Find Topics Across Twitter for All Known State Sponsored Hackers

By: Randy Grant

Abstract

This attempts to discover topics for all known state sponsored hackers for which cyber security (blue team) analysts have found evidence of activity. This evidence is known to be posted to Twitter with specific hashtags related to the specific hacker.

Design

The first step was to grab twitter posts with specific hashtags. Since I'm familiar with the cyber security domain, I created a known good list based on this correlated list of state sponsored actors that many cyber security practitioners use for reference that would be used to query for hashtags on Twitter (i.e. #apt01). Using a trial membership to PhantomBuster, the hashtags were then queried across Twitter and all twitter posts with those hashtags were returned into a json file. Importing the json into a dataframe, EDA was performed, and then 2 methods of topic modeling were performed. The first method is based on creating a vector with TF-IDF to then be input into a NMF model to which the top 20 topics were visualized. The second method was similar to the first, but I used Gensim to generate a LDA model with a visualization using the pyLDAvis library.

Data

As stated in the Design section, the data used for this project were a known good list of state sponsored actors based on a correlated list that cyber security practitioners use for reference. From that, a Twitter json was returned from PhantomBuster, to which these columns were put into a dataframe:

- tweetDate
- content
- twitterProfile
- tweetUrl
- timestamp
- query

Algorithms

Models and Vectorizers

NMF and LDA were used for method 1:

- Count vectorizer was used to feed LDA
- TF-IDF was used to feed 2 versions of NMF:
    - Kullback-Leibler divergence
    - Frobenius norm (Euclidean Distance)

Gensim's TfidfModel and LDA were used for method 2. The TfidfModel created was based on the ngrams of 1-3 using id2word.

Tools

Python, pandas and spacy for data manipulation, scikit-learn and gensim for models and scores, matplotlib and pyLDAvis for plots and graphs.

Communication

These slides and visualizations were presented, and this was uploaded to github.