CS410 Tech Review - Spotting Trends in Noisy Text Data

daechoi2@illinois.edu

Introduction

Standard natural language processing (NLP) is tuned for organized and collated contents such as news articles. Detecting trending topics on noisy social media text data, with the varying spelling of words and creative use of punctuation, is challenging. Computer scientists and linguists from Cambridge University have combined with applying NLP techniques to pick out trends in discussions on an underground cybercrime site called the HackForums. They modified a method used for capturing the linguistic changes between two corpora to be used as a trending topics tool for temporal analysis of data. They evaluated the approach by comparing the results to TF-IDF using the discounted cumulative gain metric with human annotations and found their method outperforms TF-IDF on information retrieval.

Research

Social media and underground hacking forums contain an extensive collection of noisy text data around various topics, with misspellings, changing lexicons, and slang phrases. The evolving domain-specific lexicon includes homonyms, where “rat” may be identified as an animal by standard NLP tools, but is typically defined as a type of malware “remote access trojan” in this context.

TF-IDF and LDA are both commonly used, but these have limitations. Using TF-IDF on forum data would require (1) stemming or lemmatization (grouping inflected forms of a word), and (2) defining a document either as individual posts, or a thread of posts, for best performance. Similar to TF-IDF, LDA requires finding a suitable tokenization approach and representation of a document. While LDA learns a distribution of terms in topics, this is not as lightweight computationally as TF-IDF. For detecting trend topics, burst and dynamic topic models have been commonly used.

The research team developed a technique called “weighted log-odds ratio” (WLOR), built upon a weighted log-odds ratio tool (Monroe et al., 2008) with an informative Bayesian prior (Silge et al., 2020). The team adapted the Bayesian prior approach, created for comparing two corpora, to detect trending tokens. Instead of selecting corpora by pre-existing classes, the team chose prior and target time windows, to find terms that are more likely to appear in the target period. To detect trend topics, the team used a different approach similar to “two-point trends” discussed by Kleinberg (2016) with “rising” and “falling” words. The team also used a Bayesian approach instead of measuring absolute change.

The team tested WLOR on HackerForums posts, referencing the spread of the WannaCry ransomware, a type of ransomware that encrypts data until the victim pays a ransom. The team used the same tokenization and pre-processing approach as the original log-odds and compared WLOR to TF-IDF for topic ranking and used. WLOR achieved better results than “term-frequency inverse-document-frequency” (TF-IDF). For the WannaCry event, the WLOR tool scored 0.979 compared to the TF-IDF of 0.877. For the random event, the log-odds tool scored 0.978 compared to the TF-IDF of 0.753. WLOR had a greater discounted cumulative gain score for both events than the TF-IDF approach, finding the ranking of terms provided by WLOR produced more relevant salient terms than the TF-IDF method.

Conclusion

Detecting trends in an underground hacking forum's noisy text data is an interesting use-case for the log-odds based WLOR technique. This technique can have many other uses cases around social media texts and identify a new trend. Future research can include supporting non-English texts and incorporating multiword and named entity recognition. Many cyber-crime forums are not using English, which can add complexity to the analysis. While the WLOR technique is statistical and language-independent, further research can assess the technique’s performance on foreign language data. Also, incorporating the detection of multiword expressions and named entity recognition techniques for noisy language is likely to be useful in analyzing language use in social media forums.

References

Detecting Trending Terms in Cybersecurity Forum Discussions by Jack Hughes, Seth Aycock, Andrew Caines, Paula Buttery, Alice Hutchings, 2020

https://www.cl.cam.ac.uk/~joh32/papers/2020trending.pdf