Mining the Dark Web: A Novel Approach for Placing a Dark Website under Investigation

Article *in* International Journal of Modern Education and Computer Science · October 2019

DOI: 10.5815/ijmecs.2019.10.01

Randa Basheer
Syrian Virtual University
2 PUBLICATIONS **2** CITATIONS

I.J. Modern Education and Computer Science, **2019, 10, 1-13**
Published Online October 2019 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2019.10.01

Mining the Dark Web: A Novel Approach for Placing a Dark Website under Investigation

Bassel Alkhatib
Syrian Virtual University, Damascus, Syria
Al-Sham Private University, Damascus, Syria
Email: t_balkhatib@svuonline.org, b.alkhatib.foit@aspu.edu.sy

Randa S. Basheer
Syrian Virtual University, Damascus, Syria
Email: randa_82880@svuonline.org

Received: 15 July 2019; Accepted: 11 September 2019; Published: 08 October 2019

***Abstract*—**In the last two decades, illicit activities have dramatically increased on the Dark Web. Every year, Dark Web witnesses establishing new markets, in which administrators, vendors, and consumers aim to illegal acquisition and consumption. On the other hand, this rapid growth makes it quite difficult for law and security agencies to detect and investigate all those activities with manual analyses. In this paper, we introduce our approach of utilizing data mining techniques to produce useful patterns from a dark web market contents. We start from a brief description of the methodology on which the research stands, then we present the system modules that perform three basic missions: crawling and extracting the entire market data, data pre-processing, and data mining. The data mining methods include generating Association Rules from products' titles, and from the generated rules, we infer conceptual compositions vendors use when promoting their products. Clustering is the second mining aspect, where the system clusters vendors and products. From the generated clusters, we discuss the common characteristics among clustered objects, find the Top Vendors, and analyze products promoted by the latter, in addition to the most viewed and sold items on the market. Overall, this approach helps in placing a dark website under investigation.

***Index Terms*—**Dark Web, Dark Web Market, Data Mining, Dark Web Mining, Data Pre-Processing

I. Introduction

Since the rise of the Dark Web in the beginning of 2000s, researches have analyzed activities on the Dark Web using several methodologies and perspectives. Most of these studies focused on terrorists and extremists' groups, though least of the researches considered analyzing dark web markets, where most of the illicit activities and malicious trading take place. Dark Web witnesses launching new markets frequently in many different majors. Products and services on these markets vary in a very wide range, from drugs, weapons, pornography and child abuse, malwares, software exploits and hacking tutorials, renting Botnets for security breaches, in addition to trading documents, fake IDs and stolen credit cards, renting hitmen, and many others. [1]

Researches that analyzed web resources in general considered utilizing data mining and machine learning techniques to explore hidden patterns in massive web data. The discovered hidden patterns lead to new information that helps in understanding different aspects about web resources. In criminology, which is our research area and specifically e-markets in the dark web, data mining plays a significant role in discovering new advantageous information about malicious products promoted in such markets, finding relationships between products and vendors, and the trading size of those vendors. These statistics benefit in inferring whether the vendor is an individual or a company, and the latter may refer to the existence of an organized crime [2]. Furthermore, it helps in understanding the structure of the malicious and criminals' communities in social or psychological views [3]. Information produced by Data Mining methods on dark web markets helps security agencies of different majors in investigations and studying crime and its stewards.

In general, to produce valuable information from a dark web resource, it needs to pass through several stages starting from accessing the dark web, extracting and structuring data, data cleaning and transformation, and finally mining the data.

In the next sections, we present the mining methodology on which our suggested approach depends. The methodology includes a brief description about data mining and web mining, data pre-processing, and the utilized data mining methods. We introduce our approach, which demonstrates a system of three modules: a crawler that accesses a dark web market and extracts data from the entire contents of the market, a cleaner that pre-processes data through cleaning and transforming procedures, and finally a miner. We illustrate in details how the miner analyzes products and vendors' data in three basic sides: generating Association Rules from products titles, clustering vendors, and clustering products. Furthermore, we discuss the results gained from the applied methods. This study discusses a method for law and security agencies to put a dark web market under investigation.

II. Related Work

Baravalle et al. [2] developed a miner that collects information about products and vendors on Agora market, producing statistics about products ranks according to the deliberated amounts of products, the most trading geographical areas, and names (or nicknames) of their vendors. They stated how such information calls attention to the type and size of in a dark web market, as the size products may refer to the existence of organized crimes in the market.

Robertson et al. [4] discussed supervised and semi-supervised approaches in studying dark web markets, by determining the qualitative areas of the products exclusively promoted by a specific vendor. The approach uses manual labeling in addition to clustering techniques to group products in separated categories, utilizes K-Means algorithm to generate centroids and clusters, and then evaluates generated clusters using measures like Rand-index and Entropy.

Quan Le et al. [5] introduced an approach of Malware classification for non-experts. They discussed the importance of Machine Leaning techniques in reducing the manual effort and the time it consumes. The approach starts from data pre-processing, classifies elements into benign and malicious software, and classifies the latter into types according to a previously defined labeling using CNN algorithm.

Celestini et al. [6] introduced an approach for crawling and processing data from markets on the dark web. They discussed drugs trading on the dark web, like AlphaBay, and the importance of monitoring activities on crypto-markets to produce early warnings on how drugs' trading evolves around the world. The study discussed data pre-processing techniques and preparing data for further analysis.

Marin et al. [7] discussed an approach of mining communities of malware and exploit vendors, focusing on understanding the vendors and their common characteristics. They used clustering methods and different similarity measures to find correlations among vendors' profiles from several markets, common categories they promote in, and the number of products they have in each category. They discussed how this method helps security agencies in understanding and tracking the hacking-related communities.

Furthermore, researchers studied communities' structures of the dark markets. The studies included analyzing the community and interactivity of its members, and crime and violence such trading lead to [3], in addition to collecting information to conclude the generative mechanisms that enable a market on the dark web to operate and survive [8]. Others discussed the teamwork and social involvement of members of such communities, the discussions addressed the usage of some tools that help in understanding and predicting members' behaviors, and how these markets are vulnerable to disruption [9].

Some studies discussed, from a geographical perspective, the considerable changes in revenues and trading traffic after severe changes happen to the market community, such as arresting a significant vendor or shutting down a market [10]. In additions, studies discussed analyzing geographical trafficking of products and exploring a potential relationship between the virtual world and the physical world, highlighting the role of specific parties from different countries in the international and domestic trafficking [11].

III. Methodology

In this section, we briefly describe the basic concepts our approach based on. The concepts include Data Mining, Web Mining, Data and Text Pre-processing, and the employed Data Mining methods, which include Association Rules and Clustering.

A. Data Mining and Web Mining

Data Mining is the automatic study, which a computer system performs on a huge database in order to extract new hidden relational and statistical information. Data Mining helps in discovering invisible relationships and patterns among raw data elements, and concluding future predictions for new elements, or exploring the common features in order to distribute elements into groups of the same features. Researches also describe Data Mining as Knowledge Discovery in Databases (KDD). [12]

Web Mining on the other hand, is the implementation of data mining techniques on the content, structure and usage of web resources, to identify valid unknown patterns with potential interest from massive amounts of web data. [13]

B. Data and Text Pre-Processing

Cleaning and preparing data aim to acquire higher efficiency and effectiveness in the mining process. Researches call this stage Pre-processing, and indicate that pre-processing methods can reduce from 50% to 80% of the total mining process. [14]

Text pre-processing solves the problem of the feature space high dimensionality, where features (or terms) can exceed tens or hundreds of thousands in count. It also makes text analysis more accurate, and saves time and space [14]

Text enters a series of steps including some or all of the following: [15,16]

1. Extraction, which tokenizes the text
2. Lowercase Conversion
3. Special Characters Removal
4. Stopwords Removal
5. Stemming and Lemmatization

6. Pruning rare words (as they lead to noise in data), using Document Frequency (DF).
7. TF-IDF Weighting

C. *Data Mining Methods*

This section includes brief descriptions of mining methods we used in our approach.

1) *Association Rules*

Association Rules Mining is the process of discovering patterns, relationships and information structures inside groups of objects in transactional databases and other information repositories. [17]

The most common use of association rules is in marketing, by analyzing purchasing traffics, which aims to identify groups of products that consumers buy often together, in addition to other fields such as telecommunications, risk management, stock control, agriculture and others. [18]

Generating association rules depends on two important measures: Support and Confidence, by generating all rules that fulfill support and confidence greater than user-defined thresholds named *minsup* and *minconf* respectively. [18]

One of the most common and efficient methods of association rules is FP-Tree.

2) *Clustering*

Clustering, or Cluster Analysis, is the process of dividing a huge amount of unordered data in a small number of consistent and meaningful groups, by assigning similar elements to one group, which is represented by a center. These groups have three main characteristics: 1) they are relatively homogeneous within themselves, 2) heterogeneous among each other; 3) measuring homogeneity and heterogeneity are according to pre-defined parameters, and using similarity and distance measures, such as Euclidean Distance and Cosine Similarity, to compare the elements to the center. These groups form the clusters, while the centers of the clusters represent the centroids. [19]

In Documents Clustering on the other hand, the process starts by preparing the documents through Feature Extraction, which links each document with a brief representation of its topic using Vector Space Model or Bag of Words Model. The process then calculates centroid vectors, which represent the average weight of documents in the cluster, and measures similarity between two documents using measures like Cosine Similarity that calculates the angle between two vectors. [20]

In our study, we employ two methods of Non-Hierarchal Clustering, which are K-Means and K-Medoids.

IV. System Architecture

In this section, we demonstrate the suggested system modules that perform three basic missions: crawling a dark web market, pre-processing extracted data, and mining the processed data. The market under study (which we consider not mentioning for security reasons) uses a *Tor*[1] service as a host, thus it needs a connection to Tor network, in addition to a pre-registration and login. The market is specialized in promoting and selling illicit and malicious products of several different types, distributed over 12 categories.

A. *System Components*

Our system consists of three basic modules:

1. The dark crawler *Darky*
2. The *Cleaner*
3. The *Dark Miner*

In the crawling phase, we employ our crawler *Darky* [21] to access and crawl the website, and extract data from products and vendors' pages. In the second phase, we developed a cleaning module that applies pre-processing techniques to clean and transform data. Finally, we perform data mining methods on clean data using different strategies to extract patterns. These patterns can be quite useful for various agencies in inferring information about products, vendors and their promoting styles. Fig.1 illustrates the three modules:

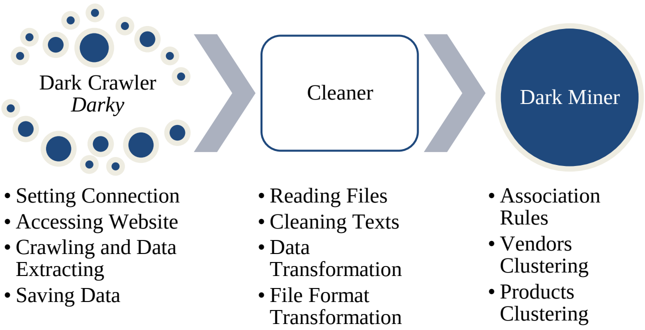


Fig.1. System Modules

B. *Data Collection*

In earlier stage, we developed a dark crawler *Darky*[2] using *Scrapy*[3]. The crawler starts by setting a connection to *Tor* network through *Tor* software integrated with *Privoxy* [4]. After setting the connection, the crawler simulates a user login with credentials we previously created on a dark web market (which we consider not mentioning for security reasons), crawls the entire website pages extracting and directly structuring data without downloading any page, and saving data in separated *json* files.

The data includes records of 6387 products and 179 vendors shown in Table 1.

[1] Tor Project, https://www.torproject.org

[2] Refer to a previous published article for details about Darky and crawling the dark web, available at: https://doi.org/10.6025/jdim/2019/17/2/51-60

[3] Scrapy, https://scrapy.org

[4] Privoxy, https://www.privoxy.org

Table 1. Number of Vendors and Products Extracted from the Dark Web Market

Vendors' Total = 179		
Products Total = 6387		
	Category	Number of Products
1	Carded Items	27
2	Counterfeit Items	92
3	Digital Products	2179
4	Drugs & Chemicals	1569
5	Fraud	887
6	Guides & Tutorials	987
7	Jewels & Gold	15
8	Other Listings	243
9	Security & Hosting	50
10	Services	139
11	Software & Malware	185
12	Weapons	14

C. *Data Pre-Processing*

Due to the difficulty of reading data in its current format, we transform data into another format that is easy to read and mine. In the pre-processing module, we developed the *Cleaner*. The main mission of the Cleaner we summarize as follows:

1) Reading elements from *json* files, and for the element "Title", the Cleaner processes it by removing non-alphabetic (or meaningless) characters, then removing Stopwords
2) Adding the read elements to a new *json* file
3) Transforming *json* file to a new file with *xlsx* format ready for mining

As we previously designed our crawler, here we need two versions of the cleaner as well, one to process products data, and the other to process vendors' data.

Fig.2 illustrates how the Cleaner works on products data:

The Cleaner processes products data as follows:

1. Starts from a list of files names that contain the products
2. Reads the elements from each file
3. Processes the Title by transforming it to lowercase, then removing the non-alphabetical characters. Using *re* library, we defined a regular expression that contains all removable characters used in titles. Some of the characters are common (like periods, commas and brackets), and others are rarely used or with strange shapes (like the symbols ❥✪✮♛), as we noticed vendors often use many of these characters in products titles to draw attention to their products. We excluded some characters that may affect the meaning of the title (like – and %). Furthermore, the cleaner removes Stopwords, we used *nltk.corpus* library for this mission, which contains a previously defined list of Stopwords in English.
4. Reads the rest of the element attributes with simple cleaning of some brackets
5. Adds the processed elements to a new data array
6. Moves to the next file
7. Empties the array in a new *json* file
8. Transforms data from the last json file to a spreadsheet with xlsx format, here we used pandas library

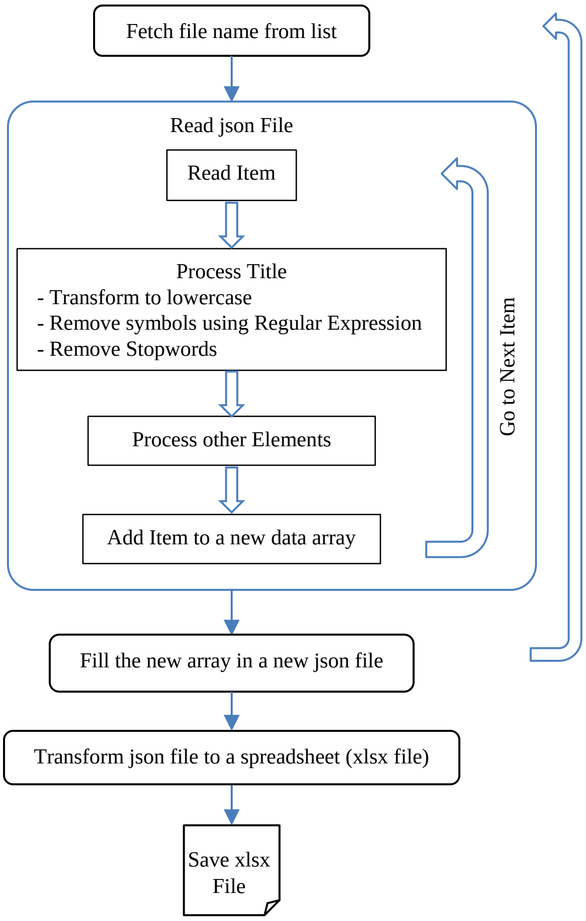

Fetch file name from list
Read json File
Read Item
Process Title
- Transform to lowercase
- Remove symbols using Regular Expression
- Remove Stopwords
Process other Elements
Add Item to a new data array
Go to Next Item
Go to Next File
Fill the new array in a new json file
Transform json file to a spreadsheet (xlsx file)
Save xlsx File

Fig.2. The Cleaner Workflow

With similar but simpler steps, the Cleaner transforms the vendors' data file.

From this stage, we obtained two spreadsheets files: products and vendors.

Table 2 shows examples of products titles before and after processing:

D. *Data Mining Strategies*

In this stage, we used RapidMiner[1] to design the processes we performed on the extracted and cleaned data, experimenting our methods to study the entire website content.

1) *Association Rules Process Design*

As mentioned before, the most common use of association rules is in the field of analyzing shopping traffic. However, we employed this technique in a different purpose, which is analyzing promotion styles. In our approach, we apply the concept of association rules

[1] RapidMiner, https://rapidminer.com/

on words, i.e. studying the words that frequently occur together in illicit products titles. This method aims to show the most used words by vendors when they promote their products on the market. We accomplished this approach by considering the “Title” as the transaction and the words as the items.

Table 2. Examples of Products Titles before and after Processing

Title before processing	Title after processing
★LSD-25 Blotter 250µg★ Qty: 250	lsd-25 blotter 250µg qty 250
♛♛ BeatsMusic Premium Account - FOR ONLY 7$!!! ♛♛	beatsmusic premium account - 7
★★ Identity Finder -- Best Data stealer ★★	identity finder -- best data stealer
★ Ketamine R High Quality 87% ★ Full Crystal's★€17.5★ EC PUREST LABTEST EVER 97%!	ketamine r high quality 87% full crystal €17.5 ec purest labtest ever 97%

The patterns of the frequently used words may give the analyzer an idea about the most promoted products or the most used combinations of words in titling products. These rules can also help in finding conceptual compositions.

The process represents documents by vectors of words, using TF-IDF to weight words. The process then performs two basic steps: First, it generates the FP-Growth tree to extract frequent words patterns, using Support threshold to determine the desired minimum number of patterns recurrence. Second, it creates association rules from the produced tree using Confidence threshold to determine the desired power of the rule.

Fig.3 illustrates the process design of generating association rules out of products titles, and Fig.4 illustrates the sub-process of documents processing and it consists of a Tokenization operation:

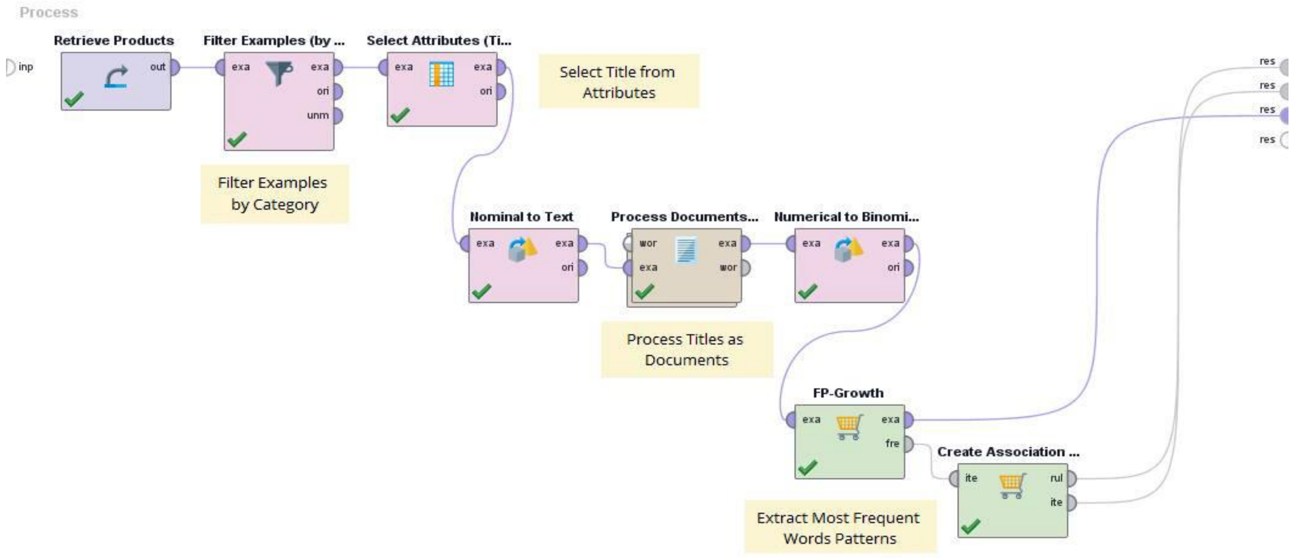


Fig.3. Association Rules Process Design

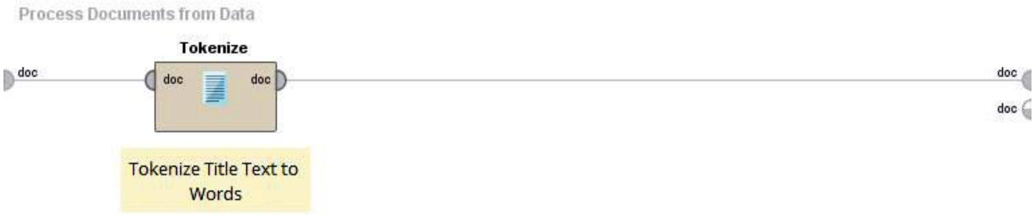


Fig.4. Document Processing Sub-Process

i. *Association Rules Results*

The market has products distributed into categories. Therefore, we divide the data to examples according to the corresponding category.

Table 3 shows the number of frequent itemsets and the maximum size of an itemset according to different values of Support and Confidence (for the other categories, they did not show results):

Table 3. Number of Generated Frequent Itemsets According to Different Support and Confidence Values

Category	No. of Examples	No. of Items	Sup.	Conf.	No. of Itemsets	Maximum Itemset Size	No. of Rules
Counterfeit Items	92	264	0.12	0.8	15	4	43
			0.13	0.85	15	4	43
Digital Products	2179	2038	0.2	0.65	14	3	13
			0.3	0.7	5	2	2
Jewels & Gold	15	48	0.2	0.65	79	6	569
			0.3	0.7	8	2	3
Other Listings	243	546	0.5	0.85	23	4	88
			0.55	0.95	9	3	10
Security & Hosting	50	133	0.1	0.65	23	4	46
			0.2	0.7	5	2	1
Services	139	256	0.15	0.65	17	4	44
			0.2	0.75	3	2	1
Software & Malware	185	331	0.15	0.7	4	2	1
			0.2	0.75	4	2	1

To explain these results, we will take "Counterfeit Items" category. Table 4 shows the generated itemsets from "Counterfeit Items" category with the values 0.13 and 0.85 for Support and Confidence respectively, and sorted by the highest Support, and Table 5 shows examples of association rules generated from these itemsets:

Table 4. Itemsets Generated from "Counterfeit Items" Category

Size	Support	Item 1	Item 2	Item 3	Item 4
1	0.413	forged			
1	0.13	labels			
1	0.13	prescription			
1	0.13	rx			
2	0.13	forged	labels		
2	0.13	forged	prescription		
2	0.13	forged	rx		
2	0.13	labels	prescription		
2	0.13	labels	rx		
2	0.13	prescription	rx		
3	0.13	forged	labels	prescription	
3	0.13	forged	labels	rx	
3	0.13	forged	prescription	rx	
3	0.13	labels	prescription	rx	
4	0.13	forged	labels	prescription	rx

Table 5. Examples of Association Rules Generated for "Counterfeit Items" Category

Rule	Confidence
[labels] --> [forged]	(confidence: 1.000)
[prescription] --> [forged]	(confidence: 1.000)
[labels] --> [forged, prescription]	(confidence: 1.000)
[forged, labels] --> [prescription]	(confidence: 1.000)
[forged, prescription] --> [labels]	(confidence: 1.000)
[forged, labels] --> [rx]	(confidence: 1.000)
[prescription] --> [forged, rx]	(confidence: 1.000)
[forged, prescription] --> [rx]	(confidence: 1.000)
[forged, rx] --> [prescription]	(confidence: 1.000)
[prescription, rx] --> [forged]	(confidence: 1.000)
[labels, rx] --> [prescription]	(confidence: 1.000)
[labels] --> [forged, prescription, rx]	(confidence: 1.000)
[forged, labels] --> [prescription, rx]	(confidence: 1.000)
[prescription] --> [forged, labels, rx]	(confidence: 1.000)
[forged, prescription] --> [labels, rx]	(confidence: 1.000)
[labels, prescription] --> [forged, rx]	(confidence: 1.000)
[rx] --> [forged, labels, prescription]	(confidence: 1.000)
[forged, rx] --> [labels, prescription]	(confidence: 1.000)
[forged, labels, rx] --> [prescription]	(confidence: 1.000)

Taking a sample word, Fig.5 illustrates association rules related to the element (word) "prescription" with the Support and Confidence of each rule:

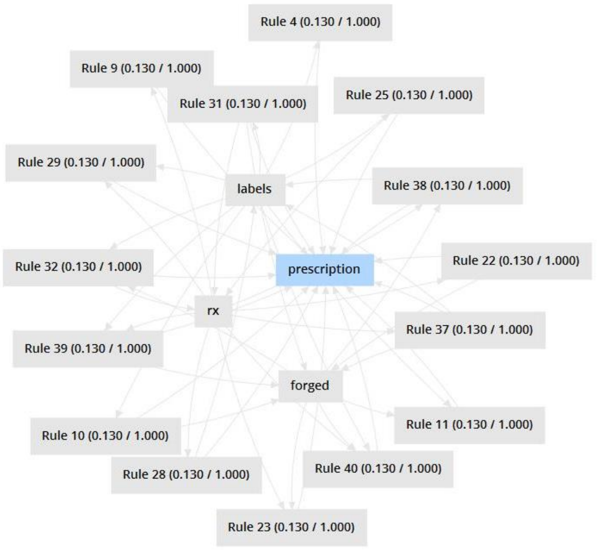


Fig.5. Association Rules Linked with the Word "prescription"

From this approach, we can also conclude the conceptual compositions used when promoting products in each category. Table 6 shows the conceptual compositions resulted from association rules in each category. We choose the compositions with size greater than 2:

2) *Clustering Processes Design*

We apply clustering concepts on two aspects: clustering vendors and clustering products.

1. Clustering Vendors

The available data about vendors are the following:

- Vendor Name
- Vendor Level
- Trust Level
- Positive Feedback
- Date of Membership
- Number (presented between two parentheses without any prefix or explanation)

From the above, we use the three attributes “Vendor Level”, “Trust Level”, and “Number” to cluster vendors. We chose these attributes due to the distinguished disparity of values we noticed among different vendors, and a potential relationship may exist between the Number and both Vendor Level and Trust Level.

Table 6. The Conceptual Compositions Concluded from the Generating Association Rules from Products Titles

Category	Conceptual Compositions
Counterfeit Items	forged, labels, prescription
	forged, labels, rx
	forged, prescription, rx
	labels, prescription, rx
	forged, labels, prescription, rx
Digital Products	account, porn, premium
	account, premium, warranty
Jewels & Gold	audemars, royal, oak
	audemars, stainless, steel
	audemars, piguet, royal, oak
	audemars, piguet, stainless, steel
	audemars, piguet, oak, stainless, steel
	piguet, royal, oak, stainless, steel
	audemars, piguet, royal, oak, stainless, steel
Other Listings	account, lifetime, premium
	account, lifetime, porn
	account, premium, porn
	lifetime, premium, porn
	account, lifetime, premium, porn
Security & Hosting	vpn, account, premium
	premium, warranty, lifetime
	premium, warranty, accounts
	premium, lifetime, accounts
	warranty, lifetime, accounts
	premium, warranty, lifetime, accounts
Services	famous, get, way
	famous, get, easy
	famous, way, easy
	get, way, easy
	famous, get, way, easy

Before starting the clustering process, we apply a performance evaluation to help in choosing the suitable clustering algorithm and the number of clusters to generate. Fig.6 illustrates the steps of performance evaluation process for both K-Means and K-Medoids using Davies–Bouldin Index[1], with different numbers of clusters and the Euclidean Distance measure.

Table 7 shows the resultant values of Davies-Bouldin Index in performance evaluation of both algorithms:

Table 7. Davies-Bouldin Index Values for both K-Means and K-Medoids Algorithms in Vendors’ Clustering with Different Numbers of Clusters

Number of Clusters	K-Means	K-Medoids
2	0.310	0.275
3	0.389	0.376
4	0.312	0.493
5	0.387	0.798
6	0.364	0.499

From this table we can see that clustering using K-Medoids with k=2 gives the least value of the index thus the best clustering. Relying on this result, we design the vendors’ clustering process illustrated in Fig.7. In the end of this process, it stores Top Vendors’ names for later use (explained in the next section):

i. Clustering Vendors Results

By completing the above process, we have the following distribution of elements on clusters:

```
Cluster Model:
Cluster 0: 172 items
Cluster 1: 7 items
Total number of items: 179
```

Table 8 includes values of the clusters centroids, and Table 9 shows Top Vendors’ data in “cluster_1”. According to the three attributes, we notice relatively high values of the Number against values of the Vendor Level and Trust Level, which they range between 2 and 5. This number may refer to the level of interactivity or professionalism of the vendors inside the market community; therefore, we considered them as Top Vendors:

Table 8. Centroids Values in the Generated Clusters of Vendors

Attribute	cluster_0	cluster_1
number	0	1243
tlevel	1	4
vlevel	1	5

Table 9. Top Vendors’ Data

name	cluster	number	tlevel	vlevel	member	positive
Lepricon	cluster_1	1045	2	3	Fri Mar 16 2018	95.77
Underworld	cluster_1	742	3	4	Sat Apr 07 2018	92.61
DrunkDragon	cluster_1	1919	1	2	Sat Feb 03 2018	93.06
rvaska	cluster_1	1316	1	2	Thu Feb 22 2018	95.71
GodsLeftNut	cluster_1	1022	1	2	Fri Apr 20 2018	97.98
TikTokCC	cluster_1	901	3	4	Wed Apr 18 2018	98.15
savastano	cluster_1	1243	4	5	Mon Apr 16 2018	98.75

[1] Davies-Bouldin Index factor stands on the principle that "the algorithm which produces clusters with minimum intra-distances, and maximum inter-distances will have minimum value of Davies-Bouldin Index", i.e. it is the percentage of internal distances of clusters elements on external distances of clusters. Therefore, the smaller its value, the better the clustering is. [22]

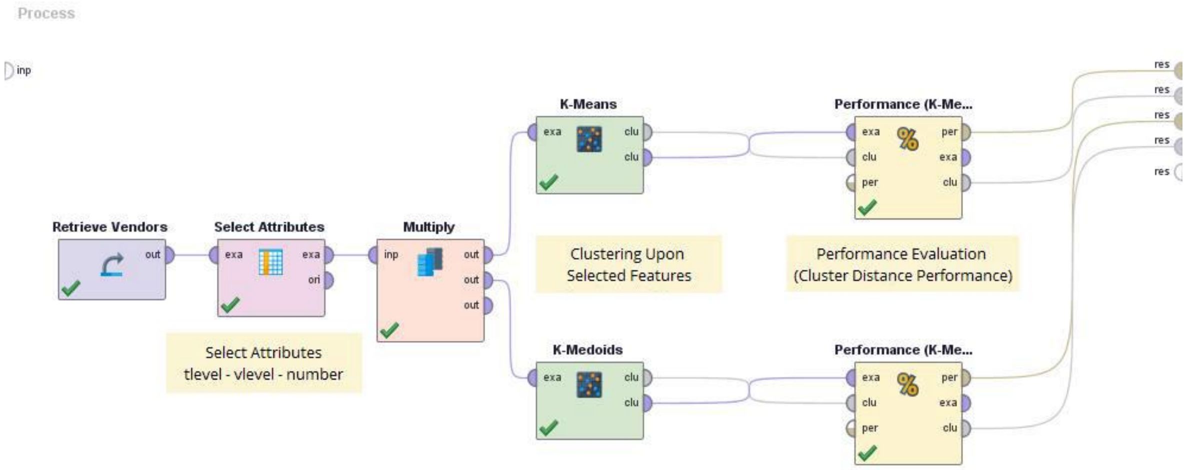


Fig.6. Process Design for Evaluating Performance of both K-Means and K-Medoids Algorithms in Vendors' Clustering

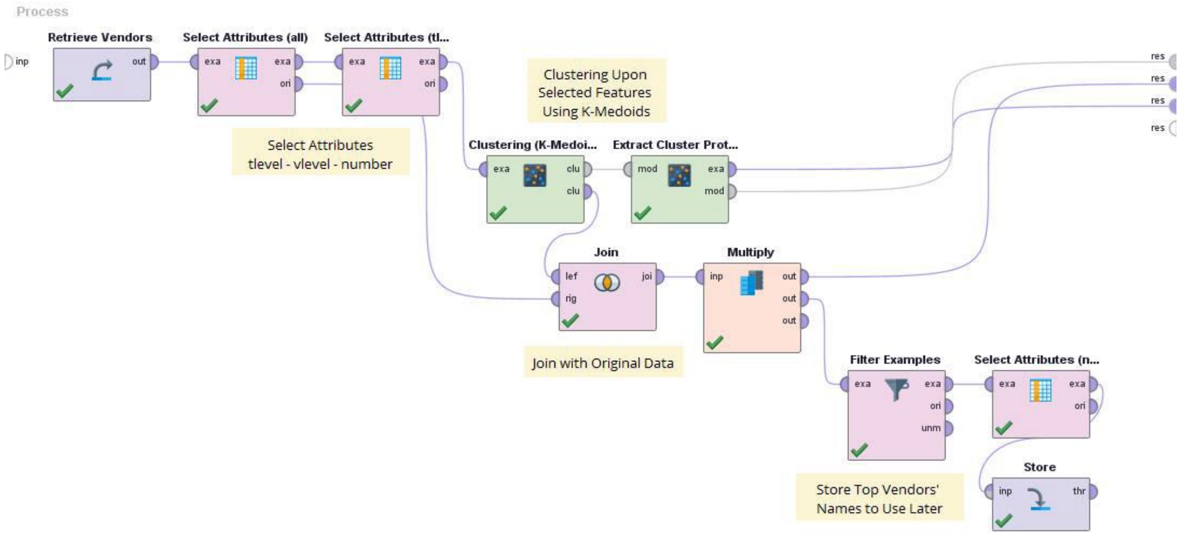


Fig.7. Process Design of Vendors Clustering Using K-Medoids Algorithm

Fig.8 illustrates the distribution of clusters elements in a 3D-Space created from the three selected attributes:

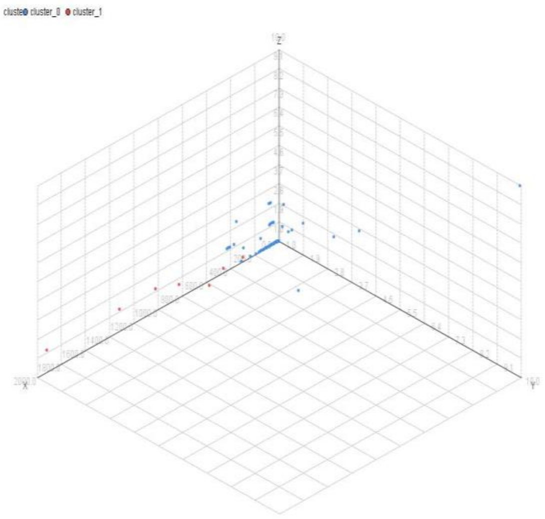


Fig.8. Vendors' Clusters Distribution in a 3D-Sapce

The extracted information may help in inferring that the mentioned vendors are the most professional, or their products are the most popular. It may also refer to the fact that they are companies or trade organizations due to their wide spreading products, as the latter (as we will see in the next section) form 1365 of 6387 items, i.e. 21.37% of the market stock.

2. Clustering Products

The available data about products are the following:

- Product title
- Vendor name
- Origin
- Shipment destination
- Price (in USD)
- Sales amount (since a specific date)
- Views number (since the same date)
- Since: the date when the mentioned sales and views had started
- Category

In products clustering, we apply clustering concepts on two aspects as well: clustering products titles by considering them as documents and applying Documents Clustering, and the other is by taking advantage of the numerical attributes of the products.

In clustering products titles, we can divide the dataset by several ways, such as clustering products belonging to a particular category, or the products sold by specific vendors, or others. Here we found it is useful to cluster products sold by the Top Vendors, the approach that may introduce important information about the common characteristics among those vendors and among the products they promote.

In a similar way, we evaluated clustering performance with K-Means to choose the best possible number of clusters[1].

Table 10 shows Davies-Bouldin Index values for different numbers of clusters:

Table 10. Davies-Bouldin Index Values for K-Means Algorithm in Products Titles Clustering According to Different Numbers of Clusters

Number of Clusters (k)	Davies Bouldin
2	2.183
3	2.309
4	2.149
5	2.212
6	2.217
7	3.439

From this table, we notice approximate values, which clustering can depend on any. In our experiment, we chose the number 4, with Cosine Similarity for measuring.

Fig.9 illustrates the process design for clustering products titles of the Top Vendors. The process represents documents with vectors of words and their weights using TF-IDF. In additions, we use DF equals to 3 to remove extremely rare word appearing in less than 3 documents. This step forms a sub-process illustrated in Fig.10, and it consists of three operations: Tokenization, stemming with Porter Stemmer, and filtering tokens by length to filter words with length less than 2. These operations help in additional reduction of words amount and gaining more accurate clustering.

i. Clustering Products Titles Results

The above process produces the following distribution of elements upon clusters:

Cluster Model:
Cluster 0: 612 items
Cluster 1: 323 items
Cluster 2: 209 items
Cluster 3: 221 items
Total number of items: 1365

These elements represent vectors of words, which their count equals to 324 words, and their weights, and the centroids calculations are according to these weights. Table 11 shows the number of weighted words for each cluster (as the rest of the words have weights equal to 0), and Table 12 lists the first twenty words with the highest weights in each cluster:

Table 11. Number of Weighted Words in each Cluster

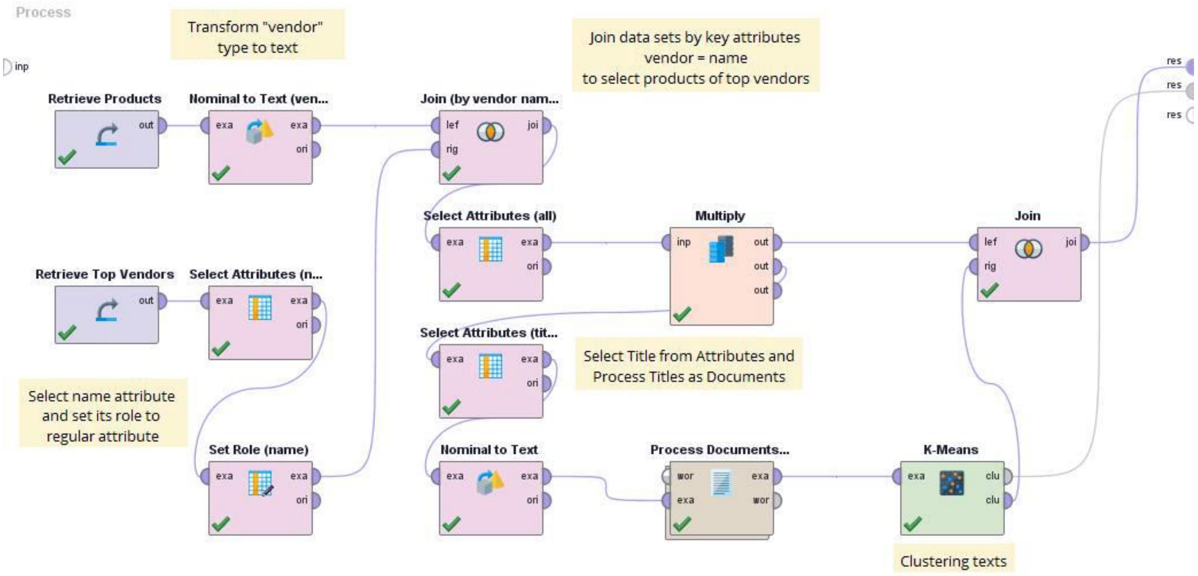
Cluster	Number of Weighted Words
cluster_0	289
cluster_1	52
cluster_2	67
cluster_3	155

Table 12. The Top 20 Words with the Highest Weights in each Cluster

cluster_0		cluster_1		cluster_2		cluster_3	
Attribute	TF-IDF	Attribute	TF-IDF	Attribute	TF-IDF	Attribute	TF-IDF
crypter	0.073	databas	0.396	config	0.360	get	0.278
make	0.060	record	0.345	mba	0.246	free	0.266
crack	0.058	plaintext	0.266	sentri	0.246	card	0.212
checker	0.042	com	0.231	snapshot	0.243	product	0.076
brute	0.039	million	0.119	snipr	0.240	amazon	0.040
monei	0.032	net	0.044	leagu	0.036	gift	0.037
keylogg	0.031	buycraft	0.039	legend	0.036	samsung	0.023
paypal	0.030	log	0.037	storm	0.032	iphon	0.021
phish	0.029	forum	0.032	domino	0.011	video	0.020
account	0.027	org	0.019	com	0.011	rdp	0.020
page	0.027	account	0.018	uk	0.009	eat	0.018
video	0.025	lifetim	0.017	cb	0.008	credit	0.017
dai	0.024	porn	0.017	chaturb	0.008	drop	0.016
pro	0.022	instant	0.013	crunchyrol	0.008	googl	0.016
hack	0.020	deliveri	0.013	deezer	0.008	cloth	0.015
cc	0.018	game	0.009	deviantart	0.008	laptop	0.015
cvv	0.018	btc	0.009	expressvpn	0.008	method	0.014
cashout	0.018	gamerzplanet	0.008	fitbit	0.008	game	0.014
easi	0.016	ultim	0.007	gamestop	0.008	item	0.014
ultim	0.015	minecraft	0.006	hulu	0.008	smart	0.013

In another perspective, and to take advantage of the numerical attributes of the products, we apply clustering on “views”, “sales” and the date when the views and sales had started determined by “since”. Here we calculate the duration in which the recorded views and sales took place, starting from “since” date until the date in which the crawler extracted the data from the website.

[1] We excluded applying K-Medoids algorithm due to the extreme time it consumes and without getting the desired results, therefore we settled for K-Means algorithm.



Transform "vendor" type to text

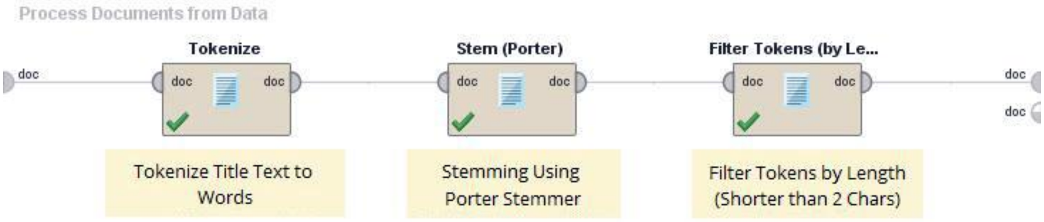
Join data sets by key attributes vendor = name to select products of top vendors

Select name attribute and set its role to regular attribute

Select Title from Attributes and Process Titles as Documents

Clustering texts

Fig.9. Process Design for Clustering Products Titles of the Top Vendors Using K-Means Algorithm



Tokenize Title Text to Words

Stemming Using Porter Stemmer

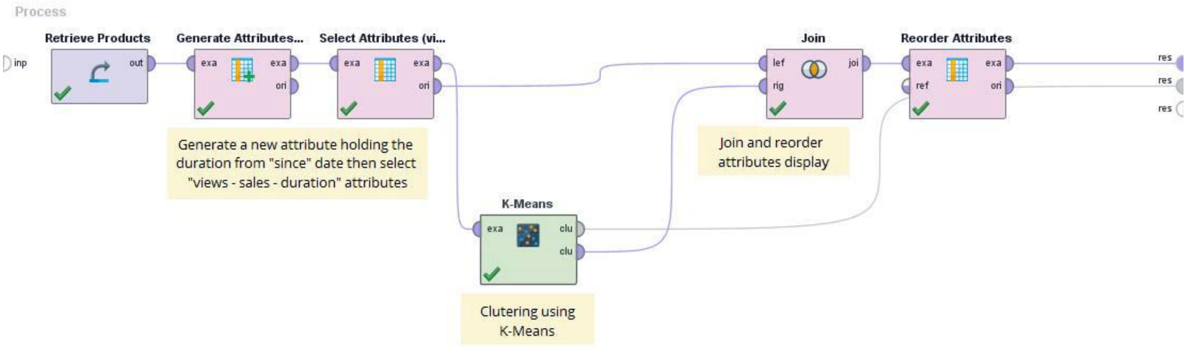
Filter Tokens by Length (Shorter than 2 Chars)

Fig.10. Processing Products Titles as Documents

Here also we found approximate values of Davies-Bouldin Index as shown in Table 13, we chose the number 7, and the Euclidean Distance for measuring. Fig.11 illustrates the process design of products clustering according to views, sales and their durations.

Table 13. Davies-Bouldin Index Values in Evaluating Performance of Products Clustering using K-Means Algorithm

Number of Clusters (k)	Davies Bouldin
2	0.505
3	0.470
4	0.528
5	0.567
6	0.552
7	0.501



Generate a new attribute holding the duration from "since" date then select "views - sales - duration" attributes

Join and reorder attributes display

Clutering using K-Means

Fig.11. Process Design of Products Clustering According to View, Sales and their Occurring Durations

ii. Clustering Products Results

The previous process produces the following distribution of the elements over clusters:

Cluster Model:
Cluster 0: 143 items
Cluster 1: 2657 items
Cluster 2: 1 items
Cluster 3: 7 items
Cluster 4: 39 items
Cluster 5: 3055 items
Cluster 6: 485 items
Total number of items: 6387

Table 14 lists the centroids values in the generated clusters, and Table 15 shows the elements in “cluster_2” and “cluster_3”, which we notice that they include the products with the most views and sales:

Table 14. Centroids Values in the Generated Clusters of Products

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5	cluster_6
sales	33.909	0.849	367	267.714	76.872	0.204	7.381
views	1206.86	196.526	13457	7446.429	2780.821	35.314	524.315
duration	148.287	123.841	196	157.429	160.436	28.704	147.757

Table 15. Items Details in “cluster_2” and “cluster_3”

cluster	category	title	vendor	origin	shipto	price	views	sales	since	duration
cluster_2	Fraud	super fresh high quality united kingdom fullz	RangeRovers	World Wide	World Wide	5	13457	367	Sun Mar 04 2018	196
cluster_3	Counterfeit Items	counterfeit 20 usd notes-high quality-full escrow minimum order 5	swimnotes	United States	United States	3.99	8146	203	Sat Mar 03 2018	197
cluster_3	Fraud	usa hq cc 100% live known balance 1 - 5k 17	TikTokCC	World Wide	World Wide	17	9174	484	Fri May 04 2018	135
cluster_3	Fraud	100 us paypal accounts balance paypal credit	Rambe	World Wide	World Wide	6	6655	128	Fri Mar 16 2018	184
cluster_3	Fraud	cvv usa high quality best quality 100%live known balance 1000 5k	savastano	World Wide	World Wide	15	5444	277	Mon Apr 16 2018	153
cluster_3	Fraud	live usa cvv known balance 1k 5k 5k 20k 20k 50k buy2get1free	friskmint	World Wide	World Wide	10	7679	336	Sun Jul 08 2018	70
cluster_3	Fraud	usa cvv best quality market hq live 100% known balance 5k 15k	savastano	World Wide	World Wide	18	9288	407	Tue Apr 17 2018	152
cluster_3	Guides & Tutorials	video proofs wanna make 77 400-158 000 day	MrMillionaire	World Wide	World Wide	499	5739	39	Sat Feb 17 2018	211

V. Conclusion and Future Work

In this paper, we discussed how to infer useful information from data extracted from a dark web market. This information helps security and law agencies in their investigations about activities that take place on the dark web.

We introduced our approach of a system that consists of three modules. The first module is The Dark Crawler *Darky*, which accesses a dark web market, by simulating a user login to the market, scans the whole website, and extracts and directly structures the extracted data. The second module is The Cleaner, which applies pre-processing procedures on extracted data, including cleaning and transformation, to prepare data for mining. The third and final module is The Dark Miner, which consists of mining processes in the fields of Association Rules and Clustering, illustrating all gained results. We applied text mining on products titles by considering each title as a document. In the field of Association Rules, we extracted frequent words used by vendors when they promote their products, and from the generated rules we inferred conceptual compositions used in titling products. In the field of Clustering, we employed Clustering techniques to find a relationship among some specific attributes in the vendors’ data, which are the Vendor Level, Trust Level, and the bracketed (mysterious) Number. The process clusters vendors according to these three attributes, and inferring that they might have a correlation, leading to the Top Vendors (or vendors with the most professionalism and interactivity). Next, we applied clustering on products titles of the Top Vendors to find the common characteristics they share according to the products they promote, achieving this through the highest weighted terms. Lastly, we performed clustering on products according to views, sales and the duration of when the mentioned views and sales have occurred.

This approach can apply several data mining techniques on extracted data. The design of the crawler differs from a dark website to another, as well as data mining operations may differ according to the website nature and quality of the data contained in the website. And for the same website, results will vary from a period of time to another, as products may significantly increase with various promotion styles, in addition to increase in vendors' number. This approach puts a dark website under investigation and in any time.

In its future prospects, we develop the system to cover other methods of data mining, such as Classification integrated with Clustering, by using the generated clusters as labels, or integrated with Association Rules in a methodology of Associative Classification. Also we expand our experiments to include markets in other languages, and involve techniques in the field of pre-processing, including corrections of the misspelled words and words containing purposely-altered characters with symbols or letters from mixed languages.

References

[1] B. Hawkins, "Under The Ocean of the Internet - The Deep Web," 15 5 2016. [Online]. Available: https://www.sans.org/reading-room/whitepapers/covert/ocean-internet-deep-web-37012. [Accessed 13 December 2018].

[2] M. S. L. S. W. L. Andres Baravalle, "Mining the Dark Web: Drugs and Fake Ids," in Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference, Barcelona, Spain, 2016.

[3] J. A. F. A. R. W. Monica J. Barratt, "Safer Scoring? Cryptomarkets, Social Supply and Drug Market Violence," International Journal of Drug Policy, vol. 35, pp. 24-31, 2016.

[4] A. D. ,. E. M. ,. E. N. ,. V. P. ,. J. S. ,. P. S. John Robertson, Darkweb Cyber Threat Intelligence Mining, Cambridge: Cambridge University Press, 2017.

[5] O. B. M. N. S. Quan Le, "Deep Learning at the Shallow End: Malware Classification for Non-Domain Experts," Digital Investigation, vol. 26, pp. S118-S126, 2018.

[6] G. M. M. M. Alessandro Celestini, "Tor Marketplaces Exploratory Data Analysis: The Drugs Case," in International Conference on Global Security, Safety, and Sustainability, 2017.

[7] M. A. E. N. P. S. Ericsson Marin, "Community Finding of Malware and Exploit Vendors on Darkweb Marketplaces," in In 2018 1st International Conference on Data Intelligence and Security (ICDIS), 2018.

[8] F. C. B. B. Paolo Spagnoletti, "An Investigation on the Generative Mechanisms of Dark Net Markets," in Proceedings of the 13th Pre-ICIS Workshop on Information Security and Privacy, 2018.

[9] D. L. N. A. S. A. M. P. M. S. Ben R. Lane, "The Dark Side Of The Net: Event Analysis Of Systemic Teamwork (East) Applied To Illicit Trading On A Darknet Market," in Proceedings of the Human Factors and Ergonomics Society 2018 Annual Meeting, 2018.

[10] I. Ladegaard, "Crime Displacement in Digital Drug Markets," International Journal of Drug Policy, vol. 63, p. 113–121, 2019.

[11] D. R. M. M. L. S. Q. R. Julian Bros éus, "A Geographical Analysis of Trafficking on a Popular Darknet Market," Forensic science international, vol. 277, pp. 88-102, 2017.

[12] H. P. Petar Ristoski, "Semantic Web in data mining and knowledge discovery: A comprehensive survey," Web Semantics: Science, Services and Agents on the World Wide Web, vol. 36, pp. 1-22, 2016.

[13] P. R. A. S. P. S. Praveen Kumari, "Web Mining - Concept, Classification and Major Research Issues: A Review," Asian J. Adv. Basic Sci, vol. 4, no. 2, pp. 41-44, 11 May 2016.

[14] L. H. P. P. C. Pritam C. Gaigole, "Preprocessing Techniques in Text Categorization," in National Conference on Innovative Paradigms in Engineering & Technology (NCIPET-2013), 2013.

[15] J. I. N. S. Vijayarani, "Preprocessing Techniques for Text Mining - An Overview," International Journal of Computer Science & Communication Networks, vol. 5, no. 1, pp. 7-16, 2015.

[16] S. K. I. A. K. M. A. A. Rida Hafeez, "Does Preprocessing Really Impact Automatically Generated Taxonomy," in 2017 13th International Conference on Emerging Technologies (ICET), 2017.

[17] J. T. P. Surbhi K. Solanki, "A Survey on Association Rule Mining," in 2015 Fifth International Conference on Advanced Computing & Communication Technologies, 2015.

[18] J. T. P. Surbhi K. Solanki, "A Survey on Association Rule Mining," in 2015 Fifth International Conference on Advanced Computing & Communication Technologies, 2015.

[19] T. Vishal, "Cluster Analysis for Market Segmentation," 3 February 2015. [Online]. Available: https://www.slideshare.net/vishtandel1991/cluster-analysis-for-market-segmentation. [Accessed 9 March 2019].

[20] N. P. M. P. Jasmine Irani, "Clustering Techniques and the Similarity Measures used in Clustering: A Survey," International Journal of Computer Applications, vol. 134, no. 7, pp. 9-14, January 2016.

[21] R. B. Bassel AlKhatib, "Crawling the Dark Web: A Conceptual Perspective, Challenges and Implementation," Journal of Digital Information Management (JDIM), vol. 17, no. 2, pp. 51-60, April 2019.

[22] "Davies-Bouldin Criterion Clustering Evaluation Object - MATLAB - MathWorks Benelux," [Online]. Available: https://nl.mathworks.com/help/stats/clustering.evaluation. daviesbouldinevaluation-class.html. [Accessed 19 April 2019].

Authors' Profiles

Bassel Alkhatib is the web sciences master director at the Syrian Virtual University and the head of Artificial Intelligence department at Information Technology Faculty at Damascus University. He holds PhD degree in computer science from the University of Bordeaux-France, 1993. Dr. Alkhatib supervises many PhD students in web mining, and knowledge management. He also leads and teaches modules at both BSc and MSc levels in computer science and web engineering in Syrian Virtual University, Damascus University, and Al-Shem Private University.

Randa S. Basheer is a web sciences master student at Syrian Virtual University. She has a Bachelor degree in information technology engineering, Damascus University, Damascus, Syria, 2006. Ms. Basheer co-authored a published article with Dr. Alkhatib: "Crawling the Dark Web: A Conceptual Perspective, Challenges and

Implementation", Journal of Digital Information Management (JDIM), April 2019. Her current research of interest is the Dark Web.

How to cite this paper: Bassel Alkhatib, Randa S. Basheer, " Mining the Dark Web: A Novel Approach for Placing a Dark Website under Investigation", International Journal of Modern Education and Computer Science(IJMECS), Vol.11, No.10, pp. 1-13, 2019.DOI: 10.5815/ijmecs.2019.10.01