Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7765859

GRAPHICAL ANALYSIS OF DARK DATA USING NATURAL LANGUAGE PROCESSING

Anand A. Chaudhari[1], Dr. M. A. Pund[2]

Department of Computer Science & Engineering, Prof. Ram Meghe Institute of Technology and Research, Badnera, Amravati-444701, Maharashtra, India

Corresponding Author Email:anand07chaudhari@gmail.com

Abstract

Dark data refers to all of an organization's underused, unidentified, and unexplored data generated in the course of a consumer's everyday online communication with different devices. Dark data is anything from machine-generated data to unstructured data retrieved from social media. Although dark data is a subclass of big data, it accounts for the majority of the entire volume of big data acquired by businesses each year. Businesses very rarely process or handle dark data for a variety of reasons, but this does not diminish its significance in terms of business impact. There are two perspectives on the significance of dark data. Unanalyzed data, according to one viewpoint, includes hidden key insights and represents a missed opportunity. From the other side of the argument, unanalyzed data that isn't handled properly can cause a lot of problems, including legal and security problems. This work also discusses the simple mathematical model of analyzing dark data and ways to convert unstructured data into structured data with example.

Keywords: Analytics, Big Data, Dark Data, Natural Language Processing

1. Introduction

We are all around data everywhere. The data that is generated is not only of a single type or from a single person. The type of data may be text, audio, video, etc. The generator of data may be a person, a system, surveys or anything. It is not only generating but propagating data that is also one of the favorite tasks of human beings. Data generation and sharing have become pocket-friendly. When data was small, only one system could handle it; now, data is generated rapidly and in a variety of forms. As we enter into the era where massive amounts of data are being generated, the issues of dealing with it still exist. Even though information in the current generation, where data has evolved in an enormous way, can only become even bigger to be filled up into a single machine, it also means that almost all conventional mining algorithms or business intelligence formed for a centralized process of data analysis may not have been readily adaptable straightforwardly to big data. Researchers point out that big data means the data is unable to be handled and processed by most current information systems or methods [1].

The researchers in the field proposed a really good concept (also referred to as the 3Vs) for what constitutes "big" data: volume, velocity, and variety. Volume indicates that the size

of data is enormous. Velocity indicates that the data is being generated quickly. Variety means that the data exists in several formats and that data will be obtained from various origins, etc. [2]. Current research finds that the 3Vs concept is inadequate to describe the current state of big data. To provide some supplemental description of big data, some more V's, i.e., veracity and value are added to it. Veracity indicates the authenticity of information added to make data big. After considering the four V's, there is one more V to consider, i.e., Value! The vast majority of useless data is useless to the organization until it is converted into something useful [3] [4] [5]. Figure 1.1 broadly gives the types of V's that convey the "big" of big data.

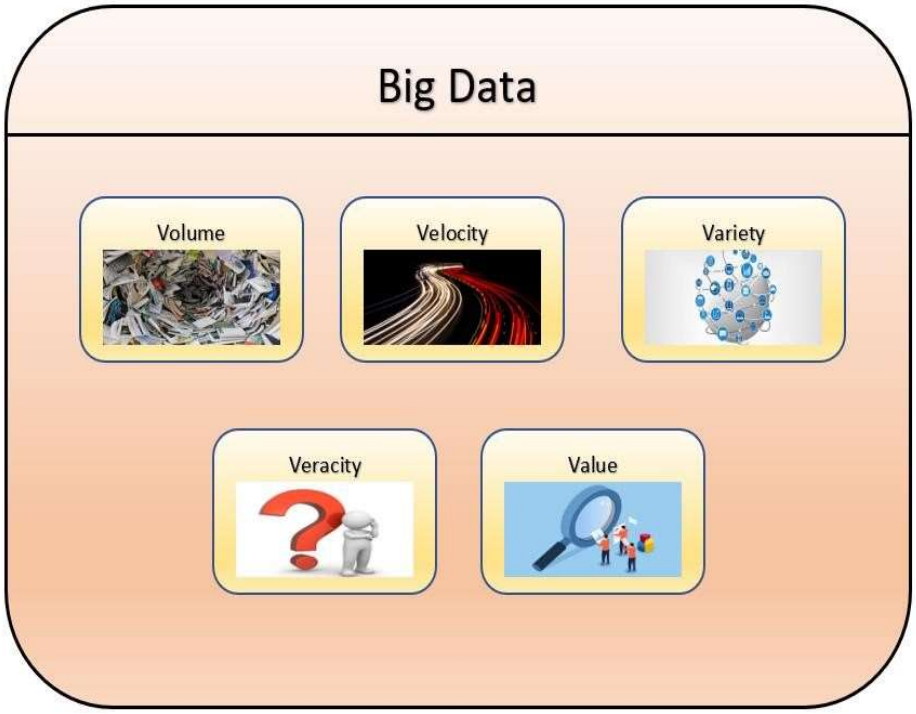

Figure 1.1: Five Vs of Big Data

The following case study gives a glimpse of dark data in the era of big data:

Case study 1: Corporate organizations acquire and keep massive volumes of information about employees and other stakeholders for compliance purposes. Once the employee leaves the company, their data is stored for some years, but after several years, it is eventually thrown into the database; neither is evaluated nor processed. That is dark data in big data related to corporations.

Case study 2: Educational institutions gather all the data from students while admission or registration, such as date of birth, Aadhar id, basic student information, additional information including SSC marks, HSC marks, JEE score, each subject marks etc. All this data is stored in ERP. This data is also required by different accreditation bodies. In accordance with the suggestions of the National Education Policy of 1986, the All-India Council for Technical Education (AICTE) formed an autonomous accreditation committee, the National Board of Accreditation (NBA), and the University Grants Commission (UGC) accreditation committee

is the National Assessment and Accreditation Council (NAAC) [6][7]. When several batches of data are admitted, the previous decade of stored data becomes useless; it goes untapped. No inference is drawn from this data, though it has the potential to draw insights related to students' analytics. All that unexplored big data becomes dark data.

Case Study 3: Utilizing surveillance video footage, enormous volumes of data are collected and warehoused. It is investigated for a period of time in accordance with the criteria and requirements of data. After that, the data is scraped from the database. If such data is analyzed, it can provide valuable information on consumer sentiments, which can be a thought-provoking step in the business world.

Case Study 4: Customer recordings are another massive amount of data stored in telecommunications for quality purposes. It is also processed for a period, but after a decade, the data becomes dead as per its importance but lives in the database. This is according to one group of researchers. Another group states that data can not be dead. It provides helpful insights that can in turn be used to create various applications such as analyzing human behavior, future actions etc.

Figure 1.2 gives the state of "dark" data in "big data" with respect to the three case studies.

Figure 1.2 State of “Dark” Data in “Big” Data

2. Literature Review

Dark data exists due to the enormous amount of data being generated and not utilized. If it is processed and used to get inference from it, then it may have a positive impact on the business domain and improve the value of it. Although very few researchers did research on dark data due to the unavailability of the required datasets, this section outlines what has been done so far with aspects of dark data.

Trajanov et al. outlined the awareness of said possibilities, which are being investigated through the use of dark data in businesses and institutions, by providing a brief summary of the theoretical underpinnings, presenting a methodology, and demonstrating examples of dark information sets in the field of the IoT domain. Data should be accessible as well as available in order to take full advantage of data-driven science and machine learning. A huge amount of data is needed so that it can be fed into a machine learning algorithm. The author suggested two methods for it. The initial method is to design software innovations that gather all of the data needed to train machine learning, then publish the app, gather data for a period of time, and finally use the dataset to train machine learning techniques. The second technique is to identify all existing records, collect them, and then attempt to construct machine learning algorithms depending on the statistics currently collected and gathered [8].

Cafarella et al. organized an honest discourse regarding Dark Data and the right data management issues in coping with it. Schema design, schema retrieval and development, knowledge extraction methods, constant data filtering, data integrity reasoning, and machine learning are all aspects of data management difficulties. The author sheds light on a new logical and reasonable difficulty that the organization must address in order to play a leading and influential role in the area of dark data [9].

Rahul et al. explain the reasons for the importance of dark data to scientific progress, along with some of the features of dark data and the technological challenges of managing it effectively. As per the researcher's study, several potentially valuable institutional and technical solutions are being developed. However, these approaches are primarily theoretical and need further investigation due to a shortage of resources, according to their study [10].

Kevin et al. gave firms advice about how to use the corporate intelligence tools and applications to uncover heretofore unrecognized or neglected data. This is accomplished via analyzing secondary information and professional accounts in depth to gain a better understanding of the many methods and theories required for discovering significant relationships and correlations in an organization's information systems. The author also focused on the strategy of business intelligence. The goal of developing a business intelligence strategy is to assist a company with long-term planning, operational monitoring, and day-to-day deliberation in order to run the company efficiently. The aim of the method is to provide employees with the knowledge that they have to accomplish their respective activities more effectively [11].

Angelo et al. tried to bridge the gaps and provide essential assistance to operational efficiency and productivity by recognizing the existence of dark data in business circumstances, identifying it, and promoting dark data utilization campaigns. The research also gave scientists a way to look into information systems in businesses to make them more creative and efficient. The suggested description and accompanying analysis of dark data in production hybrid features from a review of literature referenced with information gleaned from interviews with industry partners. The source and format of dark data, as well as their worth, were described. Using a balanced approach and combining data from intellectual and industry sources, multiple types of facts were used to learn more about "dark data" and come up with a good way to describe it [12].

Md Ajis et al. suggested Dark Information Management Framework includes one basic core regulatory element, four major procedures, and a few tasks. Because the procedure should indeed be performed continually, not just to acquire new data but also to store the data, the paradigm was offered as a frame for handling dark data.
The suggested model doesn't include dark data mining algorithms because it focuses on safety precautions that can be taken with existing dark data, depending on how each company values its data [13].

Munot et al. discussed dark data, its significance, the threats it poses, and the need for and ways of structuring the unorganized component of it. It also goes over how dark data is used in the banking industry and what percentage of a bank's data is unorganized. The study demonstrated the use of Natural Language Processing technologies such as Named Entity

Recognition and Topic Modelling to detect textual input (natural language) parts and the subject that the text revolves around. The author stumbled upon NER that can be used to identify entities and keywords, and topic modelling could be used to discover hidden trends and purposes [14].

Schembera et al. focused on the importance of dark data in the era of big data. Many big data studies have concentrated on the applications of material relevant to the research problem, omitting treatment of information that is stored on computers but is unknown to research groups. This is known as "dark data," and it is presented and explained with aspects of high-performance computing capabilities. For this purpose, the author presented data from the High-Performance Computing Center Stuttgart, a prominent HPC center in the European region. They also suggested a new role specifically designed to deal with dark data and data management in general. The role of scientific data officer is to distinguish itself among common HPC positions like chief data officers, system administrators, and security officers [15].

According to the above literature, very few academics have conducted research on dark data. Many people have used big data, but only on a subset of it. Concepts are evolving theoretically, and little effort needs to be made in practical regards. It is critical to make logical efforts to discover hidden patterns in dark data.to the above literature, very few academics have conducted research on dark data.

3. Approach to convert unstructured data into structured data

Dark data exists in an unstructured and unorganized format. It is very important to convert the unstructured and unorganized data into a structured and organized one. Structured data, when analyzed with techniques of natural language processing, data analytics, and data mining, can produce valuable hidden patterns and trends that can help stakeholders in the business domain as well.

Following steps convert unstructured data into structured and then it can be fed into a model to get insights in the form of graphs, charts, and wordcloud. As it is said, "pictures convey louder messages than text messages."

Step1: Import the dataset:

Once the problem is defined, the selection of a dataset is very important. Once it is chosen, it should be loaded into the programming environment. Once it is loaded, the shape of the dataset can be explored with the description of every attribute or feature in that particular dataset.

Step 2: Identify the missing values from dataset

Null or missing values do not have much importance in dataset analysis. They also unnecessarily create numerical confusion. If there are a greater number of null values, then it is necessary to handle them.

Step 3: Remove the null values/ fill it with some method of filling

So, once it is identified, it can be either deleted or filled with the mean value, previous or next data item value.

Step 4: Remove the punctuation

As big data is massive data, it may consist of punctuation such as full stops (.), commas (,), semi-colon (,), exclamatory marks (!), and question marks (?), which is not desirable for processing purposes. So, it can be removed with the help of the language processing function.

Step 5: Remove the emojis, special symbols

Data is generated on the web, mostly in today's era of social media. Data consists of emojis and special symbols, which are also not very helpful while drawing inferences, so they can be removed.

Step 6: Remove the HTML, web-based URL

If we consider the dark data related to social media, then it may consist of URLs or html tags that need to be ignored while processing.

Steps 4, 5, and 6 are implemented with the way of only keeping characters from a to z/A to Z/numbers from 0-9 except the other characters or symbols.

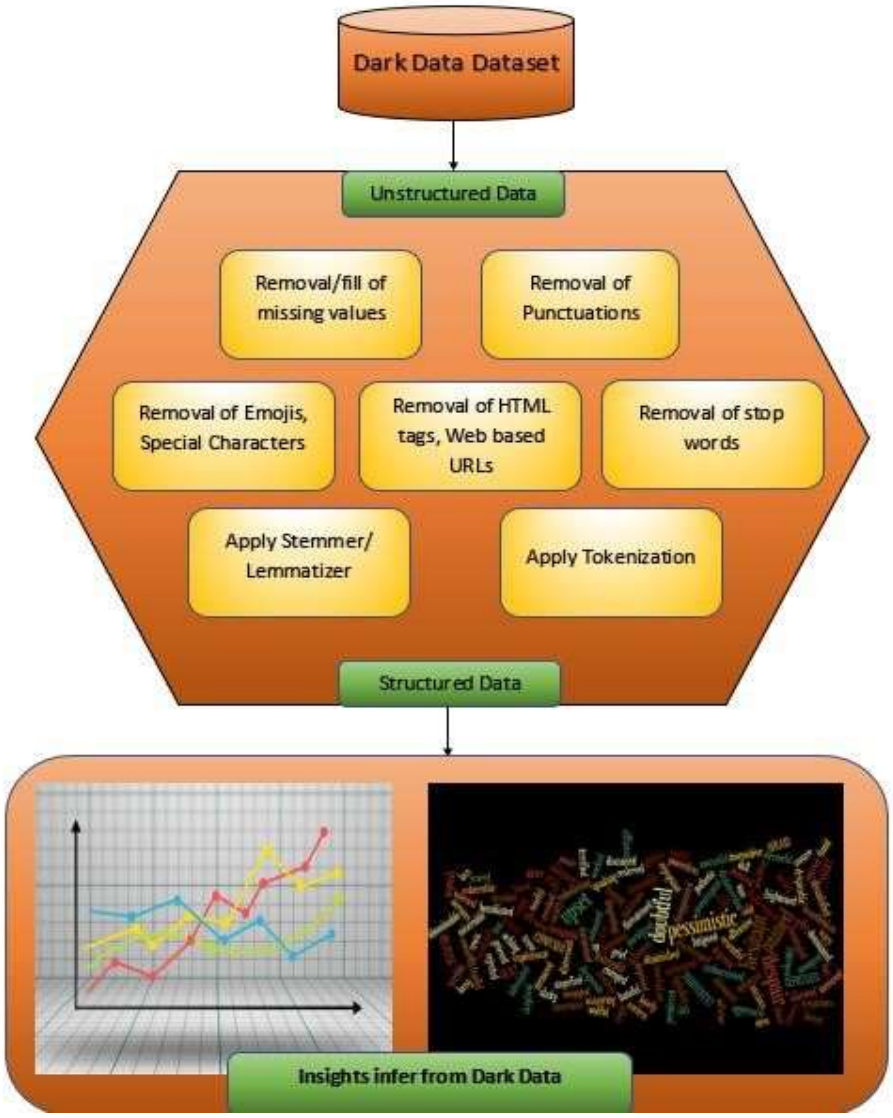

Figure 3.1 Architecture of deriving insights from Dark Data

Figure 3.1 illustrates detailed architecture of deriving insights (unknown or hidden) patterns from dark data.

Step 7: Remove the stop words
Stop words such as "am, is, are, of, the, etc." can be removed to save the computational space for analyzing the dark data.

Step 8: Apply the tokenization
Rather than analysing a complete paragraph or sentence at once, tokenization can be applied. Sentence tokenize splits the sentences into paragraphs and word tokenize splits the words in the sentences. The output of this step is tokenized words, which in turn provide a feature for generating visualizations.

Step 9: Apply the stemmer or lemmatizer
Tokenized words can be converted into simpler forms by applying a stemmer or lemmatizer. The stemmer generates the stem word from inflected words, whereas the lemmatizer infers lemma from actual words.

Step 10: Fed above step output features into model to understand analytics in view of creating visualization

Features extracted are now fed into the model, which helps to create charts and graphs using visualization libraries and tools that derive insights from them. Visualization analytics help to identify hidden patterns from dark data. Dark data is unused data and manually extracting insights is not possible for large amounts of data. Artificial Intelligence (AI) based techniques help to extract meaningful patterns from it.

4. Mathematical Model for Analyzing Dark Data

To form the mathematical model for analyzing dark data in ocean of big data, there are various variables (properties of dark data) that converts analyzing into concluding remarks.

$$\int_0^{\infty} (x1, x2 \ldots . . xn)\, dx = [\mathrm{C}]_n \qquad (Equation\ 4.1)$$

In Equation 4.1, x1, x2 … xn denotes data variables whereas C is concluding remark. Here, iterating over data from inception of data to the infinity (last maxima of the data), and after careful investigation reaching to conclusion.

5. Examine Dark Data using Natural Language Processing

Agora's data set is considered here to comprehend the hidden pattern in dark data. This is a data parsing of market information that was taken from Agora, a dark web marketplace, over the time frame between one particular year and the next. Drugs, guns, literature, services, and more are all included. Duplicate listings have been deleted, and prices for any remaining ones have been averaged. There are nearly 100,000 unique listings in the csv file that contains all of the data [16]. The majority of transactions on darknet markets feature illegal goods since Agora has a perilous aspect. The analysis is not meant to promote irresponsible darkweb

browsing or whatsoever way advocate criminal activity. Instead, it seeks to analyse one of these market platforms that is the largest in order to determine its traits and actions.

Figure 5.1 gives screenshot of dataset considered, 5.2 gives detailed information consisting column values of dataset.

	Vendor	Category	Item	Item Description	Price	Origin	Destination	Rating	Remarks
0	CheapPayTV	Services/Hacking	12 Month HuluPlus gift Code	12-Month HuluPlus Codes for $25. They are wort...	0.05027025666666667 BTC	Torland	NaN	4.96/5	NaN
1	CheapPayTV	Services/Hacking	Pay TV Sky UK Sky Germany HD TV and much mor...	Hi we offer a World Wide CCcam Service for En...	0.152419585 BTC	Torland	NaN	4.96/5	NaN
2	KryptykOG	Services/Hacking	OFFICIAL Account Creator Extreme 4.2	Tagged Submission Fix Bebo Submission Fix Adju...	0.007000000000000005 BTC	Torland	NaN	4.93/5	NaN
3	cyberzen	Services/Hacking	VPN > TOR > SOCK TUTORIAL	How to setup a VPN > TOR > SOCK super safe enc...	0.019016783532494728 BTC	NaN	NaN	4.89/5	NaN
4	businessdude	Services/Hacking	Facebook hacking guide	. This guide will teach you how to hack Faceb...	0.062018073963963936 BTC	Torland	NaN	4.88/5	NaN
5	Hackyboy	Services/Hacking	DDOS ATTACK SERVICE	New service available : Take down all websites...	0.030109432452830192 BTC	Torland	NaN	4.92/5	NaN
6	businessdude	Services/Hacking	ATM HACKING TUTORIAL	Step by Step guide and Manuals to hack ATMs ...	0.03317625980769233 BTC	NaN	NaN	4.88/5	NaN
7	CheapPayTV	Services/Hacking	CALL/SMS verification service!	Need to register an account with sms verificat...	0.012616985714285715 BTC	Torland	NaN	4.96/5	NaN
8	toysoldiers	Services/Hacking	Mac & windows address changer	- Comes with complete database to randomly cho...	0.009335870999999999 BTC	Torland	NaN	4.94/5	NaN
9	cyberzen	Services/Hacking	WiFi Hacking	-Hacking WEP/WPA/WPA2 the glory of WPS -Hacki...	0.027091319209809266 BTC	NaN	NaN	4.89/5	NaN

Figure 5.1 Screenshot of Dataset

```
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Vendor            109689 non-null  object
 1    Category         109689 non-null  object
 2    Item             109687 non-null  object
 3    Item Description 109662 non-null  object
 4    Price            109684 non-null  object
 5    Origin           99807 non-null   object
 6    Destination      60528 non-null   object
 7    Rating           109674 non-null  object
 8    Remarks          12616 non-null   object
dtypes: object(9)
memory usage: 7.5+ MB
```

Figure 5.2 Detailed Information of Dataset

Details regarding the item's sources and destinations are shown in the following visualization 5.3. According to the dataset, the majority of dealers claim to be from the United States, and the majority of dealers have access to international distribution. The following figure 5.4 shows the boxplot representing insights from rating. It extracts meaningful information related to reviews. The figures 5.4 conveyed those positive reviews are very common.

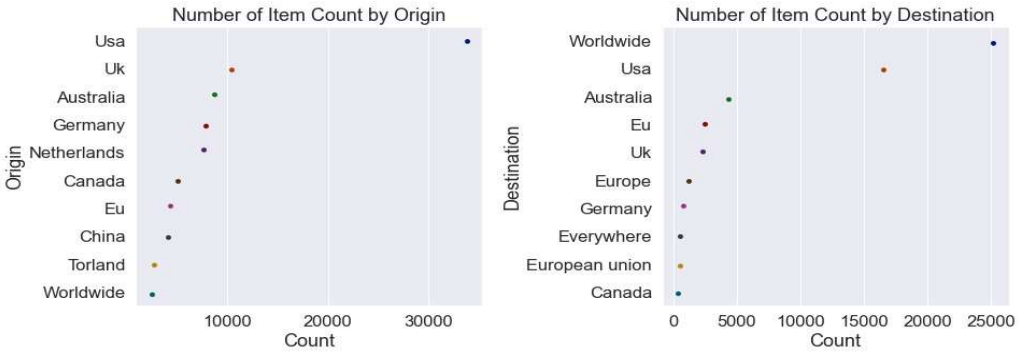

Figure 5.3 Number of item count by origin and destination

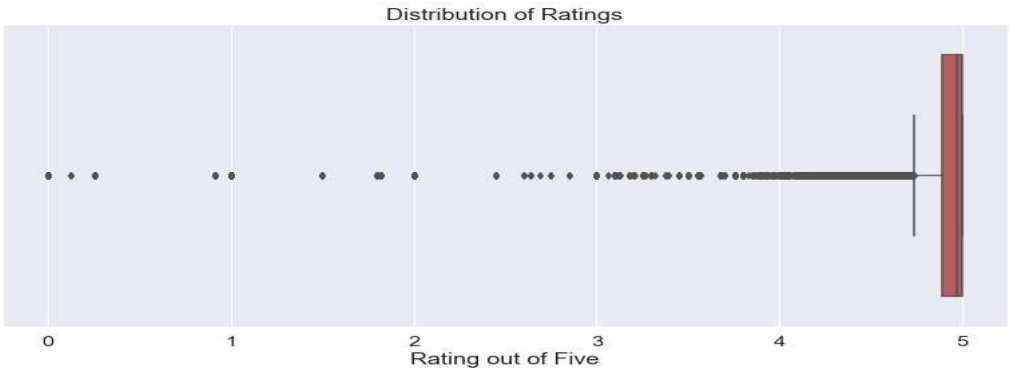

Figure 5.4 Insights from Rating variable

The graph 5.5 demonstrates how a certain individual dominates particular classes of illegal narcotics.

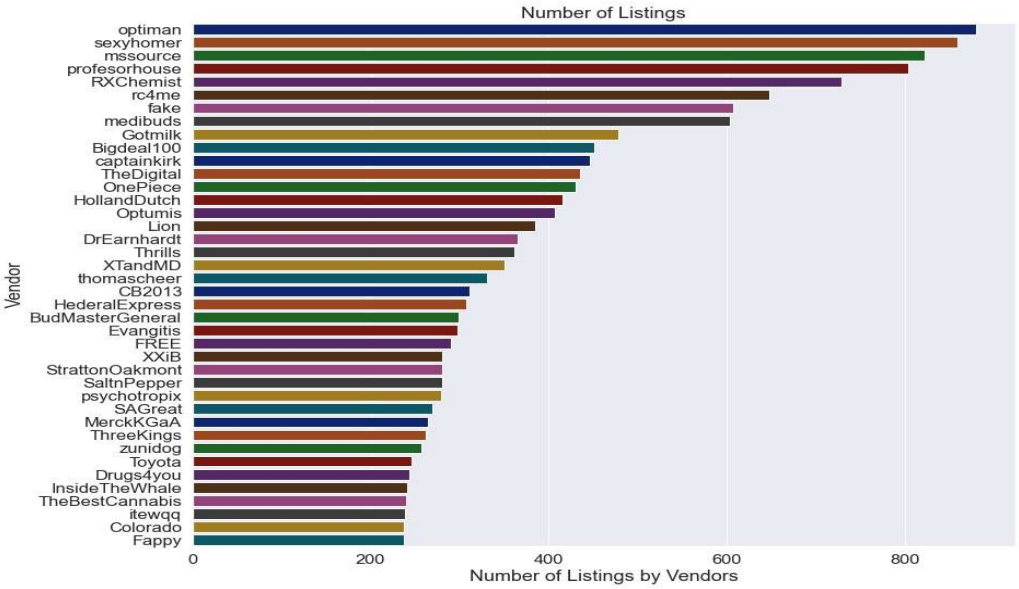

Figure 5.5 Vendor Listing

6. Conclusion and Future scope

Dark data refers to unstructured data that is not evaluated. That's the data online storage and in data warehouses as a consequence of multiple network activities that either sits around enough to fulfill the company's statute of limitations or is preserved since data warehousing is so inexpensive. Dark data produces a lot of potential to get useful insights into hidden patterns or trends that can help business to grow. This paper discussed the different case studies about how dark the data in in "big" data. This paper also suggested approach to convert unstructured data into structure data as structured data helps to derive the insights and meaningful patterns. Further aims to create a working prototype of suggested approach that can be used to generate inferences from dark data.

REFERENCES

[1] D.Fisher, R. Deline, M. Czerwinski, S. Drucker, "Interactions with big data analytics" *Interactions*. 2012

[2] D. Laney, "3D data management: controlling data volume, velocity, and variety", *Tech. Rep.* 2001. [Online].

Available:http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-ManagementControlling-Data-Volume-Velocity-and Variety.pdf

[3] Van Rijmenam, "Why the 3v's are not sufficient to describe big data, BigData Startups", *Tech. Rep.* 2013. [Online]. Available: http://www.bigdata-startups.com/3vs-sufficientdescribe-big-data/.

[4] K. Borne, "Top 10 big data challenges a serious look at 10 big data v's", *Tech. Rep.* 2014. [Online]. Available: https:// www.mapr.com/blog/top-10-big-data-challenges-look-10-bigdata-v

[5] Chun-Wei Tsai , Chin-Feng Lai , Han-Chieh Chao and Athanasios V. Vasilakos, "Big data analytics: a survey", *Journal of Big Data*, 2015

[6] S. Kaul, *Higher education in India: Seizing the opportunity, Indian Council for Research on International Economic Relations (ICRIER)*, 2006

[7] A. Stella, "External quality assurance in Indian higher education: Developments of a decade" *Quality in Higher Education*, 10(2), pp 115-127, 2020

[8] Dimitar Trajanov, Vladimir Zdraveski, Riste Stojanov and Ljupco Kocarev, "Dark Data in Internet of Things (IoT): Challenges and Opportunities" *in Proceedings of the 7th Small Systems Simulation Symposium Serbia*, 12th-14th February 2018

[9] M. Cafarella, I. F. Ilyas, M. Kornacker, T. Kraska and C. Ré, "Dark Data: Are we solving the right problems?," *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pp. 1444-1445, 2016,

[10] Rahul P & Ganeshan M "Extract the Analyzed Information from Dark Data" *Published in International Journal of Trend in Scientific Research and Development (ijtsrd)*, ISSN: 2456- 6470, Volume-4, Issue-4, pp 26-29, June 2020

[11] N.M.Kevin & J.K.Ngatia, "Dark data: Business Analytical tools and Facilities for illuminating dark data", *Scientific Research Journal (SCIRJ),* Volume IV, Issue IV, April 2016,

[12] A. Corallo, A. M. Crespino, V. D. Vecchio, M. Lazoi and M. Marra, "Understanding and Defining Dark Data for the Manufacturing Industry," in *IEEE Transactions on Engineering Management*, doi: 10.1109/TEM.2021.3051981.

[13] A. F. Md Ajis and S. Hajar Baharin, "Dark Data Management as frontier of Information Governance," *2019 IEEE 9th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, 2019, pp. 34-37, doi: 10.1109/ISCAIE.2019.8743915

[14] K. Munot, N. Mehta, S. Mishra and B. Khanna, "Importance of Dark Data and its Applications," *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, 2019, pp. 1-6, doi: 10.1109/ICSCAN.2019.8878789.

[15] B. Schembera, J.M. Durán, "Dark Data as the New Challenge for Big Data Science and the Introduction of the Scientific Data Officer" *Philos. Technol.* ,Volume 33, pp 93–115, 2020

[16] Dataset extracted from https://www.kaggle.com/datasets/philipjames11/dark-netmarketplace-drug-data-agora-20142015