pyDNetTopic: A Framework for Uncovering What Darknet Market Users Talking About

Jingcheng Yang[1], Haowei Ye[2], and Futai Zou[1(✉)]

[1] School of Cyper Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
{amiya_yang,zoufutai}@sjtu.edu.cn
[2] School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China
beicheng@std.uestc.edu.cn

Abstract. Although Dark Net Market (DNM) has attracted more and more researchers' interests, we found most works focus on the markets while ignore the forums related with them. Ignoring DNM forums is undoubtedly a huge waste of informative intelligence. Previous works usually utilize LDA for darknet data mining. However, traditional topic models cannot handle the posts in forums with various lengths, which incurs unaffordable complexity or performance degradation. In this paper, an improved Bi-term Topic Model named Filtered Bi-term Model, is proposed to extract potential topics in DNM forums for balancing both overhead and performance. Experimental results prove that the topical words extracted by FBTM are more coherent than LDA and DMM. Furthermore, we proposed a general framework named pyDNetTopic for content extracting and topic modeling uncovering DNM forums automatically. The full results we apply pyDNetTopic to Agora forum demonstrate the capability of FBTM to capture informative intelligence in DNM forums as well as the practicality of pyDNetTopic.

Keywords: Darknet market forums · Topic modeling · FBTM · pyDNetTopic

1 Introduction

With the rapid development of Internet based applications, Darknet Markets (DNMs), which are online marketplaces hosted on anonymity networks and are not indexed by any search engine, has become great hotbeds for illicit transactions. DNMs provide escrow services between buyers and sellers transacting using Bitcoin or other cryptocurrencies, usually for illegal and unregulated goods such

This work is supported by the National Key Research and Development Program of China (No. 2017YFB0802300).

N. Park et al. (Eds.): SecureComm 2020, LNICST 335, pp. 118–139, 2020.
https://doi.org/10.1007/978-3-030-63086-7_8

as drugs, weapons and so on. Former researches mostly focus on the services, products on sale (especially drugs) or traders in DNMs [6,9,11] or the functionality and mechanism of DNMs [3,8], from both technical and legal perspectives. But few pay attention to the forums related with markets. In fact, DNM related forums are nonnegligible informative treasuries for researchers. The collecting, mining and analyzing of forum data can reveal valuable information about corresponding DNMs, as well as its users. Researchers can have a quick understanding about the transaction types or goods on sale by finding users in forums talks about "drugs", "cocaine" or "bitcoin", "Monero" frequently. Also, researchers can extract some security-crucial intelligence from discussion between forum users. Recent works explore approaches to extract Cyber Threat Intelligence (CTI) from open or dark web [2,16,24].

According to the value of information hidden in forum data and the lack of relevant researches, we propose to use topic modeling for data mining and topic extraction. Given a large, unstructured collection (or corpus) of documents, topic modeling is an unsupervised machine learning algorithm that estimates the main topics of discussion [4]. As far as we know, former researches simply use LDA for topic modeling and the topical items have poor interpretability and coherence. The purpose of this work is to presenting an effective topic model for darknet market users, as well as an automatic general framework for information mining in DNM forums. Filtered Bi-term Topic Model is presented in this paper not only for generating more coherent and interpretable topical items, but also to reduce consumption and complexity for practical application. The details and experimental results of FBTM will be elaborated in Sect. 3 and 5 respectively. We also propose a general topical mining framework for DNM forums called pyDNetTopic. The functionality of pyDNetTopic lies in forum data extraction and analysis to provide a profile of main items most DNM forum users concern about. Moreover, pyDNetTopic provides an auxiliary and preliminary process for further analysis on darknet. The architecture and implementation of pyDNetTopic will be detailed in Sect. 4.

The contribution of this paper are as follows:

- We propose FBTM, a modified Bi-term Topic Model applicable in large dataset consist of documents with various lengths. FBTM not only handles the sparsity problem when facing short text, but also reduces computational and memory consumption for large corpora, which makes it appropriate for darknet environment. Experimental results show that FBTM has better coherence measures than baseline models LDA and DMM. Comparison of extracted items demonstrates the accuracy and coherence of FBTM surpass baseline models, which indicates FBTM captures the tools used by darknet users and the hot issues visitors discuss that are usually ignored by baseline models.
- We propose pyDNetTopic, a python-based automatic data extraction and topic modeling framework for DNM forums. pyDNetTopic integrate standard pipelined preprocessing steps with FBTM and two alternative baseline models. The functionality of pyDNetTopic is identifying and extracting textual

content from files scraped by web crawler, as well as hot topic detection in darknet. pyDNetTopic aims to provide a general framework for data mining and intelligence extraction for security researchers quicikly understanding underground communities.
– We utilize pyDNetTopic to reveal the latent topics with representative terms from Agora market related forums. Items of interest including the state of darknet markets as well as the security concern of darknet users provide insight to darknet or security researchers.

2 Related Work

The exploration of informative assets beneath darknet have been attracted the interests of security researchers. Topic modeling, as an unsupervised topic derivation method for key information extraction, has been used extensively to gather intelligence directly from darknet informative sources. Grisham et al. [10] analyzes the Alphabay underground marketplace - an anonymous trading grounds for illicit goods and services. This poster uses LDA to determine the providing illicit items and top sellers. Samtani et al. [22,23] proposed a general semi-automatic framework to identify and classify informative assets from underground hacker forums. The hacker assets can be categorized into attachments, tutorials and source code according to Samtani. LDA was utilized in this paper aiming for code analysis by extracting topics and functions of source code, as well as understanding the topics of attachments and tutorial postings. In terms of code analysis, Samtani train a SVM classifier to classify source codes by coding language to provide insight into their implementation. Similar to Samtani, Deliu [7] presents a two-stage, hybrid process to collect CTI from underground hacker forums. Deliu achieves CTI extraction from forums by using a hybrid machine learning model that automatically searches through hacker forums posts, identifies the posts that are most relevant to cyber security and then clusters the relevant posts into estimations of their topics. The first identification stage and second clustering stage use SVM and LDA, respectively. Kyle [19] utilize LDA to reveal states of darknet market and tools users use mostly. Kyle doesn't apply LDA directly on original materials collected from darknet, but on subreddit "DarkNetMarket" in a monthly manner.

Lots of works have been done in topic modeling to improve its accuracy and interpretability, especially in short text environment such as Twitter. Some remarkable modifications include the auxiliary of external knowledge. [17,18] propose to incorporate hidden topics discovered from large-scale external document collections into short sparse documents. Jin et al. [12] propose a novel topic model - Dual Latent Dirichlet Allocation (DLDA) model, which jointly learns two sets of topics on short and long texts and couples the topic parameters to cope with the potential inconsistency between data sets. Recently neural networks are used to learn the prior distribution of document-topic or topic-word from corpus, which are imposed to be Dirichlet distribution in LDA and its modified versions. Deep hierarchical priors, generated from Replicated Softmax Model

[21], Neural Autoregressive Density Estimator [13] or variational autoencoders [26,30] have been developed to generate hierarchical document representations as well as discover interpretable topic hierarchies. But it is worth noting that security studies mining DNM forums only use LDA. We suppose the reasons why security researchers ignore recent achievements on topic modeling lie in:

a) Security researchers focus on the clustering results or extracted keywords from topic modeling while ignoring the accuracy, coherence and interpretability of obtained results;
b) The lack of convenient framework of proposed algorithms hinder security researchers from adopting them into practical application. Most works using LDA may due to the fact that LDA has been integrated into gensim or sklearn packages, which makes it user-friendly.

In view of the development of topic modeling and its relatively backward application in darknet research, this paper is intended to present an algorithm more effective and accurate than commonly used baseline models in darknet forums environment to fill the gap. For ease of use, especially for security researchers who are unfamiliar with machine learning, we propose a general framework for information mining and topic extraction on DNM forums.

3 Background

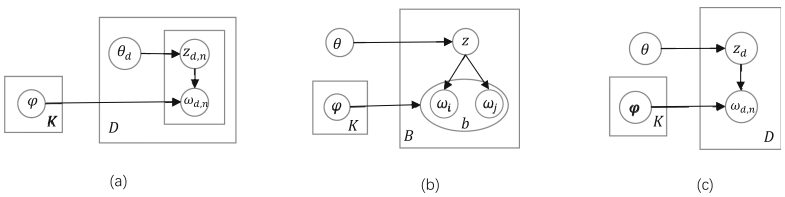

Fig. 1. Graphical representation of three topic models: (a) LDA, (b) BTM, and (c) DMM

3.1 LDA

As a classic algorithm of existing topic models, LDA [4] is a hierarchical parametric Bayesian approach to infer a low dimensional representation that captures the latent semantic topics in a large corpus. LDA models the documents as a mixture of latent topics while each topic can be described as a probabilistic distribution over words. Figure 1(a) illustrates LDA models the generation process of a word $\omega_{d,n}$ in document d within N words corpus composed of D documents and K topics.

a) Draw word mixing distribution $\varphi_k \sim Dir(\beta)$ for each topic $k \in K$
b) For each document d in D:
   - Draw topic mixing distribution $\theta_d \sim Dir(\alpha)$
   - Generate a topic assignment $z_{d,n}$ from document-topic distribution $Multi(\theta_d)$
   - Generate a word $\omega_{d,n}$ from topic-word distribution $Multi\left(\varphi_{z_{d,n}}\right)$

3.2 BTM

The reliance on word co-occurrence patterns within documents makes LDA model extremely sensitive to the length of documents. When facing short text, LDA model suffers from severe sparsity problem. To handle this, Bi-term Topic Model [28] utilizes a bi-term (a single word co-occurrence pair in a single document) rather than a single word. BTM models topic distribution over the whole bi-terms collections instead of deriving it from Dirichlet prior distribution in each document by LDA. Experimental results demonstrate that BTM exceeds traditional models such as LDA not only in short text, but also in normal text environment. Figure 1(b) illustrates the generation process of a single bi-term (ω_i, ω_j) in the whole bi-terms set B.

a) Draw word mixing distribution $\varphi_k \sim Dir(\beta)$ for each topic $k \in K$
b) Draw topic mixing distribution $\theta \sim Dir(\alpha)$ for B
c) For each bi-term b in B:
   - Generate a topic assignment z from topic distribution $Multi(\theta)$
   - Generate bi-term $b = (\omega_i, \omega_j)$ from topic-word distribution $Multi(\varphi_z)$

3.3 GSDMM

Collapsed Gibbs Sampling for Dirichlet Multinomial Mixture [29] introduces Gibbs Sampling algorithm into Dirichlet Multinomial Mixture model to probabilistically estimate the mixture component (cluster) and the probability of topic-word distribution based on two assumptions:

a) Each document is generated by a mixture model;
b) Each document originates from only one topic.

These assumptions , especially the second assumption, can indirectly enrich the word co-occurrence pattern in document level. Figure 1(c) illustrates GSDMM models the generation process of a word $\omega_{d,n}$ in document d.

a) Draw word mixing distribution $\varphi_k \sim Dir(\beta)$ for each topic $k \in K$
b) Draw topic mixing distribution $\theta \sim Dir(\alpha)$
c) For each document d in D:
   - Generate a topic assignment z_d from document-topic $Multi(\theta)$
   - Generate a word $\omega_{d,n}$ from topic-word distribution $Multi(\varphi_{z_d})$

4 Filtered Bi-Term Topic Model

4.1 Motivation

Former researches focus on the sparsity problem topic models confront in short text and solution to it. But in DNM forums, situation seems more complex. The post corpus in DNM forums cannot be regarded as short or long simply. Unlike typical short text collections, such as Tweets2011 or Trec, whose average document lengths are 5.21 and 4.94 respectively. The average document length of DNM forums are longer than 20 words typically. Although on average, DNM forum data seems to fall into the category of long text. However, DNM forum contains a large number of very short posts, while some overwhelmingly long posts increase the average length of the entire dataset. So simply using LDA may receive unsatisfying results. From this point of view, BTM seems to be a propriate choice which produce quite great performance in both short text and long text. However, when performing BTM, another serve problem emerges.

As detailed in Sect. 3, BTM extracts bi-terms from corpus to obtain global word co-occurrence patterns. But with the document length $\bar{l}$ and number $|D|$ increasing, the number of bi-terms increases in square law, which incurs unaffordable memory usage and time consumption. Table 1 illustrates the complexity analysis of three typical topic models: LDA, BTM and DMM, where M and K denotes the number of terms in corpus and the number of topics respectively. The redundancy of bi-terms in BTM makes the model cannot be applied in DNM forums directly, furthermore, in any big dataset with various document lengths.

Table 1. Complexity analysis of LDA, DMM and BTM

Method	Time complexity	Memory complexity
LDA	$O(K\|D\|\bar{l})$	$\|D\|K + MK + \|D\|\bar{l}$
DMM	$O(K\|D\|\bar{l})$	$D + (M+2)K$
BTM	$O(K\|D\|\bar{l}^2)$	$K + MK + \frac{1}{2}\|D\|\bar{l}(\bar{l}-1)$

4.2 Methodology

Since the inefficiency of BTM attributes to the redundancy of bi-terms, an intuitive solution is discarding less indicative bi-terms. Similar work has been done in [27] to discriminate topical terms using document ratio $df\,(\omega)$, which is calculated in (1).

$$df\,(\omega) = \frac{m_\omega}{M} \tag{1}$$

where m_ω represents the number of documents in which term ω is mentioned and M denotes the total number of documents. Xia et al. [27] pointed out that the term belongs to general terms when its $df\,(\omega)$ is larger than M devided by the

maximum number of documents a single topic contains. Besides, a term appears only in single document belongs to document-specific terms while remaining terms are topical terms. Xia's approach seems to have a good performance in headline-based social news clustering, but it cannot be adopted in DNM forums situations. The major drawback lies in the threshold for defining general terms. In DNM forum corpus, it is impossible that we have a prior knowledge of the approximate number of documents the largest topic contains. The "unsupervised" situation in DNM forums makes Xia's approach impractical. Inspired by that, we propose Filtered BTM to select useful bi-terms for topic extraction without any prior knowledge or assumptions.

In FBTM, we propose a novel metric, named "generality" $\gamma(\omega)$, which is formulated as (2).

$$\gamma(\omega) = \frac{f(\omega)}{df(\omega)} = \frac{\frac{n_\omega}{\sum_\omega n_\omega}}{\frac{m_\omega}{M}} \tag{2}$$

where n_ω represent the occurrence number of term ω in the whole corpus and the number of documents which contain term ω. $f(\omega)$ measures the occurrence frequency of term ω in perspective of whole corpus, whereas $df(\omega)$ reflects the document-level distribution range of term ω. So $\gamma(\omega)$ reflects how centrally the term ω distributes in documents. We believe there is a positive correlation between $\gamma(\omega)$ and the degree of how topical term ω is. Firstly, the topic-word distribution ϕ is estimated as (3) in [28].

$$\phi_{\omega|z} = \frac{n_{\omega|z} + \beta}{\sum_\omega n_{\omega|z} + M\beta} \tag{3}$$

where $n_{\omega|z}$ refers to the number of times of term ω assigned to topic z. β is a hyperparameter. The denominator seems to remain almost constant for different term ω. It is intuitive that the term with larger $f(\omega)$ tends to have larger $n_{\omega|z}$. In another word, the term occurs frequently tends to be assigned to one or multiple topics. It can be proved by the observation of output results from many topic models that some frequent words are regarded as top topical words spanning several topics. But only considering the word ratio will make words appearing frequently in most documents dominate the topical words. Although stop words filtering can exclude some commonly used words such as "the", "do" et al., some words appear frequently in darknet context such as "buy", "pay", "from" can still make results confusing. So we introduce the reciprocal of document ratio $df(\omega)$ into the calculation of "generality". If a term appears frequently in corpus, as well as widely in the whole document set, the term belongs to common words so its $f(\omega)$ and $df(\omega)$ will be large simultaneously. Dividing by a large $df(\omega)$ will reduce generality to prevent common word from ranking too high.

Secondly, in the view of Xia's classification approach, the metric generality still makes sense. Xia regards the topical terms and general terms as significant for topic modeling. Due to our observation, the topical terms always appear frequently in a few documents while general terms distribute more broadly than topical terms in document level. The top ranked terms by generality mostly consist of topical terms and general terms. It is worth noting that the so-called

document-specific terms, which appear only in one document, still has a relatively low generality, although their document ratio is quite low. Because such terms appear quite rarely and have much lower word ratios. So our proposed generality has capacity to exclude less indicative terms.

The detail of FBTM algorithm is shown in Algorithm 1. After the calculation of $\gamma(\omega)$, FBTM sorts each term in the whole corpus according to its $\gamma(\omega)$ and only retains the bi-terms consist of terms whose $\gamma(\omega)$ is in top η, which is filtering coefficient specified by researcher determining how many terms in corpus are retained. The rest bi-terms are considered as meaningful for topic modeling. The rest bi-terms set B is used by BTM to estimate the topic-word distribution ϕ, as well as some user-determined parameters. Experimental results indicate that by excluding trivial bi-terms extracted topical terms better capture the informative intelligence from darknet corpus.

Algorithm 1. Filtered Biterm Topic Model

Input: number of topics K, preprocessed document set D, hyperparameter α, β, filtering coefficient η, iterations I
Output: topic-word co-occurrence matrix Φ

for each word ω in D **do**
  Compute generality $\gamma(\omega) = \frac{f(\omega)}{df(\omega)}$
end for
Sort each word according to its generality
Select the top η words and add them to filtered word set W
for each document d in D **do**
  for each $\omega_i, \omega_j \in d$ **do**
    if $\omega_i, \omega_j \in W$ **then**
      Add bi-term (ω_i, ω_j) into bi-term set B
    end if
  end for
end for
$\Phi \leftarrow BTM(B, K, \alpha, \beta, I)$

5 Framework Architecture

Figure 2 is the overview of framework of pyDNetTopic[1]. The purpose we present pyDNetTopic is extracting the textual content from HTML file obtained by web crawler and uncovering interesting topics with representative words. Based on previous works, we integrate LDA, DMM with FBTM which presented in this paper in pyDNetTopic.

[1] The code is available in https://github.com/blade-prayer/pyDNetTopic.

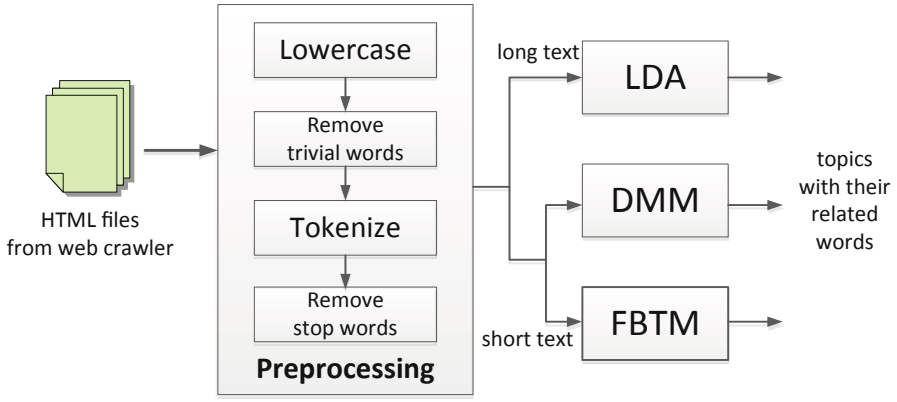

Fig. 2. Overview of architecture of pyDNetTopic

5.1 Data Extraction and Preprocessing

Based on our observation of dataset scraped from darknet, data extraction module utilizes the strategy that choosing files with "topic" in their names and selecting the "inner" elements to get textual content of posts. After filtering the nonrelevant files and finding the requisite content, here with different modes pyDNetTopic provides different text extraction strategies:

a) In long text mode, pyDNetTopic aggregate every single response within a post into one document;
b) In short text mode, pyDNetTopic write every single response into one document.

Aforementioned two modes correspond with different applicable conditions of topic models. LDA model is fit to long texts while FBTM and DMM is suited for short texts. Then pyDNetTopic performs a pipelined preprocessing procedure to the raw bulks of content to transform them into appropriate form for topic model. Such procedure is concluded by our series of experiments combined with related works [19]. We integrate these sequential steps into pyDNetTopic to make it convenient for follow-up studies.

- Lowercase texts
- Remove pure numbers with length less than 2
- Remove words with only one character or longer than 15 characters
- Remove words that appear only once
- Tokenize raw data using regular expression that match with characters a-z0-9
- Remove standard stop words in NLTK Python Library with specific stop words obtained from our observation during series experiments

During the experiments, we find that a mass of posts' topics center in drugs, which contains a lot pure numbers indicating the weights or prices of drugs.

Addition to numbers, such posts always contain a lot ill-formed words with only one character. Another phenomenon we find is that there are many quite long meaningless words stemming from the URL or PGP public keys in generated corpus. These words appearing in extracted topics will make the output confusing. pyDNetTopic removes these meaningless words out of consideration for efficiency and performance.

The selection of localized stop words is inspired by how Kyle did in [19]. When we just use the standard stop words list, we find the results of topic model output revolved around curse and common words, such as "would", "shit", "good". To alleviate this we perform FBTM repeatedly on our generated corpus and choose the top-20 most relevant words within each topic as representative word set. When we find words that add no insight to the potential topics in typical word set, we remove such words from raw corpus by adding them to the stop word list. We repeat this operation until in the top-20 typical word set there are no obviously meaningless words. Appendix A shows our final result of stop words list by performing in typical DNM forums. We believe this stop words list we concluded can be directly applied in future related researches.

5.2 Topic Models

Although FBTM has shown its good performance, we still have some reasons to retain LDA and DMM in pyDNetTopic. Firstly, although after filtering, FBTM still costs longer process time than DMM and LDA when facing a level of tens of thousands data volume. Secondly, we assume that researchers may confront some weird forums consist of abnormally long or short posts. We integrate LDA and DMM as alternative models in pyDNetTopic. The implementation of LDA is LDAMulticore in genism library. For hyperparameters α and β, we simply use the default setting in gensim as the reciprocal of topic number K. pyDNetTopic implements DMM in python. The prerequisite assumption that each document is originated from only one topic makes DMM deal with short text both rapidly and efficiently, however, at the expense of performance and accuracy in DNM forums conditions. We utilize the same default settings as in LDA for model parameters. The default and recommended topic model in pyDNetTopic is our proposed FBTM with short text mode. We set the default values of hyperparameters the same as that in [28]. The default value of filtering coefficient η which controls how many words are retained for topic modeling, is 0.5. The hyperparamter settings in pyDNetTopic are shown in Table 2.

5.3 Relevance Metric

In pyDNetTopic, we choose relevance of each term rather than its topic-word probability to rank the top topical word list of each topic. Relevance is a topical term weighted metric proposed in [25]. The definition of relevance makes a trade-off between the topic-word probability with its corresponding lift (calculated by the

Table 2. Hyperparameter settings in pyDNetTopic

Model	Parameter	Value
LDA	α	$\frac{1}{K}$
	β	$\frac{1}{K}$
	Iterations	100
DMM	α	$\frac{1}{K}$
	β	$\frac{1}{K}$
	Iterations	100
FBTM	α	1
	β	0.01
	Iterations	100
	Filtering coefficient η	0.5

topic-word probability with its marginal probability). The formulation of relevance of term ω belonging to topic t given a weight parameter ε is shown in (4).

$$r\left(\omega, t|\varepsilon\right) = \varepsilon \log\left(\phi_{t\omega}\right) + \left(1-\varepsilon\right) log\left(\frac{\phi_{t\omega}}{p_{\omega}}\right) \tag{4}$$

By introducing the lift term, relevance decrease the rankings of globally frequent terms. On the other hand, the topic-word probability term balance the noise introduced by lift term. Parameter ε determines the weight given to the probability of ω under topic t relative to its lift. We set it 0.01 empirically in pyDNetTopic.

6 Experiment

6.1 Evaluation Metrics

In order to perform a quantifying analysis, we choose a series of automatic and unsupervised evaluation metrics, coherence scores, summarized by Roder [20] for topic quality evaluation. Coherence measures do not rely on any human annotators or extrinsic reference collections. The insight of coherence is based on the observation of human expert annotations that pairs of words belong to the same topic clustered by the model tend to co-occur in the same document, while word pairs belonging to different topics have little co-occurrence tendency.

We choose UMass coherence [14], UCI coherence [15] and centroid coherence [1] for topic quality evaluation. The evaluated topic UCI coherence takes the set of top words of a topic and sum a confirmation measure over all word pairs, while UMass coherence uses an asymmetrical confirmation measure between top word pairs (smoothed conditional probability). The formulations of UMass and UCI coherence are shown in (5) and (6) respectively, where $V^{(t)} = \left(v_1^{(t)}, \ldots, v_M^{(t)}\right)$ is a list of the M most probable words in topic t.

$$C_{UMass}\left(t;V^{(t)}\right)=\sum_{m=2}^{M}\sum_{l=1}^{m-1}log\frac{D\left(v_m^{(t)},v_l^{(t)}\right)+1}{D\left(v_l^{(t)}\right)} \tag{5}$$

$$C_{UCI}\left(t;V^{(t)}\right)=\sum_{m=1}^{M-1}\sum_{l=m+1}^{M}log\frac{D\left(v_m^{(t)},v_l^{(t)}\right)+1}{D\left(v_m^{(t)}\right)D\left(v_l^{(t)}\right)} \tag{6}$$

The calculation of centroid coherence is a bit more complicated. Centroid coherence is formulated based on context vectors for every topic top word. The j-th element of the context vector $\vec{v_i^{(t)}}$ of topical word $v_i^{(t)}$ is normalized pointwise mutual information (NPMI) and calculated as:

$$v_{ij}^{(t)}=NPMI\left(v_i^{(t)},v_j^{(t)}\right)^{\gamma}=\left(\frac{log\frac{D\left(v_i^{(t)},v_j^{(t)}\right)+1}{D\left(v_i^{(t)}\right)D\left(v_j^{(t)}\right)}}{-\log\left(D\left(v_i^{(t)},v_j^{(t)}\right)+1\right)}\right)^{\gamma} \tag{7}$$

For topic t, the centroid vector $\vec{v_c^{(t)}}$ is the sum of each context vector of topical words:

$$\vec{v_c^{(t)}}=\sum_{v_i^{(t)}\in V^{(t)}}\vec{v_i^{(t)}} \tag{8}$$

The final coherence is computed as average cosine similarity between top word context vectors and their centroid $\vec{v_c^{(t)}}$.

$$C_{cen}=\frac{1}{M}\sum_{v_i^{(t)}\in V^{(t)}}cos\left(\vec{v_i^{(t)}},\vec{v_c^{(t)}}\right) \tag{9}$$

6.2 Performance Comparison

In order to make a comprehensive comparison of the clustered topic quality between three topic models: LDA, FBTM and DMM, we choose Agora forum dataset in the Darknet Market Archives[5]. Agora forum dataset is organized in the subfolders whose names indicate the dates when Branwen employed his crawler tools. Valuable intellegence can be obtained from forum user profiles and threads, which began when a forum user created a post that was followed by numerous responding posts. For time limitation, we randomly choose 1000 files from a part of subfolders, which span a whole year in 2014, for topic models in pyDNetTopic respectively and calculate the evaluation metrics. Our setting of topic number is 3 for all models. Another hyperparameters use default settings in pyDNetTopic.

The UMass, UCI and centroid coherence of three topic models in different dates of Agora forum dataset is shown in Figs. 3, 4 and 5 respectively. The average UMass coherence of FBTM fluctuating around 50 is remarkably higher than LDA and DMM. FBTM achieves almost 150% and 250% higher measures than

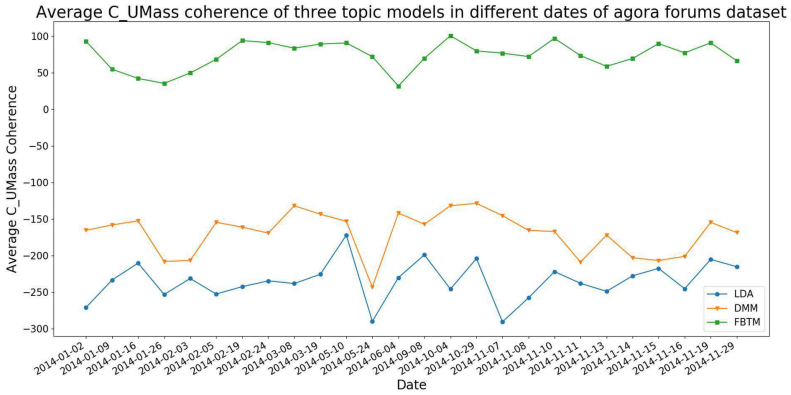

Fig. 3. Average UMass coherence of three topic models in different dates of Agora forums

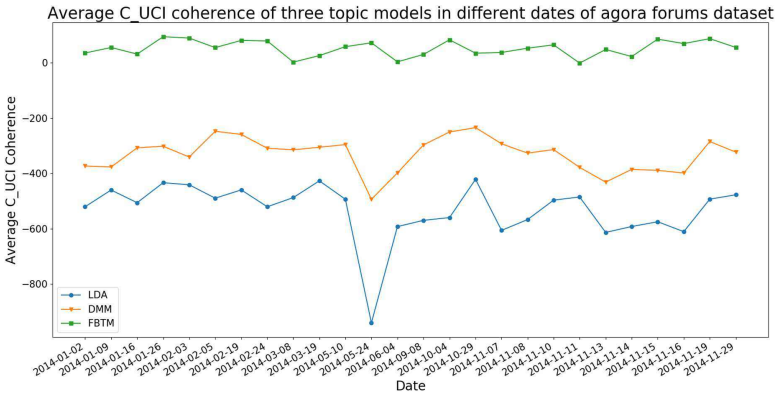

Fig. 4. Average UCI coherence of three topic models in different dates of Agora forums

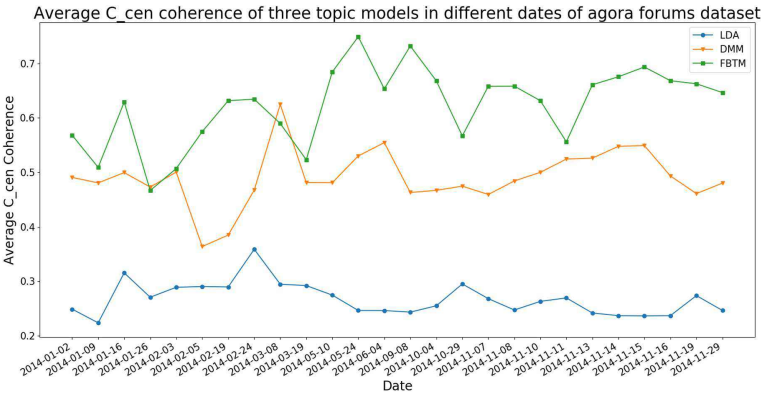

Fig. 5. Average centroid coherence of three topic models in different dates of Agora forums

LDA and DMM. UCI coherence gives a similar result: measures of FBTM is significantly higher than LDA and DMM. Due to our observation of traning files, we attribute the dramatic drop of performance of LDA in "05–24" to the high proportion of short text in dataset, especially some overly short documents. The severely negative affection on performance demonstrates the sparsity problem of LDA, which makes it the worst in all three topic models whereas extensively used in former researches. On the contraty, the stable measures of FBTM demonstrates its capability of handling various length files. About centroid coherence, situation seems a little different. The coherence scores of DMM are slightly lower than FBTM, while LDA still performs worst. We suggest the assumption made by DMM that each document originates from only one topic make the topical words selecting by DMM model much more similar in semantic space.

Table 3 lists some typical topics with their related words selected by three topic models in 02–03. The italic and bold words are related to the transaction and security concerns respectively. As shown in Table 3, FBTM can capture the typical terms of researchers' interests better than DMM and LDA. The topics FBTM extracts can be concluded as topics about "Markets&Drugs", "Trust&Security", "General conversation about markets". According to comparison, the topics FBTM generates consist of more specific words. The word "KUSH", "SDB" and "Sativa", which are kinds of drugs, and the keyword "ddos", "BitcoinFog" are selected only by FBTM. According to that we can have an quick understanding about the majority drugs darknet users prefer, the bitcoin mixing up platform prevalent in darknet(BitcoinFog), as well as the concerns against DDOS attack or business fraud from darknet users. The topics LDA generated is much more ambiguous with a high repetition rate of topical words. Only thing we can see is topic 2 related to the trust issues in darknet market due to some words like "scammer". In general, we can conclude that FBTM can provide a more specific informative intelligence from darknet, which makes it a better choice than LDA that used mostly in previous works Table 3 .

6.3 Result Analysis

Here we give a analysis of topical results listed in detail in Appendix B to show how topic model can help darknet intelligence mining. Owing to space constraints, we choose some of results generated by FBTM according to the date when Branwen applied his crawler in Darknet Market Archives. We show the top twenty most relevant words in each topic. The italic and bold words are related to the transaction and security concerns respectively for more visually appealing results.

Discussion About Transaction and Drug Knowledge. Results in Appendeix B supported the role of Agora market serving as a social platform that enabled users to exchange their goods or reviews for promoting underground transactions. Similar to traditional underground forums, many vendors tend to publish advertisements of their products with introductions and attached links

Table 3. Topics with related words in 02–03

Topic	DMM	FBTM	LDA
Markets& Drugs	shipping item need days5	party *SDB* REAL *KUSH*	*SDB* follows represent number
	add order lt gt	whole traditional generations **TOR**	dealer deposit long either
	last smoke address *MeO*	*Markovich* **AGORA** process *USD*	place invite say status
	new *SDB* bond listing	guide reason funds dedicated	ran people maligan vend
	another last smoke address	priority banks active **SR1**	scammed market chemicals stock
Trust& Security	**onion** market hash vendors	less try let established	either place say invite
	weed fucking back **Tor**	**BitcoinFog ddos** days change	trust index stock along
	squidgy going forum money	around scammer trust work	scammed scammer market still
	far send almost thing	described amp confident clean	chemicals people maligan register
	cannabis lol re black	provide allways *cannabis* squidgy	true wrong tormarket kxevkwmxhe7mby
General conversation about market	**onion** BEGIN register SIGNATURE	REAL *KUSH SDB* Nite	000 dealer number represent
	http LINK **GnuPG** ETw6LBrqv6http	login address **MingW32** hand	long follows deposit vend
	Version END account **MingW32**	Market administrative **TOR** successful	status popular single cool
	v1 REFERRAL please THE	link *Markovich* best outdoors	however **TOR** tormarket wrong
	team top MESSAGE NEW	**GNU** END **GnuPG** *Sativa*	none ksxq wanted topic

in the forums, especially new coming vendors. This can be reflected by the topic with keywords such as "Welcome" paired with the vendors' names. The topical words reveal some active vendors such as "Markovich", "Trim", "SQK". Another fact is active vendors always have business among multiple markets. For example, vendor "SQK" always appearing with another market "Pandora" seems that he does business in both two markets, which can be demonstrated by the content of posts.

An important functionality of underground Agora forums is it build a reputation system as a complement to the market. Forums provide a platform for exchanging evaluation and experience in transaction or drugs with freedom. That makes the forum a best place for customers to receive honest reviews and identify scammers. The common topical words "smell", "Feedback", "QUALITY" come from posts buyers exchange their usage experience or recommendation on the selling drugs. Prevalence of "scam", "scammer"or "trust" reflects the fraudulent conduct in DNM. FBTM also lists some prevalent goods, which are not captured by other models. "delta9", "x25", "KUSH", "meth", "DMT", "Lamas" are all chemical drugs sold in market. "cannabis" is also a prevalent goods while "Sativa" is the most popular species.

Another interesting findings are the transaction means used by darknet users. Topical words such as "address", "drugs", "days", "pack", and "time" seems to indicate discussions focused on purchase and delivery of drugs. The common choice for goods delivery is USPS owing to its high occurence rate in topics around transaction or delivery. Cryptocurrency is another big issue dis-

cussed frequently in darknet forums. Mentioned escrow services or tools are bitcoin exchanges or wallets such as "LocalBitcoins", "Multibit", "bitcoinamory", "brainwallet", "Mycelium". Although bitcoin is the most commonly used cryptocurrency in darknet market, the usage of credit card "Pasmo" become prevalent among darknet users after being recommended in "02–24".

Security Concern and Risk Management. Security concern is always a significant issue in darknet forums. Thereinto, anonymity is the most frequently discussed property due to the appearance of terms "anonymous", "anonymizing" or "stealth". We find the most prevalent tool darknet users used for confidential communication is "pgp" (Pretty Good Privacy), which is implemented by a free open source software "GnuPG". A substantial part of posts contains a large list of PGP keys at the end. Another important communication service mentioned in the forums are emailing indicated by words such as "Tormail", "Vmail", "URSSmail", "Thunderbird" and "SMTP". Besides, darknet users also use "Pidgin" for instant messaging. For secrecy, Pidgin is utilized through Tail, a living operating system where all software is configured to connect to the internet through Tor, which is also reflected in our results. As for anonymity of transaction, mixing is a service which attempts to increase the anonymity of cryptocurrency transactions, where a group of users exchange cryptocoins with each other to increase the difficulty in tracing transactions. "Bitcoinfog" is the most prevalent coinjoin platform appearing in the topic about security.

Darknet users also focus on the security of their wallet, account or the market itself. According to the topics, "DDoS" attack from "hacker" are the most probable attack the darknet website may come under. Apart from the external threat, scam in the market is also an important issue. To prevent business fraud, darknet users take measures which can be seen from the terms such as "escrow" and "multisig" which aims to prevent markets scamming users into purchasing items that they will never receive ("exit scamming"). "FEing", which is a transaction form meaning "finalized early" where the buyer has to release the fund kept in the marketplace's wallet to the vendor before the goods were actually shipped, becomes prevalent since "02–24" to protect vendors.

7 Conclusion

We have presented a general framework named pyDNetTopic, which is integrated with prevalent topic models, enabling automatically extracting textual content and uncovering hot topic issues in Darknet Market forums for researchers to have a general understanding on darknet. Especially, in order to adapt to big data scenarios in corpus with diverse text lengths such as Darknet Market forums datasets, we propose an improved topic model with less resource consumption and computational complexity named FBTM. Better coherence measures indicates better quality and interpretability of topical terms generated by FBTM than baseline models. The full results and analysis on real world darknet dataset demonstrate how pyDNetTopic can help by data mining and extracting informative intelligence in darknet researches.

Appendix A List of Additional Stop Words

The listing words are some common words among all topics that provide no useful information. We regard such words as general stop words in pyDNetTopic and remove them in preprocessing.

fuck, get, got, shit, see, u0e2a, would, use, think, like, xa0, sr, know, u0e3f, good, tquot, u2591, u25ac, make, fe, day, although, ands, soooo, yet, favs, So, ll, went, br, en, often, knowing, liking, one, get, thinking, even, could, go, going, fucking, fuck, shit, also, use, using, much, got, good, make, making, really, see, want, need, sure, right, still, take, taking (Tables 4, 5, 6, 7, 8, 9, 10, 11, 12, 13) .

Appendix B Full Topic Results of Agora Forums in 2014

Table 4. Result of 2014-01-09

Topic	Terms
Vendor review	Products stupid *vaporizer* referred fees start growing giving roof reading ur day5 lost book started terms members refer bank closed
Trust & Security	*TorBay* point products mean months thread fees sad scammer trustable described impossible ve buyers selling comes option crooked wasn hand
Transaction issue	Points absolutely *vaporizer* access More *Pandora* kind *SQK* bottom legit *TorBay* crooked strain sell *Tranzcentral* cloning start *Sativa* server week

Table 5. Result of 2014-02-03

Topic	Terms
Markets & Drugs	Party *SDB* REAL *KUSH* whole traditional generations **TOR** *Markovich* **AGORA** Process USD guide reason funds dedicated priority banks active **SR1**
Trust & Security	Less try let established **BitcoinFog ddos** days change around scammer trust work described amp confident clean provide allways *cannabis* squidgy
General conversation about market	REAL *KUSH SDB* Nite login address **MingW32** hand Market administrative **TOR** successful link *Markovich* best outdoors **GNU** END **GnuPG** *Sativa*

Table 6. Result of 2014-02-24

Topic	Terms
General conversation about market	Partner began **Tormail code** Legit try suddenly buzz browser genetics Suite **USB** 2004 Vendors Instruments either Guide Please Requires Do
Anonymity & Security	**URSSMail VPN Gateway** blend *Passmo* **SMTP** log privacy Using **TrueCrypt** **bitmessage FEing** Audio green encrypt **Tormail** items compound anonymous
Vendor review	**www** *Passmo* items BEST recommended testing protect Walther genetic Japan Information anonymous *Indica* x4 larger pretty trusted straight person **Tormail**

Table 7. Result of 2014-03-19

Topic	Terms
Transaction tools and items	**Tormail LocalBitcoins** *meth* protects *USPS* DrEarnhardt **Tails Pidgin** House shared paper parental deeper experiences delivery Interways called NOW *Enanthate Purps*
General conversation about transaction	Located **bitcoinarmory** integrity replaced *Trim* Japan pack smells Interways thick Feedback spread contains genetics lovely via forums worth *Blotters* Dawg
Anonymity	Use **VPN Vmail** Information interest Cert **Tormail** consider **bitmessage** anonymous additional **Pidgin** hidden given part admin right **SMTP** began integrity

Table 8. Result of 2014-05-10

Topic	Terms
Unknown	Obviously patients consistently wouldn commissions writing ad_listing pages user inch generate **Pidgin** low Sunday Sign March *Morrissons* hardly Grape situation
General conversation about market	*Lotus* won flooring anymore nlm quick Vendors usb previous words id Wallets **Thunderbird** specific asked beautiful geometric p1http recognized *CBG*
General conversation about transaction	Boguu inch commissions **LocalBitcoins** offers brand almost improved *cannabinoids* privacy sensitive bred Wallets seeking priority pain posting stealth partner known

Table 9. Result of 2014-06-04

Topic	Terms
Security	Id *blotters* numerous **agora** comedown security package sells priced **Vmail** clear price b1t **ddos VPN Tormail** privacy *blotter* kind **truecrypt**
Privacy	Thirty hour unsigned Liquid **truecrypt** finished crushed Use nowhere privacy trying id headers deal ocbruins encryption register webmail client basically
General conversation about markets	*Lotus* p1http bizarre nowhere comedown activity *x25* unfortunately **Linux** labor Download low confirmed opportune kinda secure worried transaction park combined

Table 10. Result of 2014-09-08

Topic	Terms
Drugs	Laid trusted activity asserted smooth *delta9* tv *cannabisrelief* crystal matter welcome eaten set mcg 110ug bag March writing system found
General conversation about markets	*Lotus* universe asked flooring news weeks regularly *Blotters* American bizarre tv crystal writing Vendors trusted confirmed online situation activity stable
Vendor review	Boguu numb box previous *cannabinoids* indeed outside messaged parents July deeper pages brand preview coin seconds moment community known index

Table 11. Result of 2014-10-29

Topic	Terms
Vendor advertisement	*Lamas* Vendors ups thought focus opinion inch white truly asked dark presences *drugs* setting bizarre weeks offer **php** forum Jim
Vendor review	Words *Xanax* partner leave work answer old quick unsigned outside perception universe comedown lower comeup sbizarre information feeling truly USD
Drug	*Lamas* anyone white reccomend contains quotes *drugs* info thoroughly amount Other *Sativa Shamrock* trusted *DMT* stable allows fun line looked

Table 12. Rresult of 2014-11-11

Topic	Terms
Transaction tools	Public *Paypal* **Multibit Mycelium** buys really cash Using someone **Blockchain** signed *silver* certain Love lower amazing Offline watching buyer blotter
Vendor advertisement	**LocalBitcoins** welcome anyone unsigned however understanding put refund prejudice page three forums connects tried **brainwallet** funds longer big Jim untested
Drug review	Password *blotter* lectures packs left paper nature *BMR* wall want documentaries set jar site ready world Any Vendors Guide White

Table 13. Result of 2014-11-19

Topic	Terms
Vendor review	Numb indeed *USPS* container snoopy recommended *Sativa* Even *Floyd* Go Low hot March **Pidgin** Some re Twins Cheap Vendors solution
Mailing discussion	**Tormail** asked deeper Jim crazy **VPN Vmail** actually ask truly writing **SMTP** yesterday old spread spice hacked disk network **bitmessage**
Unknown	Won bizarre overall soft inch **Linux** hope writing bonus words comeup yesterday connected action posts spread tripped unsigned universe numb

References

1. Aletras, N., Stevenson, M.: Evaluating topic coherence using distributional semantics, pp. 13–22, March 2013
2. Almukaynizi, M., Grimm, A., Nunes, E., Shakarian, J., Shakarian, P.: Predicting cyber threats through hacker social networks in darkweb and deepweb forums, pp. 1–7, October 2017. https://doi.org/10.1145/3145574.3145590
3. Biddle, P., England, P., Peinado, M., Willman, B.: The darknet and the future of content protection. In: Feigenbaum, J. (ed.) DRM 2002. LNCS, vol. 2696, pp. 155–176. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-44993-5_10
4. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993 (2013)

5. Branwen, G., et al.: Dark net market archives, 2011–2015. www.gwern.net/Blackmarket%20archives (2015)
6. Christin, N.: Traveling the silk road: a measurement analysis of a large anonymous online marketplace, pp. 213–224, May 2013. https://doi.org/10.1145/2488388.2488408
7. Deliu, I., Leichter, C., Franke, K.: Collecting cyber threat intelligence from hacker forums via a two-stage, hybrid process using support vector machines and latent dirichlet allocation, pp. 5008–5013, December2018. https://doi.org/10.1109/BigData.2018.8622469
8. Dittus, M., Wright, J., Graham, M.: Platform criminalism: The 'last-mile' geography of the darknet market supply chain, pp. 277–286, April 2018. https://doi.org/10.1145/3178876.3186094
9. Eimer, T., Luimers, J.: Onion governance: Securing drug transactions in dark net market platforms, August 08 2019
10. Grisham, J., Barreras, C., Afarin, C., Patton, M.: Identifying top listers in alphabay using latent dirichlet allocation, p. 219, September 2016. https://doi.org/10.1109/ISI.2016.7745477
11. Hout, M.C., Bingham, T.: 'Surfing the silk road': a study of users' experiences. Int. J. Drug Policy **24**, 524–529 (2013). https://doi.org/10.1016/j.drugpo.2013.08.011
12. Jin, O., Liu, N., Zhao, K., Yu, Y., Yang, Q.: Transferring topical knowledge from auxiliary long texts for short text clustering, pp. 775–784, October 2011. https://doi.org/10.1145/2063576.2063689
13. Larochelle, H., Lauly, S.: A neural autoregressive topic model. In: Advances in Neural Information Processing Systems, vol. 4, pp. 2708–2716, January 01 2012
14. Mimno, D., Wallach, H., Talley, E., Leenders, M., Mccallum, A.: Optimizing semantic coherence in topic models, pp. 262–272, January 2011
15. Newman, D., Lau, J., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence, pp. 100–108, January 2010
16. Nunes, E., et al.: Darknet and deepnet mining for proactive cybersecurity threat intelligence, July 2016
17. Phan, X., Nguyen, L., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceedings 17th International Conference on World Wide Web, pp. 91–100, February 2020
18. Phan, X.H., Nguyen, C.T., Le, D.T., Nguyen, L., Horiguchi, S., Ha, Q.: A hidden topic-based framework toward building applications with short web documents. IEEE Trans. Knowl. Data Eng. **23**, 961–976 (2011). https://doi.org/10.1109/TKDE.2010.27
19. Porter, K.: Analyzing the DarkNetMarkets subreddit for evolutions of tools and trends using LDA topic modeling. Digit. Invest. Int. J. Digit. Forensics Incid. Response **26**, S87–S97 (2018)
20. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining, pp. 399–408, February 2015. https://doi.org/10.1145/2684822.2685324
21. Salakhutdinov, R., Hinton, G.: Replicated softmax: an undirected topic model. pp. 1607–1614, January 2009
22. Samtani, S., Chinn, R., Chen, H.: Exploring hacker assets in underground forums, pp. 31–36, May 2015. https://doi.org/10.1109/ISI.2015.7165935
23. Samtani, S., Chinn, R., Chen, H., Nunamaker, J.: Exploring emerging hacker assets and key hackers for proactive cyber threat intelligence. J. Manag. Inf. Syst. **34**, 1023–1053 (2017). https://doi.org/10.1080/07421222.2017.1394049

24. Sapienza, A., Bessi, A., Damodaran, S., Shakarian, P., Lerman, K., Ferrara, E.: Early warnings of cyber threats in online discussions, January 2018
25. Sievert, C., Shirley, K.: Ldavis: A method for visualizing and interpreting topics, June 2014. https://doi.org/10.13140/2.1.1394.3043
26. Srivastava, A., Sutton, C.: Autoencoding variational inference for topic models, March 2017
27. Xia, Y., Tang, N., Hussain, A., Cambria, E.: Discriminative bi-term topic model for headline-based social news clustering. In: FLAIRS Conference (2015)
28. Yan, X., Guo, J., Lan, Y., Cheng, X.: A biterm topic model for short texts. pp. 1445–1456, May 2013. https://doi.org/10.1145/2488388.2488514
29. Yin, J., Wang, J.: A dirichlet multinomial mixture model-based approach for short text clustering. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2014. https://doi.org/10.1145/2623330.2623715
30. Zhang, H., Chen, B., Guo, D., Zhou, M.: Whai: Weibull hybrid autoencoding inference for deep topic modeling, March 2018