Machine Learning Use In OSINT

Giorgi Iashvili - *Caucasus University / SCSA*

DeepSec 2022

DEEPSEC

SCIENTIFIC CYBER SECURITY ASSOCIATION

კავკასიის უნივერსიტეტი
CAUCASUS UNIVERSITY

Use of OSINT in different sectors

Open source intelligence together with ML elements can be used in different fields like everyday users, offices, and industrial or corporate sectors.

Key fields in data gathering

- Cyber Intelligence groups
- Law firms
- IT security personnel
- Special investigators
- IT oriented corporations
- Financial and insurance sector
- Ethical hackers
- Black hat hackers
- Hacktivist groups

Open data: social media and sources

- Potential hosts;
- Information about domains;
- Media lookup;
- Contact details;
- Files online;
- Location information;

Data collection automated methods

Automated data collection mechanisms are used very widely in different spheres today. Even information technology giants like **Google use AI** in their searching mechanisms to build the system with predicted search results.

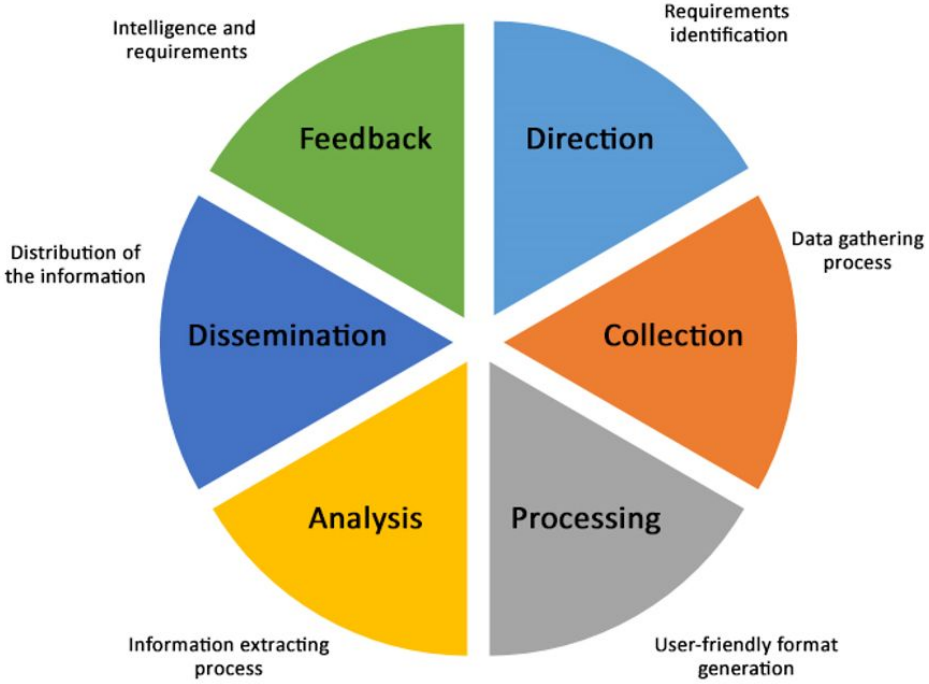
Platforms like **TensorFlow** use artificial intelligence to analyze a huge amount of data, such as approach is used on Gmail, and **Google translate** platforms

Data collection automated methods

There are a lot of different approaches used in social networks, with friends suggestion modules and **video streaming** platforms with recommendations of the clips.

The data about user activities and preferences in such cases is collected and processed by AI algorithms, this provides the end user with **better recommendations**

Intelligence cycle


Intelligence and requirements
Feedback
Direction
Requirements identification
Data gathering process
Collection
Distribution of the information
Dissemination
Analysis
Processing
Information extracting process
User-friendly format generation

Intelligence process: **requirements identification**

Clarification of the requirements and search area. Choosing the **right direction**, working with the requirements is an important step to start the process correctly and not lose time to back to the beginning repeatedly.

Intelligence process: data gathering

Defining the vectors of the work and planning activities. On this stage should be defined the appropriate informational sources. Data collection process also can be performed in different manner of the activities, such ***active*** and ***passive*** data gathering.

Intelligence process: **user-friendly** format generation

Gathered information needs to be sorted and represented in ***understandable format***. It can be special spreadsheet or graphical representation, or even database with prepared data

Intelligence process: information extracting process

The collected data needs to be processed and exploitation activities should be performed based on the ***relevant information***

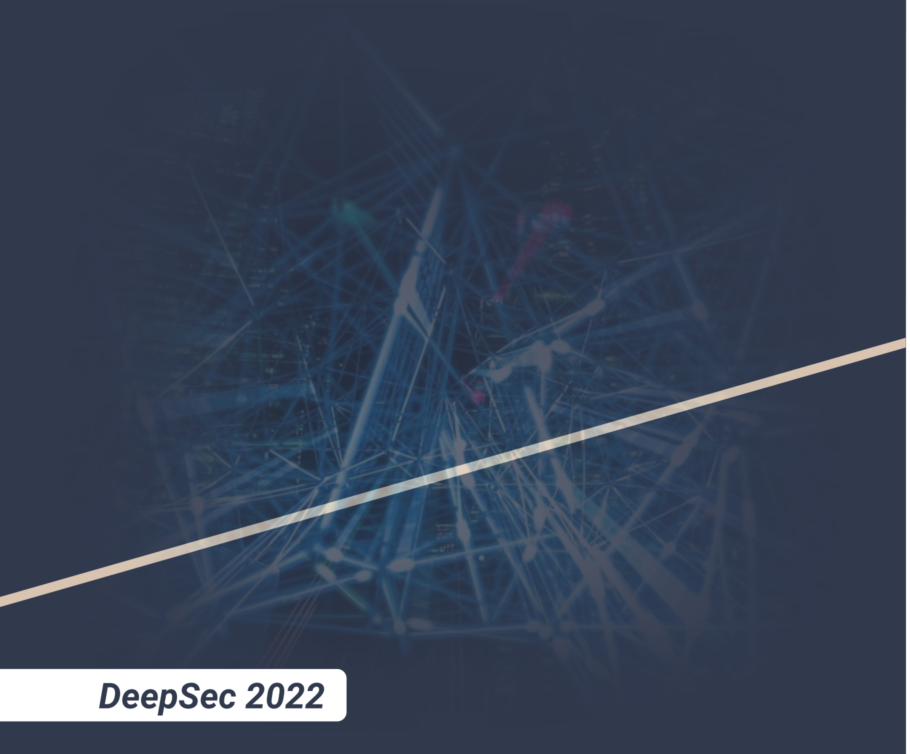
Intelligence process: **distribution of the information**

Gathered information should be shared with the trusted parties to achieve more ***global results*** in the frame of the OSINT activities

Intelligence process: **intelligence and requirements**

The final feedback should be check to meet the requirements. The **accuracy** of the entire process should be assessed.

Possibilities and limitations



Together with the possibilities of artificial intelligence today, there are different limitations when it goes to data gathering process.

During such activities, we need to take into account the fact, that ***not everything is reachable*** using open sources and in some cases ***on the internet*** itself.

Possibilities and limitations

During the building of AI-based OSINT working system, we need to consider the ***limitations and scenarios*** every time.

Together with the technical limitations, the ***ethical*** part is also should be considered.

Use of AI in OSINT

Use of artificial intelligence in OSINT activities can increase the efficiency and quality of the search process.

Using an artificial intelligence a lot of processes like ***web crawling***, data collection, analysis of the patterns can be ***automate***.

Use of AI in OSINT

As artificial intelligence algorithms learn based on the ***previous experience***, it might have an impact on the entire process.

Recommendation works based on ***different parameters***, like user-preferred categories, music genre or timing.

Use of AI in OSINT

The process of the open source intelligence works as follows:

1. Work with sources of information to get the **relevant ones**;
2. Data gathering process using appropriate **informational sources**;
3. Processing of the information based on the sources and **search requirements**;
4. Data collection and analysis based on **multiple sources**;
5. Generation of the results and **reporting** activities;

AI use in the intelligence cycle

Switching form a manual processes to automation machine learning-oriented analysis is extremely important especially when we work with ***real-world operations***.

These processes relies on ***massive data collection*** and analysis to artificial intelligence powered mechanisms and simulations

AI use in the intelligence cycle

The intelligence cycle can be improved on the following stages:

- **automation** of the data collection mechanisms;
- **structuring** of the data;
- automated alerts and **reports**;
- **dissemination** of the collected data;

DeepSec 2022

AI use in the intelligence cycle

Based on this process the human involvement is needed on the **first** and **last** steps.

The first the step: definition of the requirements

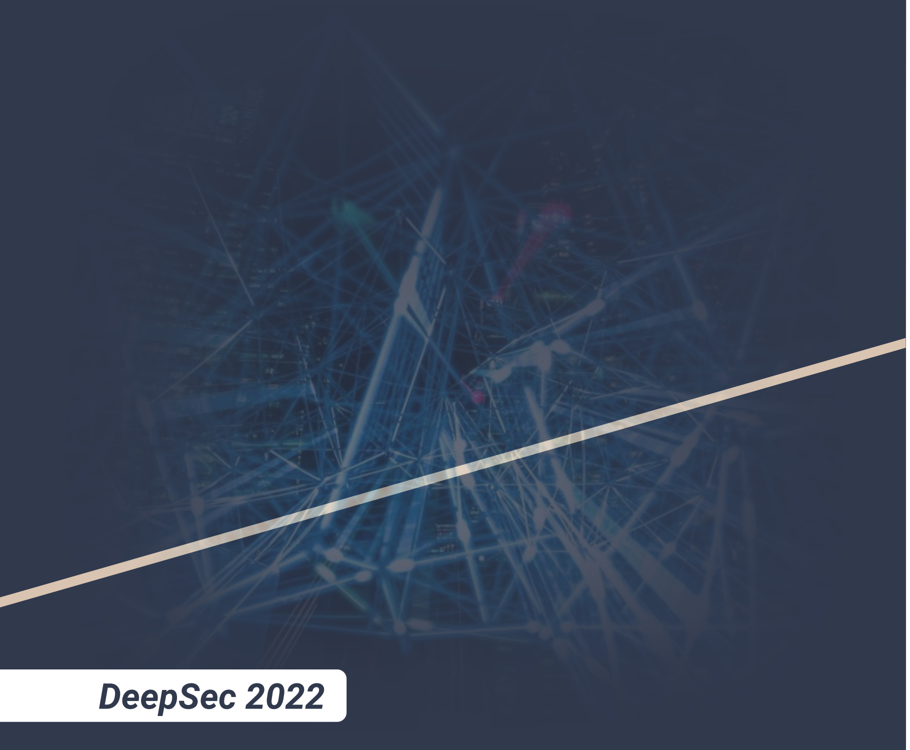
The last step: make sure, that the process was performed correctly

AI use in the intelligence cycle

OSINT is a great candidate to be improved with artificial intelligence because of a ***huge amount*** of the data to be filtered and sorted.

Manually this process may take too much time, but if some main processes are automated, this job can be done in very ***short time***.

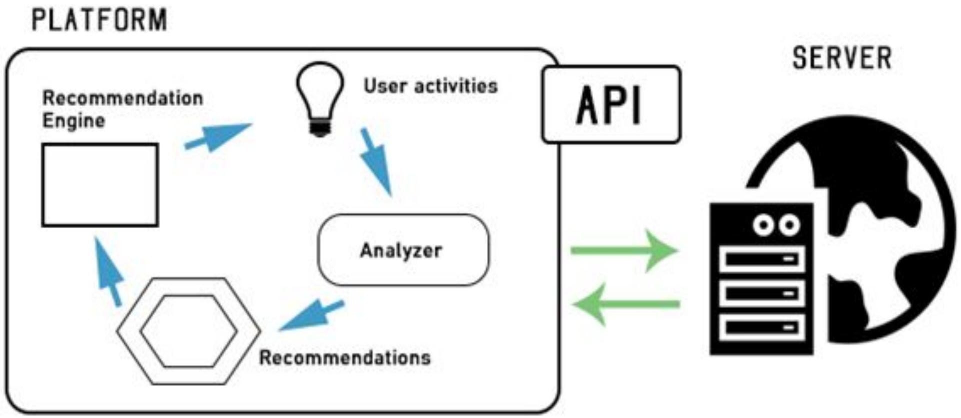
Content-based approach



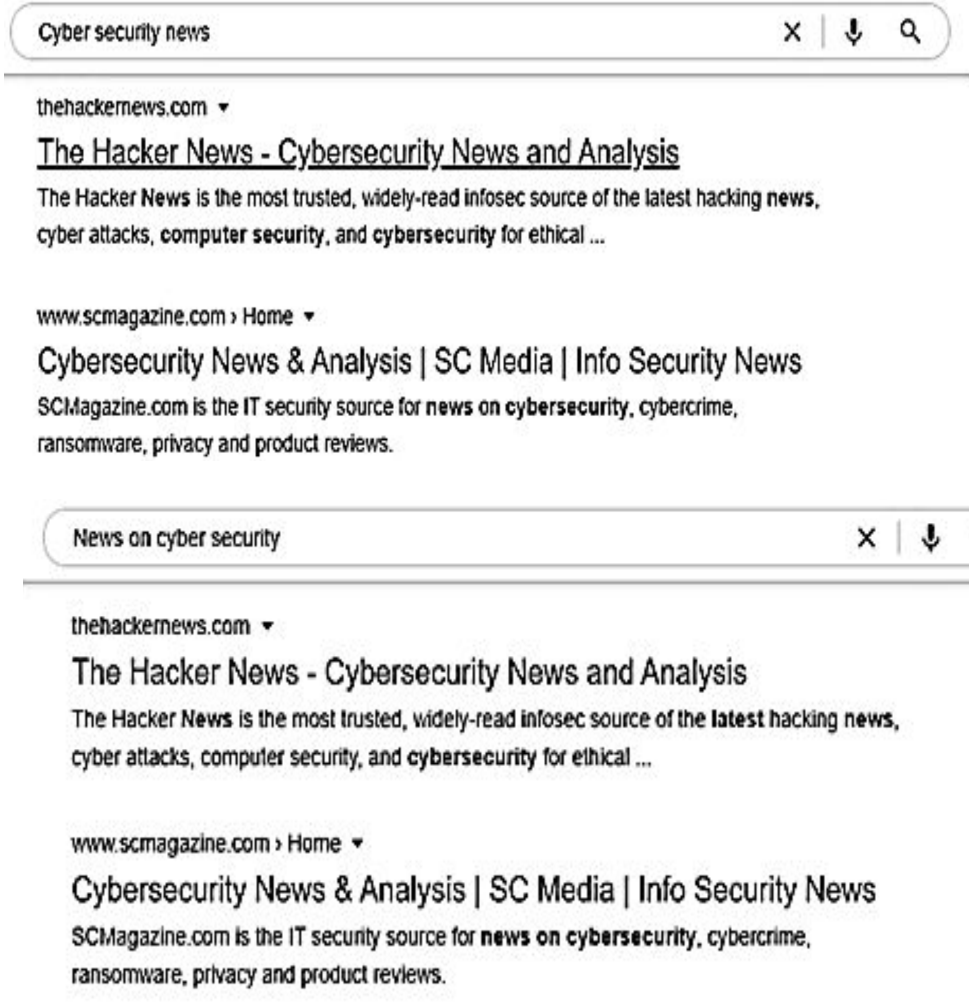
Content filtering systems help the users to find the most ***relevant content***, based on their needs and interests.

The security of the users is extremely important for any modern system today.

Content-based approach: **working process**



Content-based approach: **terms difference**



Cyber security news

thehackernews.com ▾

The Hacker News - Cybersecurity News and Analysis

The Hacker **News** is the most trusted, widely-read infosec source of the latest hacking **news**, cyber attacks, **computer security**, and **cybersecurity** for ethical ...

www.scmagazine.com › Home ▾

Cybersecurity News & Analysis | SC Media | Info Security News

SCMagazine.com is the IT security source for **news on cybersecurity**, cybercrime, ransomware, privacy and product reviews.

News on cyber security

thehackernews.com ▾

The Hacker News - Cybersecurity News and Analysis

The Hacker **News** is the most trusted, widely-read infosec source of the **latest** hacking **news**, cyber attacks, computer security, and **cybersecurity** for ethical ...

www.scmagazine.com › Home ▾

Cybersecurity News & Analysis | SC Media | Info Security News

SCMagazine.com is the IT security source for **news on cybersecurity**, cybercrime, ransomware, privacy and product reviews.

Term frequency matrix

The result is achieved using **text mining** techniques.

The most obvious technique for the text mining is ***TF Matrix*** (term frequency matrix).

Search engine uses TF Matrix to analyze the **frequency of an every word** in the search query

DeepSec 2022

Term frequency and weighting

Two main mechanisms are used in recommender systems:

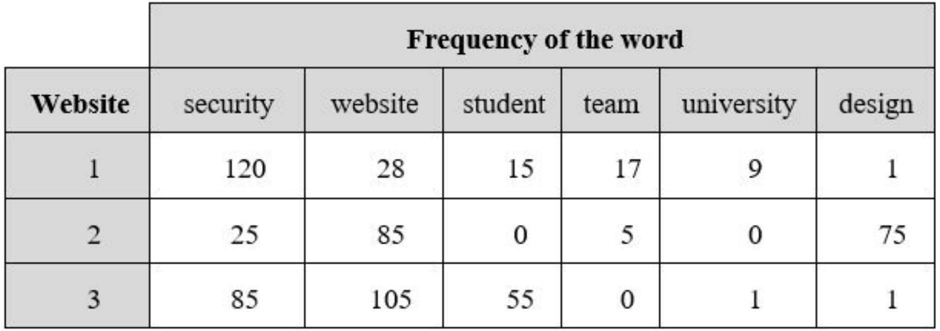
Term Frequency *(TF)*

Inverse Document Frequency *(IDF).*

We can identify the frequency of the words used for websites with the **different content**.

Term frequency and weighting

The results of counting the different words frequency of each website’s content



	Frequency of the word					
Website	security	website	student	team	university	design
1	120	28	15	17	9	1
2	25	85	0	5	0	75
3	85	105	55	0	1	1

TF and IDF

Word “X” occurs in first website 20 times and in the second website four times, does not mean that the word “X” is five times more relevant in the first website than in the second website.

*The difference in this case is **much less**.*

Together with the term frequency we must take into account the **inverse document frequency** (IDF) - measures how important is the concrete case

TF and IDF

Term frequency (TF) is usually divided by the length of the document and shows the total number of the **concrete terms** in the content:

TF(A) = number of times term **A** appears in the content and is ***divided by the total number*** of the terms in this document

IDF and importance

As ***inverse document frequency*** (IDF) measures the importance of the concrete case: “the”, “of”, “or”, “is” - **less important**.

We find the balance using the following formula:

IDF(A) = log_e (***number of all documents*** divided by the ***number of documents with A in the content***).

DeepSec 2022

TF-IDF calculation

If we have **ten million** documents in total and term A appears **in 1000 documents** the document frequency is calculated:

log(10 000 000 / 1 000) = 4

Weighted TF

Assigning the weight ***to be equal*** to the number of the concrete term’s occurrence

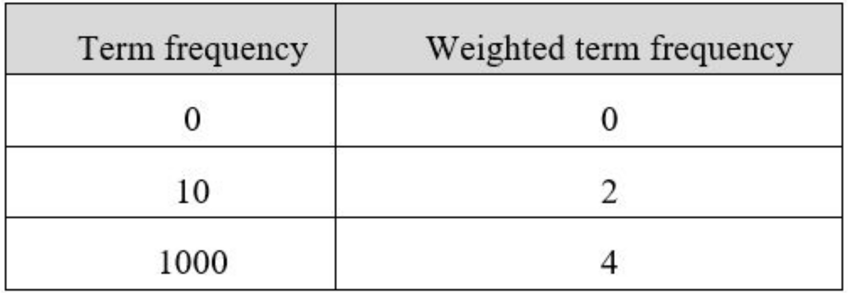
$$w_{t,d} = \begin{cases} 1+log_{10}tf_{t,d}, & if\ tf_{t,d} > 0 \\ 0, & otherwise \end{cases}$$

in weighted TF the terms frequency is **dampened**.

Weighted TF

The ***weighted TF values*** are more ***comparable*** to each other than the values for original term frequency.

Here we can see the transformation of original term frequencies:



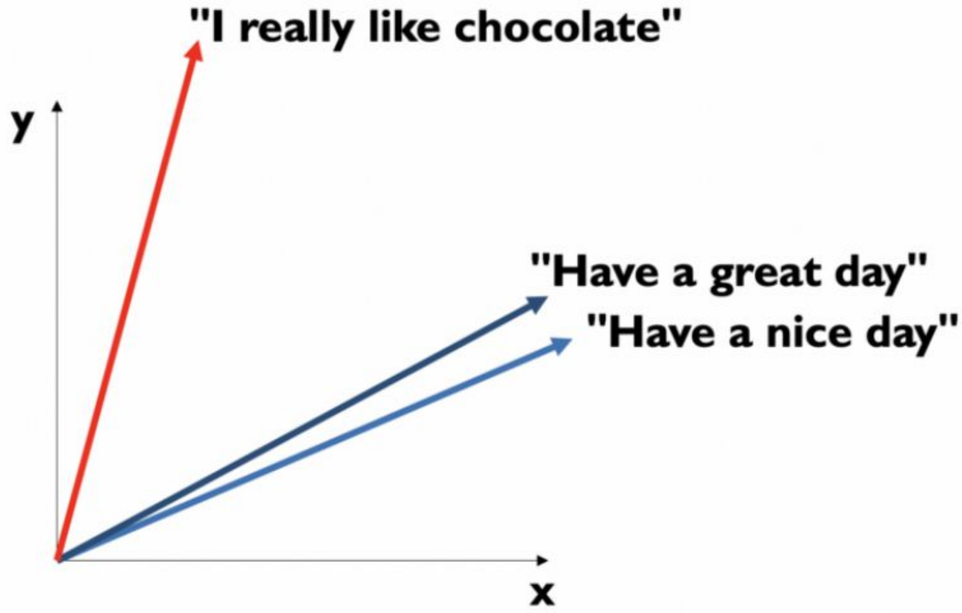
Term frequency	Weighted term frequency
0	0
10	2
1000	4

Vector space model

Once we calculate TF-IDF, we need to determine which items are ***closer to each other***. This can be done by means of the Vector Space Model.

By means of VSM we can compute the proximity which relies on the ***angle between the vectors***.

Vector space model



Vector space model

To determine and have the better accuracy we calculate ***the cosine*** of the angle ***between the vectors***

After **normalization** of the vectors, the length become **equal to one**, that means that calculation of the cosine is simply the sum-product of these vectors

(cos(*“have a great day”*, *“have a nice day”*))

Future of AI in OSINT

Use of artificial intelligence in open source intelligence activities can power up the assistant mechanisms like **Amazon echo**, or **Siri** to make it possible to collect more relevant and rich data about the subject based on the requirements.

Conclusions

Use of mechanisms such is content-based filtering in OSINT activities may ***power up*** the entire process of getting the relevant information with the higher accuracy:

- better speed;
- better accuracy;

Thanks!

Giorgi Iashvili
Caucasus University / SCSA
giiashvili@cu.edu.ge

DEEPSEC

SCIENTIFIC CYBER SECURITY ASSOCIATION

კავკასიის უნივერსიტეტი
CAUCASUS UNIVERSITY