Digital Investigation xxx (2018) 1–11

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

DFRWS 2018 USA — Proceedings of the Eighteenth Annual DFRWS USA

Analyzing the DarkNetMarkets subreddit for evolutions of tools and trends using LDA topic modeling

Kyle Porter

Department of Information Security and Communication Technology, NTNU, Gjøvik, Norway

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Topic modeling
Latent dirichlet allocation
Web crawling
Datamining
Semantic analysis
Digital forensics
Surface-web monitoring

ABSTRACT

Darknet markets, which can be considered as online black markets, in general sell illegal items such as drugs, firearms, and malware. In July 2017, significant law enforcement operations compromised or completely took down multiple international darknet markets. To quickly understand how this affected the markets and the choice of tools utilized by users of darknet markets, we use unsupervised topic modeling techniques on the DarkNetMarkets subreddit in order to determine prominent topics and terms, and how they have changed over a year's time. After extracting, filtering out irrelevant posts, and preprocessing the text crawled from the subreddit, we perform Latent Dirichlet Allocation (LDA) topic modeling on a corpus of posts for each month from November 5th, 2016 to November 5th, 2017. Our results indicate that even analyzing public forums such as the DarkNetMarkets subreddit can reveal trends and keywords related to criminal activity and their methods, and that users of the dark web appear to be becoming increasingly more security-minded due to the recent law enforcement events.

Introduction

The *dark web*, websites which can only be reached through anonymity networks such as Tor (Dingledine et al., 2004) and are not indexed by any search engine, is well known for hosting criminal marketplaces. These *darknet markets* sell illicit items such as drugs, weapons, and hacking tools. Recent research shows that the markets and forums on the dark web have been valuable sources of cyber threat intelligence (Deliu, 2017; Nunes et al., 2016; Samtani et al., 2015), as well as general sources of intelligence for law enforcement agencies (Van Buskirk et al., 2016) to monitor the state of darknet markets to identify emerging trends.

In July 2017, two of the most popular darknet markets, AlphaBay and Hansa, were shut down by various law enforcement agencies (Gibbs and Beckett, 2017). From this, one may conjecture that the state of darknet markets and their customers may be going through a more tumultuous period than usual. To analyze the effect of these real world events, and to identify changes in behavior by darknet customers, we gather data from public sources on Reddit.[1]

Reddit is a social media platform with specific interest oriented forums called subreddits, and in this work we extract intelligence from the subreddit called DarkNetMarkets. We crawl this subreddit for a year's worth of posts and data to obtain a corpus for each month between November 5th, 2016 to November 5th, 2017. Ultimately, 15,400 posts were gathered from the subreddit, and since manually analyzing the corpora is exceedingly time consuming, we utilize Latent Dirichlet Allocation (LDA) (Blei et al., 2003)) unsupervised topic modeling to extract month to month information pertaining to the state of the darknet markets, the security and anonymity tools used by visitors to the darknet markets, and the cryptocurrency and related services used when purchasing items over the darknet.

The primary contribution of this work are the topics and terms produced by performing LDA topic modeling on the data from the DarkNetMarkets subreddit, wherein we can relatively quickly observe how the tools and the trends in markets, security, and cryptocurrency have changed from November 5th, 2016 to November 5th, 2017. From analyzing this data, following the July 2017 busts we can see an increase of uncertainty on the part of the users of the darknet markets, as well as an increase in security-mindedness. We note that law enforcement agencies are already monitoring this subreddit, so this information may already be known, but we none-the-less empirically show the effectiveness of LDA on criminally associated subreddits to quickly come to understand important content of a year's worth of subreddit posts.

The following is an outline of the paper. First we describe background information regarding details of Reddit, how darknet markets generally operate, and a brief overview of Latent Dirichlet

E-mail address: kyle.porter@ntnu.no.
[1] https://about.reddit.com/.

https://doi.org/10.1016/j.diin.2018.04.023

Allocation. The next section describes our experimental methodology, including our methods for extracting, filtering, and cleaning our dataset before applying topic modeling. Afterwards, we describe our topic modeling results for each month of posts on the DarkNetMarkets subreddit and analysis. Finally, we describe the related work, and conclude with a summary and discussion.

Background

In this section, we discuss attributes of Reddit, the darknet market community, and aspects of Latent Dirichlet Allocation so the topics produced by our experiments are more understandable.

The corpus: Reddit and subreddits

Reddit is a news aggregation and discussion website, where posts are organized into subreddits of specific interests. Subreddits are much like a standard "board" on a forum, and for our purposes the posts on these subreddits serve as the typical "threads". Each post has a title, potentially self-posted information by the author, and comments in response to the title or what was said by the author. Oftentimes, posts have a "flair", which is put in place by the author of the post or moderator of the subreddit that classifies the type of post being made. Furthermore, every post has a Unix timestamp associated with it, and therefore the corpus can be analyzed with respect to any given timeframe. The subreddit used for our experiments is "www.reddit.com/r/DarkNetMarkets".

DarkNetMarkets subreddit

The subreddit "DarkNetMarkets" is a public subreddit, where users discuss the goods and services of the black markets that can only be reached via an anonymity network such as Tor. Topics of conversation appear to mostly revolve around drugs, and the purchasing of drugs with cryptocurrency. More interesting topics of conversation include advice on increasing anonymity, operational security, tools used to improve stealthy financial transactions, and the state of markets and vendors. Vendors are essentially drug dealers who use the darknet markets as their platform to do business. Users purchase from vendors who they believe they can trust over darknet markets they believe are uncompromised using cryptocurrency.

The DarkNetMarkets subreddit has been a source of controversy in the past. In 2015, the FBI requested Reddit to reveal personal information regarding some of the DarkNetMarkets contributors (Knibbs, 2015). Surprisingly, the subreddit was banned on March 21, 2018 due to a new rule from Reddit administrators that forbids communities to use Reddit as a medium to exchange or perform transactions of prohibited goods or services (Franceschi-Bicchierai, 2018).

Latent Dirichlet Allocation

To perform unsupervised topic modeling on the data extracted from the Darknet Markets subreddit we use Latent Dirichlet allocation (LDA) (Blei et al., 2003)). This methodology was chosen as it is simple and is often used in a variety of sciences for topic modeling text corpora, which can be used as a type of text summarization of a large set of documents. LDA produces a model of a corpus of documents, where the model assumes that each of the documents in the corpus are derived from a generative process where each document consists as a distribution of a finite set of topics, each topic is a multinomial distribution of the vocabulary of words in the corpus, and each word of the document is drawn from one topic in the generative process.

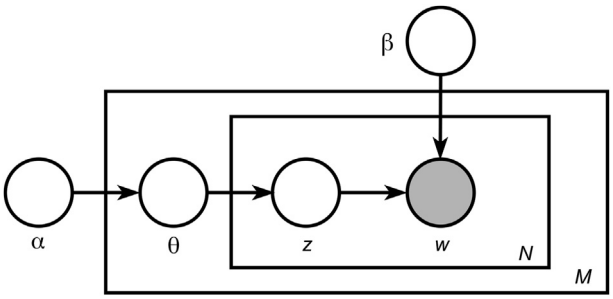

Fig. 1. Latent dirichlet allocation graphical model (Blei et al., 2003)).

Fig. 1 shows a graphical model of LDA, where the value w represents a vector of N words in document i of a total of M documents. A topic z is assigned to each word w_j of a document i, and therefore makes each document a composition of topics represented by some topic distribution θ over the document i. High α values represent that each document has a relatively even distribution of the topics, whereas low values of α indicate that the documents have a sparse distribution of all topics. Similarly, a high β value represents if topics are a relatively even distribution of the vocabulary of words, versus a low value of β which represents a sparse distribution of words per topic. Both α and β are set to default values of 1 divided by the number of topics for our experimentation (0.1). The latent elements we learn from running the LDA algorithm are the distributions of topics per document, and the distribution of words per topic.

A common issue regarding topic modeling via LDA is that the topics generated are not always interpretable or coherent by humans (Chang et al., 2009). To increase the certainty of being capable of classifying our generated topics, we use a relevancy metric introduced by Sievert and Shirley (2014). Typically, topics output a ranked list of the most probable terms in a topic, but this is often problematic as common terms in the corpus generally rank highly in the lists of multiple topics. This can make distinguishing topics difficult. The equation for relevance is given below.

$$\text{rel}(\text{term w} \mid \text{topic t}) = \lambda * p(w|t) + (1-\lambda) * p(w|t)/p(w) \qquad (1)$$

After generating the topic model of a corpus, we can adjust the weight λ to influence the word ranking per topic according to relevance. When $\lambda = 1$, the standard ranking is returned as it is simply the conditional probability of the word w given the topic t. As λ approaches 0, the weight of the ratio of the word-topic probability $p(w|t)$ to overall word probability $p(w)$ increases. In this fashion, words with high probability $p(w)$ are ranked lower as λ approaches 0.

Experimental methodology

In our experiment we wish to extract tools and trends as well as changes in tools and trends in the DarkNetMarkets subreddit from the topic models generated by the LDA algorithm with the subreddit posts as input. Of specific interest is to observe how these items have changed after the July 2017 busts. To accomplish this, we create a corpus of subreddit posts for each 12 months of the year, where we began to extract subreddit posts from 00:00 November 5th, 2016, and limit our data extraction to 00:00 November 5th, 2017 (UTC). From this corpus, we compose smaller corpora consisting of posts from the 5th of each month to the 5th of the next month, where then we preprocess each monthly corpus and prepare it as input into

Please cite this article in press as: Porter, K., Analyzing the DarkNetMarkets subreddit for evolutions of tools and trends using LDA topic modeling, Digital Investigation (2018), https://doi.org/10.1016/j.diin.2018.04.023