What's Trending in the Darknet?

Wall Street Market Topic Modeling and Sentiment Analysis

What the Authorities Don't Tell You

Darknet markets are abstract to the public. Some of us may have watched Mr. Robot. Others may have read about the arrest/jail sentencing of Ross Ulbricht AKA "Dread Pirate Roberts", the creator/operator of Silk Road, which is estimated to have averaged $15 million in annual sales.

If one were to visit a darknet market, he/she will find a thriving economic ecosystem that resembles eBay. Simultaneously, these obscure economies are solving complex technical challenges like ensuring anonymity and creating trust among thieves.

Up until 2019, one of the most prosperous darknet markets was the Wall Street Market (WSM). The illegal virtual bazaar included approximately 5,400 vendors, tens of thousands of items/services for sale, and more than 1.1 million user accounts. At its apex, WSM was the second largest darknet market in the world. In May 2019, in a covert collaboration among investigators from the United States, Germany, and Netherlands, ten individuals were arrested; the offenders were in the U.S. and Europe. These persons were responsible for creating/operating the WSM. Upon arrest, the authorities seized more than 550k euros.

Project Motivation and Overview

To improve the efficacy of cybercrime operations, authorities should examine and learn the views/habits of the marketplace participants. How do they communicate? How do the vendors transact and transport their products? What are the common critiques and trending topics?

Like "standard" e-commerce sites, darknet market users communicate via well-organized forums. A collection of these forum posts can serve as a corpus to train natural language processing (NLP) models. Applying topic modeling could help answer the previously mentioned questions. With such insight intelligence agencies would be better equipped to imitate a darkweb market participant.

After authorities shut down WSM, the National Science Foundation (NSF) funded the Data Infrastructure Building project which was led by the University of Arizona's Artificial Intelligence Lab. The corpus is comprised of more than 42,000 posts written by over 5,500 unique members. The posts were published on the WSM forum between October 2016 and November 2018.

Metrics

Topic modeling is an unsupervised machine learning exercise. For this project, I assumed that the documents/forum posts do not have corresponding labels.

Although various topic model methodologies were performed, this summary focuses on Latent Dirichlet Allocation (LDA) and K-Means Clustering (Kmeans).

To assess the performance of the two LDA models, I used perplexity. Perplexity is a statistical measure that compares a document's theoretical word distribution, which is represented by document's topic allocation, to the document's actual words.

I will discuss the LDA algorithm in greater detail but for now let us accept that the model is used to develop probability distribution of "hidden" topics for each WSM forum post. Consequently, I will use perplexity to assess the model's ability to predict another sample.

I fitted the LDA model using two transformations of the dataset – I will elaborate in greater detail. The two outputs allowed me to compare each model's performance.

I would be remiss to omit mentioning perplexity's limitations. Research has demonstrated that perplexity and human judgement are often not correlated and even sometimes slightly anti-correlated.

To measure the performance of the second model, Kmeans, I used the silhouette score. Kmeans clusters/groups each data point into **"k"** number of groups – more on this later. Consequently, an appropriate metric for the model is one that evaluates the similarity of objects within the same group and the difference among objects of different groups/clusters.

The silhouette score ranges from -1 to +1. A high value signifies that an object is well matched to the members of its own cluster and possess weak ties to the objects in different clusters. If the score is low or negative, then the model may be assuming an inappropriate number of clusters. Lastly, the silhouette score can be calculated with any distance metric, I used Euclidian distance.

Benchmark

To ameliorate lack of correlation between perplexity and human inferenced in the case of LDA, I used a naïve human-focused benchmark for this project. The WSM had subforums. Examples include Services, Counterfeits, Introductions, and Psychedelics. These pre-defined subforum topics provide a well-established and concise point of reference.

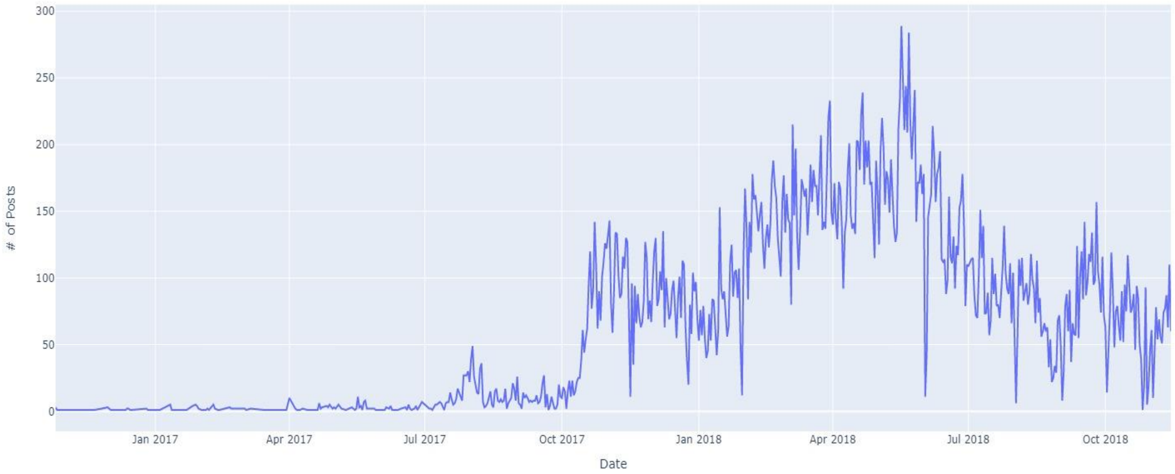
I also believed the subforums served as a reasonable benchmark for the Kmeans results.

Exploring Parts Unknown

The line chart below shows the daily number of posts during the life of the dataset. The visualization displays a significant uptick in October 2017. Given lack of transparency of the darknet, I can only speculate what triggered the increase in participation.

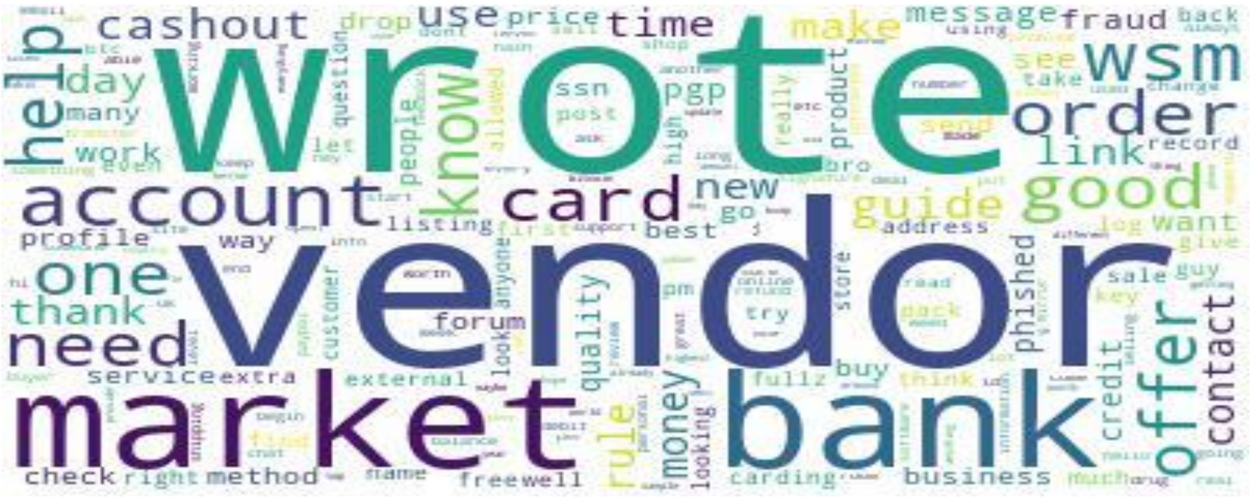
In July 2017, AlphaBay, which was one of the largest darknet markets at the time, was shut down by U.S., Canadian and Thailand law enforcement. It is possible that AlphaBay users/vendors migrated to WSM. The lag can partially be explained by the application process required for vendors. It is common practice to require vendors to prove their reputation and/or provide cash deposits via cryptocurrencies.

Wall Street Market Forum Daily Post Frequency



As previously mentioned, WSM was a vibrant marketplace comprised of prosperous entrepreneurs. Consequently, the most frequently used words are nearly identical to the words - e.g. vendor, market, bank, and account - that one would expect to be most prevalent in a legal marketplace.

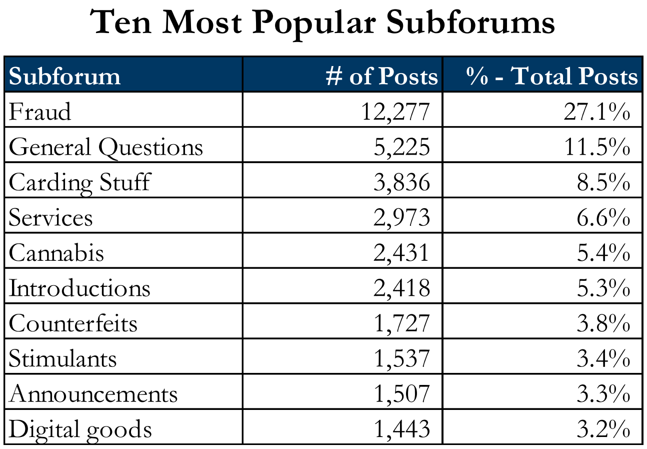
Wall Street Market Forum Word Frequency



Data and Data Processing

The initial format of the data was a MySQL file, which I transferred to Python using a combination of MySQL Workbench and the "sql" Python package. There were more than 45,000 posts.

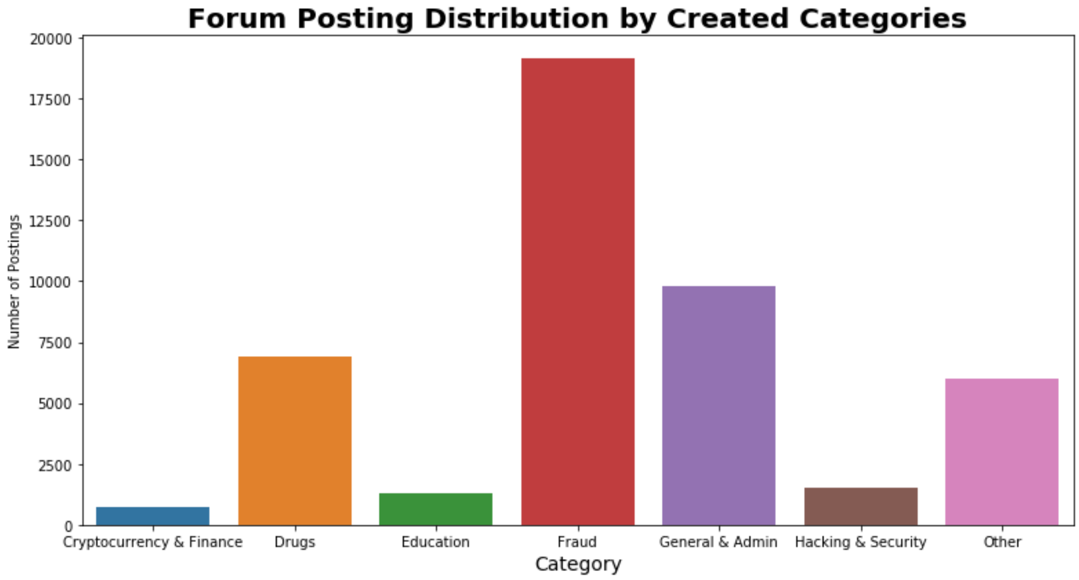
Technically, the dataset had predefined labels, i.e. subforums. There was a total of 40 subforums. I excluded these topics when performing the topic modeling. The table below displays the ten most active subforums.



Ten Most Popular Subforums

Subforum	# of Posts	% - Total Posts
Fraud	12,277	27.1%
General Questions	5,225	11.5%
Carding Stuff	3,836	8.5%
Services	2,973	6.6%
Cannabis	2,431	5.4%
Introductions	2,418	5.3%
Counterfeits	1,727	3.8%
Stimulants	1,537	3.4%
Announcements	1,507	3.3%
Digital goods	1,443	3.2%

To develop a rough benchmark of potential topic groupings, I used the subforum names to manually coalesce the subforums into topic groupings. The bar plot below shows the results.



The forum posts were initially in html which required parsing and removing of HTML vestige, e.g. "\n". Additionally, most of the writing was informal and serial numbers and websites were referenced throughout the documents. The serial numbers and websites were erased.

I removed the punctuations and expanded contracted works, e.g. "who's" became "who is". In addition to the default “stop words”, various "noninformative" words were omitted, e.g. "thing" and "could".

The documents were then lemmatized, which differs from stemming. Both lemmatization and stemming generate the root form of the documents’ words. The difference is that a lemma is an actual language word whereas a stem may not be a coherent word. After lemmatization, the documents were tokenized.

I then transformed the lemmatized tokens using both Bag-of-Words (BoW) or Term Frequency-Inverse Document Frequency (Tf-Idf) vectorizer. The conversion resulted in numeric vectors that were used as inputs into the NLP models.

Topic Modeling Algorithms

Latent Dirichlet Allocation

As previously mentioned, LDA is a statistical approach to topic modeling. Latent means hidden; therefore, LDA constructs a "hidden" layer, that represents imaginary topics. The documents are then randomly defined as a distribution of the imaginary topics. The next phase in the model is to express each imaginary topic by a distribution of words included in the corpus.

Once the topic-word pairs are created, the model identifies each word in the documents and reviews whether each word is assigned to the correct topic, eventually the model converges on the number of selected components/groups. Since the algorithm feeds itself posteriori, LDA is a Bayesian approach.

Given that the dataset is a relatively informal marketplace in which esoteric products are sold, LDA is well suited iteratively learns from its priori. Moreover, topics are most likely to be a permutation of various topics.

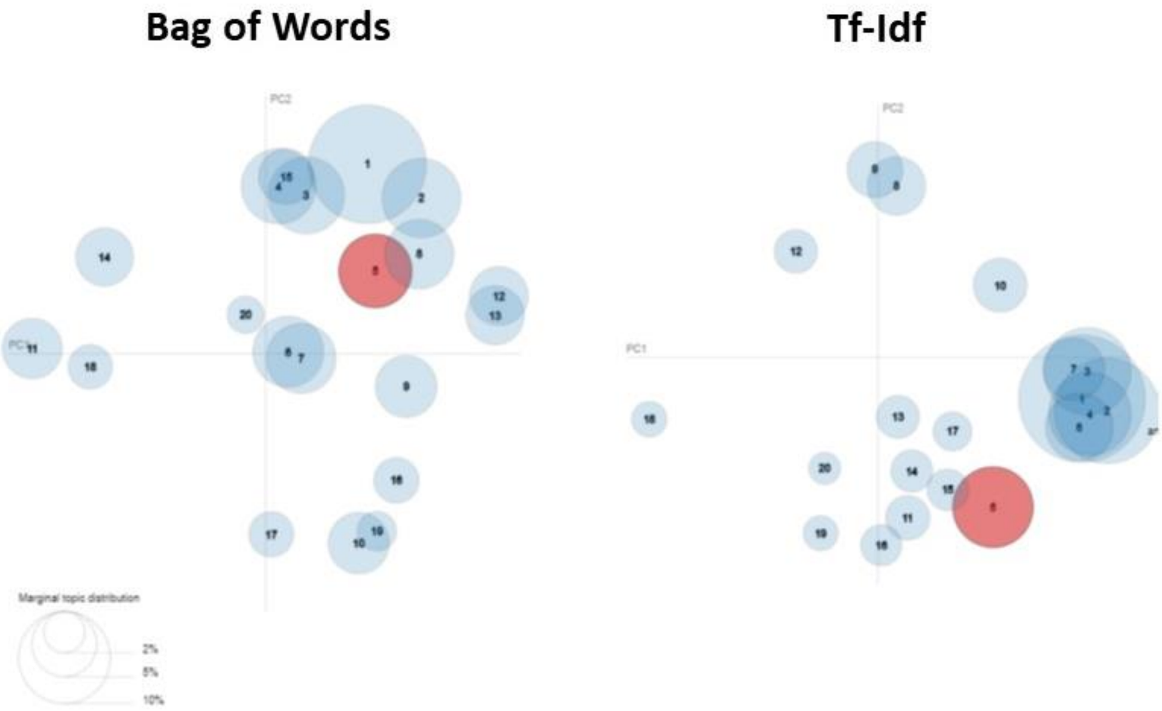
For example, see the following post from ExpectUS on July 14, 2018:

> “Scusi wrote: metropolis wrote: And you were called out as a scammer on Hansa too. You spam the forums and tell people to finalize and leave feedback right away so that you can offer additional support then you block them and they have no way to change feedback so they come to the forums and expose you...SEARCH THE FORUMS BEFORE DOING BUSINESS WITH THIS LAME Correct. This is EXACTLY was this scammer does.lolsitcks and stones. Sticks and stones.
>
> I just mashed out two orders while reading your posts.So you can say whatever you want, I have many repeat buyers and a very high reputation so your comments meaning nothing ^_^Maybe you should buy from me and see what kind of serious serious moolah will come your way.Enjoy!Your Pal!ExpectUs
>
> offer link: LEVEL 9 Early Retirement Pack Sale 100x SSN Bank Records 10 Cashout Guides $500 offer link: 10 Packs SSN/Bank Records with Cashout Guides $80 Extras! Many Extras! All New Cashout Guides! Personal Carding Fraud Lessons/ EMV X2 5.1 Debit/Credit Card Software only $300Highest Balance Bank Accounts Checking/Business/Money Market/ HIGHEST NET WORTH SSNs.”

The first paragraph of ExpectUS's post is about WSM compliance and user report. Meanwhile, the second paragraph relates to trusted commerce. And the last paragraph describes fraud products, digital assets, and service offerings. This post demonstrates the use of being able to categorize one document/forum post into several groups.

Before fitting the model, one must determine the number of components/clusters that will be used to represent the dataset. I selected 20 components, an arbitrary value that was primarily elected for ease of visualization. In future revisions, I will identify the optimal value"

I also applied two LDA methods. The approaches differed in how I transformed/vectorized the documents into numerical values. I used BoW and Tf-Idf vectorization. The charts below demonstrate that BoW was more successful in grouping the documents within the elected number of components.

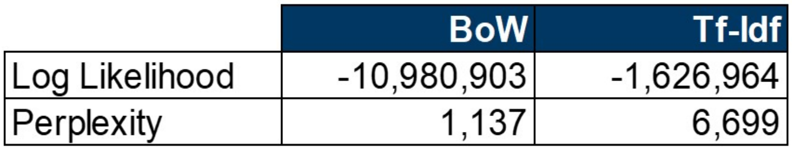


Bag of Words

Tf-Idf

LDA Performance Assessment

As previously discussed, I used perplexity as the metric to evaluate the models' performances. Perplexity is used in various probabilistic models. A lower score denotes that the probability distribution succeeded at predicting the document text.



	BoW	Tf-Idf
Log Likelihood	-10,980,903	-1,626,964
Perplexity	1,137	6,699

.

In hindsight, I should have used perplexity to determine the optimal number of components. This approach is like plotting distortion and inertia for Kmeans, which I discuss in the next section.

Given perplexity's shortcomings, i.e. poor correlation with human interpretation, coherence score would have been better suited for this task. In this project's second iteration, I will use Gensim instead of Sklearn to implement the LDA model, allowing for a straightforward evaluation of the two models based on their coherence scores.

LDA Justification

When compared to the subforum topics, the LDA model results shared a couple of topics like finances, see topic #16. respectively. However, other topics l overlap or are condensed. For example, the LDA model groups all drug-related terms into one component, see Topic #12; meanwhile, the WSM subforums are broken down in topics like Cannabis, Stimulants, and Ecstasy.

The following are a couple of topics developed by the LDA model that was fitted using the BoW-vectorized documents.

Topic #1: order utopia quality kush refund best product reship weed price

Topic #12: mdma profile cocaine gram weed sample coke hash quality market

Topic #16: account bank drop paypal transfer need money cash btc service

As I discuss later in the paper, I believe that the model still requires additonal work. There is too much overlap in the model with regards to commerce and financial transaction. Although, I suspect a revised LDA model will still struggle in differentiating posts that reference stolen/hacked financial information from post explaining payment procedures.

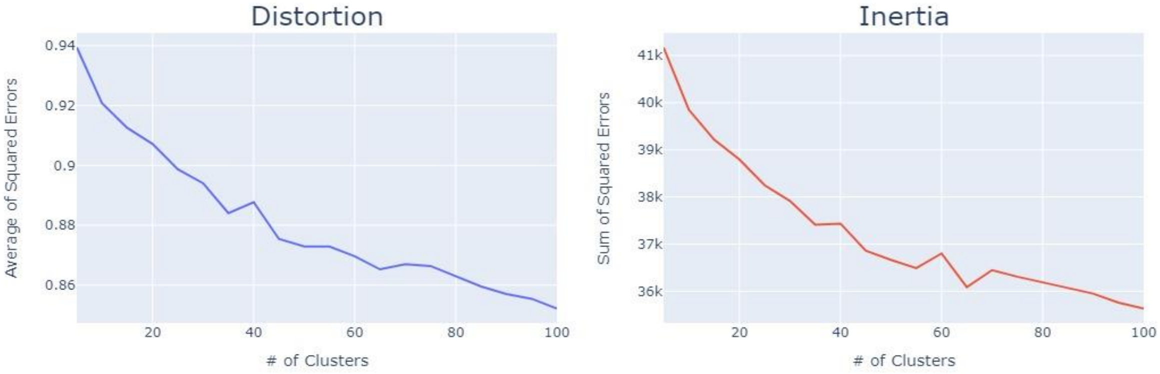
K-Means Clustering

I also applied KMeans to obtain a different perspective in how the documents could be organized. KMeans is a more rudimentary model and consequently possesses (potentially inappropriate) assumptions. For example, the method defines each cluster by its center; to do so, the model assumes that clusters are convex, distinct, and approximately equal in size.

I used Kmeans because it measures the distance between the center centroids and their corresponding datapoints. Prior to fitting the Kmeans model to the text documents (forum posts), I transformed the posts into vectors using the Tf-Idf vectorizer. Theoretically, these vectors can be plotted on a multidimensional graph. The Kmeans model calculates the datapoints distances to create the optimal groupings based on the selected number of clusters.

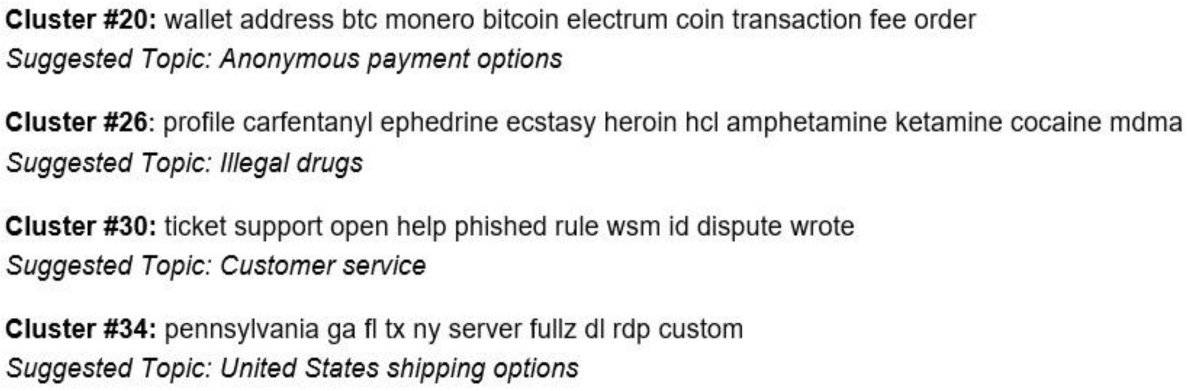
Using numerous variations, I attempted to reduce the dataset's dimensionality by applying Linear Discriminant Analysis or Incremental Principal Component Analysis (PCA) prior to training the KMeans model. Unfortunately, the memory-intensive sparse matrix generated by the Tf-Idf vectorizer caused memory errors when transforming the dataset that I would have like to use as the input to the Kmeans model.

KMeans' most important hyperparameter is the number of clusters. This value is somewhat arbitrary. To identify the optimal value, I calculated the sum and average of the squared errors (intra-cluster variances) - distance among the clusters' centroids and the corresponding datapoints - to see the number of clusters that minimizes the error while being cognizant of overfitting.



Based on the distortion and inertia plots, which are commonly referred to as the "Elbow Methods", I selected 65 clusters. The excerpt below are four selected clusters and their respective ten most common terms. I created the "Suggested Topics" based on the clusters' terms.

Selected Clusters - Top Ten Terms



Cluster #20: wallet address btc monero bitcoin electrum coin transaction fee order
Suggested Topic: Anonymous payment options

Cluster #26: profile carfentanyl ephedrine ecstasy heroin hcl amphetamine ketamine cocaine mdma
Suggested Topic: Illegal drugs

Cluster #30: ticket support open help phished rule wsm id dispute wrote
Suggested Topic: Customer service

Cluster #34: pennsylvania ga fl tx ny server fullz dl rdp custom
Suggested Topic: United States shipping options

To assess the KMeans model's performance, I used the silhouette score. The KMeans' silhouette score on its own does not provide insight; I required a benchmark. Consequently, I applied another clustering method referred to as DBSCAN. DBSCAN is an interesting comparison because, unlike KMeans, the algorithm does **not** assume that the clusters are of convex shapes **nor** require a priori, i.e. a pre-defined number of clusters.

KMeans Performance Assessment

DBSCAN has two principal hyperparameters: epsilon and min_samples. min_samples is a threshold to determine which samples can qualify as **core samples**. I will omit a deep explanation of these hyperparameters as DBSCAN's sole purpose in this project is to establish a baseline silhouette score. I assumed 0.5 and 5 for epsilon and minimum samples, respectively.

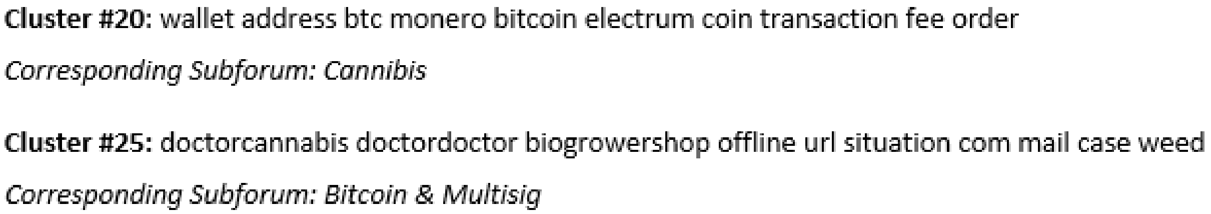
The KMeans model resulted in a silhouette score of **0.06** in contrast to a score of **-0.14** for the DBSCAN. In conclusion, neither of these models performed well. KMeans' silhouette score that is close to zero represents, on average, that the clusters are not clearly divided. This reconciles with the outcomes from the LDA models.

Kmeans Justification

I struggled to compare Kmeans' results to the subforum benchmark. The "Elbow Methods" suggest a relatively large number of clusters. I selected 65 as a compromise. However, increasing the number would have led to substantial decrease in the squared error metrics used to assess the difference between the model and actual values.

As I mentioned earlier, I believe that using PCA or LSA would remove a significant portion of the noise in the dataset improving Kmeans' performance.

Kmeans generated a few results that were in alignment with the benchmark:



Cluster #20: wallet address btc monero bitcoin electrum coin transaction fee order

Corresponding Subforum: Cannibis

Cluster #25: doctorcannabis doctordoctor biogrowershop offline url situation com mail case weed

Corresponding Subforum: Bitcoin & Multisig

Conclusion

Unfortunately, neither LDA nor KMeans Clustering were helpful in organizing the WSM forum posts into clear and distinct groups allowing better understanding of the operation and/or primary topics of the darkweb market participants. In the next iteration, I will examine how the removal of generic and ecommerce terms, e.g. vendor, people and sell - affects the outcome. I suspect that omitting these words should allow for the models to categorize the documents more cogently.

I will also back-solve for the optimal number of components to use in LDA. And, I will use BoW-vectorized documents in the Kmeans model.

I also applied Non-Negative Matrix Factorization (NMF). Unfortunately, the approach's results were unnoteworthy; therefore, I excluded them from the project summary.

It seems that among the three primary topic modeling methods - LDA, KMeans, and NMF - that LDA is best suited for this project. I intuitively agree with this conclusion as I feel a Bayesian approach is better suited in working with a less conventional corpus. Therefore, in this project's next iteration, I will employ the Gensim package to optimize the LDA model.