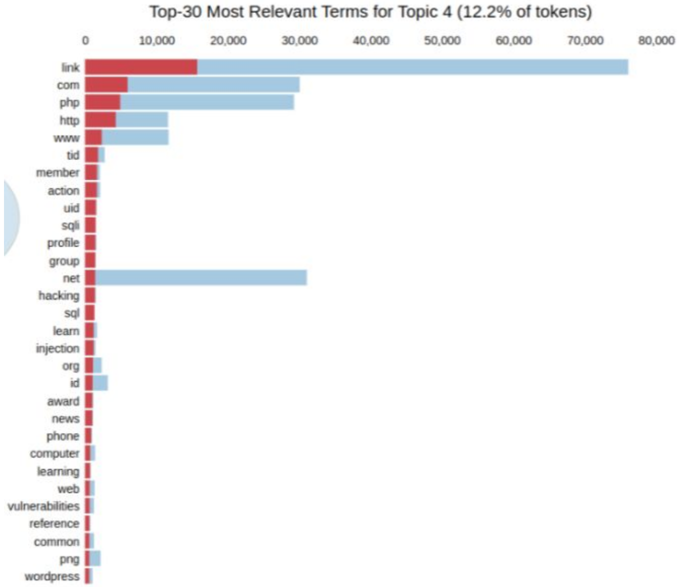
Exploitation Topic: Using only the **12.7%** of tokens, the most relevant words are "security", "trillium" (the user who cite the highest quantity of CVE codes), "multisploit", "tool", "exploit", etc.

We saw some intersections of words (e.g., "code", "cve", "exploit", "link", and "vulnerability") between the topics Weaponization and Exploitation.



Topic 4

Others Topic: Using only the **12.2%** of tokens, the most relevant words are "link", "com", "php", "member", "profile", "learn", etc.



Conclusions

Conclusions

- We show that applying topic modeling techniques to the study of exploitation in the wild offers valuable insights and benefits in the field of cybersecurity.

- By analyzing textual data related to vulnerabilities, exploits, and real-world attack scenarios, topic modeling contributes to a deeper understanding of the exploitation landscape.

- By extracting latent topics from data sources such as vulnerabilities, exploit forums, or security incident reports we understand topics about exploit trends.

- This understanding helps security professionals, and researchers stay informed about emerging threats and prioritize their defense strategies accordingly.

-

Thanks! Any questions?

felipe.moreno@ppgi.ufrj.br

THANKS!

Any Questions?