

Analysis of dark web pages using GoW and Topic Modelling



Advisors:

Prof. Sunil Bhirud

Prof. Ashwini Dalvi

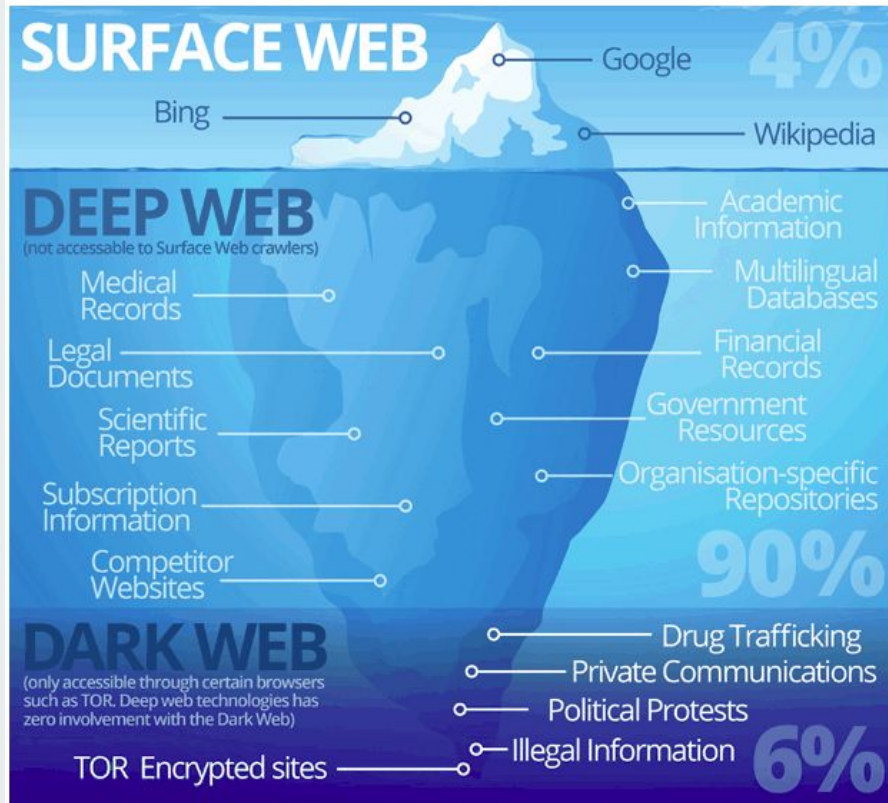
Team:

- Saurabh Raut - 171070005
- Nirmitt Joshi - 171070010
- Dhairya Bhuta - 171070011
- Saikumar Nalla - 171070030

Quick Overview

- Scraping dark web pages using custom crawler
- Data Cleaning
- GoW vs BoW
- Comparing GoWs with word cloud
- Extracting keyphrases from GoWs
- Latent Dirichlet Allocation (Topic Modelling)
- Results and Analysis
- Conclusion

Dark Web



Introduction & Problem

Darkweb is a platform to carry out:

- Drug Rackets
- Contract Killing
- Political Protests

Exponential growth in Darkweb data.

Text analysis of the content of these dark webpages will impact many domains:

- Cybercrime Prevention
- Detective Agencies
- Expose Political Blunders....

Dataset collection by scraping the dark web pages

- Used custom dark web crawler to successfully extract **44,407** HTML web-pages
- Data processing to perform meaningful analysis
- Data comprised of both: English as well as Non-English web pages
- Languages other than English :- Japanese, Spanish, Russian

Dataset Cleaning



Steps Involved:

1. In order to perform textual analysis on dark-web content, we extract only the webpage contents by removing HTML tags and Javascript code
2. As this analysis was confined to English language, we had to remove Non-English web-pages
3. Removing Facebook posts

Evolution of the dataset

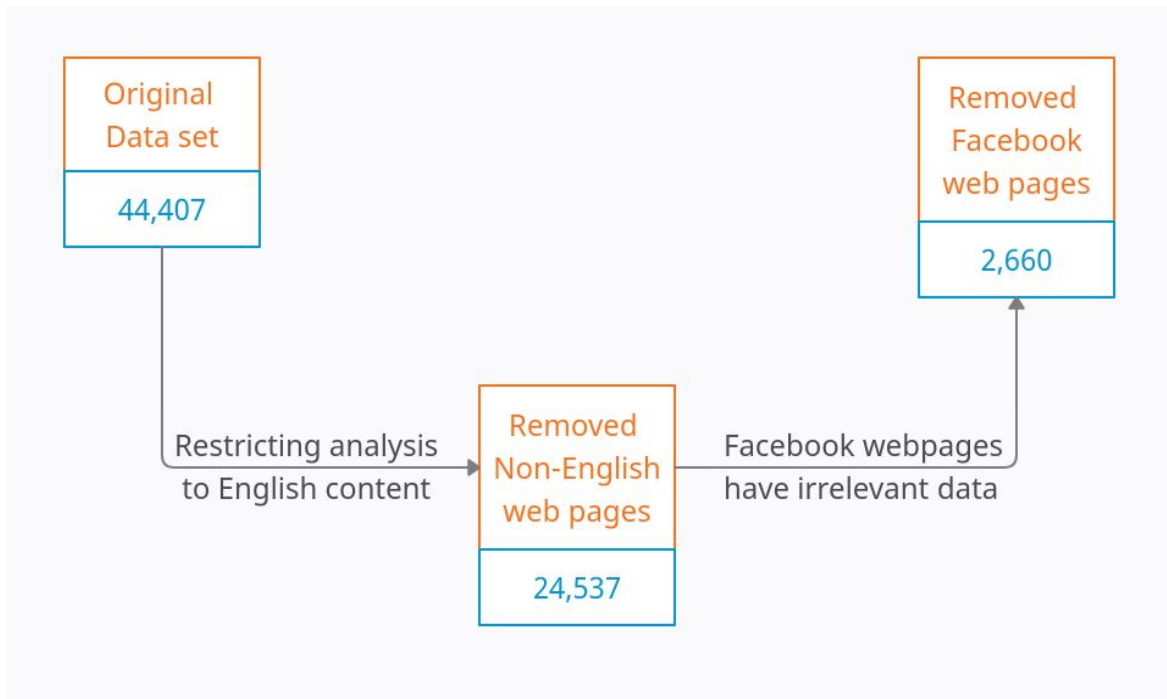
Original document

```
<html>
  <head>
    <title>Buy Paypal accounts</title>
  </head>
  <body>
    <h1>Reliable and Secure</h1>
    <p>Best place to buy PayPal accounts with balance $100, $1000, etc</p>
    <h2>Click here to know more</h2>
  </body>
</html>
```

After removing HTML tags

Buy Paypal accounts
Reliable and Secure
Best place to buy PayPal accounts with balance \$100, \$1000, etc
Click here to know more

Evolution of the dataset



Dataset Properties

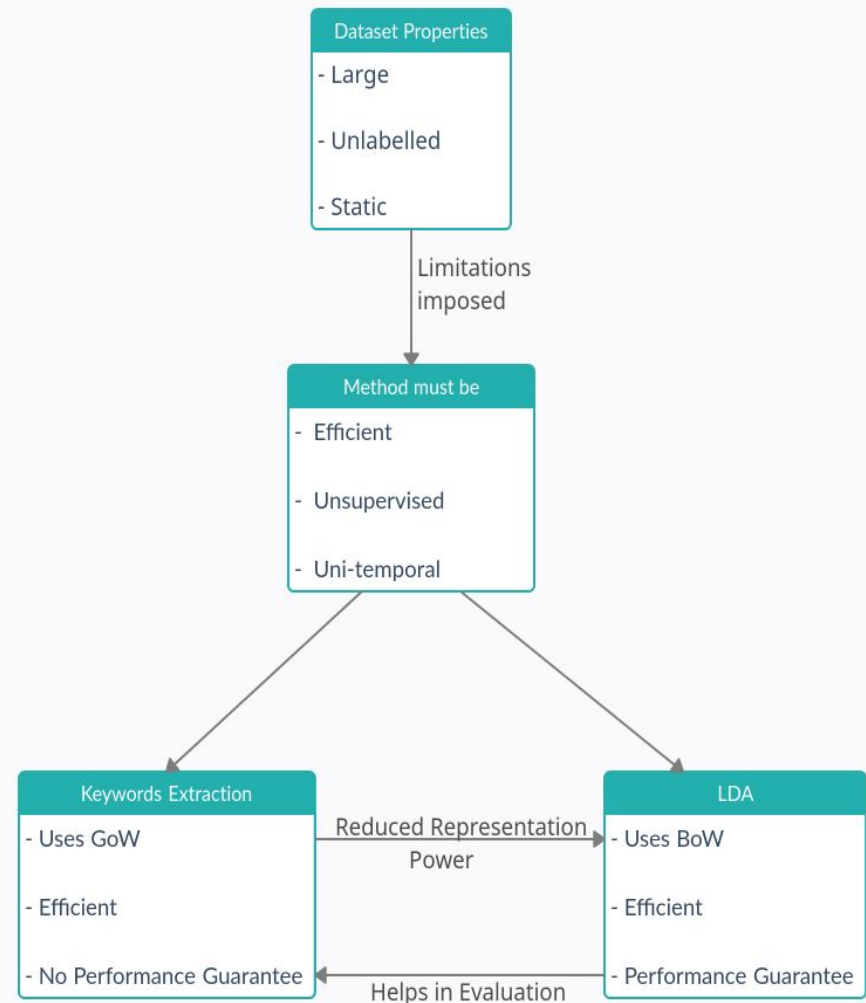
- **Large:** Restricts the use of models having polynomial or greater than linear time complexity
- **Unlabelled:** No scope for supervised machine learning models
- **Static:** Temporal analysis of data cannot be performed

Bag of Words (BoW) :

- BoW describes the occurrence of words within a document
- Does not capture any order/structure amongst the words in the document

Graph of Words (GoW) :

- Nodes are words
- Edges represent co-occurrence of words within a specific window
- Graph properties capture the relationship between words



Developing the GoW Construction Algorithm:



Primitive algorithm:

- Consider all words of document for GoW Construction
- Drawback :- Dense GoWs generated

Improvement:

- Include words having frequency greater than 2
- Intuition :- Words occurring more than once are significant

Constructing GoWs for analysing the documents



GoW construction algorithm:

Input: Document D

Output: Graph of words G

Procedure:

```
wordList = D.split()
```

```
for word in wordList:
```

```
    if len(word) >= 3:
```

```
        don't remove word
```

```
endfor
```

```
G = DiGraph()
```

```
for word in wordList:
```

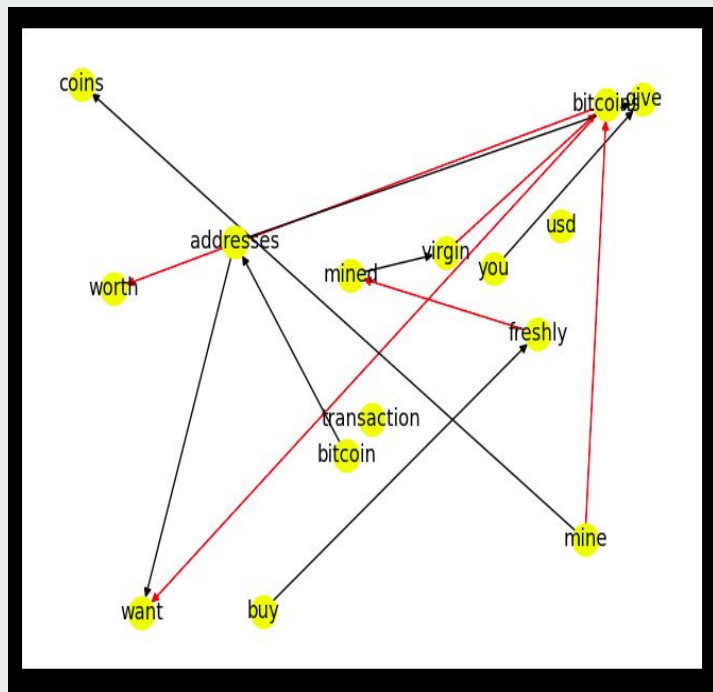
```
    G.addNode(word)
```

```
    G.addEdge(word, nextWord)
```

```
endfor
```

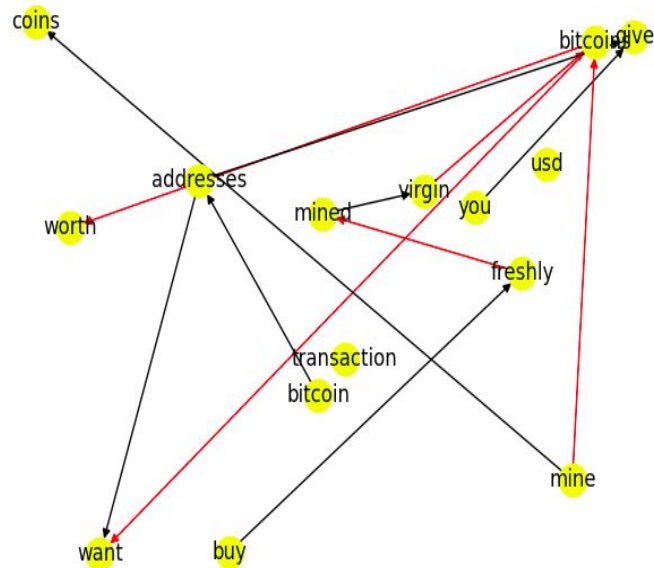
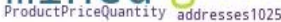
Remove all nodes (words) that occur less than 2 times in G and also the corresponding edges

Sample GoW for a document

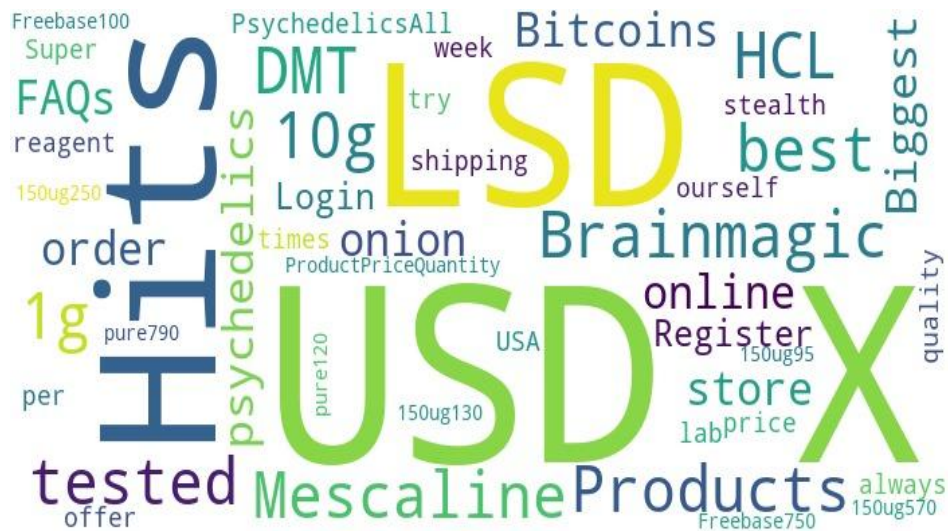
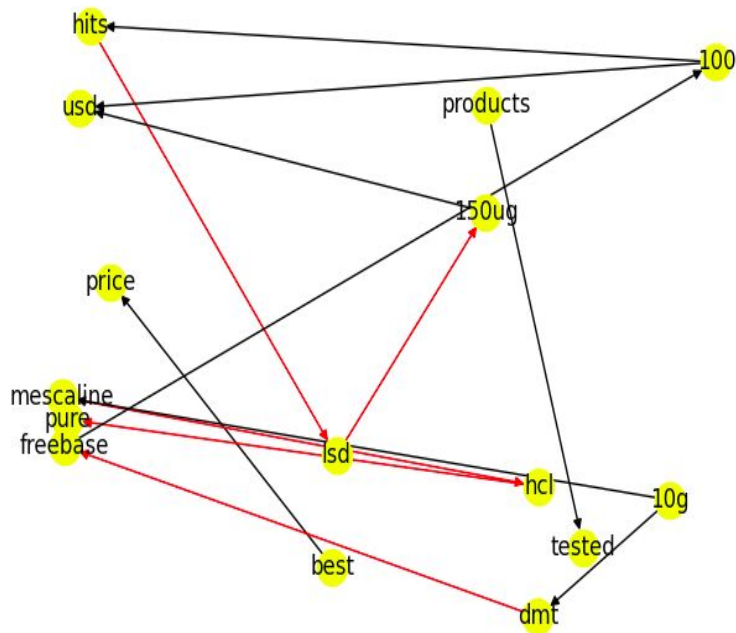


The adjoining GoW is for a web-page that is related to selling bitcoins (marketplace) and the same can be understood by analysing the graph.

Note: The red edges represent that these pair of words occur more than once in the document together and are thus of more importance than other edges.



Comparing GoWs and Word Clouds



Keywords Extraction using Graph-of-Words

- Unsupervised
- Efficient

Intuition



- Central nodes make good keywords.
- Nodes with **high centrality** in GoW representations usually correspond to the keywords that a human would pick for the documents.
- To conclude, we need to understand some **Centrality Measures**.

Centrality Measures



- **Closeness:** The closeness $C_C(v)$ of a node v is defined as the inverse of the total distances from every other node to v in the graph.

$$C_C(v) = \frac{1}{\sum_{u \neq v} d(u, v)}$$

- **Betweenness:** The betweenness $C_B(v)$ of a node v is defined as the fraction of shortest paths from all vertices (except v) to all others (except v) that pass through v .

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Graph Degeneracy

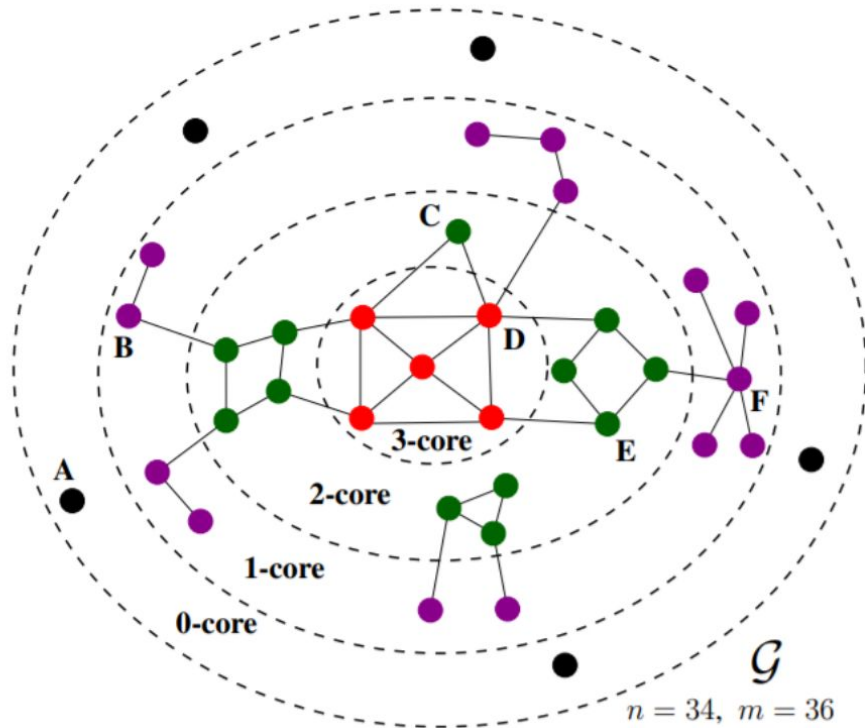


- Communities of central nodes make better keywords.

Definition

A subgraph $H = (V_H, E_H)$, induced by the subset of vertices $V_H \subseteq V$, is called a k -core or a core of order k **if and only if** $\forall v \in V_H, \deg_H(v) \geq k$ and H is the **maximal subgraph** with this property. I.e. it cannot be augmented without losing this property.

Explanation with figure



Algorithm for learning Corewords



- ① For each Graph-of-Words representation:
 - Find the k -core with less than 30 words.
 - This is a “rich” community words.
 - Now, using **closeness** centrality measure, find the best four vertices with the maximum closeness $C_C(.)$.
 - Return the vertices correspond to those vertices.

Time Complexity



- A linear $O(n + m)$ time algorithm for the k -core decomposition.
- The main idea is to remove the vertex of lowest degree (in the remaining subgraph) at each step and decrease the degree of its adjacent neighbors by one.
- The vertices are initially sorted in linear time using bin sort since there are at most $\Delta(G) + 1$ distinct values for the degrees and $\Delta(G) < n$.
- Closeness can be computed in $O(1)$ time on a graph of 30 vertices.
-

$$\text{Time Complexity} = O(n + m)$$

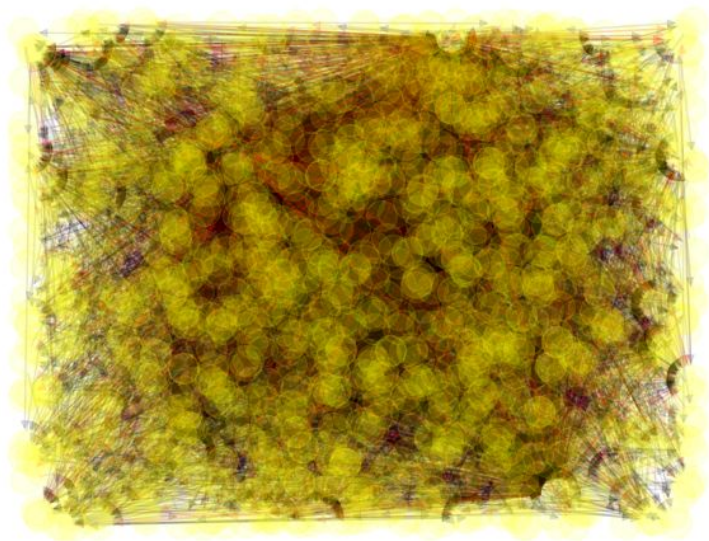
Results



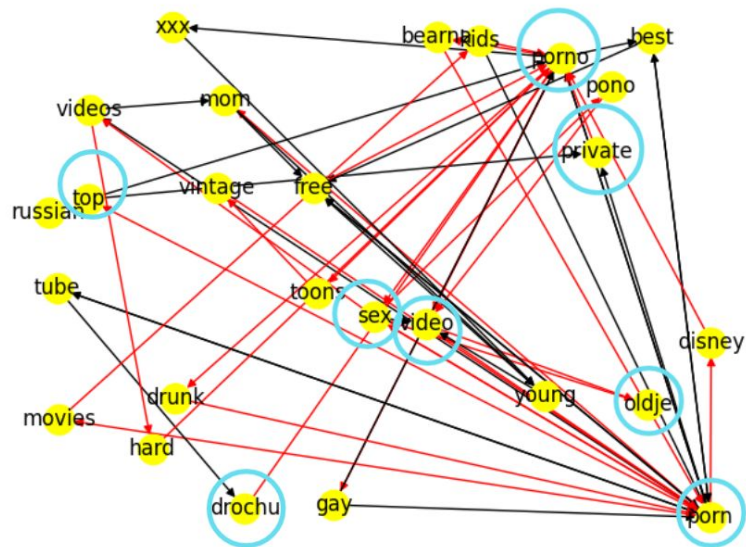
- ① Document 1: **Keywords:** jasmine sex, girls, fun camera, school adult
- ② Document 2: **Keywords:** tls, start-tls, ssl, authentication
- ③ Document 3: **Keywords:** cash atm, cloned, fast shipping, cards

Example:

```
[('drochu porno breana porn', 0.07903937347253942), ('private porno', 0.06769265264163632), ('video oldje sex porn top', 0.0377745788037254), ('marga porn', 0.03633337329545605)]
```



Original GoW



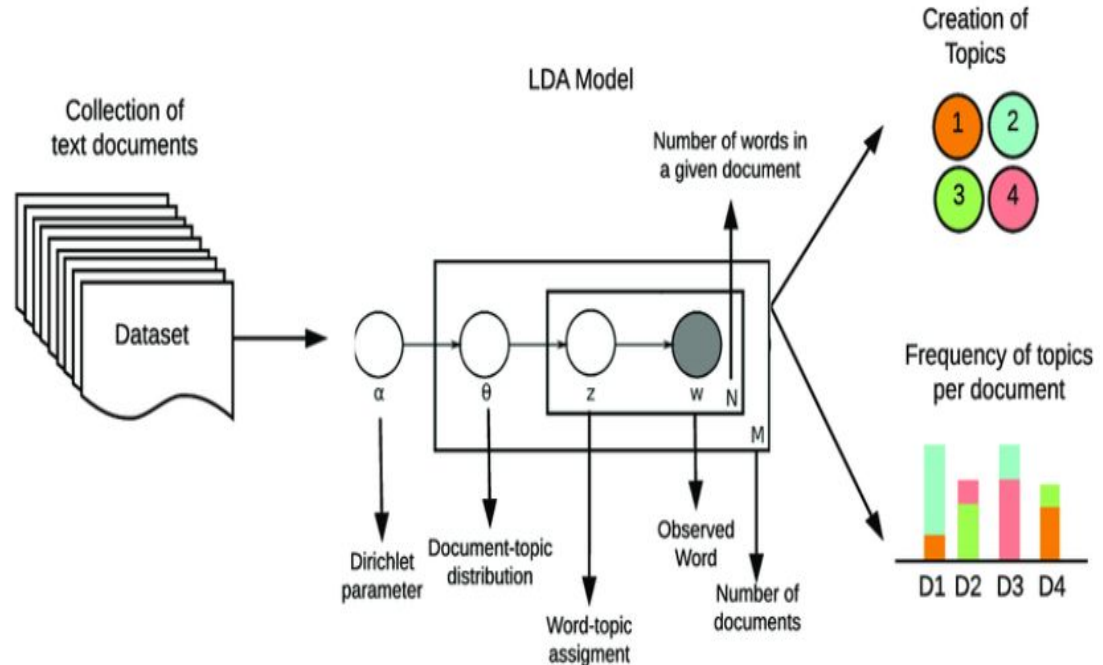
K-Core Graph with keywords highlighted

Topic Modelling: Latent Dirichlet Allocation (LDA)

- Unsupervised
- Uses Bag-of-Words

Latent Dirichlet Allocation (LDA)

- **Topic modeling** is a type of statistical modeling
- **Latent Dirichlet Allocation (LDA)** is an example of topic model.
- Generative probabilistic model
- Builds a topic per document model and words per topic model



Processing the input for LDA

studying
studies
study

Lemmatization

study
study
study

Lemmatization of the words studying, studies, and study

Lucy said that the car's engine was running all night



Lemmatization



Lucy say that the car engine be run all night

- Removed punctuation marks
- Lemmatization on the documents
- The goal of lemmatization was to break down the word into its simplest form.

Dictionary & document term matrix



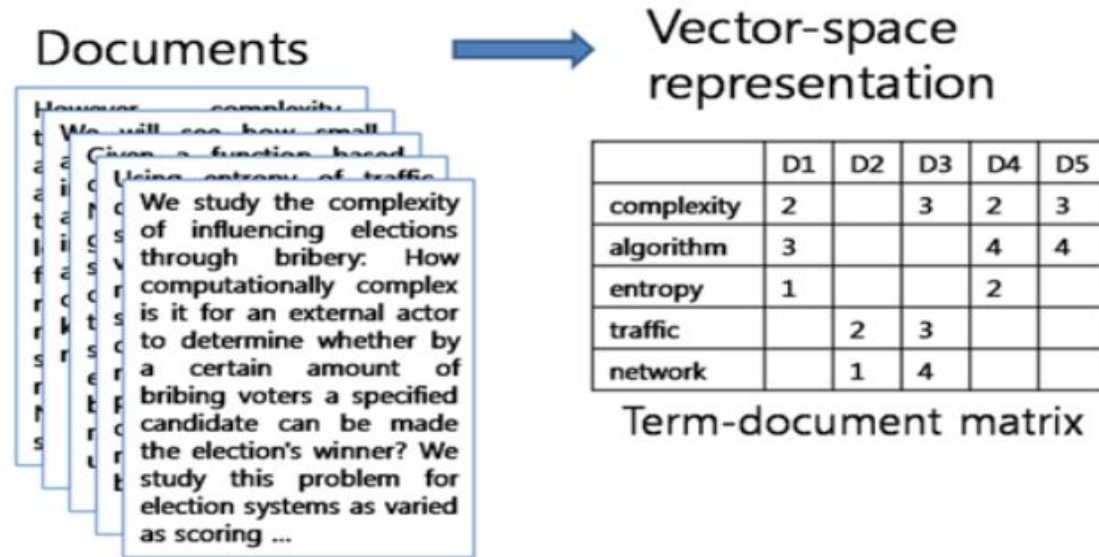
To generate an LDA model, we need to understand how frequently each term occurs within each document.

```
from gensim import corpora, models  
  
dictionary = corpora.Dictionary(texts)
```

The Dictionary() function traverses texts, assigning a unique integer id to each unique token while also collecting word counts and relevant statistics.

```
corpus = [dictionary.doc2bow(text) for text in texts]
```

Document term matrix (Bag of Words)



- Corpus, is a list of vectors equal to the number of documents.
- In each document vector is a series of tuples.

```
>>> print(corpus[0])  
[(0, 2), (1, 1), (2, 2), (3, 2), (4, 1), (5, 1)]
```

Gensim implementation of LDA

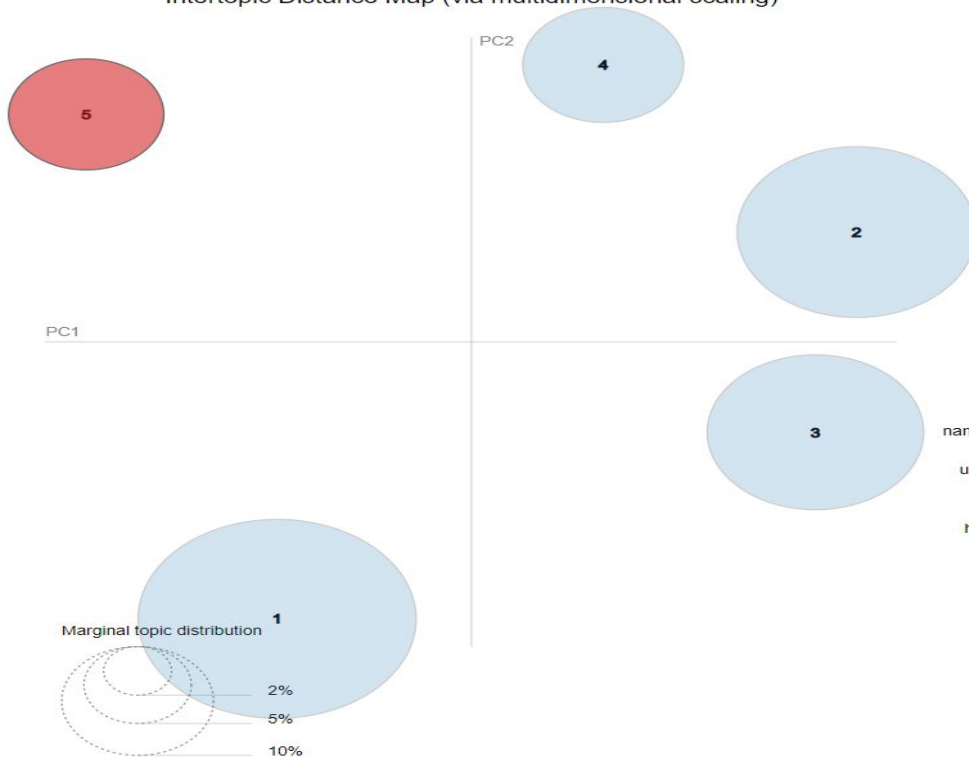
```
▶ # Creating the object for LDA model using gensim library
LDA = gensim.models.ldamodel.LdaModel

# Build LDA model
lda_model = LDA(corpus=doc_term_matrix, id2word=dictionary, num_topics=5, chunksize=5)
```

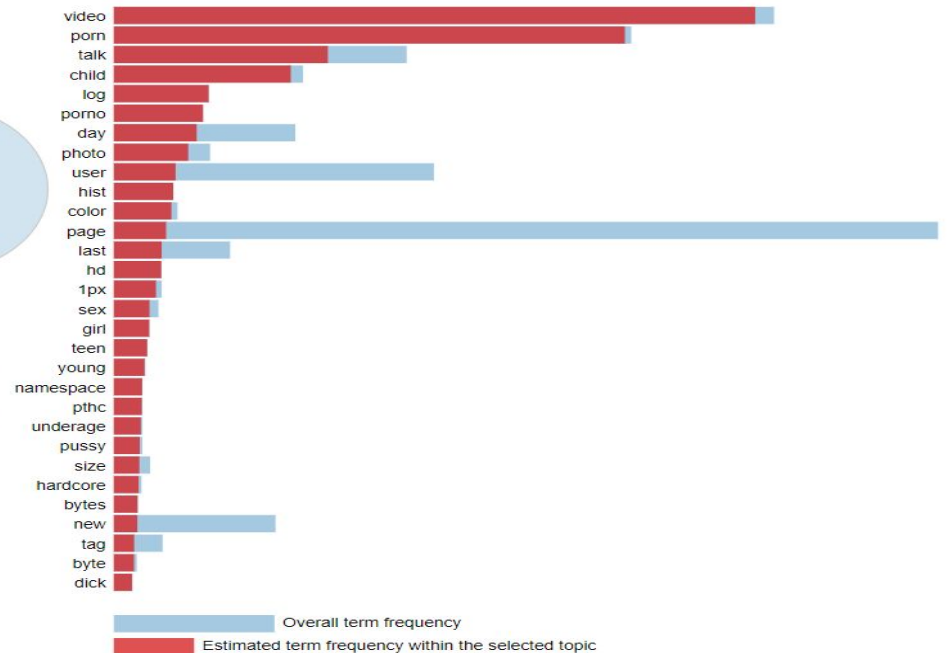
Parameters:

- **id2word:** (required) The LdaModel class requires dictionary to map ids to strings.
- **corpus:** (required) Stream of document vectors or sparse matrix of shape
- **num_topics:** (required) An LDA model requires the user to determine how many topics should be generated. Analysing the dataset and trying different values, we figured out that we should set this to 5.

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 5 (10.4% of tokens)



1. $sallency(\text{term } w) = \text{frequency}(w) * [\sum_t p(t | w) * \log(p(t | w)/p(t))]$ for topics t ; see Chuang et. al (2012)
 2. $relevance(\text{term } w | \text{topic } t) = \lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Sievert & Shirley (2014)

```
lda_model.print_topics()
```

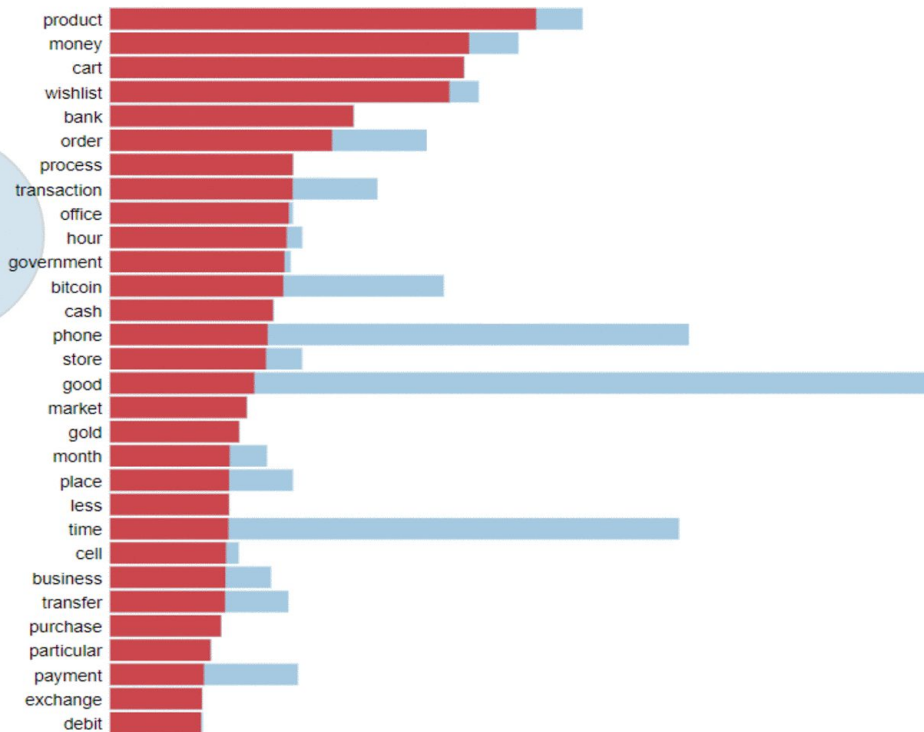
```
[ (0,
  '0.012*"people" + 0.009*"many" + 0.008*"way" + 0.008*"right" + 0.008*"person" + 0.008*"real" + 0.007*"possible" + 0.007*"thing" + 0.007*"year" + 0.007*"phone"'),
  (1,
  '0.044*"page" + 0.016*"search" + 0.016*"tool" + 0.016*"site" + 0.015*"category" + 0.015*"name" + 0.014*"file" + 0.013*"user" + 0.012*"action" + 0.012*"good"'),
  (2,
  '0.017*"card" + 0.013*"web" + 0.012*"use" + 0.012*"-" + 0.012*"key" + 0.011*"server" + 0.010*"service" + 0.010*"password" + 0.009*"network" + 0.009*"email"'),
  (3,
  '0.023*"product" + 0.019*"money" + 0.019*"cart" + 0.018*"wishlist" + 0.013*"bank" + 0.012*"order" + 0.010*"process" + 0.010*"transaction" + 0.010*"office" + 0.010*"hour"'),
  (4,
  '0.116*"video" + 0.093*"porn" + 0.039*"talk" + 0.032*"child" + 0.017*"log" + 0.016*"porno" + 0.015*"day" + 0.013*"photo" + 0.011*"user" + 0.011*"hist"') ]
```

Visualization of the LDA model output

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 4 (11.2% of tokens)



Overall term frequency

Estimated term frequency within the selected topic

1. $\text{saliency}(\text{term } w) = \text{frequency}(w) * [\sum_t p(t | w) * \log(p(t | w)/p(t))]$ for topics t ; see Chuang et. al (2012)

2. $\text{relevance}(\text{term } w | \text{topic } t) = \lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Sievert & Shirley (2014)

Labelling documents with topics using LDA

- Documents are given as input to the LDA model and the probability of that document being in each topic is being calculated.
- The topic with highest probability is assigned to that document

Documents

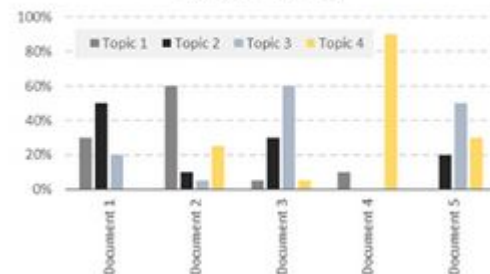


LDA

	Weight	Words
Topic 1	3%	Page
	2%	Privacy
	1%	File

Topic 3	2%	Server
	1%	Network
	1%	Password

Topics allocation to documents



Topics decided for labelling the documents



After analysing the nature of our dataset, we came up with the following topics for labelling the documents:

1. Hidden Wiki
2. Marketplace
3. Porn
4. Software
5. Others

Hidden Wiki: This comprises a major portion of our dataset and is similar to the wikipedia of the surface web as it acts as a storehouse for onion links of different services

Marketplace: This consists of all webpages selling different products such as drugs, paypal accounts, passports, unregistered guns, bitcoins, etc

Porn: This includes different porn related web-pages

Software: This comprises of blogs, articles, discussions related to security, software updates, etc

Others: This covers every other topic not included in the above categories

Application of LDA Model on a sample document

[27] documents[76]

```
"Little whore CP porn Home Login 400 GiB amateur child porn video 1985 2019 video photo shooting The latest update: may 2019 The main cp child porn baby The main section website mode  
rn photos videos CP Retro cp child porn loli 1970s/1980s photos video clips CP LOGIN Video cp child porn pedo Amateur video clips excellent quality CP Barbara Wolfgang CP porn One man  
y video files published portal Amateur video Miranda Leo CP PORN One many video files published portal Amateur video Emma Thomas CP PORN One many video files published portal Home vid  
eo Melanie Aaron CP PORN One many video files published portal Home video LOGIN Many videos... Get access Send Bitcoins specific address get access photo video files To Bitcoin addres  
s: 1MGV475mzJUF1AMTJbpqx6PDdZvScBmaoi Pay amount: 0.004 Instructions As soon receive payment server send small amount BTC Wallet. This small amount used password. For example receive  
d BTC 0.0002115 password would 2115 (password must least 4 characters). Your username automatically generated login page . This widely used method. Our server identify BTC address imm  
ediately send small amount back wallet. The whole process fully automated. *The password sent automatically three payment confirmations. News I love brother CP porn child Added Admin  
Mar. 11 2019 Family BDSM CP porn child Added Admin Mar. 29 2019 Rohan Killawala (Indian 16 y.o. 8 y.o.) CP porn child Added Admin Apr. 04 2019 Little friends CP porn child Added Admin  
Apr. 19 2019 First time sister CP porn child Added Admin May. 03 2019 Outdoor sex CP porn child Added Admin May. 18 2019 LOGIN Contact us If want upload content website call email adu  
lt-webcp@secmail.pro If questions regarding website access fee call email pay-webcp@secmail.pro Copyright 2017-2019 little whore | Child Porn DarkNet "<
```



```
print(tokenized_documents[76])
```



```
, 'child', 'porn', 'video', 'photo', 'late', 'update', 'main', 'cp', 'child', 'porn', 'baby', 'main', 'section', 'modern', 'photo', 'video', 'child', 'porn', 'loli', 'photo', '
```

[26] print(lda_model.get_document_topics(dictionary.doc2bow(tokenized_documents[76])))

```
[(0, 0.1317646), (1, 0.22423206), (2, 0.14666602), (3, 0.03711828), (4, 0.4602191)]
```

Comparison of LDA output with labels obtained using keywords



- We manually labelled each document and used the key-phrases for reference
- These labellings were compared with labellings generated by the LDA model
- Here is an example of the comparison

Doc No.	LDA labellings	Keyword Labellings	Yes/No
1	wiki	wiki	1
2	software	software	1
3	wiki	wiki	1
4	wiki	unk	0
5	porn	porn	1
6	marketplace	marketplace	1

Final Result



- The LDA model successfully labelled most of the documents with the same topic names, as they were assigned manually using keyphrases.
- The success of the model was determined using the classification model evaluation technique of **accuracy**.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

- The accuracy of the model was found to be **78.05%** which implies that out of the total 2660 webpages, approximately 2100 documents are being labelled correctly.

Result analysis



1. 78% of the document labellings obtained by keyword extraction method and LDA model agree with each other
2. Trade-off between the representational power and computational power of **Graph of Words (GoW)** and **Bag of Words (BoW)**
3. GoW has superior representation power and also captures better patterns in the document.
Trade-off : Requires very high computational power
4. GoW finds **intra-document** structure
5. BoW is given as input to LDA and is computationally efficient
6. LDA uses **inter-document** structure to make educated guesses

Conclusion



- Graph of Words (GoW) to find intra-document structure
- Graph of Words (GoW) compared with Word Clouds
- Keywords extraction from GoW
- LDA for document classification
- GoW and LDA are two completely different unsupervised learning approaches and we were successful in analysing the darknet data using them

References



1. Samtani, S., Zhu, H., and Chen, H. (2020). Proactively identifying emerging hacker threats from the dark web: A diachronic graph embedding framework (D-GEF). *ACM Trans. Priv. Secur.*, 23(4):21:1–21:33
2. Broséus, J., Rhumorbarbe, D., Mireault, C., Ouellette, V., Crispino, F., and Décary-Héту, D. (2016). Studying illicit drug trafficking on darknet markets: Structure and organisation from a canadian perspective. *Forensic science international*, 264.
3. Deb, A., Lerman, K., and Ferrara, E. (2018). Predicting cyber-events by leveraging hacker sentiment. *Inf.*, 9(11):280.
4. Faisal Khan (2018). The enigma of the ‘dark web’. [Online; accessed November 21, 2020].
5. Faizan, M. and Khan, R. A. (2019). Exploring and analyzing the dark web: A new alchemy. *First Monday*, 24(5).

Thank You !!