

Topic Modeling for data discovery

A Cybersecurity use case

Harini Kannan (@jarvision__)

Data Scientist @[Capsule8](#)

Applied ML in cybersecurity

- Application of ML in cybersecurity can be broadly classified into two classes:
 - defending against malicious activity | Eg.
 - learning and detecting specific attacks
 - user behavior analysis
 - network analysis
 - aiding attack mechanisms | Eg.
 - adversarial models to circumvent detections
 - data poisoning
 - reconnaissance activity to capture leaked credentials

Red Team Vs Blue Team - What do they do ??



First step for red teams - Information Gathering

- Red teams consist of ethical hackers who evaluate system security in an objective manner
- Once a target is set, gather information on the attack target
- The required information could be anything:
 - names, phone numbers, email addresses, organizations
 - social media accounts
 - open ports, domain names, IP addresses and other network traffic data
 - insecure files, programs and server misconfiguration
 - leaked passwords, passcodes for different apps

Welcome to a messy log dataset

- Various system data dumps, most of them unstructured or semi structured
- Where do you start ?
- This is one of the most time consuming tasks for a red team
- One of their main goals is to recover leaked passwords
- Simple tools like "grep" could be useful, but only if the resulting filter gives us a readable amount of data -- and if you already know what you seek
- Classic NLP problem

Challenges

- Unstructured data
- Data from multiple logs
- No labels
- Lack of context

Enter Topic Modeling

- Statistical modeling technique for discovering “topics” that occur in a collection of documents
- Derive hidden patterns exhibited by a text corpus
- Useful for document clustering

Examples of applied Topic modeling

- NYT uses Topic Modeling to boost their user-article recommendation engines
- Recruitments industries extract latent features of job descriptions and map them to the right candidates
- In Marketing - Clustering social media profiles, emails and customer reviews
- SEO ! Google confirms they are using topic modeling to assist their ranking algorithm

More [here](#)

Common Techniques used for Topic Modeling

- LSA (Latent Semantic Analysis)
- PLSA (Probabilistic Latent Semantic Analysis)
- LDA (Latent Dirichlet Allocation)

Latent Dirichlet Allocation

Intuition behind LDA:

- Each document is a collection of topics in a certain proportion
- Each topic is a collection of keywords in certain proportion
- Once number of topics is given to the algorithm, it:
 - rearranges the topics distribution within the documents and keywords distribution within the topics to obtain a good composition of topic-keywords distribution
 - basically backtracks and tries to figure out what topics would create those documents in the first place

LDA Mechanics

- LDA is a matrix factorization technique
- In Vector space, any corpus can be represented as a document-term matrix

	W1	W2	W3	<u>Wn</u>
D1	0	2	1	3
D2	1	4	0	0
D3	0	2	3	1
<u>Dn</u>	1	1	3	0

- This matrix shows a corpus of N documents D1, D2, D3...Dn with a vocabulary size of M words- W1, W2, W3...Wn
- LDA converts this Document-Term matrix into 2 lower dimensional matrices - M1 and M2

	K1	K2	K3	K
D1	1	0	0	1
D2	1	1	0	0
D3	1	0	0	1
<u>Dn</u>	1	0	1	0

	W1	W2	W3	<u>Wm</u>
K1	0	1	1	1
K2	1	1	1	0
K3	1	0	0	1
K	1	1	0	0

- M1 is a document-topic matrix with dim (N,K)
- M2 is a topic term matrix with dim (K,M)
- N is no. of documents, K is no. of topics, M is vocab size

LDA Mechanics (cont.)

- For every topic, two probabilities P1 and P2 calculated
 - P1 $p(\text{topic } t / \text{document } d)$: proportion of words in document d that are currently assigned to topic t
 - P2 $p(\text{word } w / \text{topic } t)$: proportion of assignments to topic t over all documents that come from this word w
- Current topic-word assignment is updated with a new topic with the probability - product of P1 and P2
- After a no. of iterations, steady state is achieved where document-topic and topic-term distributions is good and is the convergence point of the model

Let's get to the problem

- We got hold of a nation state hacker's HDDs
- Access was given to all the system logs in a single hard drive for our analysis
- Request from a red team member:
 - They spent a lot of time to filter out few 100s of passwords from the logs
 - Can we build a tool that would make this faster and simpler

Let's get to it!!

LDA Topic modeling | Step by step

1. Prepare stopwords
2. Remove unnecessary characters
3. Tokenize sentences into list of words
4. Remove stop words, lemmatize
5. Create dictionary and corpus to be used for topic modeling
6. Create model and coherence scores for different N no. of topics
7. Best model selection
8. Dominant topic in each document and topic weight
9. Most representative sentence for each topic
10. Word cloud
11. TSNE clustering
12. Visualize topic model

Exercise notebook

More on this [here](#)

Interesting topics from the hacker dataset



The word cloud in Topic 3 exposes the keyword “password,” so we know the best starting point for our search. In fact, the model uncovered 162 passwords within Topic 3

Data points in topic 3 containing str "password"

```
34 [07:14:44] [X] CM109,506001727,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0506001727 password g
262 [07:32:46] [X] CM06,263351460,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0263351460 password a7
281 [07:32:59] [X] CM047,500375993,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0500375993 password 00
2024 [09:11:32] [X] CM028,205238590,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0205238590 password ww
2034 [09:13:57] [X] CM057,206699064,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0206699064 password By
2035 [09:13:58] [X] CM025,507127626,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0507127626 password 13
3102 [08:02:07] [X] CM052,206472362,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0206472362 password bu
3147 [08:03:14] [X] CM048,201079812,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0201079812 password hb
3304 [08:11:21] [X] CM056,206255014,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0206255014 password 92
3305 [08:11:22] [X] CM055,502097668,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0502097668 password gc
3343 [08:12:22] [X] CM032,201080653,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0201080653 password g3
3345 [08:12:24] [X] CM027,506304917,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0506304917 password ka
3461 [08:20:21] [X] CM046,506610962,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0506610962 password y9
3462 [08:20:22] [X] CM043,201079713,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0201079713 password 76
3465 [08:20:23] [X] CM029,238326416,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0238326416 password er
3467 [08:20:26] [X] CM044,201068368,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0201068368 password lh
3471 [08:20:30] [X] CM054,201862206,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0201862206 password 61
3473 [08:20:31] [X] CM050,201080879,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0201080879 password kr
3474 [08:20:31] [X] CM056,506610962,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0506610962 password yo
3477 [08:20:33] [X] CM054,201862206,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0201862206 password 61
3478 [08:20:33] [X] CM055,207183702,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0207183702 password q7
3487 [08:20:44] [X] CM050,206448642,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0206448642 password fx
3488 [08:20:45] [X] CM043,201079713,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0201079713 password 76
3529 [08:21:31] [X] CM060,200192963,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0200192963 password f2
3530 [08:22:01] [X] CM064,206253873,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0206253873 password xv
3533 [08:22:10] [X] CM055,506535543,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0506535543 password 11
3534 [08:22:12] [X] CM063,238391488,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0238391488 password fi
3691 [08:32:21] [X] CM045,264608290,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0264608290 password qd
3818 [08:41:47] [X] CM054,206498990,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0206498990 password pe
3836 [08:42:06] [X] CM061,206254857,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0206254857 password cy
3838 [08:42:06] [X] CM050,206713858,Access your saved data from your Computer with https://.com/CloudBackup/ Login 026713858 password j1
3839 [08:42:06] [X] CM062,201080816,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0201080816 password g9
3841 [08:42:07] [X] CM047,500375993,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0500375993 password 00
3841 [08:42:08] [X] CM057,206254063,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0206254063 password u6
3843 [08:42:09] [X] CM050,206253872,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0206253872 password 3y
3844 [08:42:11] [X] CM056,200192767,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0200192767 password j9
3846 [08:42:23] [X] CM063,502722969,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0502722969 password b1
3848 [08:42:27] [X] CM064,206052311,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0206052311 password x2
3849 [08:42:33] [X] CM066,267111760,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0267111760 password d1
3943 [08:50:17] [X] CM052,206472362,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0206472362 password bu
3946 [08:50:55] [X] CM015,206498998,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0206498998 password s7
3955 [08:51:07] [X] CM044,265105857,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0265105857 password ts
3990 [08:51:56] [X] CM043,2060601517,Access your saved data from your Computer with https://.com/CloudBackup/ Login 02060601517 password c4
4097 [08:02:13] [X] CM040,206443517,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0206443517 password hm
4099 [08:02:21] [X] CM041,206443432,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0206443432 password or
4105 [08:03:02] [X] CM025,203509311,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0203509311 password 91
4211 [08:14:21] [X] CM012,500054029,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0500054029 password gw
4218 [23:40:15] [X] CM054,201080930,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0201080930 password k1
5085 [16:13:54] [X] CM054,200656843,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0200656843 password r7
5093 [16:45:37] [X] CM08,502204505,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0502204505 password cu
5098 [16:45:42] [X] CM015,502204146,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0502204146 password e4
5099 [16:45:43] [X] CM011,502203943,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0502203943 password zv
5100 [16:45:44] [X] CM016,200649899,Access your saved data from your Computer with https://.com/CloudBackup/ Login 020649899 password w9
5101 [16:45:45] [X] CM013,502204159,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0502204159 password u6
5105 [16:45:51] [X] CM016,200656397,Access your saved data from your Computer with https://.com/CloudBackup/ Login 020656397 password 80
5109 [16:45:53] [X] CM012,502204537,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0502204537 password ku
5110 [16:45:53] [X] CM09,502204244,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0502204244 password jy
5113 [16:45:55] [X] CM014,200656830,Access your saved data from your Computer with https://.com/CloudBackup/ Login 020656830 password ge
5558 [01:47:43] [X] CM011,206643567,Access your saved data from your Computer with https://.com/CloudBackup/ Login 020643567 password iv
5559 [01:47:43] [X] CM012,206443611,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0206443611 password 91
5561 [01:47:32] [X] CM014,206498900,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0206498900 password j9
5577 [01:47:40] [X] CM013,206499288,Access your saved data from your Computer with https://.com/CloudBackup/ Login 020649288 password gv
5586 [01:47:44] [X] CM015,506452514,Access your saved data from your Computer with https://.com/CloudBackup/ Login 0506452514 password c7
```

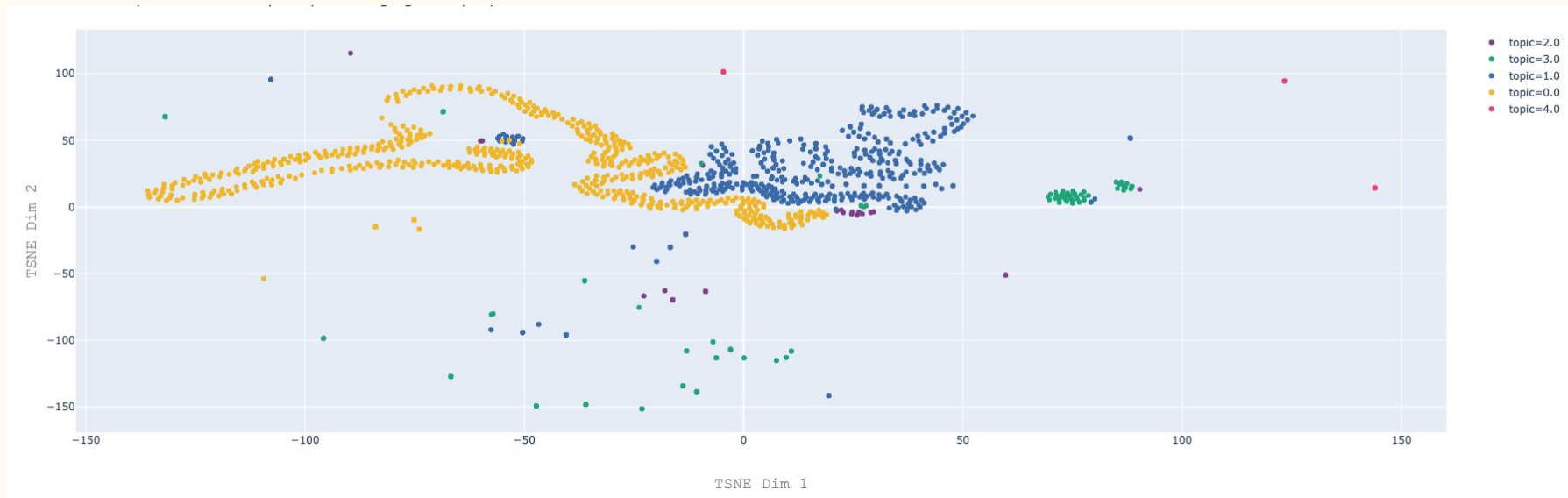
Topic 0 also looks promising with beguiling keywords such as “verification”, “verify”, “account”, and “code”

Data points in topic 0 containing str “verification”

```
topic0[topic0['Text'].str.contains("verification")]['Text']
```

67	[07:15:10]	[REDACTED]	COM11,206689795	verification code (8547) may only be used once to verify mobile number. For account safety, d
76	[07:15:13]	[REDACTED]	COM2,266266639,	verification code (2295) may only be used once to verify mobile number. For account safety, d
82	[07:15:16]	[REDACTED]	COM13,266290564	verification code (2858) may only be used once to verify mobile number. For account safety, c
89	[07:15:19]	[REDACTED]	COM9,205253118,	verification code (6673) may only be used once to verify mobile number. For account safety, d
90	[07:15:19]	[REDACTED]	COM6,503464996,	verification code (2068) may only be used once to verify mobile number. For account safety, d
91	[07:15:25]	[REDACTED]	COM8,503469360,	verification code (0059) may only be used once to verify mobile number. For account safety, d
95	[07:15:28]	[REDACTED]	COM3,206473997,	verification code (0936) may only be used once to verify mobile number. For account safety, d
96	[07:15:29]	[REDACTED]	COM12,204578749	verification code (4929) may only be used once to verify mobile number. For account safety, c
97	[07:15:29]	[REDACTED]	COM4,503867177,	verification code (9235) may only be used once to verify mobile number. For account safety, d
105	[07:15:39]	[REDACTED]	COM14,262683284	verification code (4704) may only be used once to verify mobile number. For account safety, c
109	[07:15:41]	[REDACTED]	COM32,206321439	verification code (4765) may only be used once to verify mobile number. For account safety, c
112	[07:15:42]	[REDACTED]	COM34,275421452	verification code (7264) may only be used once to verify mobile number. For account safety, c
117	[07:15:46]	[REDACTED]	COM30,500375998	verification code (0931) may only be used once to verify mobile number. For account safety, c
121	[07:15:49]	[REDACTED]	COM5,505781215,	verification code (6916) may only be used once to verify mobile number. For account safety, d
124	[07:15:50]	[REDACTED]	COM15,206474150	verification code (7014) may only be used once to verify mobile number. For account safety, c
125	[07:15:50]	[REDACTED]	COM38,204542343	verification code (3743) may only be used once to verify mobile number. For account safety, c
126	[07:15:50]	[REDACTED]	COM7,206472852,	verification code (5684) may only be used once to verify mobile number. For account safety, d
127	[07:15:51]	[REDACTED]	COM33,204422657	verification code (9614) may only be used once to verify mobile number. For account safety, c
128	[07:15:51]	[REDACTED]	COM27,200274447	verification code (7494) may only be used once to verify mobile number. For account safety, c
129	[07:15:51]	[REDACTED]	COM37,278460336	verification code (3443) may only be used once to verify mobile number. For account safety, c
130	[07:15:51]	[REDACTED]	COM42,207453870	verification code (7907) may only be used once to verify mobile number. For account safety, c
132	[07:15:54]	[REDACTED]	COM29,503866699	verification code (5625) may only be used once to verify mobile number. For account safety, c
133	[07:15:54]	[REDACTED]	COM39,503469165	verification code (8096) may only be used once to verify mobile number. For account safety, c
144	[07:16:07]	[REDACTED]	COM26,503867260	verification code (4175) may only be used once to verify mobile number. For account safety, c
145	[07:16:08]	[REDACTED]	COM36,206689325	verification code (5543) may only be used once to verify mobile number. For account safety, c
147	[07:16:09]	[REDACTED]	COM41,204894822	verification code (6444) may only be used once to verify mobile number. For account safety, c
148	[07:16:09]	[REDACTED]	COM1,204422657	verification code (4400) may only be used once to verify mobile number. For account safety, d

The trusted TSNE



Explore ML in InfoSec & Topic Modeling further !!

References

- <https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/>
- <https://blog.marketmuse.com/topic-modeling-for-seo-explained/>