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	Wn
D1	0	2	1	3
D2	1	4	0	0
D3	0	2	3	1
Dn	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
Dn	1	0	1	0

	W1	W2	W3	Wm
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(topic t / document d): proportion of words in document d that are currently assigned to topic t
  - P2 p(word w/ 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!!

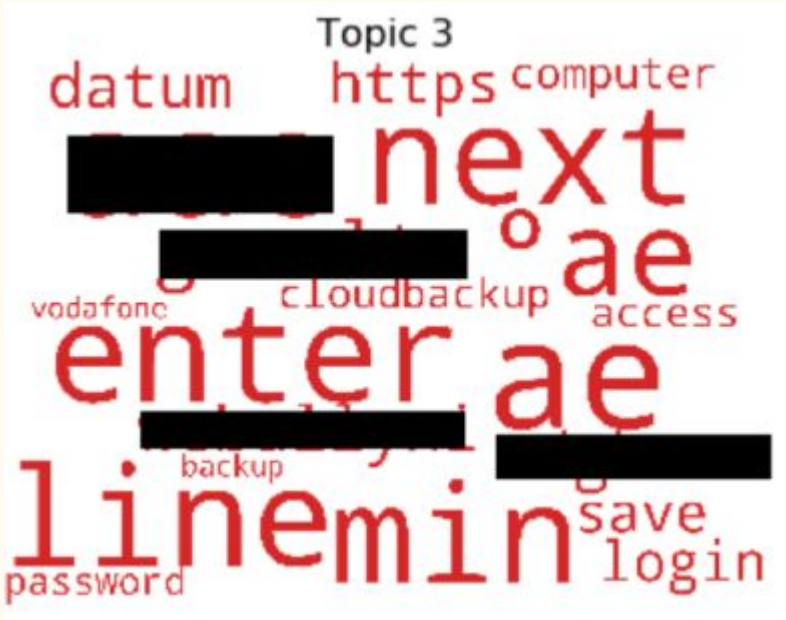
LDA Topic modeling | Step by step

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

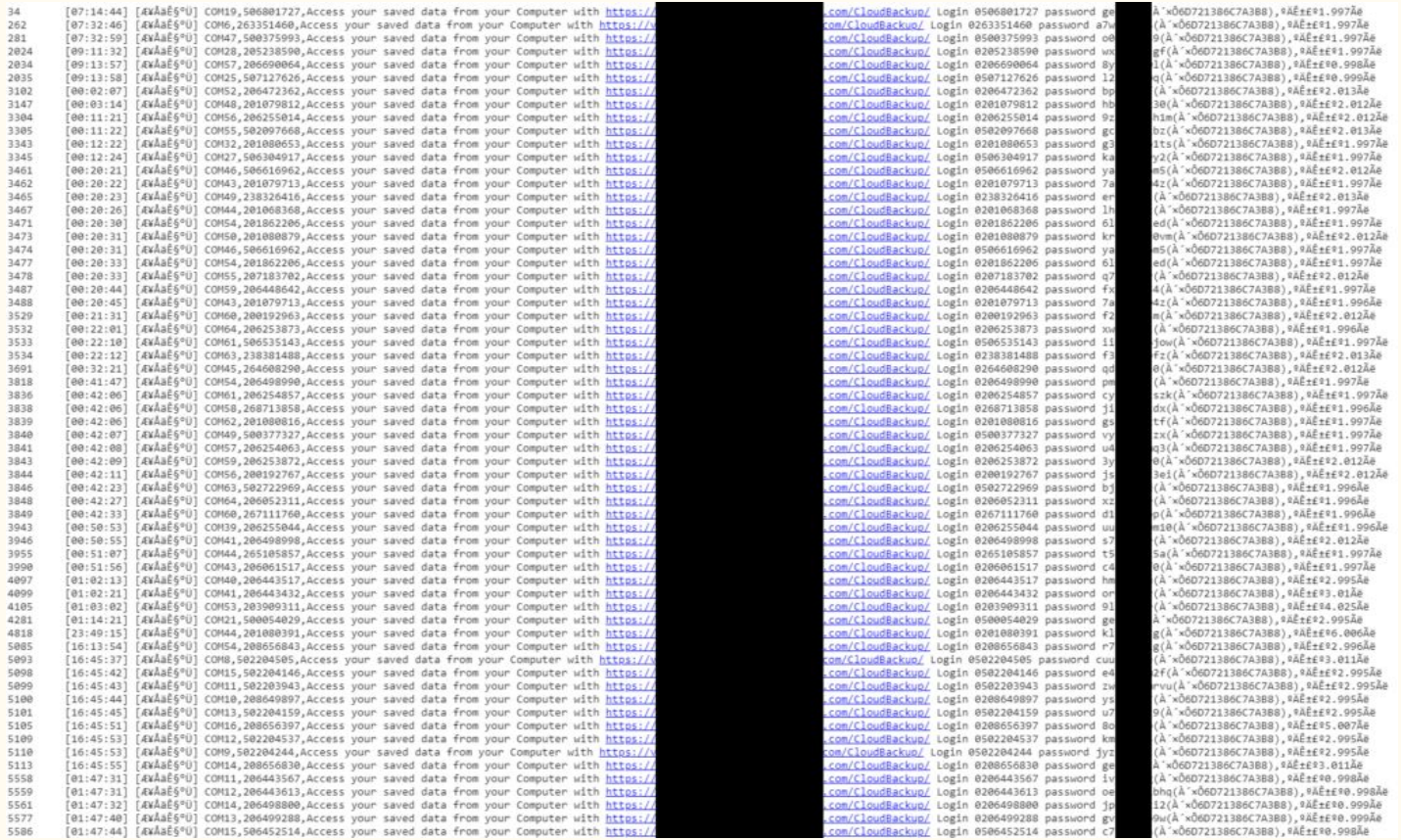
More on this here

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”



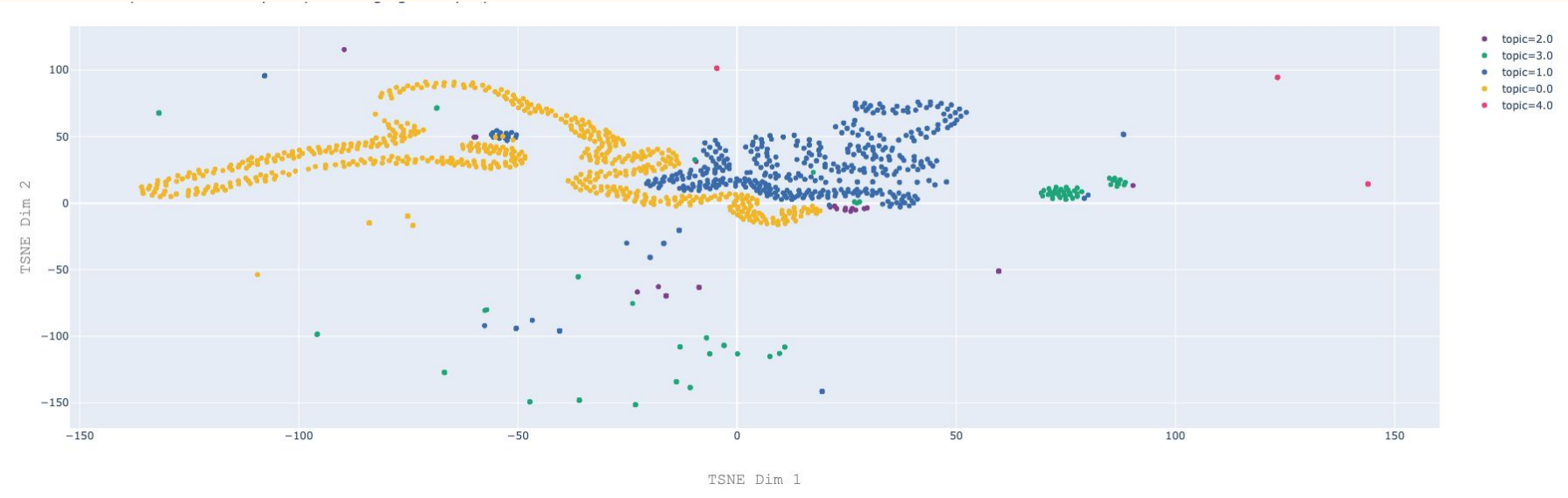
Topic 0 also looks promising with beguiling keywords such as “verification”, “verify”, “account”, and “code”

Data points in topic 0 containing str “verification”

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

67	[07:15:10] [Æ¥Åä³É¹¦] COM11,206689795	verification code (8547) may only be used once to verify mobile number. For account safety, d
76	[07:15:13] [Æ¥Åä³É¹¦] COM2,266266639,	erification code (2295) may only be used once to verify mobile number. For account safety, dc
82	[07:15:16] [Æ¥Åä³É¹¦] COM13,266290564	verification code (2858) may only be used once to verify mobile number. For account safety, d
89	[07:15:19] [Æ¥Åä³É¹¦] COM9,205253118,	erification code (6673) may only be used once to verify mobile number. For account safety, dc
90	[07:15:19] [Æ¥Åä³É¹¦] COM6,503464996,	erification code (2068) may only be used once to verify mobile number. For account safety, dc
91	[07:15:25] [Æ¥Åä³É¹¦] COM8,503469360,	erification code (0059) may only be used once to verify mobile number. For account safety, dc
95	[07:15:28] [Æ¥Åä³É¹¦] COM3,206473997,	erification code (0936) may only be used once to verify mobile number. For account safety, dc
96	[07:15:29] [Æ¥Åä³É¹¦] COM12,204578749	verification code (4929) may only be used once to verify mobile number. For account safety, d
97	[07:15:29] [Æ¥Åä³É¹¦] COM4,503867177,	erification code (9235) may only be used once to verify mobile number. For account safety, dc
105	[07:15:39] [Æ¥Åä³É¹¦] COM14,262683284	verification code (4704) may only be used once to verify mobile number. For account safety, d
109	[07:15:41] [Æ¥Åä³É¹¦] COM32,206321439	verification code (4765) may only be used once to verify mobile number. For account safety, d
112	[07:15:42] [Æ¥Åä³É¹¦] COM34,275421452	verification code (7264) may only be used once to verify mobile number. For account safety, d
117	[07:15:46] [Æ¥Åä³É¹¦] COM30,500375998	verification code (0931) may only be used once to verify mobile number. For account safety, d
121	[07:15:49] [Æ¥Åä³É¹¦] COM5,505781215,	erification code (6916) may only be used once to verify mobile number. For account safety, dc
124	[07:15:50] [Æ¥Åä³É¹¦] COM15,206474150	verification code (7014) may only be used once to verify mobile number. For account safety, d
125	[07:15:50] [Æ¥Åä³É¹¦] COM38,204542343	verification code (3743) may only be used once to verify mobile number. For account safety, d
126	[07:15:50] [Æ¥Åä³É¹¦] COM7,206472852,	erification code (5684) may only be used once to verify mobile number. For account safety, dc
127	[07:15:51] [Æ¥Åä³É¹¦] COM33,204422657	verification code (9614) may only be used once to verify mobile number. For account safety, d
128	[07:15:51] [Æ¥Åä³É¹¦] COM27,200274447	verification code (7494) may only be used once to verify mobile number. For account safety, d
129	[07:15:51] [Æ¥Åä³É¹¦] COM37,278460336	verification code (3443) may only be used once to verify mobile number. For account safety, d
130	[07:15:51] [Æ¥Åä³É¹¦] COM42,207453870	verification code (7907) may only be used once to verify mobile number. For account safety, d
132	[07:15:54] [Æ¥Åä³É¹¦] COM29,503866699	verification code (5625) may only be used once to verify mobile number. For account safety, d
133	[07:15:54] [Æ¥Åä³É¹¦] COM39,503469165	verification code (8096) may only be used once to verify mobile number. For account safety, d
144	[07:16:07] [Æ¥Åä³É¹¦] COM26,503867260	verification code (4175) may only be used once to verify mobile number. For account safety, d
145	[07:16:08] [Æ¥Åä³É¹¦] COM36,206689325	verification code (5543) may only be used once to verify mobile number. For account safety, d
147	[07:16:09] [Æ¥Åä³É¹¦] COM41,204894822	verification code (6444) 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/