НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
НАВЧАЛЬНО-НАУКОВИЙ ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ
Кафедра інформаційної безпеки

До захисту допущено
Завідувач кафедри

_____ **Дмитро ЛАНДЕ**

(підпис)

«_____»_____2023 р.

Дипломна робота

на здобуття ступеня бакалавра

за освітньо-професійною програмою «Системи, технології та математичні методи кібербезпеки»
спеціальності 125 «Кібербезпека»

на тему: Пошук аномалій пов'язаних із кіберінцидентами в технології OSINT

Виконав: здобувач вищої освіти **IV** курсу, групи ФБ-94
Васюченко Георгій Сергійович

(підпис)

Керівник професор, д.т.н., завідувач кафедри Інформаційної безпеки
Ланде Дмитро Володимирович

(підпис)

Рецензент старший викладач Тітков Дмитро Валерійович
(посада, науковий ступінь, вчене звання, науковий ступінь, прізвище, ім'я, по батькові)

(підпис)

Засвідчую, що у цій дипломній роботі немає запозичень з праць інших авторів без відповідних посилань.
Здобувач вищої освіти

(підпис)

Київ – 2023 року

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
НАВЧАЛЬНО-НАУКОВИЙ ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ
Кафедра інформаційної безпеки

Рівень вищої освіти – перший (бакалаврський)
Спеціальність – 125 «Кібербезпека»
Освітньо-професійна програма «Системи, технології та математичні методи кібербезпеки»

ЗАТВЕРДЖУЮ
Завідувач кафедри

_____ Дмитро ЛАНДЕ
(підпис)

«___»_____2023 р.

ЗАВДАННЯ
на дипломну роботу здобувачу вищої освіти

Васюченко Георгій Сергійович

1. Тема роботи: «Пошук аномалій пов'язаних із кіберінцидентами в технології OSINT»
керівник роботи к.н., доцент кафедри Інформаційної безпеки
Ланде Дмитро Володимирович,
затверджені наказом по університету від «26» травня 2023 р. № 2028-С

2. Термін подання здобувачем вищої освіти роботи 15 червня 2023 р.

3. Вихідні дані до роботи: література по методології OSINT, статті по операціями в кіберпросторі, методи аналізу великих даних, документація систем Telegram, Elasticsearch та Kibana.

4. Зміст роботи: проведення збору, обробки та аналізу великих даних з відкритих джерел за методологією OSINT для пошуку аномалій пов’язаних із кіберінцидентами

5. Перелік ілюстративного матеріалу (із зазначенням плакатів, презентацій тощо): Презентація до захисту дипломної роботи.

6. Дата видачі завдання 3 квітня 2023 р.

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів дипломної роботи	Примітка
1	Визначення напрямку дослідження та формулювання теми	01.10.2022 - 24.10.2022	виконано
2	Визначення рекомендованих джерел інформації та їх опрацювання	25.10.2022 - 01.02.2023	виконано
3	Дослідження використання методології OSINT, пошук джерел інформації	01.03.2023 - 15.03.2023	виконано
4	Дослідження методів збору та обробки даних, написання власних інструментів	16.03.2023 - 21.05.2023	виконано
5	Вивчення систем для роботи з даними, їх порівняння	23.03.2023 - 21.05.2023	виконано
6	Вивчення Elasticsearch та Kibana, робота з даними за допомогою даних інструментів, аналіз даних	01.04.2023 - 10.06.2023	виконано
7	Проходження переддипломної практики	14.04.2022 - 21.05.2023	виконано
8	Передзахист дипломної роботи	16.06.2023	виконано
9	Доопрацювання	16.06.2022 - 23.06.2023	виконано
10	Захист дипломної роботи	23.06.2023	виконано

Здобувач вищої освіти _____ Георгій ВАСЮЧЕНКО

Керівник роботи _____ Дмитро ЛАНДЕ

РЕФЕРАТ

Дипломна робота складається з 3 розділів, містить 42 рисунки, 2 додатки та 29 джерел посилань, загальний обсяг роботи — 75 сторінок.

Об’єкт дослідження: моніторинг відкритих джерел інформації на предмет аномалій, пов’язаних з кіберінцидентами.

Предмет дослідження: дослідження методів збору даних з відкритих джерел, їх обробки та аналізу отриманих для побудови моніторингу аномалій пов’язаних з кіберінцидентами.

Мета дослідження: визначення релевантних джерел для збору інформації про операції в кіберпросторі, її збір, обробка та її аналіз з метою виявлення аномалій пов’язаних з кіберінцидентами

Методи дослідження: аналіз літературних джерел, наукових робіт, електронних ресурсів та посібників на досліджувану тему, вивчення публікації з порівняльним аналізом інструментів та практичних досліджень.

Результат роботи: аналіз існуючих методів збору даних з відкритих джерел, їх обробки та аналізу. Використання методів аналізу для проведення пошуку аномалій у кіберпросторі на прикладі платформи Telegram.

Ключі слова: OSINT, кіберпростір, моніторинг, соціальні мережі, API, Elasticsearch, Kibana.

ABSTRACT

The work consists of 3 chapters, 42 illustrations, 2 appendixes and 29 references, the total volume of the work is 75 pages.

Object of research: monitoring of open sources of information for anomalies related to cyber incidents.

Subject of research: study of methods for collecting data from open sources, processing and analyzing them to build monitoring of anomalies related to cyber incidents.

Purpose of the study: to identify relevant sources for collecting information about operations in cyberspace, its collection, processing and analysis in order to identify anomalies related to cyber incidents

Research methods: analysis of literary sources, scientific papers, electronic resources and manuals on the topic, study of publications with a comparative analysis of tools and case studies.

Results: analysis of existing methods of collecting data from open sources, processing and analyzing them. The use of analysis methods to search for anomalies in cyberspace on the example of the Telegram platform.

Keywords: OSINT, cyberspace, monitoring, social networks, API, Elasticsearch, Kibana.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

OSINT – Open Source Intelligence

Кіберпростір — інформаційне середовище, що функціонує за допомогою комп'ютерних систем.

УІБ — управління інформаційною безпекою

ШПЗ — шкідливе програмне забезпечення

ПЗ — програмне забезпечення

TTP – Tactics, techniques and procedures

API – Application Programming Interface

REST API – Representational state transfer API

JSON – JavaScript Object Notation

СКБД — система контролю базами даних

ELK stack – стек інструментів Elasticsearch, Logstash, Kibana

ВСТУП

У сучасну цифровому епоху кіберпростір став невід'ємною складовою частиною нашого повсякденного життя. За допомогою різноманітних пошукових систем, веб ресурсів, електронних видань, соціальних мереж та інших джерел ми отримуємо ту чи іншу інформацію, ціна якої неухильно росте.

Постає питання: що робити з цією величезною кількістю даних, що генерується в кіберпросторі, як зручніше її зберігати та аналізувати. У контексті пошуку аномалій, пов'язаних із кіберінцидентами, невід'ємною складовою є використання методології пошуку інформації з відкритих джерел, відомої як Open Source Intelligence або ж просто OSINT, що передбачає пошук та збір інформації, що представляє оперативний інтерес, із загальнодоступних джерел, та її подальший аналіз.

У цій роботі пропонується провести усі кроки методології OSINT щодо конкретних кіберінцидентів та більш детально зупинитись на етапі ручного аналізу зібраної інформації з використанням інструменту великих даних — пошукової системи Elasticsearch для збереження, пошуку та обробки даних і Kibana як інструменту для візуалізації та інтерфейсу до Elasticsearch.

Актуальність роботи полягає в потребі розробки інструменту моніторингу відкритих джерел інформації для аналізу в кіберінцидентів.

Мета і завдання дослідження полягають в визначенні релевантних джерел для збору інформації про операції в кіберпросторі, її збір, обробка та її аналіз з метою виявлення аномалій пов'язаних з кіберінцидентами.

На шляху до виконання поставленої мети стоять наступні завдання:

1. Аналіз методології OSINT та її застосування відповідно до предмету дослідження.
2. Визначення джерел для збору інформації на платформі Telegram.

3. Створення інструментів для збору даних з платформи Telegram та їх обробка.
4. Аналіз існуючих систем введення, зберігання, пошуку та опрацювання даних.
5. Проведення аналізу зібраних даних для пошуку аномалій, пов’язаних з кіберінцидентами.
6. Порівняння розробленого підходу та інструментів з існуючими системами.

Методи дослідження: аналіз літературних джерел, наукових робіт, електронних ресурсів та посібників на досліджувану тему, вивчення публікації з порівняльним аналізом інструментів та практичних досліджень.

Наукова новизна одержаних результатів полягає в дослідженні відповідних джерел інформації, створенні та використанні ефективних інструментів для виявлення аномалій у кіберпросторі, забезпеченні аналізу інформації використовуючи вільне програмне забезпечення.

Практичне значення одержаних результатів: результати дослідження матимуть практичне значення для виявлення та оцінки проведених у кіберпросторі операцій за допомогою підходу OSINT.

1 ТЕОРЕТИЧНІ ДАНІ ЩОДО OSINT ТА КІБЕРІНЦИДЕНТІВ. МЕТОДИ ПОШУКУ, ЗБОРУ ТА ОБРОБКИ ДАНИХ З СОЦІАЛЬНИХ МЕРЕЖ

1.1 Сфера OSINT

OSINT — це методологія розвідки, що передбачає пошук, збір, обробку та аналіз розвідувальної інформації із широкого спектру загальнодоступних джерел, таких як веб ресурси, соціальні мережі, форуми, блоги, електронні видання новин, пошукові системи, журнали, публічні записи та інші відкриті джерела, тобто найчастіше мова йде про комп'ютерні джерела, але не обмежується ними.

Серед інших методологій розвідки OSINT виділяється двома основними ознаками:

- Використання відкритих джерел інформації, які надають інформацію, що не має обмежень на загальний доступ до неї. Тобто сама інформація не є захищеною від публічного розкриття, а це означає, що вона не має обмежень у доступі для зацікавлених осіб і може вільно використовуватися для тих чи інших цілей.
- Використання загальнодоступної інформації, яка навмисне надана для широкого використання користувачами. Тобто сама інформація від початку призначена для поширення і використання загальною публікою без необхідності в отриманні спеціального дозволу. Вільність перегляду, отримання, використання та поширення відповідно до правил та законів регулювання цієї інформації — основні властивості загальнодоступної інформації.

Основною особливістю даної методології варто відзначити те, що вона дозволяє ефективно зібрати значну кількість необхідної інформації за

короткий період часу та з використанням мінімуму ресурсів, при цьому, як інформація, так і інструменти OSINT зазвичай є безкоштовними.

Роль OSINT у розвідці визначається багатьма аспектами, тому в процесі планування і підготовки до проведення OSINT враховуються такі чинники:

- Вартість.

  Безкоштовною вважається приблизно 40% всієї інформації в Інтернеті. Такий великий відсоток усім доступної інформації зумовлено тим, що веб сайти та платформи надають безкоштовний доступ до своїх ресурсів з метою залучення користувачів та просування власних послуг.

  Комерційні ж ресурси, зазвичай методами стягування плати за підписку або певний обсяг інформації, можуть пропонувати ‘ексклюзивний’ та більш високоякісний матеріал.
  Безліч безкоштовних OSINT інструментів, доступних для використання. Їх кількість та якість може суттєво різнитися, функціоналу може не вистачати для забезпечення повного інформаційного покриття, тому розробка власних інструментів під конкретні задачі є частим сценарієм.
  Тобто вартість збору інформації напряму залежить від сервісу, який використовується в якості джерела та інструментів пошуку, збору, обробки та аналізу даних.

- Актуальність. Загальнодоступна інформація забезпечує актуальність, доступність та широкий обсяг інформації без необхідності залучення додаткової техніки та спеціалістів.

- Якість та надійність. Загальнодоступна інформація викликає менше сумнівів ніж інформація, яка взята з таємних джерел або подана ними ж у спеціальних звітах. Це впливає на суб’єктивність зібраної інформації та її ангажованість.

- Обсяг. OSINT передбачає збір сукупності якісних даних з великої кількості доступних джерел, що зумовлює отримання великих обсягів

значимої інформації, що, однак, потребує ретельної обробки та аналізу.

- Швидкість. OSINT передбачає швидкий збір актуальної оперативної інформації. Час на пошук самої інформації різко скорочується завдяки використанню відкритих джерел та зв'язкам між різними джерелами. Також варто враховувати, що актуальні події швидко з'являються на різних платформах та у звітах. Щодо залучених людських та технічних ресурсів — тут все обмежується поставленими задачами, але у порівнянні з іншими методологіями розвідки OSINT є менш затратним.
- Простота у використанні. Відсутність окремих регулювань (окрім як правил та законів регулювання інформації в Інтернеті загалом), на перегляд, отримання, використання та поширення даних, що можуть сповільнити або унеможливити їх збір, себто відсутність режимів доступу та певних бар'єрів.
- Детальний аналіз даних. Зібрані дані можуть використовуватися для глибинного аналізу певних тем, як наприклад загальних трендів, прихованих зв'язків між різними джерелами та самою інформацією, прогнозування та розвитку ситуацій. На основі зібраних з відкритих джерел даних можна проводити глибинний аналіз з метою вирішення тих чи інших поставлених задач.

У провідних країнах світу методологія OSINT вважається одним із найважливіших інструментів для захисту своїх інтересів та використовується для забезпечення розвідувальних, державних, приватних, комерційних та інших органів. Методологія та технології OSINT дозволяють проводити повноцінний аналіз та розвідку базуючись на зібраних даних, що може грати ключову роль у військовій розвідці та використовуватись при плануванні військових дій а також їх інформаційному та інших супроводах.

У методології OSINT виділяють наступні етапи розвідувального циклу:

1. Планування та підготовка (planning and preparations) — оцінка потреб і вимог завдання, визначення цілей та відповідних джерел, які можна використати для пошуку необхідної інформації.
2. Збір (collection) — основний та найважливіший етап збору даних та інформації з якомога більшої кількості релевантних джерел.
3. Обробка (processing) — організація і зіставлення зібраних даних.
4. Аналіз і підготовка звіту (analysis and presentation) — інтерпретація зібраної інформації для того, щоб зрозуміти що було зібрано, виявлення закономірностей та хронології процесу. Звіт являється відповіддю на розвідувальне завдання, висновки та рекомендації щодо наступних кроків.
5. Поширення або ж розповсюдження (dissemination) — актуальний в деяких сферах крок, який описує презентування та поширення результатів з відкритих джерел для зацікавлених сторін. Варто зазначити, що процес є циклічним, може перетинатись та повторюватись, якщо цього вимагає ситуація.

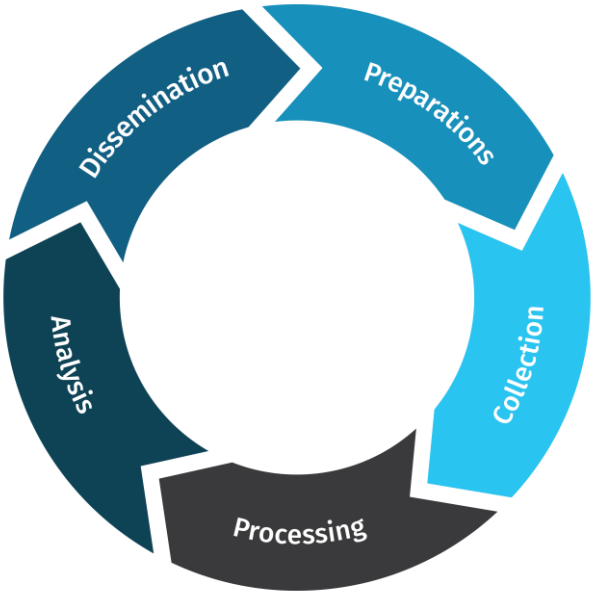


Рисунок 1.1 — Цикл, що ілюструє етапи процесів методології OSINT

Використання методології OSINT для виявлення та запобігання кіберінцидентів є ефективним підходом. Він дозволяє виявляти загрози, збирати розвідувальну інформацію та відстежувати діяльність зловмисників.

Основними способами даного підходу є:

1. Пошук актуальних вразливостей в системі, що дозволяє попередити можливі атаки та прийняти відповідні заходи для їх усунення.
2. Моніторинг загроз. За допомогою постійного моніторингу веб-сайтів, соціальних медіа, форумів тощо, можна виявляти загрози, методи атак та стратегії зловмисників.
3. Моніторинг активності. Виявлення незвичайних або підозрілих активностей у мережах або системах. Моніторинг надійних джерел дозволяє буду готовим до атаки, що планується, відбувається або відбулась. Навіть постфактум можна оцінити збитки, надані компанії за допомогою моніторингу профільних ресурсів, на яких, наприклад, торгують вкраденою інформацією.

Отже, використання методології OSINT є важливим інструментом боротьбі з кіберінцидентами в тому числі, адже вона дозволяє підвищити рівень свідомості щодо ймовірних загроз, виявити їх та прийняти відповідні заходи для запобігання атакам. Постійний моніторинг та аналіз публічно доступної інформації сприяє збереженню безпеки в цифровому середовищі.

1.2 Кіберінциденти та їх наслідки

Російсько-українська війна, що почалася в 2014 році, принесла із собою повноцінний двовимірний формат конфлікту — кібервійну, яка відбувається в кіберпросторі, та війну, власне звичну, між збройними силами країн на землі, і обидва методи ведення війни відбуваються паралельно та синхронізовано для досягнення більшого впливу на суперника.

Кіберінциденти, що викликані веденням кібервійни, пропоную розглядати відповідно до трьох нижче наведених концептів:

- стратегічні комунікації — координація, інтеграція та синхронізація органів державної влади та органів місцевого самоврядування, інтеграція та синхронізація діяльності державних органів з громадським суспільством тощо.
- ведення інформаційної війни — діджиталізація інформаційного середовища змінила умови ведення інформаційної війни. На
- найпростішому рівні інформаційну війну можна описати як навмисне маніпулювання або використання інформації тією чи іншою стороною з метою вплинути на вибір і рішення, які приймає супротивник, з метою отримання військової або стратегічної вигоди. Таким чином, сфера інформаційної війни включає в себе інформаційні та психологічні операції, кібератаки (що можуть бути спрямовані на дуже широкий вибір цілей — від військових об’єктів і до державних та приватних підприємств, засоби масової інформації тощо) та інші методи ведення інформаційної війни.
- управління інформаційною безпекою (УІБ) — невід’ємна частина стратегічної комунікації та методів ведення інформаційної війни. Управління інформаційною безпекою є цілісним підходом до забезпечення безпеки на перетині інформації, інформаційних технологій та інфраструктури, як фізичної, так і цифрової, від ворожих дій. УІБ також напряму пов’язане з тим, як держави намагаються контролювати медіатизацію війни, яку можна розділити на три етапи. До кінця 1990-х років війни медіатизувались переважно через засоби масової інформації (журналісти на телебаченні, радіо, преса тощо) з ланцюгом подачі інформації у форматі ‘держава — журналісти — широка аудиторія’. З появою соціальних медіа — а це початок 2000-х років — медіатизація стала більш хаотичною: будь-хто може документувати, архівувати та брати участь у медіатизації війни.

Нинішній же третій етап відзначається тим, що професійні медіа та військові інституції зупинили хаотичну динаміку соціальних мереж і почали більш ефективно їх використовувати для своїх цілей. Можна з впевненістю сказати, що методи другого етапу вже `віджили` своє та все менше використовуються, уступаючи методам третього етапу.

Україна є постійною мішенню кібератак щонайменше з 2014 року. Щороку відбуваються та відбиваються тисячі атак, що робить Україну своєрідною пісочницею (sandbox) для нової кіберзброї, тактик та інструментів. Пропоную розглянути періоди 2014-2015х років, 2016-2021х років, початку 2022 року і від лютого 2022 року та до сьогодення для вивчення етапів та видів кібератак проти України.

- Атаки 2014-2015х років.

  13 березня 2014 року було здійснено 8-хвилинну DDoS-атаку на українські комп'ютерні мережі задля відвернення уваги громадськості від політичних подій в країні.

  25 травня 2014 року (із заздалегідь проведеною підготовкою та розвідкою) було здійснено серію кібератак з метою маніпулювання голосуванням на президентських виборах. Хакери з групи `CyberBerkut` вторглись в мережу, видаляли системні файли та намагались змінити результати виборів, поставивши за мету руйнування систем Центральної Виборчої Комісії. Атака не вдалась, оскільки ШПЗ було видалено за 40 хвилин до початку голосування. Власне, хакерам вдалося лише затримати підрахунок голосів.

  Після цього регулярно здійснювались кібератаки на електромережі.

  23 грудня 2015 року постраждали кол-центри та мережі трьох енергорозподільчих компаній та понад 230.000 українців зазнали відключення електроенергії тривалістю від 1 до 6 годин.

  Варто зазначити, що відключення електроенергії хакери домоглись, але повністю вивести обладнання з ладу не вдалось.

- Атаки 2016-2021х років.

  Під час цього періоду кількість кібератак та їх вплив значно зріс. У червні 2017 року запуск ШПЗ NotPetya, який наразі вважається найбільш руйнівною кібератакою в історії, через бухгалетрське ПЗ вразив ЧАЕС, близько 13.000 пристроїв, що використовувались державними установами, банками, поштовими службами, газетами, транспортною інфраструктурою та іншими бізнесами. Вдалось провести знищення даних на комп'ютерних дисках із унеможливленням відновлення даних після враження ШПЗ. Загалом атака зачепила близько 65 країн і 50.000 сиситем, включачи європейські та американські фірми, завдавши неймовірно високих збитків.

  У 2018 році відбулась атака на станцію дистиляції хлору в Аулі, яка працює в 23 областях України.

  У 2021 році відбулась атака на веб-сайти Служби безпеки України, систему електронної взаємодії органів виконавчої та державної влади, однак шкоди системам завдати не вдалось.

- Атаки на початку 2022 року.

  На початку 2022 року кількість кібератак знову різко зросла.

  13 січня було помічено ШПЗ, націлене на український уряд та кілька некомерційних (у тому числі ІТ) організацій.

  14 січня близько 70 урядових веб сайтів, зокрема Кабінету Міністрів, Міністерства Оборони, Міністерства Закордонних Справ та інших державних інституцій опинилися під тимчасовим контролем хакерів.

  У середині лютого DDoS-атака вивела з ладу сайти кількох урядових відомств, банків та радіостанцій на декілька годин.

  Дані дії можна вважати як проведенням дорозвідки цілей, так і спробою оцінити власні TTP груп, що здійснювали кібератаки.

- Атаки від лютого 2022 року.

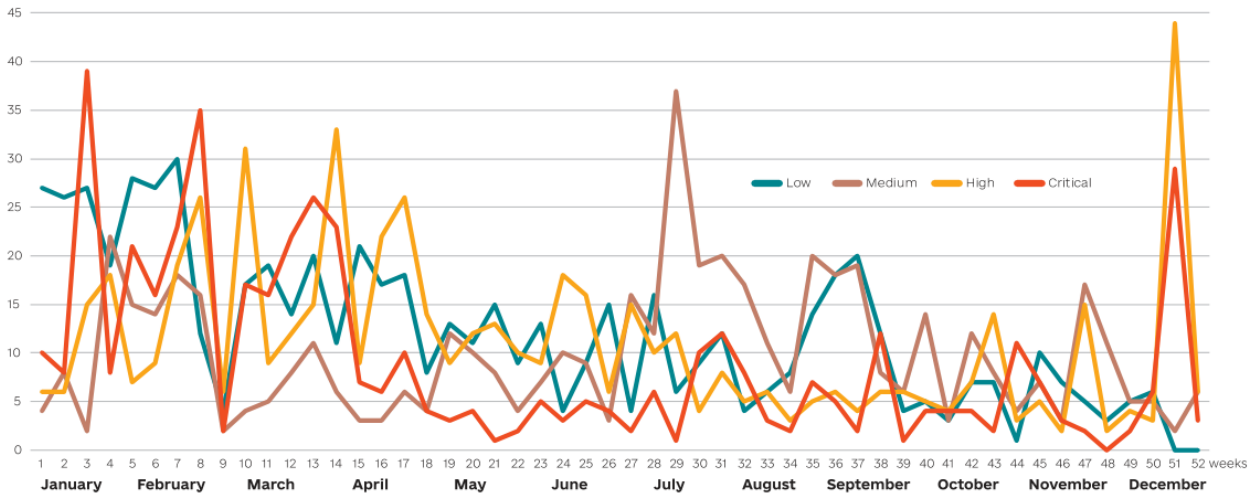


Рисунок 1.2 — Загальне представлення кіберінцидентів відповідно до їх значущості (severity) проти українських ресурсів за 2022 рік

Від лютого 2022 року по березень (включно) 2023 року, згідно з ресурсом CyberPeace Institute, на українські ресурси було здійснено понад 359 атак на 5 основних секторів:

1. Державне управління — 81 кібератака.
2. Фінансові організація — 43 кібератаки.
3. Засоби масової інформації — 34 кібератаки.
4. Інформаційно-комунікаційні технології (ІКТ) — 18 кібератак.
5. Транспорт та енергетика — 18 кібератак.

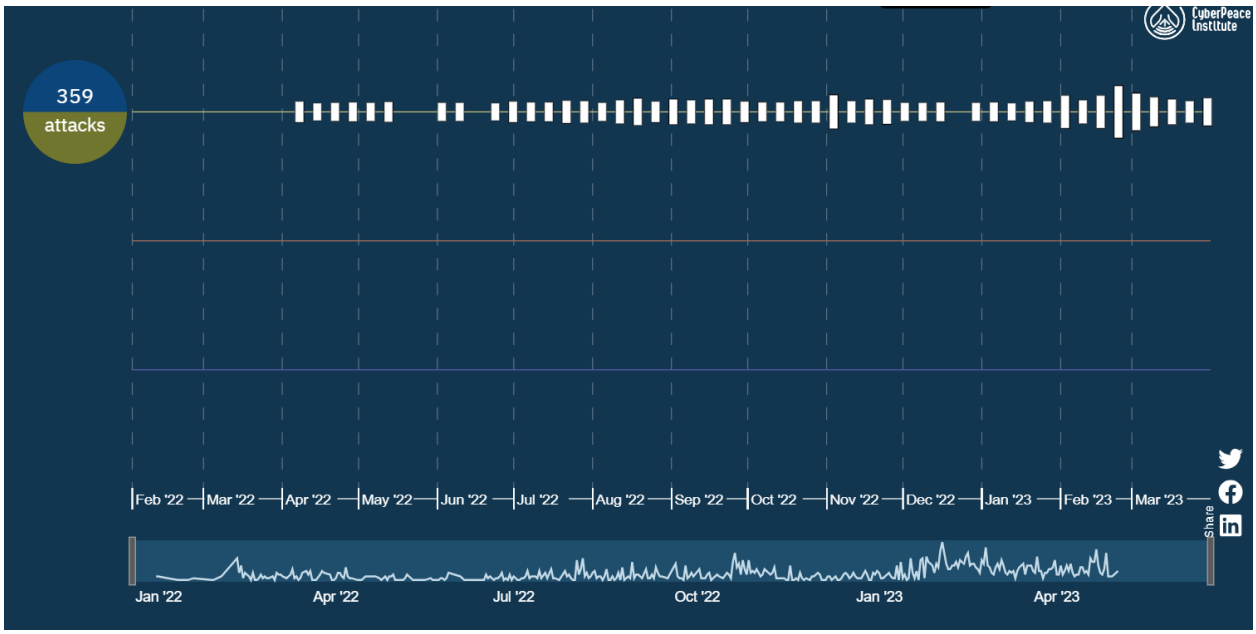


Рисунок 1.3 — Кількість кібератак на українські ресурси від 01.02.2022 по 01.07.2023

24 лютого здійснено кібератаки на системи зв'язку видання 'Kyiv Post` та супутникову мережу KA-SAT, що призвело до відключень фізичних осіб, державних та приватних організацій.

25 лютого здійснено кібератаки на урядові веб-сайти за допомогою ШПЗ IsaacWiper та станцію прикордонного контролю на кордоні з Румунієї з метою порушення в'їзду біженців.

28 лютого здійснено кібератаки на цифрову інфраструктуру України з метою блокування доступу до фінансових та енергетичних послуг.

Від березня і до сьогоднішнього дня

14 березня ШПЗ CaddyWiper було розгорнуто в організаціях урядового та фінансового секторів.

16 березня в ефірі одного з телеканалів з'явився заклик до капітуляції від президента України.

17 та 18 березня було здійснено череду фішингових атак за електронних листів проти урядових осіб та військовослужбовців.

20 березня було використано бекдор LoadEdge для встановлення шпигунського ПЗ.

Кіберінциденти стали прямо впливати на цивільних осіб, цивільні об’єкти (включно з приватними компаніями), інфраструктуру, що забезпечує цивільні об’єкти та життєво важливі послуги для цивільного населення. Державні установи, об’єкти критичної інфраструктури (енергетичні компанії, комерційні організації, логістичні компанії тощо), оборонні організації (включаючи Міністерство Оборони, ДПСУ тощо) є основним фокусом для кібератак з 2014 року і по сьогодення.

Рисунок 1.4 — Найбільш відомі кібератаки на українські ресурси від 2014 по 2022 рік

Угрупованнями, які стали найактивнішими у нападах на українські об’єкти, є:

1. People’s CyberArmyofRussia – 99 кібератак.
2. NoName057(16) – 16 кібератак.

3. Anonymous Russia – 26 кібератак.

4. Sandworm (UAC-0082) – 19 кібератак.

5. DEV-0586 — 12 кібератак.

Всього від 24.02.2022 по 31.04.23 зафіксовано щонайменш 63 російських угруповання, що приймають участь у веденні кібервійни проти України.

Так звані російські ‘хактивістські’ угруповання (Netside Group, Russian Clay, People’s CyberArmyofRussia, KillNet, Xaknet, Zarya, RaHDIt, Z-Team тощо) займаються менш затратними, комплексними та впливовими операціями, які мають за мету в основному медіа ефект в Інтернеті, як наприклад: проведення DDoS-атак, дефейс веб-ресурсів, проведення OSINT операцій відносно фізичних та юридичних осіб тощо.

Більшість з угруповань, що займаються кібератаками на українські ресурси прагнуть, з-поміж власне практичної користі від нанесених збитків, медійного ефекту за допомогою ‘демонстрації сили’ та авторитету.

До прикладу взяти російські ‘хактивістські’ угруповання, що часто ведуть блог або канал на популярних веб-ресурсах та діляться інформацією про заплановані або проведені операції, можна знайти такі подробиці щодо кібероперацій як:

- Дата та час проведення операції.
- Суб’єкти проведення операції — ідентифікація причетних угруповань.
- Об’єкт проведення операції — який саме ресурс зазнав кібератаки.
- Завдані збитки — зазвичай надаються пояснювальні медіа файли, який підтверджують ефект від завданої кібероперації або ж, наприклад при зливі даних (Hack&Leak), публікуються семпли отриманої інформації у вільний доступ задля підтвердження її автентичності та поширення між користувачами, часто задля виклику відчуття безпорадності перед

атакуючими, мобілізації громадської реакції (з метою втрати довіри до тих чи інших організацій), пошкодженні репутації або політичного тиску.

Відповідно, застосування методології OSINT з використанням в якості відкритих джерел інформації ресурси атакуючих угруповань а також агрегатори злитих у вільний доступ даних, моніторинг та аналіз їх активності може суттєво збільшити ефективність боротьби з кібероперація́ми наступним чином:

1. Виявлення потенційних загроз.
2. Виявлення вразливостей.
3. Запобігання поширенню даних.
4. Ідентифікація причетних угруповань.
5. Ідентифікація ресурсу, який став суб'єктом кібероперація, його оповіщення щодо інциденту та аналіз помилок.
6. Виявлення інсайдерів у тих чи інших організаціях, що стали посібниками витоку або поширення даних.
7. Оцінка завданих збитків.

1.3 Актуальність соціальних мереж як джерел інформації та їх моніторинг

Соціальні мережі та зростання їх популярності призвело до того, що вони стали забезпечувати не тільки комунікації та обмін думками, але й потужними інформаційними платформами, що охоплюють дуже широкий спектр різних тем та актуальних подій. На сьогоднішній, при веденні

гібридної війни, де інформаційна складова є не менш важливою у порівнянні з діями військового характеру, виникає нагальна потреба моніторингу соціальних мереж, адже вони активно використовуються з метою психологічного впливу на аудиторію. В першу чергу потрібно відзначити такі соціальні мережі, форуми та месенджери як:

1. Twitter
2. Reddit
3. Telegram
4. Discord
5. 4chan
6. LiveJournal

Самі ж сторони неофіційно використовують той же Telegram чи Discord для проведення та планування операцій, поширення інформації між учасниками тощо. Різні організації та користувачі по всьому світу діляться в них своїми думками щодо поточної військової ситуації та, відповідно, містять підтримку однієї зі сторін.

Інформація в соціальних мережах містить багато шуму, дезінформації та відвертих фейків, що потрібно враховувати при моніторингу. При цьому врахування подібних джерел є однозначно необхідним, адже інформація яку вони можуть містити включає етапи підготовки, проведення та пост-кампанії з реалізації кібератак а також висвітлення кіберінцидентів.

Використання розвідки за відкритими джерелами (якими соціальні мережі і являються) — OSINT – при реалізації моніторингу соціальних мереж є більш ніж актуальним та дієвим підходом. Практичне значення такої системи моніторингу полягає в інтеграції з іншими інформаційними системами, дослідження та аналіз контенту в мережі, що може допомогти з вирішенням

питань у сфері кібербезпеки, а саме виявляти аномалії, пов’язані з кіберінцидентами як заздалегідь, так і постфактум.

До прикладу можна взяти інцидент з витоком секретних документів (з грифом ‘Цілком таємно’) американської розвідки 28 квітня 2023 року. У документах, викладених на платформі Discord у лютому-березні 2023 року, було надано оцінки військової ситуації та детальний аналіз потенційної політики країн. Автентичність документів було підтверджено і сам матеріал витоку швидко став предметом для дискусій. Вперше про існування документів на широку аудиторію повідомила американська газета New York Times лише після того, як 5 квітня (починаючи з 21:29 за київським часом) низка російських Telegram каналів вже поширили витік з відредагованими даними. За декілька годин до поширення в Telegram, серію матеріалів з витоку було розміщено на платформі 4chan.

Ключовим моментом у цій історії є те, як ніхто не помітив витік завчасно, коли можна було запобігти поширенню даних, ідентифікувати першоджерело та зконтактувати з ним, оцінити завдані збитки та ідентифікувати самого інсайдера з розвідки, що став причиною зливу, вчасно.

Подібні задачі як раз і має вирішувати розвідка з використанням відкритих джерел. Варто враховувати, що контент моніторинг Discord каналів є майже неможливою задачею, тому на етапі поширення даних через Discord зупинити або хоча би дізнатися про витік можливості теж не було. Однак після цього витік поширився через іншу платформу — 4chan – де не було ніякої реакції на викладений матеріал. До широкого кола користувачів дана інформація дійшла вже після її публікації на найбільш популярних платформах — Telegram та Twitter, однак момент вчасного реагування на інцидент вже було втрачено.

1.4 Методи пошуку, збору та обробки даних з соціальних мереж

У цьому розділі пропонується розглянути актуальні методи пошуку, збору та обробки даних з соціальних мереж на прикладі Telegram. Дана платформа є найбільш широко використовуваною для зв'язку між представниками спільнот та, власне, спільнотами та користувачами, у ній є можливість створювати як публічні, так і приватні канали й чати, залишати коментарі — усе це робить Telegram зручним джерелом для проведення розвідки за допомогою методології OSINT. Окремо потрібно відзначити широкий вибір контенту, який зазвичай є недоступним або тим, що підлягає блокуванню, у порівнянні з іншими платформами.

1.4.1 Методи пошуку джерел даних

Telegram забезпечує користувачам кілька методів пошуку джерел отримання даних, які можуть допомогти знайти конкретну інформацію.

Приклади методів пошуку джерел даних в Telegram:

1. Повнотекстовий пошук за назвою джерела.
2. Пошук джерел за категоріями з використанням методу `get_suggested_broadcasts`.
3. Пошук джерел за ключовими словами з використанням методів `search_channel` та `search_groups`
4. Пошук джерел за хештегами груп та каналів з використанням методу `search_tags`.
5. Використання зовнішніх ресурсів, перехід з інших платформ.

Варто відзначити, що пошук у Telegram виконується локально на пристрої користувача, тому результати пошуку обмежені доступними на цьому пристрої даними, тобто потрібно від початку знати джерела, в яких пошук інформації має принести користь.

Окрім цього, Telegram надає відкрите API – прикладний програмний інтерфейс, що дозволяє розробникам створювати власні інструменти та додатки для здійснення більш точного та розширеного пошуку даних у групах та каналах.

Щодо конкретних джерел для збору даних, то тут варто враховувати особливості платформи та професійній сфері, в якій ці дані знаходяться. Telegram використовується для неофіційної комунікації багатьма угрупованнями, існує безліч каналів та груп, що є чудовим джерелом для збору інформації. Корисним також є можливість встановлювати зв’язки між різними спільнотами на основі їх взаємних репостів (тобто коли відбувається пересилання повідомлення між різними джерелами, поширюючи його вміст).

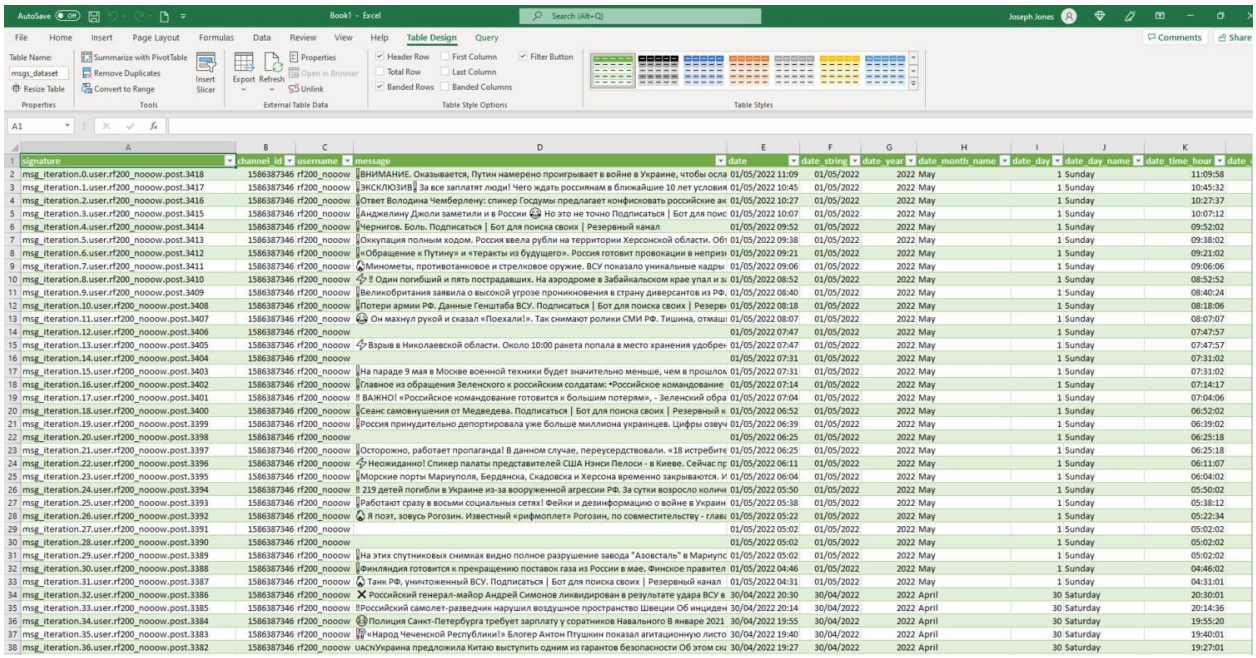
У першу чергу мною розглядаються власноруч знайдені агрегатори злитих даних, спільноти російських хакерських (у тому числі ‘хактивістських’) угруповань.

1.4.2 Методи збору даних

Приклади методів збору даних в Telegram:

1. Telegram Bot API – Telegram надає API для створення і використання ботів, які можуть збирати дані з різних джерел. Боти можуть підписуватись на певні групи та канали для моніторингу.
2. Web Scraping – якщо Telegram-клієнт не надає необхідну функціональність для збору даних, можна використовувати техніку

веб-скрапінгу, тобто аналіз HTML-коду сторінки, що відображається в Telegram, для витягування потрібної інформації. Зазвичай це робиться з використанням таких інструментів як BeautifulSoup або Selenium.



signature	channel_id	username	message	date	date_string	date_year	date_month_name	date_day	date_day_name	date_time_hour
msg_iteration.0.user.rf200_nooow.post.3418	1586387346	rf200_nooow	ВНИМАНИЕ. Оказывается, Путин намерено проигрывает в войне в Украине, чтобы осла	01/05/2022 11:09	01/05/2022	2022	May	1	Sunday	11:09:58
msg_iteration.1.user.rf200_nooow.post.3417	1586387346	rf200_nooow	ЭКСКЛЮЗИВ	За все заплатят люди! Чего ждать россиянам в ближайшие 10 лет условия	01/05/2022 10:45	01/05/2022	2022	May	1	Sunday	10:45:32
msg_iteration.2.user.rf200_nooow.post.3416	1586387346	rf200_nooow	Ответ Володина Чемберлену: спикер Госдумы предлагает конфисковать российские ак	01/05/2022 10:27	01/05/2022	2022	May	1	Sunday	10:27:37
msg_iteration.3.user.rf200_nooow.post.3415	1586387346	rf200_nooow	Анджелину Джоли заметили и в России Но это не точно Подписаться | Бот для поис	01/05/2022 10:07	01/05/2022	2022	May	1	Sunday	10:07:12
msg_iteration.4.user.rf200_nooow.post.3414	1586387346	rf200_nooow	Чернигов. Боль. Подписаться | Бот для поиска своих | Резервный канал	01/05/2022 09:52	01/05/2022	2022	May	1	Sunday	09:52:02
msg_iteration.5.user.rf200_nooow.post.3413	1586387346	rf200_nooow	Оккупация полным ходом. Россия ввела рубли на территории Херсонской области. Обі	01/05/2022 09:38	01/05/2022	2022	May	1	Sunday	09:38:02
msg_iteration.6.user.rf200_nooow.post.3412	1586387346	rf200_nooow	«Обращение к Путину» и «теракты из будущего». Россия готовит провокации в неприз	01/05/2022 09:21	01/05/2022	2022	May	1	Sunday	09:21:02
msg_iteration.7.user.rf200_nooow.post.3411	1586387346	rf200_nooow	Минометы, противотанковое и стрелковое оружие. ВСУ показало уникальные кадры	01/05/2022 09:06	01/05/2022	2022	May	1	Sunday	09:06:06
msg_iteration.8.user.rf200_nooow.post.3410	1586387346	rf200_nooow	‼ Один погибший и пять пострадавших. На аэродроме в Забайкальском крае упал и за	01/05/2022 08:52	01/05/2022	2022	May	1	Sunday	08:52:52
msg_iteration.9.user.rf200_nooow.post.3409	1586387346	rf200_nooow	Великобритания заявила о высокой угрозе проникновения в страну диверсантов из РФ.	01/05/2022 08:40	01/05/2022	2022	May	1	Sunday	08:40:24
msg_iteration.10.user.rf200_nooow.post.3408	1586387346	rf200_nooow	Потери армии РФ. Данные Генштаба ВСУ. Подписаться | Бот для поиска своих | Резерв	01/05/2022 08:18	01/05/2022	2022	May	1	Sunday	08:18:06
msg_iteration.11.user.rf200_nooow.post.3407	1586387346	rf200_nooow	Он махнул рукой и сказал «Поехали!». Так снимают ролики СМИ РФ. Тишина, отмаш	01/05/2022 08:07	01/05/2022	2022	May	1	Sunday	08:07:07
msg_iteration.12.user.rf200_nooow.post.3406	1586387346	rf200_nooow		01/05/2022 07:47	01/05/2022	2022	May	1	Sunday	07:47:57
msg_iteration.13.user.rf200_nooow.post.3405	1586387346	rf200_nooow	Взрыв в Николаевской области. Около 10:00 ракета попала в место хранения удобре	01/05/2022 07:47	01/05/2022	2022	May	1	Sunday	07:47:57
msg_iteration.14.user.rf200_nooow.post.3404	1586387346	rf200_nooow		01/05/2022 07:31	01/05/2022	2022	May	1	Sunday	07:31:02
msg_iteration.15.user.rf200_nooow.post.3403	1586387346	rf200_nooow	На параде 9 мая в Москве военной техники будет значительно меньше, чем в прошлол	01/05/2022 07:31	01/05/2022	2022	May	1	Sunday	07:31:02
msg_iteration.16.user.rf200_nooow.post.3402	1586387346	rf200_nooow	Главное из обращения Зеленского к российским солдатам: •Российское командование	01/05/2022 07:14	01/05/2022	2022	May	1	Sunday	07:14:17
msg_iteration.17.user.rf200_nooow.post.3401	1586387346	rf200_nooow	‼ ВАЖНО! «Российское командование готовится к большим потерям», - Зеленский обра	01/05/2022 07:04	01/05/2022	2022	May	1	Sunday	07:04:06
msg_iteration.18.user.rf200_nooow.post.3400	1586387346	rf200_nooow	Сеанс самовнушения от Медведева. Подписаться | Бот для поиска своих | Резервный к	01/05/2022 06:52	01/05/2022	2022	May	1	Sunday	06:52:02
msg_iteration.19.user.rf200_nooow.post.3399	1586387346	rf200_nooow	Россия принудительно депортировала уже больше миллиона украинцев. Цифры озвуч	01/05/2022 06:39	01/05/2022	2022	May	1	Sunday	06:39:02
msg_iteration.20.user.rf200_nooow.post.3398	1586387346	rf200_nooow		01/05/2022 06:25	01/05/2022	2022	May	1	Sunday	06:25:18
msg_iteration.21.user.rf200_nooow.post.3397	1586387346	rf200_nooow	Осторожно, работает пропаганда! В данном случае, переусердствовали. «18 истребите	01/05/2022 06:25	01/05/2022	2022	May	1	Sunday	06:25:18
msg_iteration.22.user.rf200_nooow.post.3396	1586387346	rf200_nooow	Неожиданно! Спикер палаты представителей США Нэнси Пелоси - в Киеве. Сейчас пр	01/05/2022 06:11	01/05/2022	2022	May	1	Sunday	06:11:07
msg_iteration.23.user.rf200_nooow.post.3395	1586387346	rf200_nooow	Морские порты Мариуполя, Бердянска, Скадовска и Херсона временно закрываются. И	01/05/2022 06:04	01/05/2022	2022	May	1	Sunday	06:04:02
msg_iteration.24.user.rf200_nooow.post.3394	1586387346	rf200_nooow	‼ 219 детей погибли в Украине из-за вооруженной агрессии РФ. За сутки возросло колич	01/05/2022 05:50	01/05/2022	2022	May	1	Sunday	05:50:02
msg_iteration.25.user.rf200_nooow.post.3393	1586387346	rf200_nooow	Работают сразу в восьми социальных сетях! Фейки и дезинформацию о войне в Украин	01/05/2022 05:38	01/05/2022	2022	May	1	Sunday	05:38:12
msg_iteration.26.user.rf200_nooow.post.3392	1586387346	rf200_nooow	Я поэт, зовусь Рогозин. Известный «рифмоплет» Рогозин, по совместительству - глав	01/05/2022 05:22	01/05/2022	2022	May	1	Sunday	05:22:34
msg_iteration.27.user.rf200_nooow.post.3391	1586387346	rf200_nooow		01/05/2022 05:02	01/05/2022	2022	May	1	Sunday	05:02:02
msg_iteration.28.user.rf200_nooow.post.3390	1586387346	rf200_nooow		01/05/2022 05:02	01/05/2022	2022	May	1	Sunday	05:02:02
msg_iteration.29.user.rf200_nooow.post.3389	1586387346	rf200_nooow	На этих спутниковых снимках видно полное разрушение завода "Азовсталь" в Мариупс	01/05/2022 05:02	01/05/2022	2022	May	1	Sunday	05:02:02
msg_iteration.30.user.rf200_nooow.post.3388	1586387346	rf200_nooow	Финляндия готовится к прекращению поставок газа из России в мае. Финское правител	01/05/2022 04:46	01/05/2022	2022	May	1	Sunday	04:46:02
msg_iteration.31.user.rf200_nooow.post.3387	1586387346	rf200_nooow	Танк РФ, уничтоженный ВСУ. Подписаться | Бот для поиска своих | Резервный канал	01/05/2022 04:31	01/05/2022	2022	May	1	Sunday	04:31:01
msg_iteration.32.user.rf200_nooow.post.3386	1586387346	rf200_nooow	❌ Российский генерал-майор Андрей Симонов ликвидирован в результате удара ВСУ в	30/04/2022 20:30	30/04/2022	2022	April	30	Saturday	20:30:01
msg_iteration.33.user.rf200_nooow.post.3385	1586387346	rf200_nooow	‼Российский самолет-разведчик нарушил воздушное пространство Швеции Об инциден	30/04/2022 20:14	30/04/2022	2022	April	30	Saturday	20:14:36
msg_iteration.34.user.rf200_nooow.post.3384	1586387346	rf200_nooow	Полиция Санкт-Петербурга требует зарплату у соратников Навального В январе 2021	30/04/2022 19:55	30/04/2022	2022	April	30	Saturday	19:55:20
msg_iteration.35.user.rf200_nooow.post.3383	1586387346	rf200_nooow	«Народ Чеченской Республики!» Блогер Антон Птушкин показал агитационную листо	30/04/2022 19:40	30/04/2022	2022	April	30	Saturday	19:40:01
msg_iteration.36.user.rf200_nooow.post.3382	1586387346	rf200_nooow	uacnУкраина предложила Китаю выступить одним из гарантов безопасности Об этом ск	30/04/2022 19:27	30/04/2022	2022	April	30	Saturday	19:27:01

Рисунок 1.5 – Приклад даних з Telegram, отриманих через скрапінг у .csv форматі

3. Telegram Client API – використання сторонніх бібліотек і інструментів для доступу до Telegram Client API. Збір даних відбувається шляхом програмного доступу до різних функції і властивостей платформи.

У процесі порівняння основних методів збору даних з Telegram, було прийнято, що використання Telegram Client API може повністю задовільнити потреби при виконанні поставленого завдання.

Для реалізації збору даних з необхідних джерел у Telegram було використано мову програмування Python v.3.10.11 для написання збирача (parser) даних і такі Python бібліотеки як telethon, pytz та datetime (ДОДАТОК А).

Сам збирач даних необхідно зареєструвати як застосунок, що буде працювати через Telegram API на офіційному сайті Telegram.

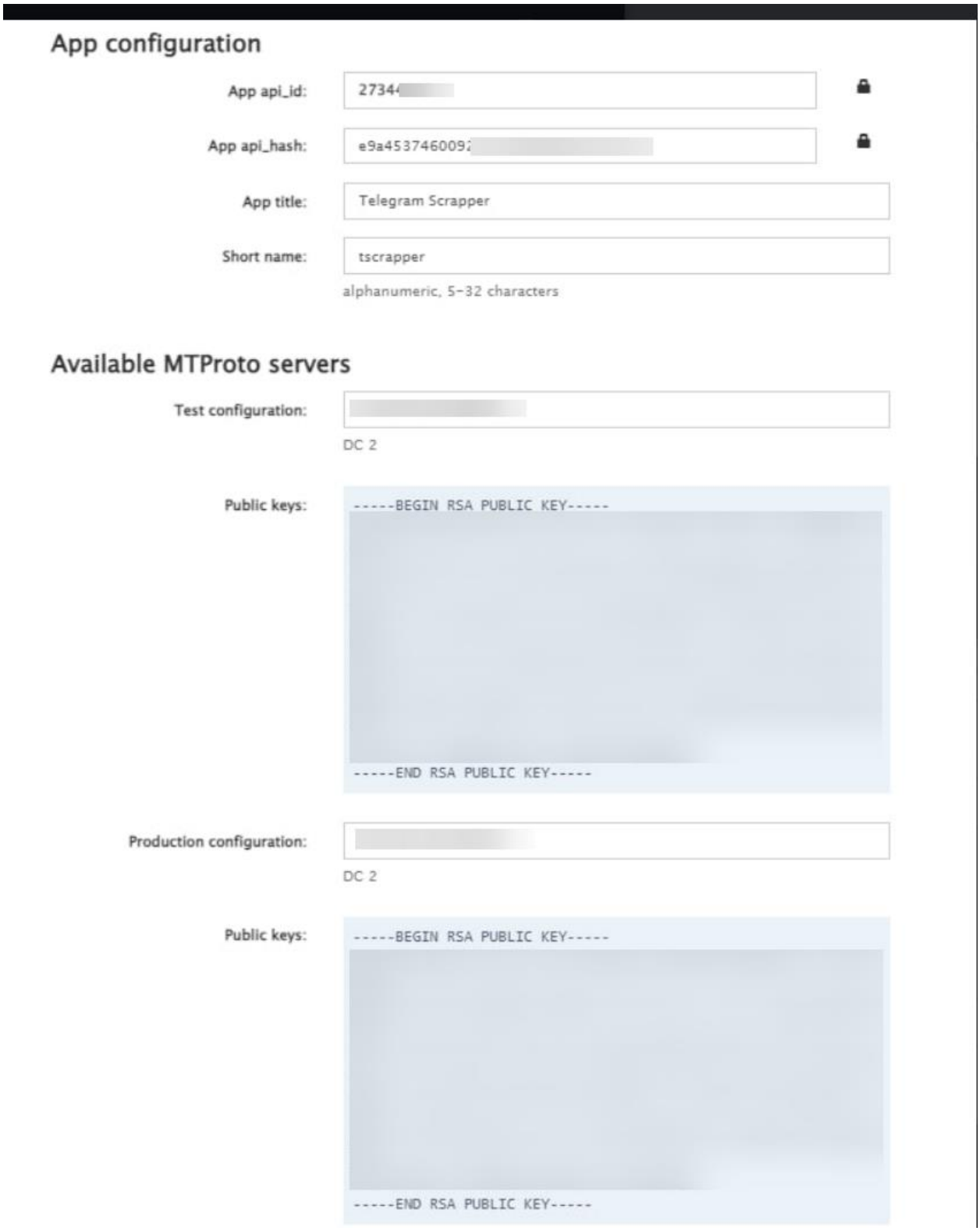


Рисунок 1.6 – Налаштування та загальна інформація про створену програму для роботи з Telegram API

Бібліотека telethon надає доступ до Telegram Client API, її основні можливості включають:

- Авторизація та автентифікація користувача в Telegram та виконання необхідних дій під їх профілем.
- Ефективне отримання даних (повідомлень) з різних джерел платформи Telegram.
- Робота з групами та канали.
- Робота з медіафайлами

Завдяки чому було отримано такі корисні дані з джерел як:

- Channel ID – унікальний ідентифікатор каналу або групи.
- Channel Name – унікальне створення ім’я каналу або групи.
- Message Text – тіло повідомлення, яке було зібрано.
- Date Created – дату публікації повідомлення.
- Attached Media – вивід інформації про додані до повідомлення медіафайли, такі як .pdf, .docx, .csv та інші, що можуть містити корисну інформацію як наприклад передані дані або матеріал зливу.

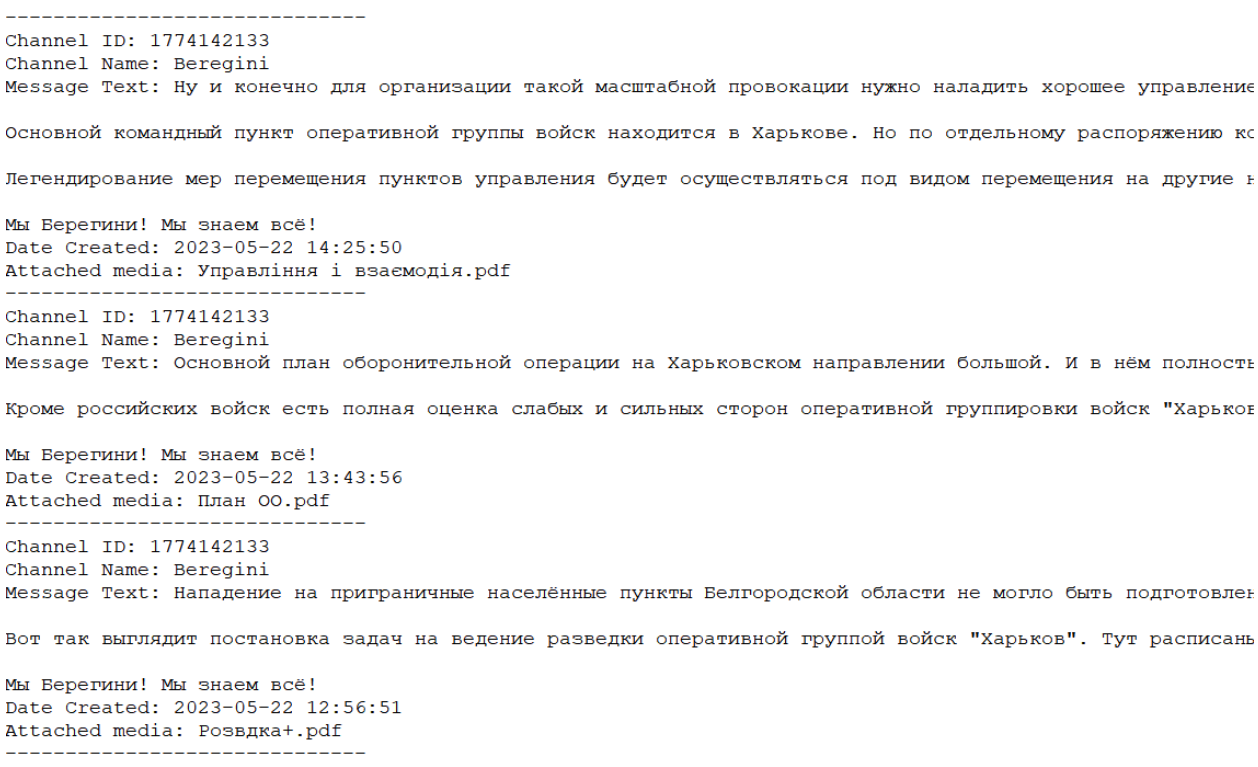


Рисунок 1.7 – Приклад зібраних парсером даних з Telegram у .txt форматі

Програма успішно збирає необхідні за заданий проміжок часу повідомлення з потрібного каналу або групи і розподіляє їх відповідно до зазначених у програмі типів даних.

Даний парсер повністю покриває задачу збору інформації, але повноцінно працювати з таким масивом даних є неможливим, адже виникають наступні проблеми:

1. При публікації одного повідомлення з декількома медіафайлами картинками (формату .jpg, .png тощо), парсер записує цю інформацію як декілька різних повідомлень.

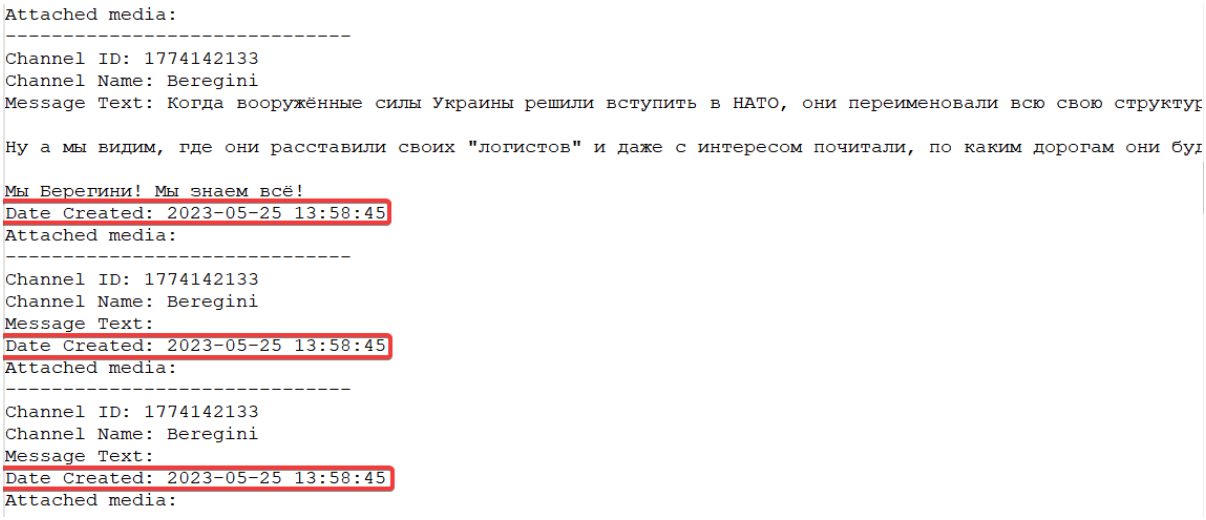


Рисунок 1.8 – Приклад зібраних парсером даних з Telegram, де видно запис пустих повідомлень, які насправді являються картинками, доданими до публікації

2. Формат даних на виході з парсеру - .txt, що унеможливлює повноцінний аналіз зібраних даних.
3. Формування повноцінних пустих блоків, що не несуть ніякої корисної інформації, якщо публікація з себе представляла лише медіафайли картинки

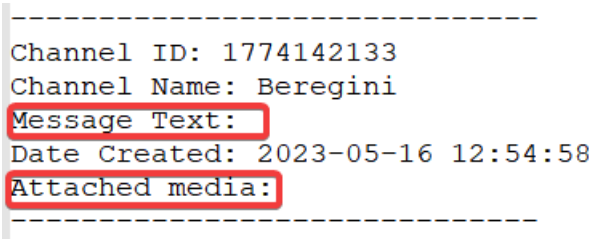


Рисунок 1.9 – Приклад пустого блоку

На виході з програми ми отримуємо великі дані, що неможливо якісно структурувати, обробити, проаналізувати та візуалізувати. Виходячи з цього, наступною задачею буде обробка отриманих даних, приведення їх до зручного формату для подальшого завантаження у системи введення, зберігання, пошуку та опрацювання даних.

1.4.3 Методи обробки даних

Обробка даних — це весь процес перетворення, тобто очищення або приведення даних у придатний для використання стан. Після проведення збору даних на попередньому етапі ми отримали так звані сирі дані, вони не мають зручного для аналізу формату. Чисті ж дані це дані, отримані після обробки сирих даних, та готові до аналізу. Непослідовні або відсутні дані виправляються або видаляються, самі дані перетворюються в більш зручний формат.

Основні задачі, які має вирішити обробка даних, це:

1. Виявлення помилок користувача на етапі введення даних або при читанні попередньо отриманого файлу з даними.
2. Перевірка на відсутність вхідних даних.
3. Проблеми, пов’язані з конкретним ПЗ.
4. Проблеми, напряму пов’язані з обробкою даних: вилучення, перетворення, форматування тощо.

Формат JSON (JavaScript Object Notation) є представленням структурованих даних у текстовому вигляді, який є читабельним для людини і комп’ютеру водночас. Він використовується для передачі даних між різними програмними мовами та платформами, оскільки сам формат не залежить від конкретної мови програмування.

JSON є зручним форматом для представлення складних структур даних. Він складається з двох основних структур:

- Набору пар назва-значення.
- Впорядкованого списку значень.

У JSON використовуються такі їхні форми:

1. Об’єкт — послідовність пар назва-значення.
2. Масив — послідовність змінних.
3. Змінна — може бути рядком, числом, логічним оператором (True\False), null, об’єктом, масивом.
4. Рядок — послідовність з нуля або більше символів юнікода.

Даний формат є зручним для завантаження в системи введення, зберігання, пошуку та опрацювання даних, що будуть розглядатися в наступному підпункті.

Для реалізації обробки даних, отриманих за допомогою парсеру з каналів платформи Telegram, було використано мову програмування Python v.3.10.11 і таку Python бібліотеки як json для переведення даних до повноцінного JSON формату (ДОДАТОК Б).

Завдання, які виникли після етапу збору даних та які було вирішення на етапі обробки даних:

1. Видалення блоків, в яких водночас поля ‘Message Text’ та ‘Attached Media’ є пустими. Наявність подібних блоків до обробки свідчить про те, що канал опублікував медіафайл картинку.

```
------------------------------
Channel ID: 1774142133
Channel Name: Beregini
Message Text:
Date Created: 2023-05-16 12:54:58
Attached media:
------------------------------
Channel ID: 1774142133
Channel Name: Beregini
Message Text:
Date Created: 2023-05-16 12:54:58
Attached media:
------------------------------
```

Рисунок 1.10 – Приклад пустих блоків, що було повністю видалено

2. Переформатування .txt у .json для подальшої роботи з більш зручним форматом.


```
[
  {
    "Channel ID": 1774142133,
    "Channel Name": "Beregini",
    "Message Text": "А сегодня мы выложим ещё п
    "Date Created": "2023-05-30 14:50:35",
    "Attached media": "СХЕМА розподілу сил та з
  },
  {
    "Channel ID": 1774142133,
    "Channel Name": "Beregini",
    "Message Text": "",
    "Date Created": "2023-05-30 14:50:35",
    "Attached media": "ПВД №27.pdf"
  },
  {
    "Channel ID": 1774142133,
    "Channel Name": "Beregini",
    "Message Text": "Мы продолжаем публиковать
    "Date Created": "2023-05-29 13:50:55",
    "Attached media": "Днепр.pdf"
  },
```


Рисунок 1.11 – Приклад переформатованих у JSON даних

3. Підрахунок кількості строк, що було очищено на етапі обробки та кількість записаних строк у новий результуючий JSON файл.

1.5 Огляд існуючих систем введення, зберігання, пошуку та опрацювання даних для поставлених задач

ElasticSearch, Manticore Search та PostgreSQL є популярними системами для роботами з великими обсягами даних, але вони мають чимало відмінностей у своїх можливостях та підходах до роботи з даними, тому вибір між цими системами має чималу роль для якісного та повноцінного виконання завдання.

1.5.1 Elasticsearch

Рисунок 1.12 – Elasticsearch

Elasticsearch — це розподілена інформаційно-пошукова система, яка призначена для швидкої обробки великих масивів неструктурованих або погано структурованих даних та орієнтована на документи, а саме CSV або JSON документи. Elasticsearch проводить початоку обробку даних, індексує їх та зберігає, що призводить до прискорення процесу пошуку інформації.

Elasticsearch використовує потужну пошукову бібліотеку Apache Lucene для повнотекстового пошуку та індексування даних. Lucene працює на основі оберненого індексу, що дозволяє швидко знаходити документи, які містять певні слова або фрази.

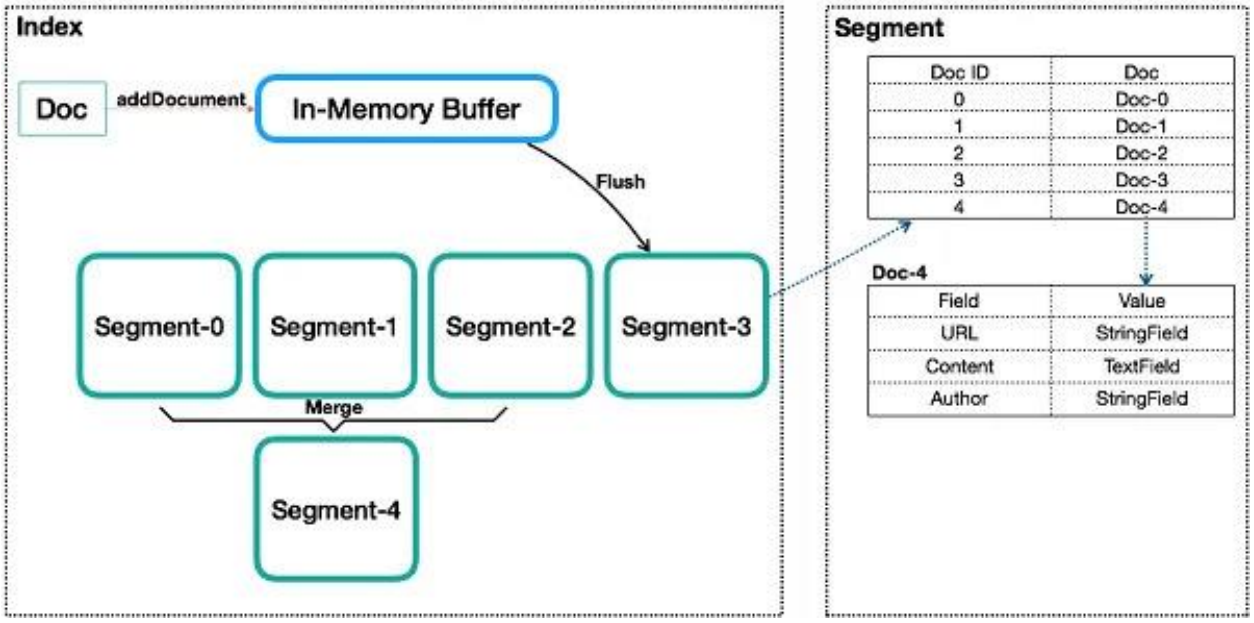


Рисунок 1.13 – Інвертована індексація у Elasticsearch

Основні особливості Elasticsearch:

1. Open-Source.
2. Швидкодія завдяки пошуковій бібліотеці Apache Lucene.
3. Можливість обробки неструктурованих даних.
4. Підтримка повнотекстового пошуку.
5. Підтримка вбудованого аналізу даних.
6. Підтримка REST API.
7. Просте керування та масштабованість системи.
8. Стійкість до внутрішніх помилок.

1.5.2 Manticore Search

Рисунок 1.14 – Manticore Search

Manticore Search є форком Sphinx Search, повнотекстової пошуково-аналітичної бази даних. Вона розроблена для забезпечення високої продуктивності як на малих, так і на великих обсягах даних без необхідності заглиблюватися в них. Основні особливості Manticore Search:

1. Open-Source
2. Швидкодія навіть у порівнянні з Elasticsearch. Завдяки оптимізованому search engine, Manticore Search може обробляти великі обсяги даних і забезпечувати швидких доступ до результатів пошуку. Manticore Search вміє розпаралелювати пошуковий запит на всі ядра за замовчуванням, що дає змогу ефективно завантажити ядра процесора. Manticore Search є істотно швидшим при завантаженні та обробці великих масивів даних, наприклад датасет на 1.7млрд документів завантажується в 25 разів швидше ніж в Elasticsearch.

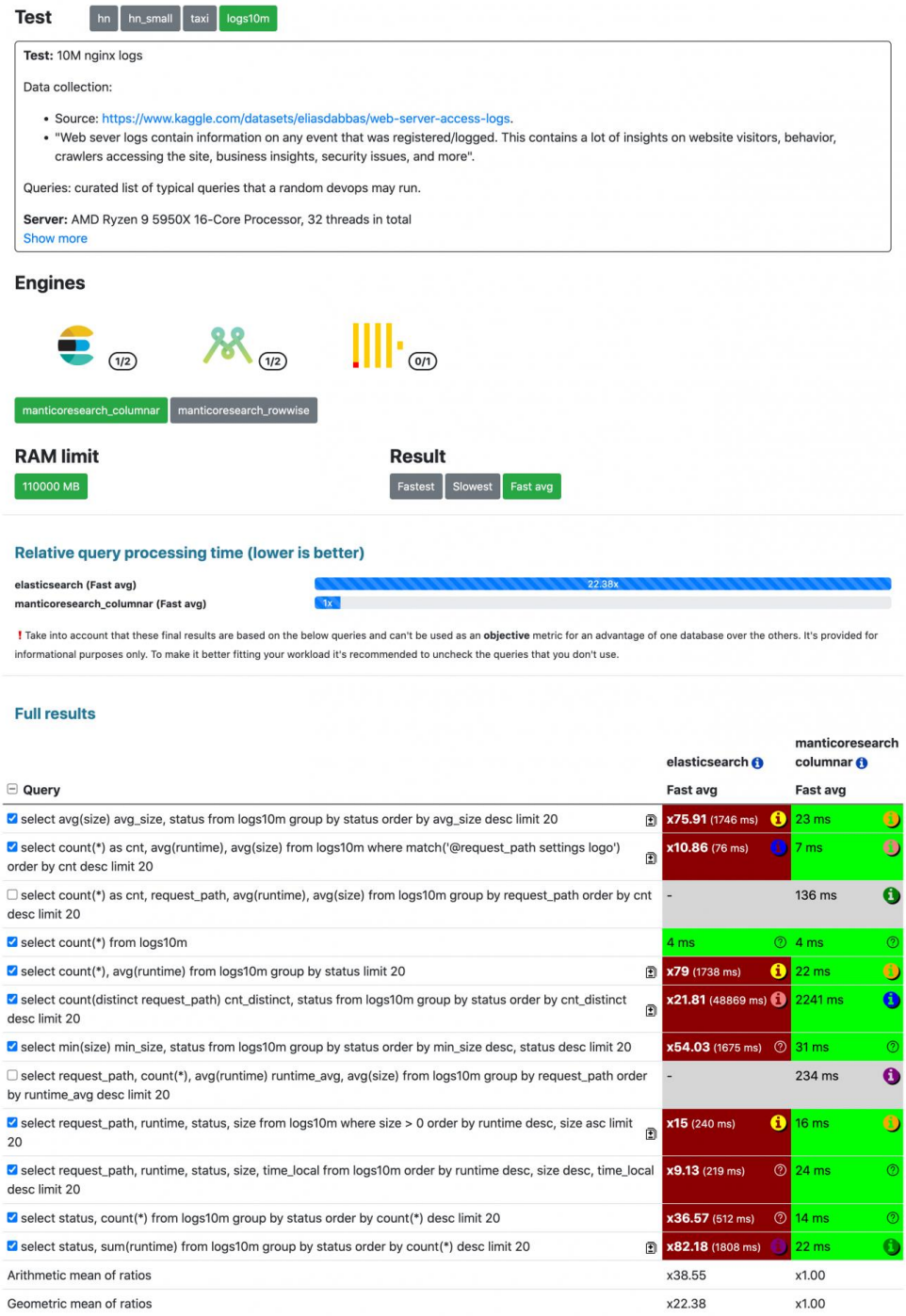


Рисунок 1.15 – Порівняння швидкодії Manticore Search та Elasticsearch на прикладі аналізу 10млн логів Nginx, Manticore Search є у 22 рази швидшим

3. Підтримка повнотекстового пошуку.
4. Підтримка розширених пошукових запитів (операторів, фільтрів, вагів, проміжок дат тощо).

5. Підтримка розподіленого пошуку, що дозволяє розподілити дані і обробки запитів між кількома вузлами для масштабованості і продуктивності системи.
6. Підтримка розширених функцій ранжування, як наприклад BM25, TF-IDF, векторого пошуку тощо.
7. Підтримка реплікації.
8. Інтеграцій з різними джерелами даних, як наприклад CSV, XML, JSON формати тощо. Наявність JSON інтерфейсу.
9. Нативна підтримка інших СКБД, як наприклад MySQL або PostgreSQL.
10. Підтримка автоматичної ідентифікації (auto id).
11. Наявність сховища документів.
12. Можливість створювати просунуті real-time індекси.

1.5.3 PostgreSQL

Рисунок 1.16 – PostgreSQL

PostgreSQL є однією з найбільш популярних реляційних систем керування базами даних. Основні особливості PostgreSQL:

1. Open-Source.
2. Повноцінна СКБД, що включає широкий спектр функції (індексація, оптимізація запитів, резервне копіювання тощо).

3. Сумісна з різними платформами та проміжним ПЗ
4. Підтримка багатоверсійного контролю.
5. Можливість створювати триггери.
6. Транзакційна цілісність — підтримка ACID (Atomicity, Consistency, Isolation, Durability).
7. Сумісність зі стандартом ANSI-SQL а також об’єкто-орієнтовність з ANSI-SQL2008.
8. Підтримка JSON формату.
9. Можливість зв’язуватись з іншими сховищами даних, як наприклад NoSQL.

1.5.4 Вибір системи введення, зберігання, пошуку та опрацювання даних

З-поміж запропонованих систем було обрано Elasticsearch оскільки вона виконує більшу кількість необхідних функції та\або відсутності потреби в таких функціях як:

1. Спеціалізація на повнотекстовому пошуку.
2. Входження до стеку ELK, можливість нативного якісного поєднання з інструментом Kibana для візуалізації та моніторингу.
3. Гнучкість та розширюваність архітектури.
4. Нативна робота з даними формату JSON та CSV.
5. Підтримка комплексів запитів та аналітики.
6. Підтримка REST API.
7. Відсутність потреби в зв’язаних даних та реляційній моделі (у порівнянні з PostgreSQL).

8. Широка підтримка командою розробників та учасників спільноти (у порівнянні з Manticore Search).
9. Відсутність потреби у надшвидкій обробці даних великих масивів даних.

Висновки до розділу 1

У першому розділі дипломної роботи було розглянуто важливість та необхідність використання методології OSINT для моніторингу даних у соціальних мережах, обґрунтовано актуальність соціальних мереж як надійних джерел інформації у контексті збору інформації щодо проведення операцій в кіберпросторі.

Було розглянуто основні кіберінциденти та їх наслідки в контексті російсько-української гібридної війни. Проведено пошук джерел якісної інформації у Telegram та розроблено інструмент для збору даних з платформи. Зібрані з Telegram дані було також оброблено й очищено іншим інструментом, сформовано дані у JSON форматі для подальшого аналізу.

У контексті вибору інструменту для введення, зберігання, пошуку та опрацювання даних було обрано інструментів Elasticsearch, адже він дозволяє виконувати поставлені задачі більш ефективно ніж інші зазначені у цій роботі.

2 ОПИС ЗАПРОПОНОВАНОГО ПІДХОДУ

2.1 Завантаження та пошук даних у системі Elasticsearch з використанням інструменту Kibana

Стек інструментів ELK (Elasticsearch, Logstash, Kibana) було створено для зберігання та обробки великих масивів даних. Однак саме ядро — Elasticsearch – можна використовувати як окремо, так і в поєданні з іншими інструментами стеку.

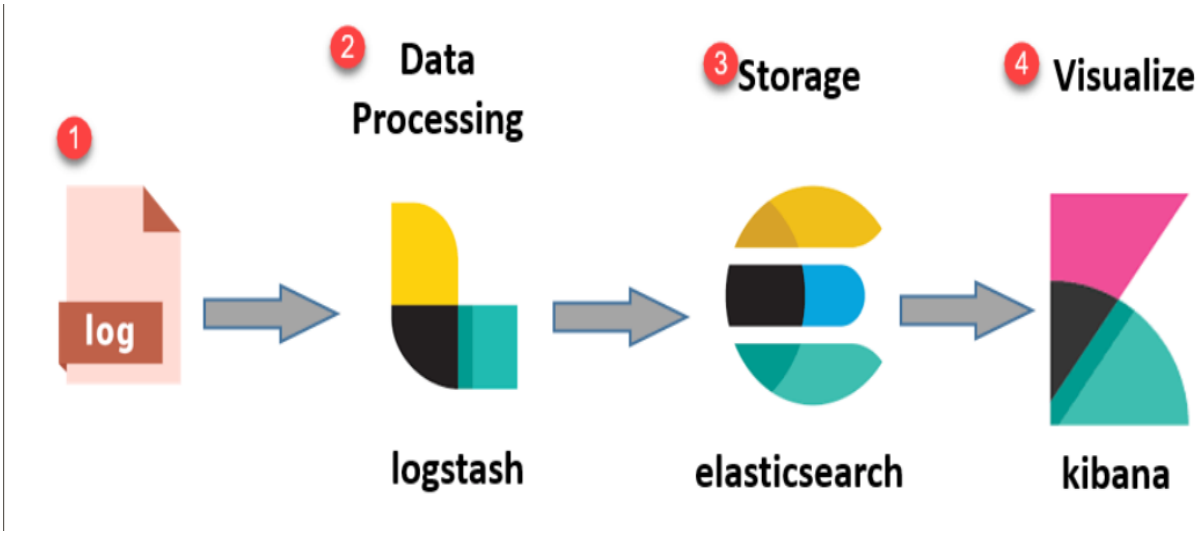


Рисунок 2.1 – Архітектура стеку ELK

Для виконання поставлених задач було обрано використання лише двох інструментів зі стеку ELK – Elasticsearch як пошукова система та Kibana як інструментів візуалізації.

Kibana використовується в ролі користувацького інтерфейсу Elasticsearch, інструменту керування та візуалізації даних. Він надає зручний інтуїтивний інтерфейс і при цьому не обрізає можливостей Elasticsearch.

Інструменти Elasticsearch та Kibana було встановлено на домашній дистрибутив Ubuntu 20.04 LTS для подальшої роботи з localhost. Вибір

зумовлено підтримкою та стабільністю дистрибутиву, широким вибором пакетів для встановлення та керування Elasticsearch та Kibana, зручним управлінням пакетів за допомогою APT (Advanced Packaging Tool) та сумісність з іншими інструментами. Після встановлення та налаштування обох інструментів можемо зайти на localhost:5601 та localhost:9200 для перегляду веб-інтерфейсу Kibana та Elasticsearch відповідно.

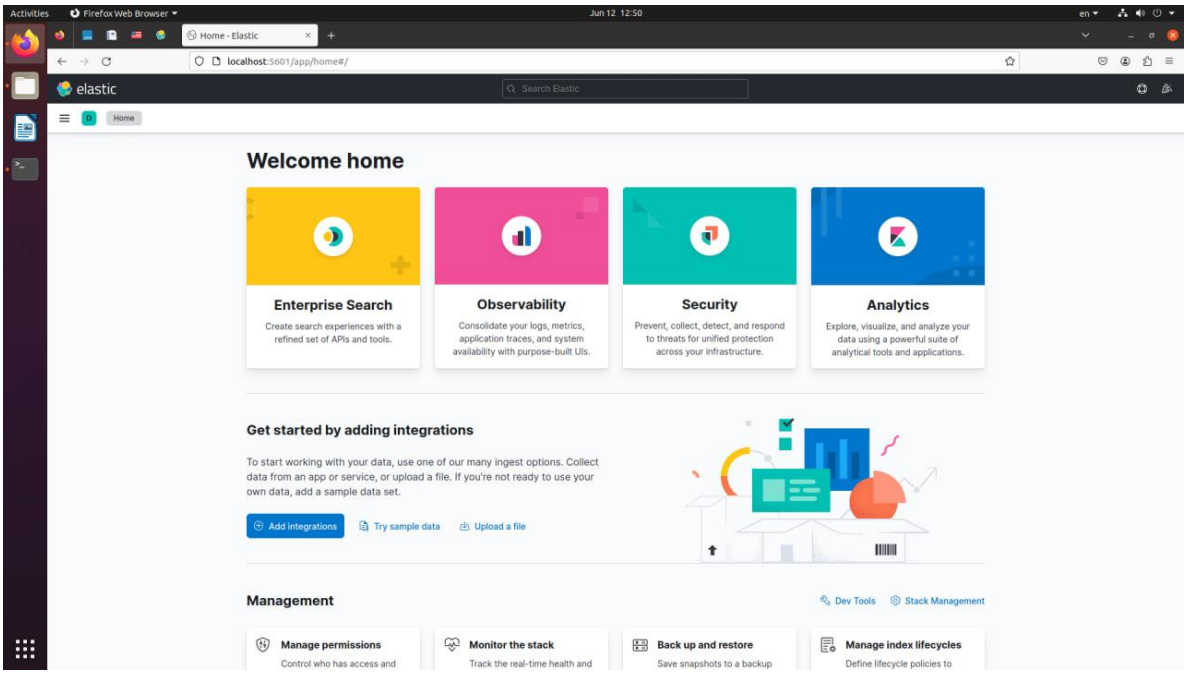


Рисунок 2.2 – Вигляд панелі ‘Home’ у Kibana

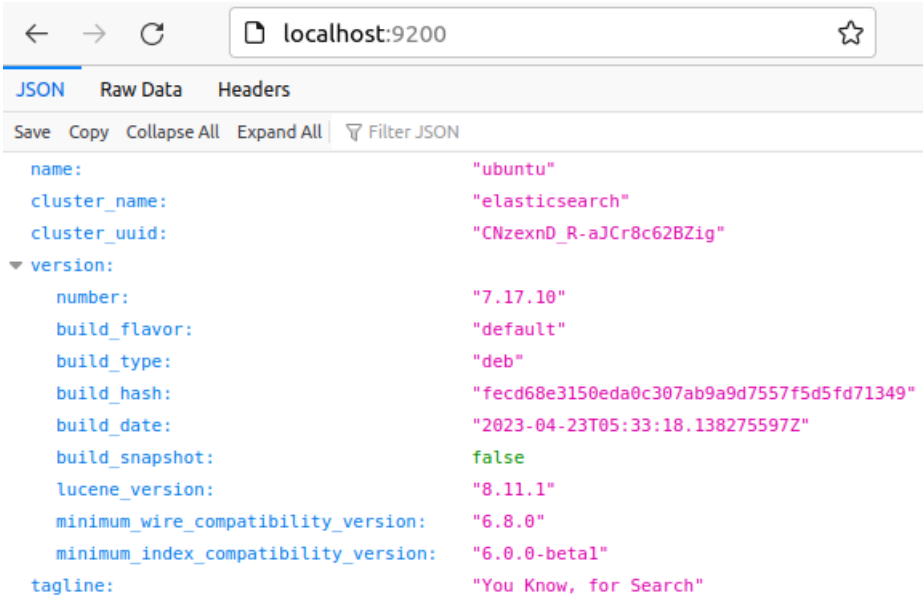


Рисунок 2.3 – Інформація від Elasticsearch про свій стан та конфігурацію

У Elasticsearch дані зберігаються у вигляді розподіленого індексу — контейнеру, в якому знаходяться документи одного типу. Кожен індекс складається з одного або кількох шардів — фізичних фрагментів даних, які розподілено по вузлах кластера Elasticsearch.

Запити до Elasticsearch можна робити напряму через Kibana Console, що знаходиться у розділі Dev Tools. Kibana Console взаємодіє з Elasticsearch за допомогою API та дозволяє виконувати різні операції, такі як створення, оновлення, видалення документів, виконання пошукових запитів тощо. Підтримуються різні HTTP-методи, такі як GET, POST, DELETE, PUT тощо.

Існує декілька способів завантаження даних у Kibana які варто розглянути:

1. Завантаження даних через панель ‘Upload File’. Підтримувані формати — JSON (з переведенням рядка), CSV, TSV, файли журналів із загальним форматом timestamp. Розмір даних до 100мб.

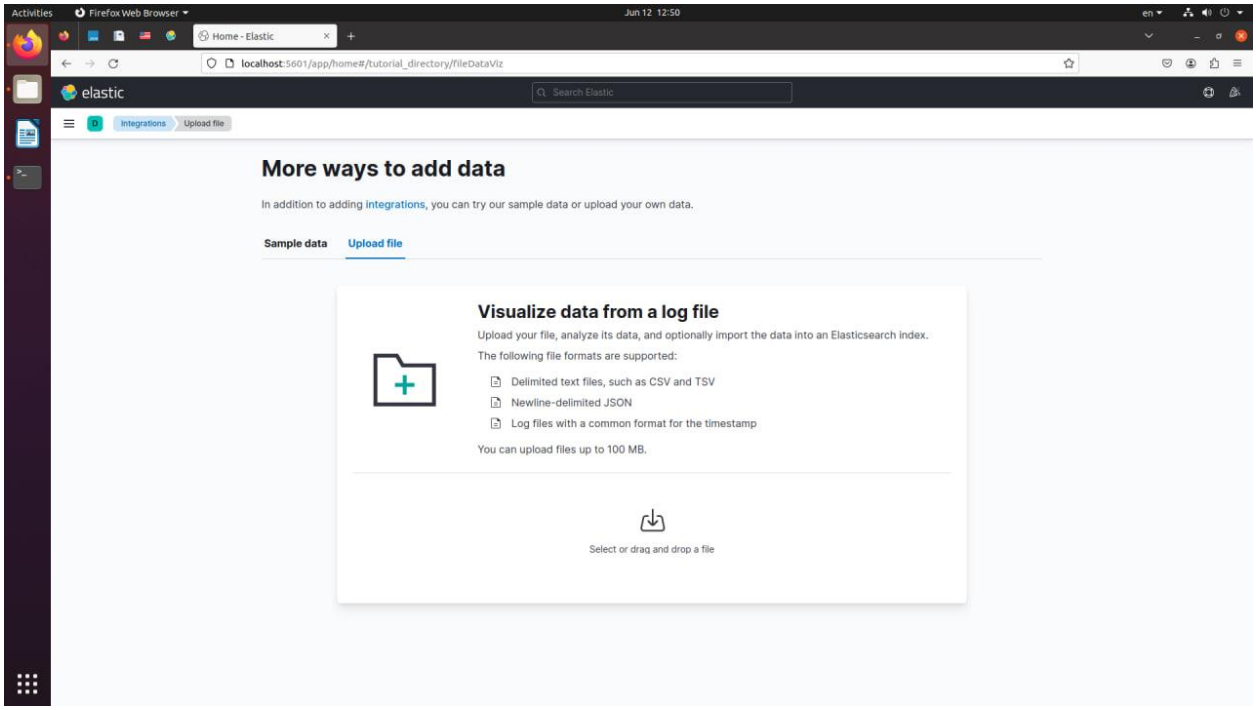


Рисунок 2.4 – Імпорт файлу через панель ‘Upload File’ у Kibana

2. Завантаження даних через cURL POST запит. *curl -XPOST "http://localhost:9200/*index_name*/_doc/_bulk" -H "Content-Type: application/json" --data-binary "@*.json"*

3. Завантаження даних через REST API, доступ до якого надається напряму через панель ‘Dev Tools > Console’ у Kibana.

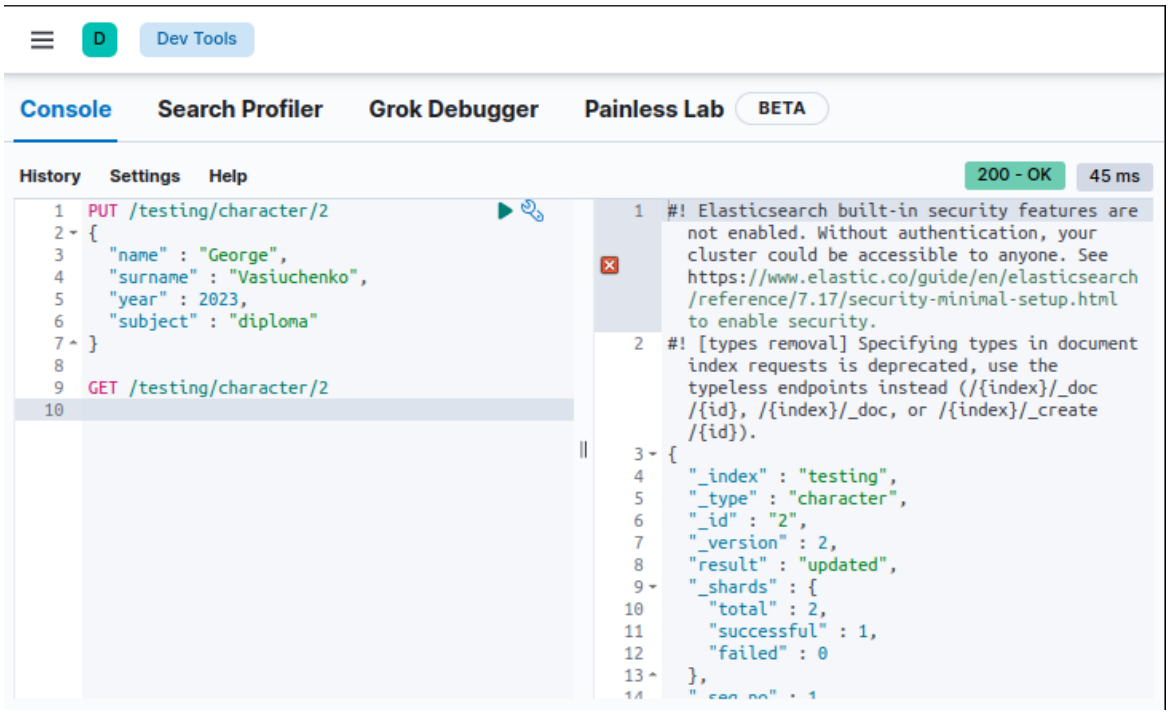


Рисунок 2.5 – Імпорт документу через консоль Kibana за допомогою методу PUT

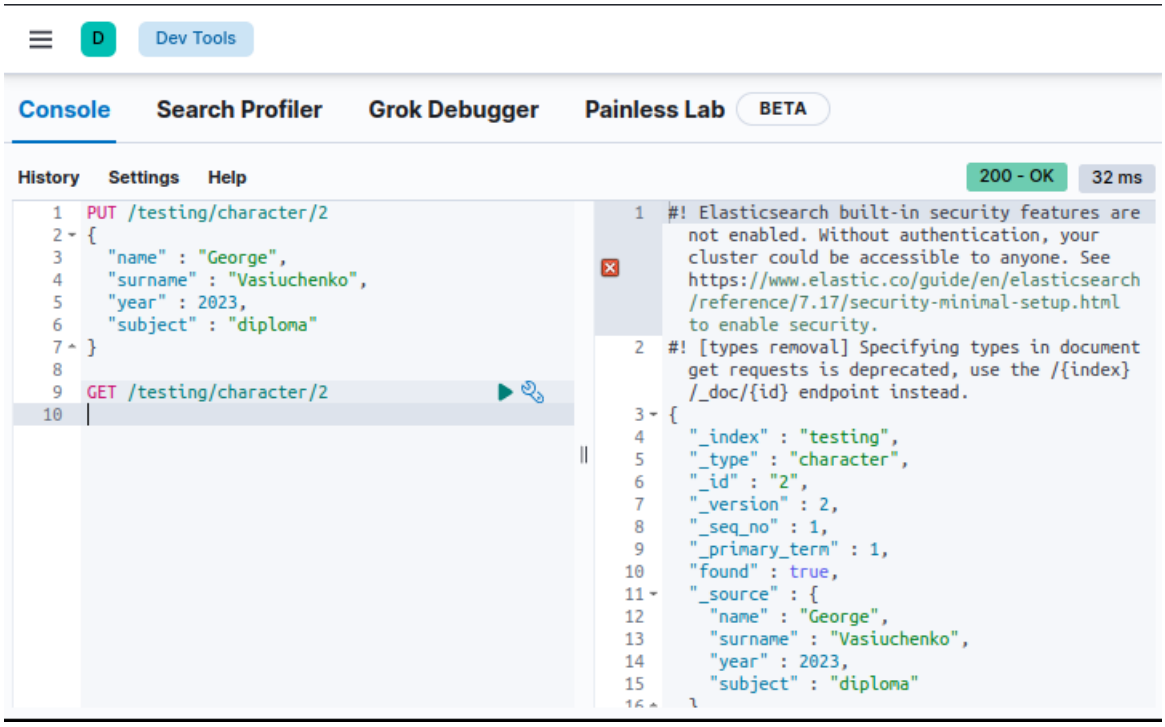


Рисунок 2.6 – Перегляд документу через консоль Kibana за допомогою методу GET

Оскільки ми використовуємо Elasticsearch та Kibana без інструменту Logstash, то за обробку та передачу різноманітних типів даних до Elasticsearch для подальшого збереження можемо використати механізм GROK.

GROK це шаблонний механізм для розбору та структурування неструктурованих даних. За допомогою нього можна будувати власні шаблони, які відповідають вхідним даним, та, відповідно до шаблону, витягувати корисну інформацію з цих даних. GROK базується на регулярних виразах та використовує вбудовані шаблони для визначення різних типів даних як наприклад IP-адреси, дати, час тощо. За допомогою GROK можна власноруч витягувати окремі поля та нормалізувати значення.

Приклад GROK шаблону, для перевірки структури масиву даних у JSON форматі. Перевірити відповідність даних до шаблону можна за допомогою інструменту для тестування GROK шаблонів у самій Kibana – GROK Debugger.

\[

\s*\{

\s*"Channel ID": %{NUMBER:channel_id},

\s*"Channel Name": "%{DATA:channel_name}",

\s*"Message Text": "%{GREEDYDATA:message_text}",

\s*"Date Created": "%{YEAR}-%{MONTHNUM}-%{MONTHDAY} %{HOUR}:%{MINUTE}:%{SECOND}",

\s*"Attached media": "%{DATA:attached_media}"

\s*\}

\s*\]

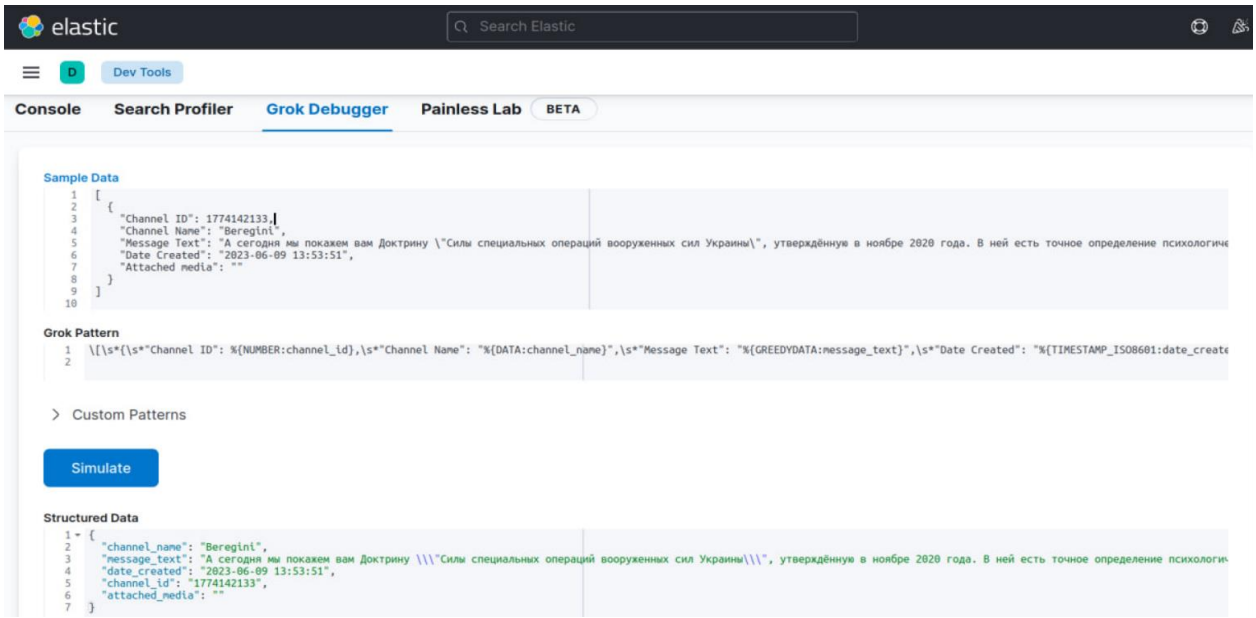


Рисунок 2.7 – Використання GROK Debugger у Kibana для перевірки структури масиву даних у JSON файлі

Меню імпорту файлу до Elasticsearch через панель ‘Upload File’:

1. Підраховується кількість строк у файлі. Визначається структура даних завантажуваного файлу, формат поля дати та часу.

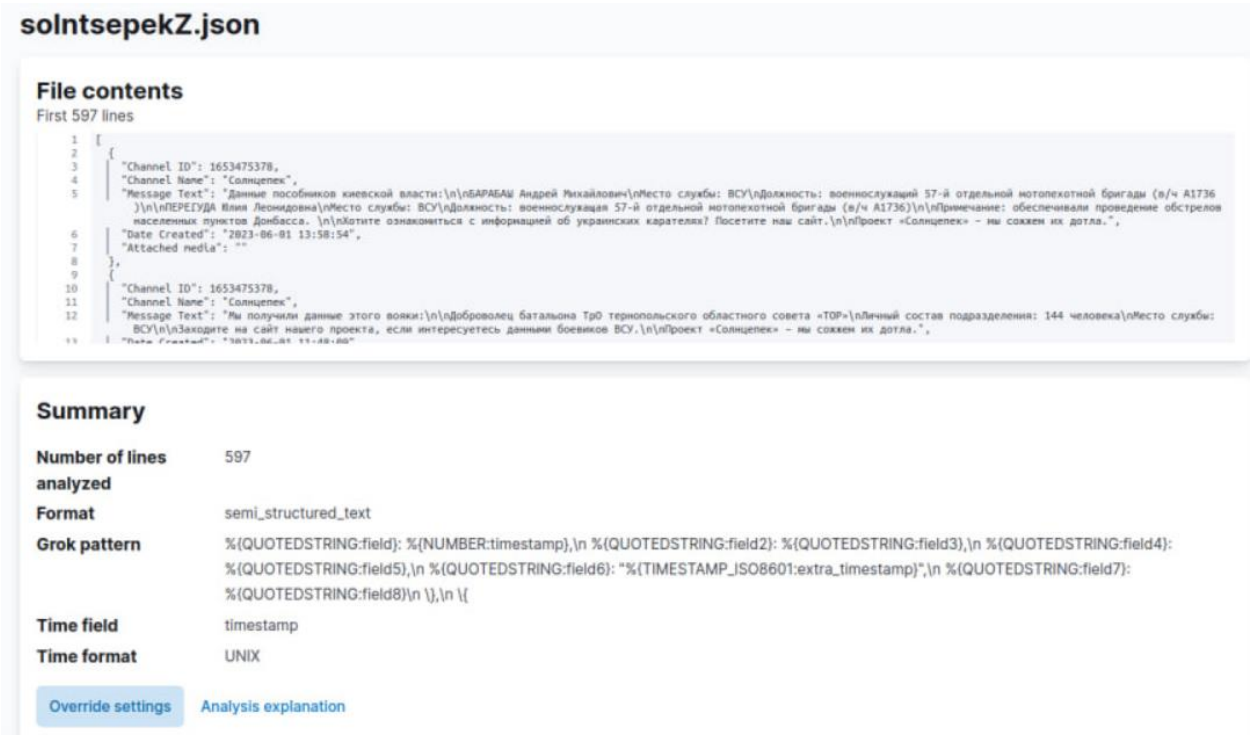


Рисунок 2.8 – Завантаження та автоматичний аналіз структури файлу

2. Здійснюється автоматичний підбір шаблону GROK. Також є можливість додавання власних шаблонів.

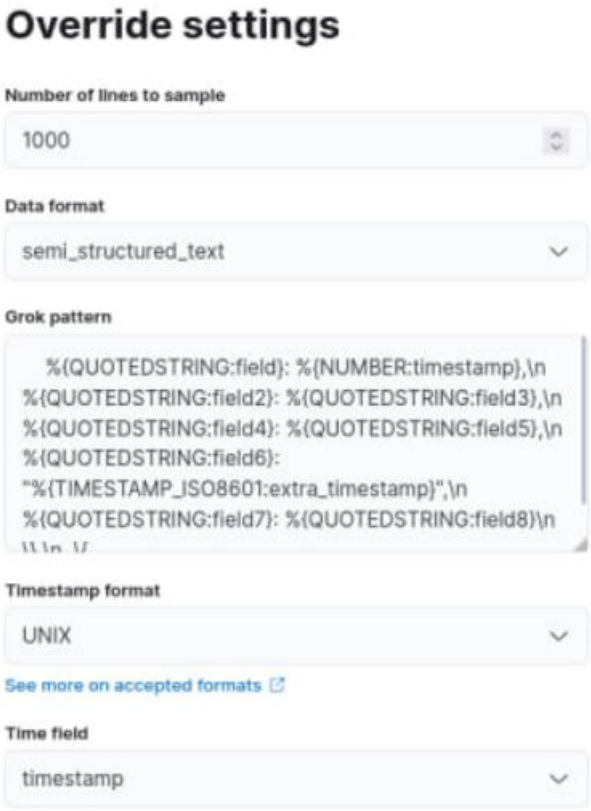


Рисунок 2.9 – Можливість редагування шаблону GROK

3. Розпізнавання структури файлу та полів у ньому.

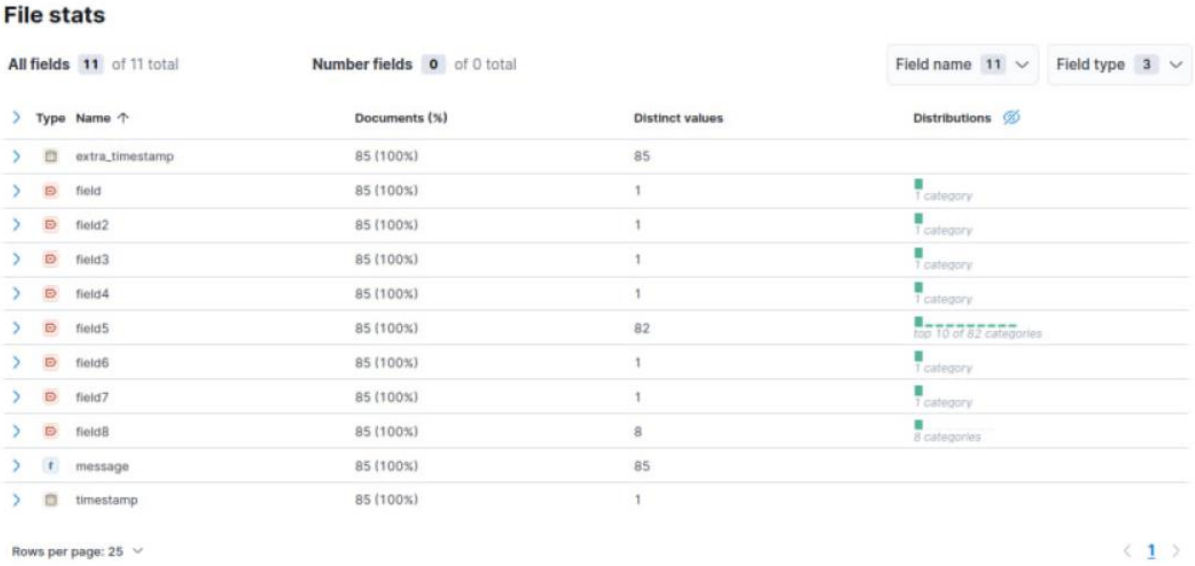


Рисунок 2.10 – Можливість редагування шаблону GROK завантажуваного файлу

Після визначення цих даних та натискання кнопки імпорту, запускається процес імпортування, який складається з наступних кроків:

1. Processing file – перетворення даних у документи NDJSON, щоб їх можна було отримати за допомогою bulk api. Кожен рядок стає окремим документом у форматі JSON.

2. Creating index – створення індексу (або обирання вже створеного індексу). Визначення параметрів індексу (кількість реплік, шардів).

3. Creating ingest pipeline – створення конвеєра для перед-обробки даних перед їх індексуванням. Загалом це парсинг, нормалізація та збагачення даних.

4. Uploading the data – завантаження даних до обраного індексу. Дані надсилаються в індекс за допомогою bulk api.

5. Creating index pattern — створення шаблону індексу Kibana (опціонально). Цей крок визначає розпізнавання та інтерпретування даних.

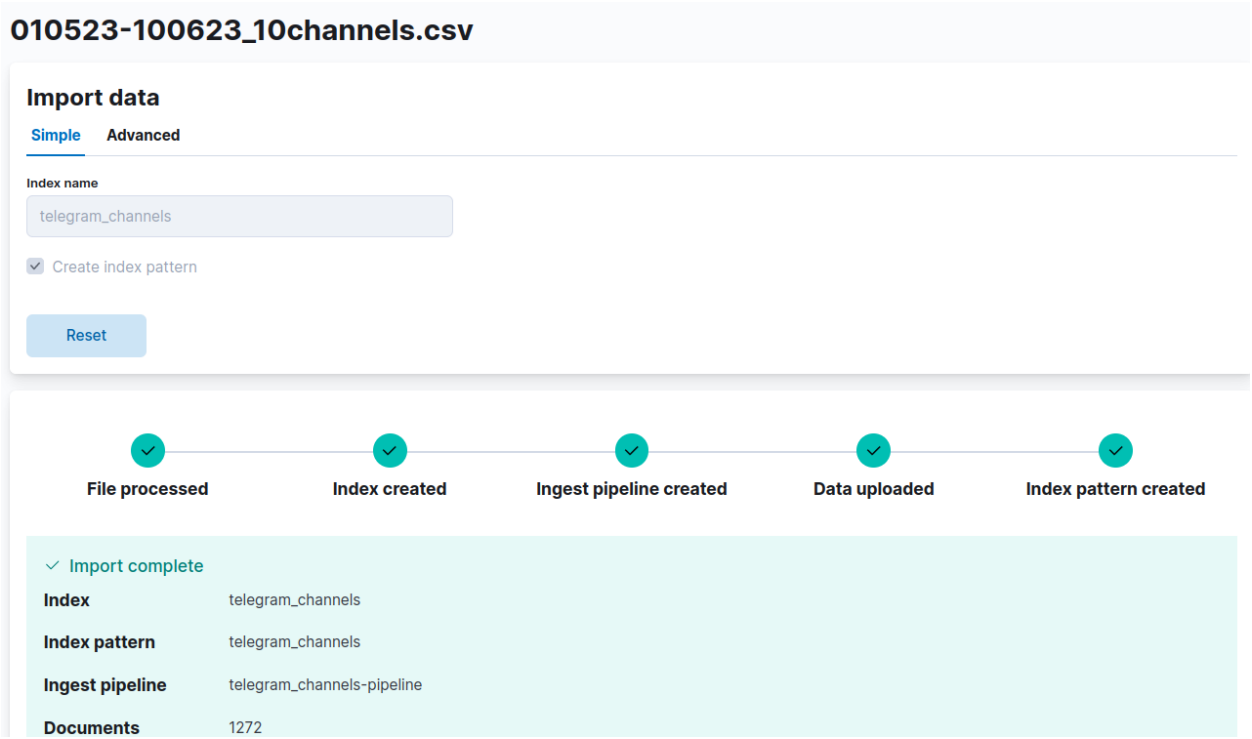


Рисунок 2.11 – 5 кроків завантаження файлу у форматі .csv до Elasticsearch через Kibana

Для інтерактивного пошуку та аналізу даних, окрім як через стандартну консоль яка приймає команди REST API, використовують інструмент Discover. Одразу після імпорту файлу можна перейти на панель Discover та почати перші операції з пошуку та аналізу завантажених даних.

Discover дозволяє проводити пошук одразу по всім полям, з фільтром, зі вказанням поля, зі вказанням значень у декількох полях та інші види пошуку. Також є можливість сортувати значення за спаданням або зростанням одразу в інтерфейсі.

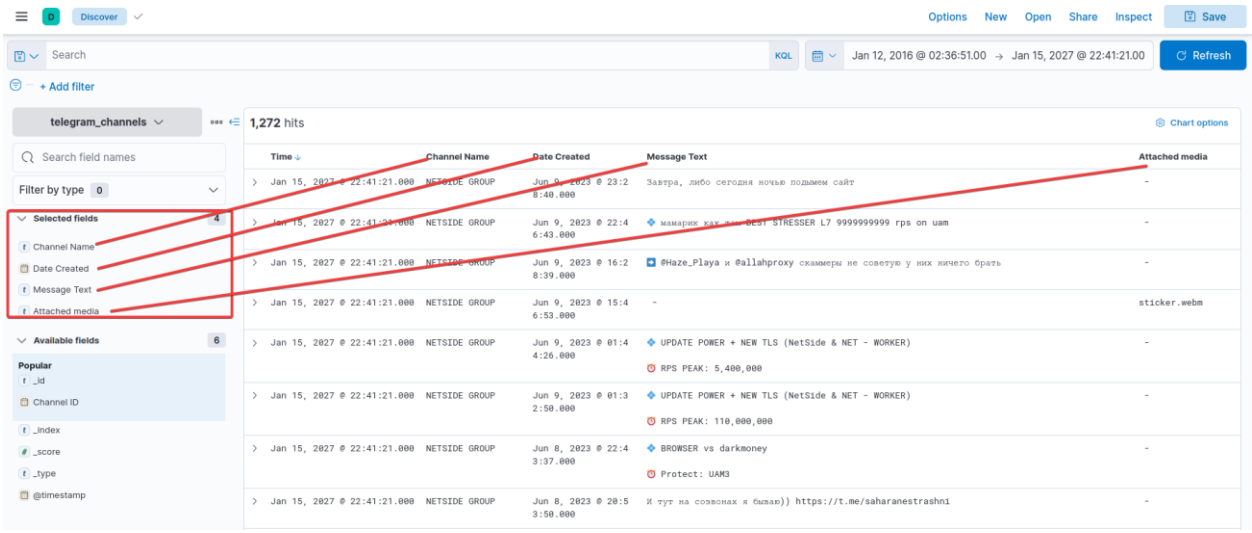


Рисунок 2.12 – Вікно Kibana Discover

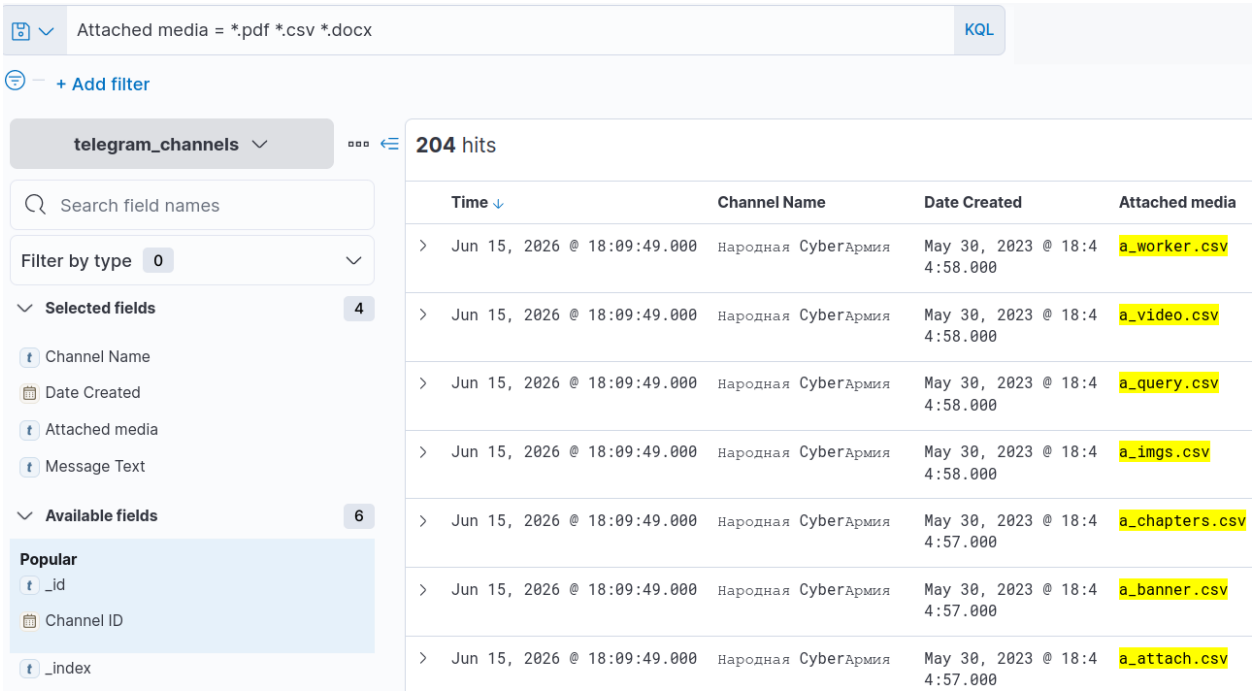


Рисунок 2.13 – Приклад результатів пошуку по полю ‘Attached media’ на предмет .pdf, .csv, .docx файлів

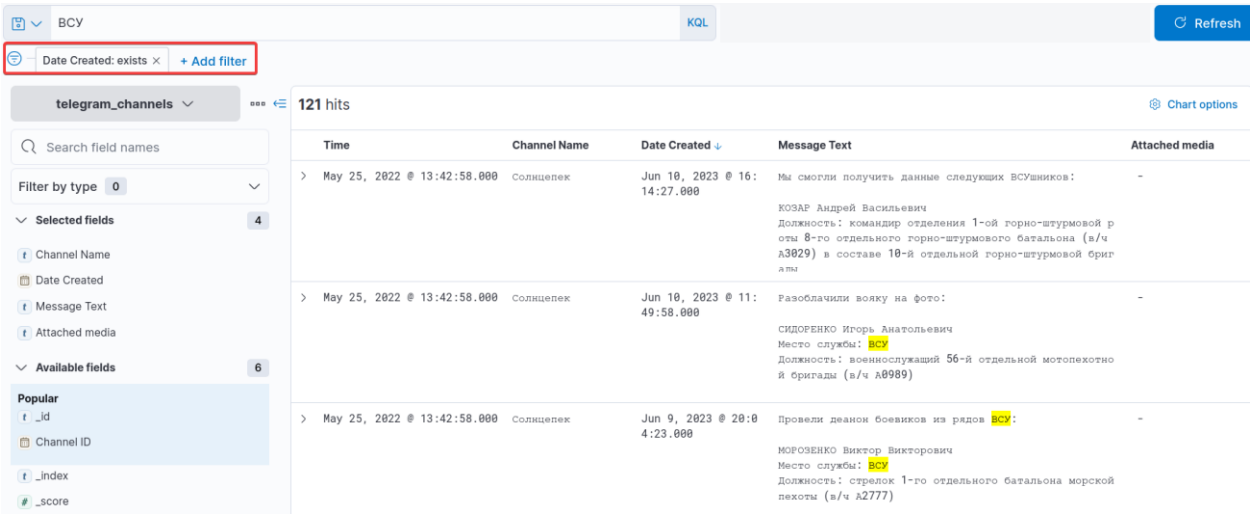


Рисунок 2.14 – Приклад результатів пошуку по всім полям за ключовим словом ‘ВСУ’ та фільтруванням за датою по спаданням

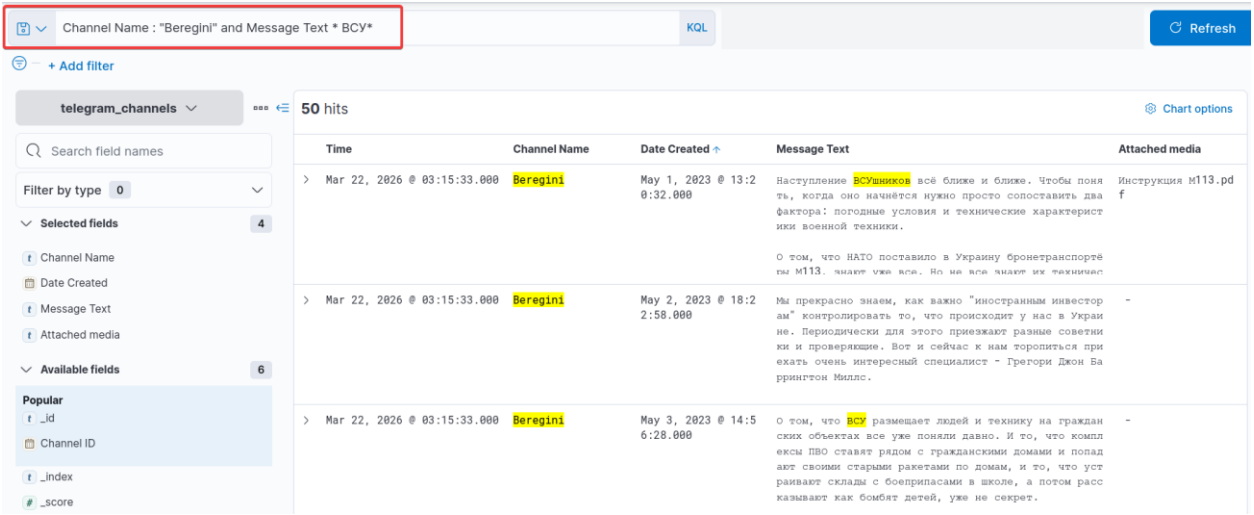


Рисунок 2.15 – Приклад результатів пошуку за двома полями водночас

Результати пошуку можна одразу експортувати в окремий файл прям з меню Discover.

2.2 Аналіз зібраних даних

Було створено тестовий набір даних за допомогою інструменту збору інформації з 10 Telegram каналів, що публікують інформацію про злами українських ресурсів, проведення DDoS атак, поширення витоків даних, OSINT операції, планування кібератак тощо. Увесь набір даних було очищено та приведено у зручний для аналізу формат.

Список 10 Telegram каналів, які виступили джерелами інформації:

1.https://t.me/nemeZ1da_ru

2.https://t.me/bg14_ru

3.https://t.me/JokerDPR

4.https://t.me/solntsepekZ

5.https://t.me/CyberArmyofRussia_Reborn

6.https://t.me/informZarya

7.https://t.me/NetSide_official

8.https://t.me/xaknet_team

9.https://t.me/data1eaks

10. приватний агрегатор витоків даних

З кожного каналу дані було зібрано від 01.05.2023 по 10.06.2023, що в сумі дало 1272 записів. В залежності від методу аналізу, наш JSON файл можна легко перевести у CSV формат та завантажити в ПЗ Excel для аналізу.

Channel Name	Count of Message Text
Beregini	50
Data1eaks | Утечки баз данных	14
LEAKS AGGREGATOR | УТЕЧКИ АГРЕГАТОР | БАЗЫ ДАННЫХ | СЛИВ |	551
NETSIDE GROUP	58
XakNet Team RU	10
Zapя	34
Джокер ДНР	16
Народная CyberАрмия	268
НемеZида	165
Солнцепек	106

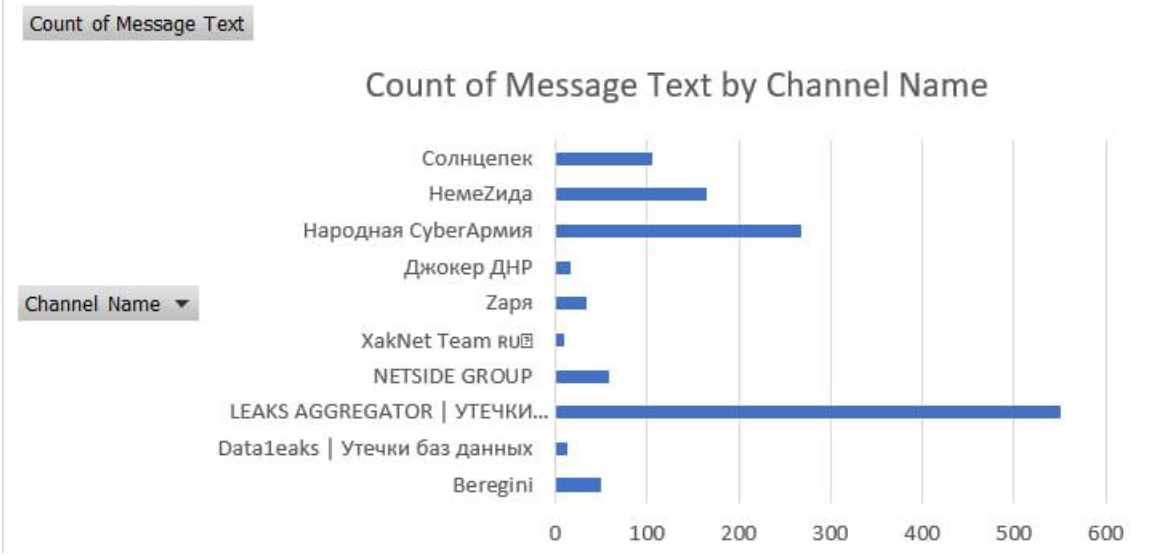


Рисунок 2.16 – Підрахунок кількості записів відповідно до джерела

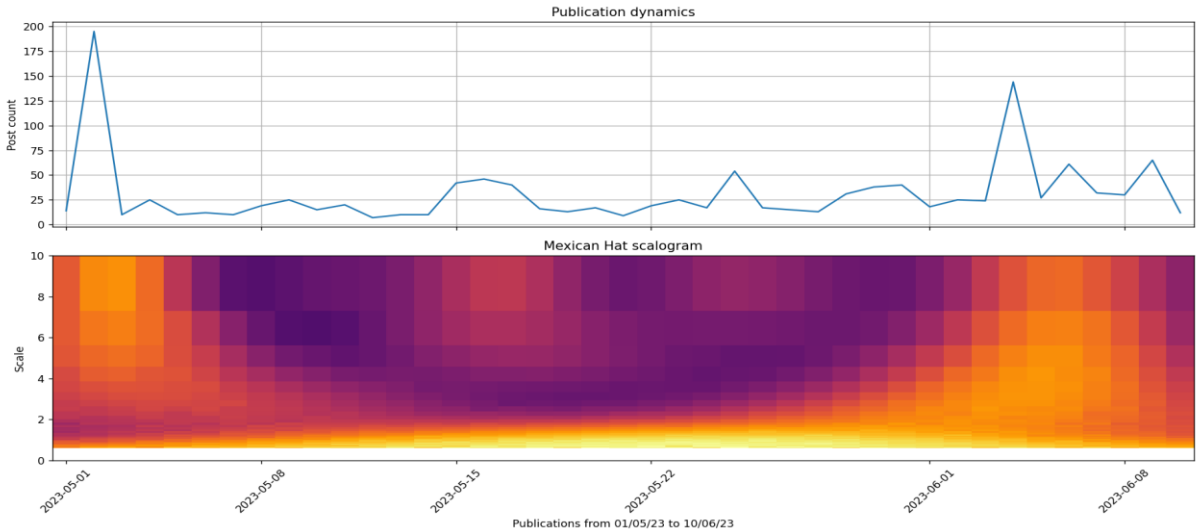


Рисунок 2.17 – Скейлограма на базі вейвлету ‘Mexican Hat’, бачимо сплески у таких датах як 2 травня та 4 червня

Зібрані дані можна аналізувати різними способами та для різних цілей. Як наприклад, використання одного з найпростіших способів — пошук перетину даних за двома словниками, або ж отримання масиву даних в результаті тематичних запитів. Обидва словники містять сутності кирилицею. Перший словник містить хакерський сленг (ддос, дамп, исходн[ик], дефейс, материал, сайт тощо) та методи дії (клад[ём], ламаем, вылож[им] тощо), а другий словник містить терміни, що означають військові терміни в українському контексті (ТрО, бригада, ЗСУ, ВСУ тошо).

В результаті ми отримуємо тематичний масив даних, що описує дії, спрямовані проти українських Сил Оборони. Також було додано пошук по полю ‘Attached Media’, тобто пошуковий запит має видати тільки ті результати, в яких було опубліковано якийсь витік інформації у тому чи іншому вигляді. Знайдено 8 перетинів, усі вони походять з каналу ‘Beregini’ та мають певні витоки Сил Оборони, а саме плани дій, оборони міст, бойові статути тощо.

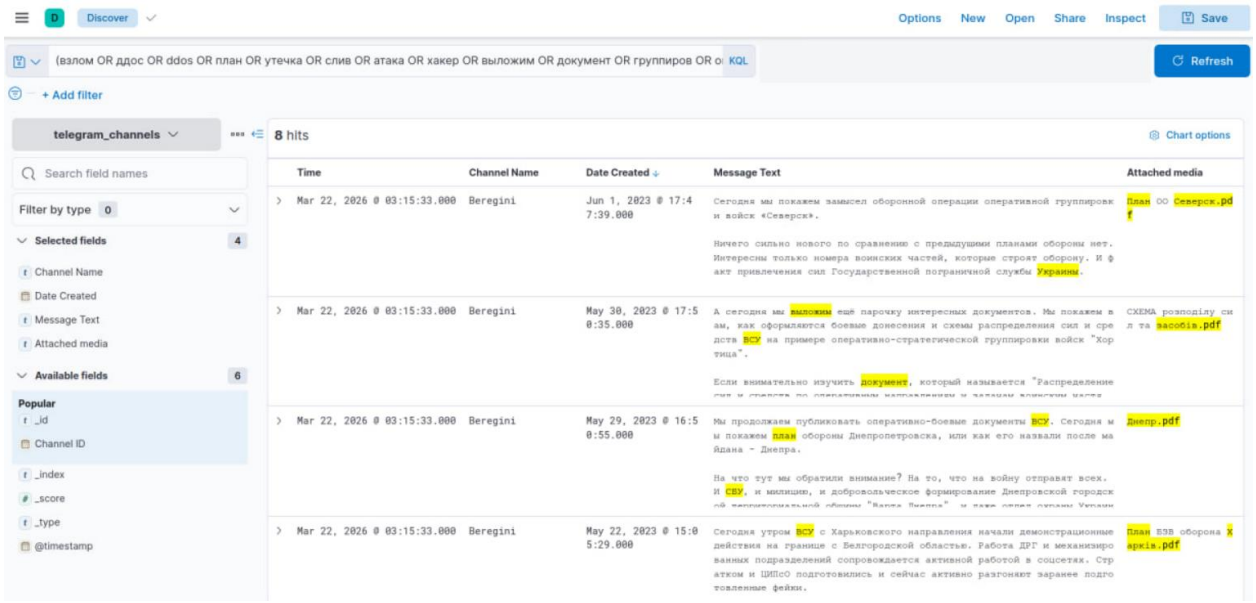


Рисунок 2.18 – Повнотекстовий пошук тексту за пересіченням двох словників із фільтром

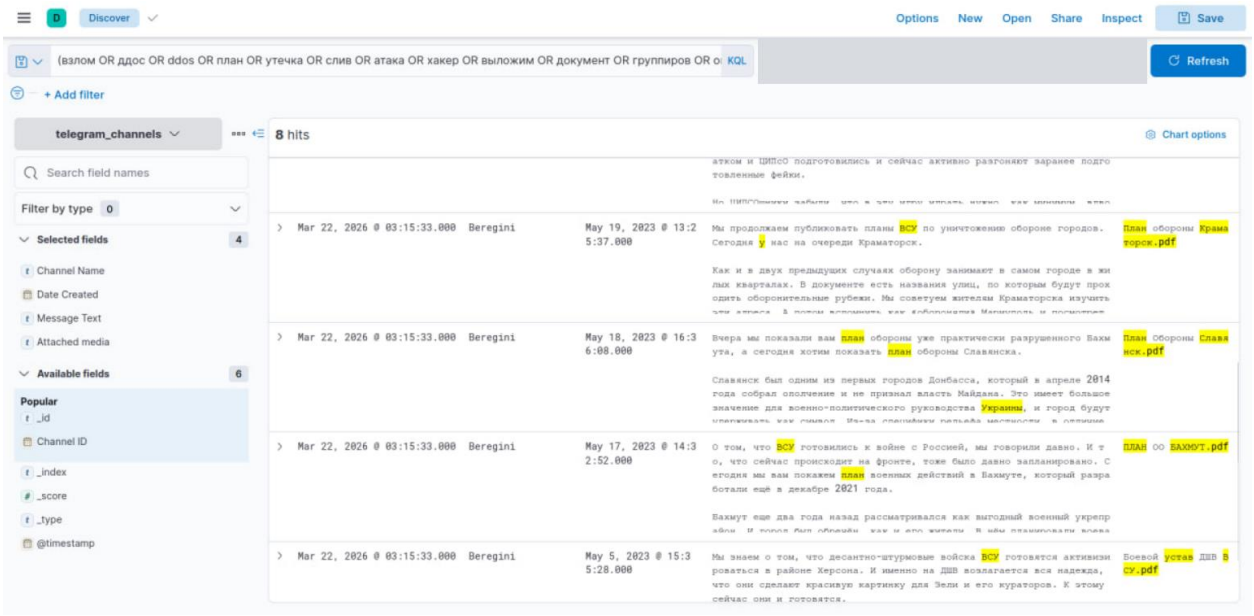


Рисунок 2.19 – Продовження попереднього запиту

Також нам може бути цікавим список IP адрес та посилань (URL) на ресурси, що зазнають атаки або плануються бути атакованими. Бачимо 81 таке повідомлення за увесь період. Можна обмежити його часовими проміжками знизу та зверху для більш актуальних результатів.

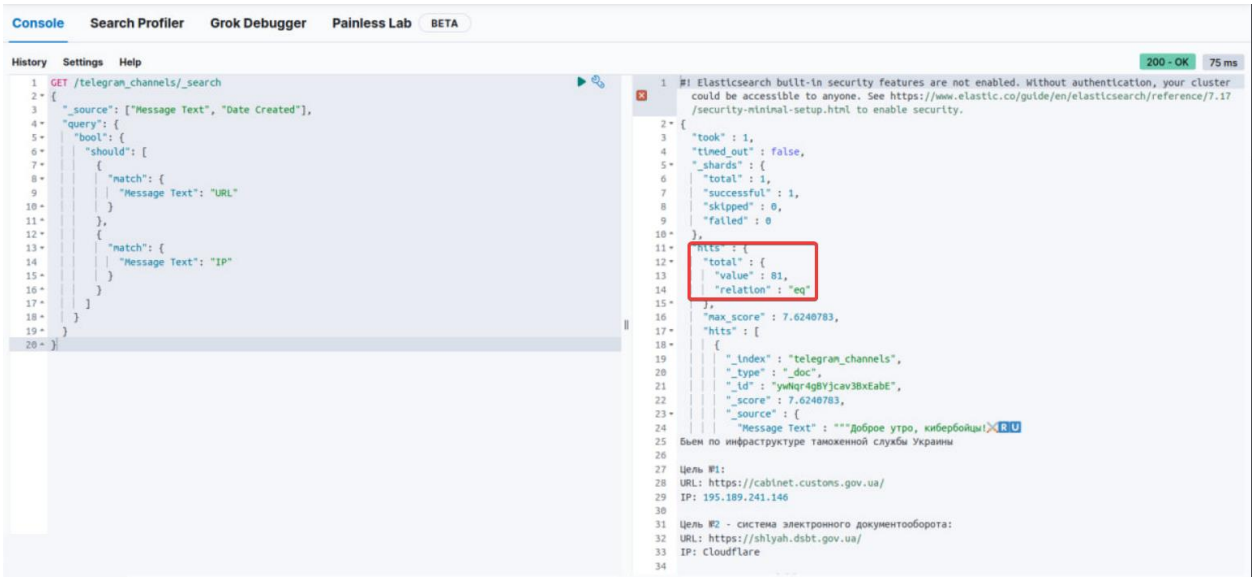


Рисунок 2.20 – Результати запиту

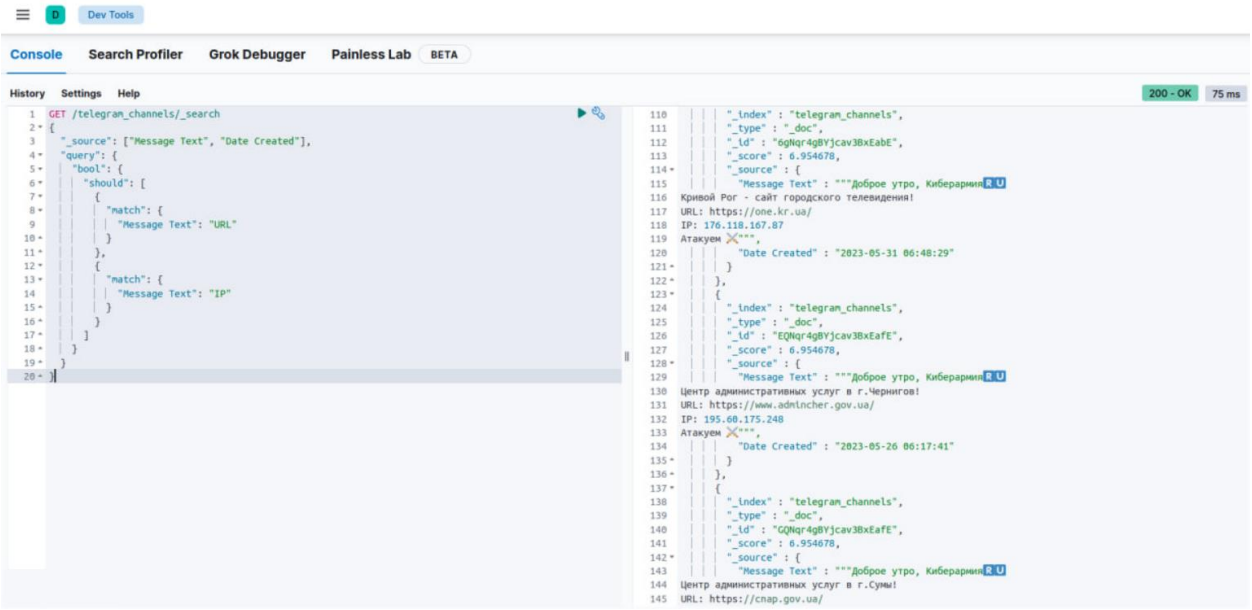


Рисунок 2.21 – Продовження попереднього запиту

Результат ми отримуємо у вигляді JSON відповіді, яку надалі можна далі аналізувати. Наприклад дістати усі IP та URL ресурсів, що були отримані. Ці ресурси і є об’єктами кібератак зі сторони угруповання.

Рисунок 2.22 – Результати екстрагування IP та URL

Також нас може цікавити саме джерело, наприклад "Народная CyberАрмия", з цілями атак (IP та URL ресурсу) на українські (*.ua) ресурси.

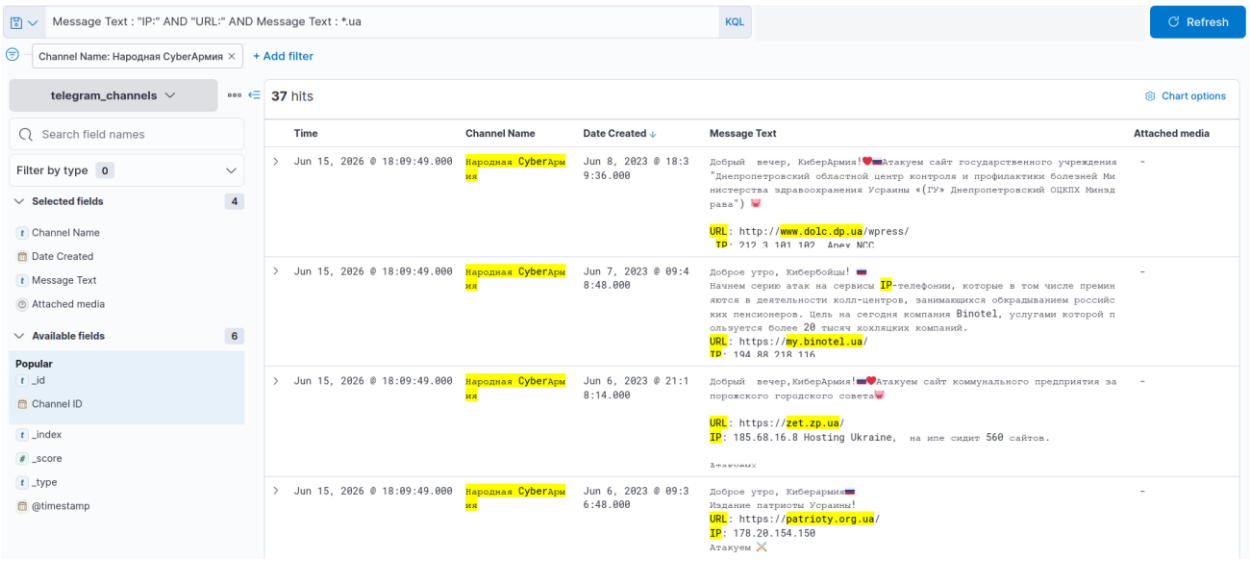


Рисунок 2.23 – Результати запиту

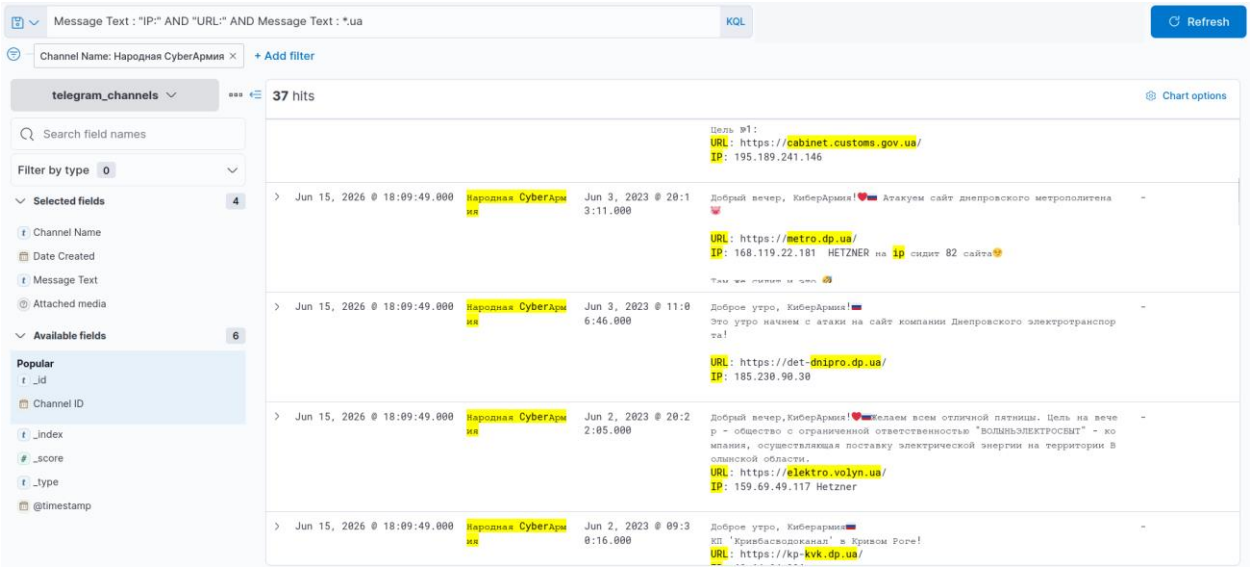


Рисунок 2.24 – Продовження попереднього запиту

З даними у JSON форматі зручно працювати, переводити у інші формати та аналізувати. Для переведення JSON у CSV можна використати навіть онлайн-інструменти. Для роботи з CSV в ПЗ Excel достатньо лише у вкладці Data імпортувати необхідний .csv файл та відредагувати тип даних за потреби.

2.3 Технологічні рішення

Мною було обрано Elasticsearch, який розроблено на основі Apache Lucene, та використовує документ-орієнтовану модель даних, яка зберігає самі документи у JSON форматі. Elasticsearch надає багато можливостей для роботи з JSON-документами, включаючи пошук, фільтрацію, агрегацію, сортування та інші операції. Використання додаткового інструменту Kibana для візуалізації та аналітики надає зручний графічний інтерфейс до даних, які зберігаються в Elasticsearch. Також з'являється можливість працювати напряму через інтерфейс окрім як через REST API Elasticsearch.

Аналіз за датою публікації наочно показав, що дати проведення кібератак та поширення витоків можуть бути приурочені до відповідних свят, як наприклад 1 травня чи 9 травня.

Аналіз даних за словниками, що передбачає виявлення певних слів або термінів у наборі даних, є ефективним для виконання поставлених задач, адже він дозволяє провести аналіз сфер, що можуть ставати об'єктами кібератак, розмежувати їх відповідно до обраних словників. Цей метод є корисним аналізу тематик на основі тексту. Ключовим фактором ефективного використання цього методу є екстрагування вагомих понять у зібраних даних.

Повнотекстовий пошук по полям 'Message Text' та 'Attached Media' може допомогти ідентифікувати ресурси, що стали джерелом витоку тієї чи іншої інформації. Відповідно до дати публікації витоку можна встановити

час витоку, наскільки він поширився та скільки джерел вже розмістили ці дані.

Аналіз тексту з використанням регулярних виразів, як наприклад для пошуку IP адрес та URL може використовуватись для ідентифікації ресурсів, що можуть зазнати, або прямо зараз зазнають кібератаки, наприклад дефейс чи DoS, адже джерело публікації часто надає ці дані для демонстрації своїх

можливостей. Цей метод може використовуватись для вчасного реагування та моніторингу кібероперацій.

Висновки до розділу 2

У другому розділі дипломної роботи було розглянути методи завантаження та пошуку даних у системі Elasticsearch з використанням інструменту Kibana. Було досліджено основні методи завантаження даних у Kibana, такі як пряме завантаження через інтерфейс Kibana та меню ‘Upload File’, використання cURL POST запитів та REST API, доступ до якого надається через Console у Kibana. Ознайомився з GROK шаблонами, які базуються на регулярних виразах, та використовуються для структурування даних відповідно до заданого шаблону.

Аналіз зібраних даних проводився на основі зібраного набору даних у 1272 записи у форматі JSON, з яким легко працювати в системі Elasticsearch і так само легко переструктуровувати в інший формат, такий як CSV, для аналізу в інших системах. Було проведено ручний аналіз даних та інтерпретовано результати аналізу відповідно до поставлених у дипломній роботі задач.

3 ПОРІВНЯННЯ З ІСНУЮЧИМИ СИСТЕМАМИ

3.1 Порівняння розробленої системи з існуючими

Особливістю сьогодення є постійне зростання обсягу інформації і складності її обробки. Велика кількість джерел інформації так і залишиться великої кількістю джерел інформації, якщо саму інформацію не зберігати, обробляти та аналізувати під конкретні задачі.

Це стало поштовхом до створення інформаційно-пошукових систем та систем контент-моніторингу, що здатні проводити аналіз інформації, виокремлення необхідних даних та їх представлення користувачу. Моніторинг інформації здійснюється за допомогою аналізу змісту потоків інформації протягом заданого часового проміжку. Використання подібних систем має забезпечувати автоматизований збір інформації, її структуризацію, класифікацію, та групування за семантичними ознаками. Окремо варто відзначити встановлення неявних взаємозв’язків між окремими поняттями понять та джерелами інформації для демонстрації взаємозалежності досліджуваної мережі.

Більшість таких систем орієнтовані на новини, блоги, телебачення, веб-сторінки та форуми як джерела інформації і виконують функцію стеження за конкурентами, визначення трендів та прогнозів, що може бути корисно для бізнесів (маркетологів, рекламістів, аналітиків тощо), які мають бути в курсі глобального настрою, інформаційного поля навколо досліджуваного об’єкта та репутації брендів, компаній, організацій тощо.

Приклади таких систем:

1. Meltwater.
2. Talkwalker.
3. Brandwatch.

4. InfoStream

5. Attack Index

Усі перераховані системи є платними та закритими, що унеможливлює розгляд детального функціонування системи, використаного системного та програмного забезпечення. У кращому випадку можна дізнатись технології підходів конкретної платформи щодо аналізу інформації та їх реалізації в теорії. Користувачу-замовнику надається лише веб-інтерфейс з автоматизованим доступом до системи або самі результати пошуку з проведеним аналізом у вигляді звіту.

Рисунок 3.1 – Технологічні процеси системи контент-моніторингу

Також існує така система агрегування інформації із соціальних мереж як ‘КіберАгрегатор’, дані про яку є у відкритому доступі, завдяки чому можна розглянути її функціонування та інфраструктуру. Важливим є її орієнтованість на соціальні мережі та месенджери, що є релевантним і для

запропонованої мною системи для роботи з даними. Система працює з даними різної структури через велику кількість джерел для збору інформації, що викликає необхідність форматування усіх даних в стандартні формати (CSV, JSON, XML). Використання вільного ПЗ дозволяє наслідувати цій системі, оскільки немає необхідності додатково вкладатися у ПЗ та інструменти для побудови аналогу.

Апаратна платформа системи включає декілька серверів (інформаційний проксі-сервер, сервер збору даних, сервер аналітики, інтерфейсний сервер), що виконують різні задачі та працюють паралельно.

Також примітним є використання таких інструментів як Gephi та Matlab для аналітики і прогнозування на основі зібраних даних і Neo4j для побудови зв’язків між поняттями.

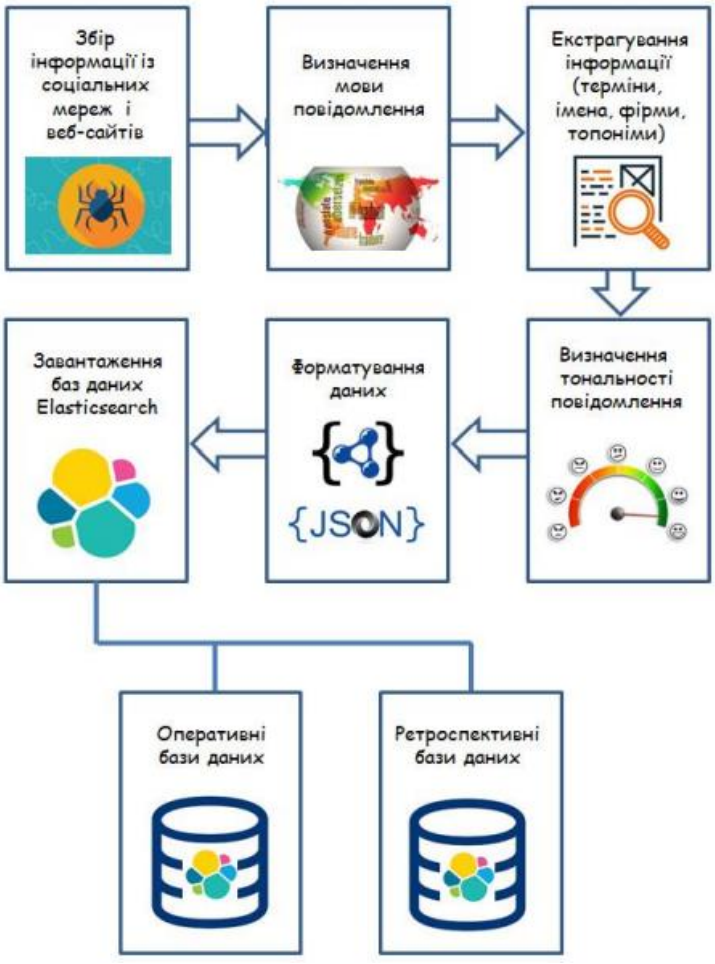


Рисунок 3.2 – Етапи обробки інформації в системі ‘КіберАгрегатор’

Незважаючи на те, що дані про систему є у відкритому доступі, скористатись нею може не кожен через закритий внутрішній доступ.

У процесі виконання цієї дипломної роботи було розглянуто деякі елементи таких систем ‘у домашніх умовах’. Основну увагу було зосереджено аналітичній діяльності щодо методології OSINT та її практичному використанні при аналізі кіберінцидентів. На прикладі проведених кібератак та витоків було визначено джерела корисної інформації, як варто моніторити та дані з яких було зібрано й оброблено власними інструментами для подальшого завантаження до системи введення, зберігання, пошуку та опрацювання даних. Як найбільш оптимальну щодо поставлених задач було обрано розподілену інформаційно-пошукову систему Elasticsearch, що призначена для швидкої обробки великих масивів даних завдяки використанню пошукової бібліотеки Apache Lucene. Унікальна документ-орієнтовна структура Elasticsearch, що зберігає документи у JSON форматі надає широкий спектр можливостей для пошуку та аналізу даних за допомогою запитів через REST API або ж через веб-інтерфейс, який надає інший інструмент — Kibana. Було обрано декілька видів ручного аналізу даних, результатом яких стала корисна інформація, що може бути застосована для відстеження атак на ресурси та моніторингу й оцінки витоків даних.

Висновки до розділу 3

Інформаційно-пошукові системи та системи контент-моніторингу відіграють важливу роль у сучасному інформаційному середовищі. Вони спрощують пошук та доступ до інформації, пропонуючи швидкий та зручний спосіб отримання саме необхідних зрізів даних. Систематичне дослідження

та оцінка вмісту інформації з різних джерел може грати чималу роль при стратегічному плануванні та прийнятті рішень.

У сфері кібербезпеки подібні системи використовуються для виявлення потенційних загроз та вразливостей, запобігання поширенню витоку даних, ідентифікації їх першоджерела, оцінці завданих збитків та виявленні інсайдерів. У маломасштабних проектах, який і розглядається у цій роботі, неабияку увагу варто приділити якісному підбору джерел інформації і їх видобутку для подальшого завантаження в систему для роботи з даними. Також варто підібрати найбільш ефективні методи аналізу даних для виконання поставлених задач і, за потреби, автоматизувати побудовані процеси. Здебільшого подібні системи є модульними, тобто додавання нових джерел інформації, в яких дані структуруються іншим чином, не має стати складністю, якщо ці моменти були враховані на етапі вибору інструментів для роботи з даними. До прикладу Elasticsearch працює з неструктурованими або погано структурованими даними, через що є популярним вибором для подібних задач.

ВИСНОВКИ

1. У роботі розглянуто пошук аномалій пов’язаних із кіберінцидентами в технології OSINT. Для цього було досліджено методологію OSINT та її використання в сфері кібербезпеки, визначено її основні особливості, етапи, підходи та результати використання.
2. Визначено релевантні загальнодоступні джерела інформації та обґрунтовано актуальність соціальних мереж як джерел інформації та доцільність їх моніторингу.
3. Було розроблено 2 інструменти на мові Python для збору та обробки даних. Структуру зібраних з Telegram даних було продумано таким чином, щоб мати можливість аналізувати джерела публікації, текст публікації, дату публікації та додані до тексту медіафайли, в яких, за задумом, можуть бути додані корисні дані. Самі дані було очищено та переформатовано до формату JSON, який є зручним для подальшого проведення аналізу.
4. Проведено огляд та порівняльний аналіз існуючих систем введення, зберігання, пошуку та опрацювання даних. З-поміж розглянутих систем було обрано інформаційно-пошукову систему Elasticsearch як ту, що може швидко працювати з великими масивами неструктурованих або погано структурованих даних, що робить її чудовим вибором для обробки, пошуку та аналізу даних з різних джерел, звідки інформація збирається в різних структурах.
5. Було створено тестовий набір даних за допомогою інструменту збору даних з 10 Telegram каналів, що публікують інформацію про злами ресурсів, проведення атак, поширення витоків даних, OSINT операції, планування кібератак тощо. Увесь набір було попередньо очищено та приведено до зручного для аналізу формату. Дані зібрано з 01.05.2023 по 10.06.2023, що в сумі дало 1272 записів.

Проведено ручний аналіз даних та інтерпретовано результати аналізу відповідно до поставлених у дипломній роботі задач. Було підраховано кількість записів відповідно до дати публікації та відстежено дати їх сплесків. За допомогою пошуку тексту в системі Elasticsearch за пересіченням двох словників із фільтром було отримано тематичний масив даних, що містить інформацію про опубліковані витоки. У результаті виконання запиту по текстовому полю даних з використанням регулярних виразів на пошук IP адрес та URL було екстраговано корисні дані, що свідчать про вже проведені або заплановані кібератаки на ресурси.

6. Відкритих та доступних до дослідження інформаційно-пошукових систем та систем контент-моніторингу існує невелика кількість, але було проведено аналіз їх роботи та порівняння з розробленим підходом та інструментами.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1. Ланде Д. В. ПРАВОВІ ПИТАННЯ КОНКУРЕНТНОЇ РОЗВІДКИ / Дмитро Володимирович Ланде., 2022. – 18 с. – (УДК 316.324.8).
2. Open Source Intelligence: The Beginners' Guide to OSINT [Електронний ресурс] – Режим доступу до ресурсу: https://www.liferaftinc.com/blog/the-beginners-guide-to-osint.
3. What is Open-Source Intelligence? [Електронний ресурс] – Режим доступу до ресурсу: https://www.sans.org/blog/what-is-open-source-intelligence/.
4. Breaking Down the OSINT Cycle [Електронний ресурс] – Режим доступу до ресурсу: https://ntrepidcorp.com/case-studies/breaking-down-the-osint-cycle/.
5. Bazzell M. Open Source Intelligence Techniques: Resources for Searching and Analyzing Online Information / Michael Bazzell. – USA, Coppell, TX, 2019. – 574 с. – (7).
6. The Role of Open-Source Intelligence in the War in Ukraine [Електронний ресурс] – Режим доступу до ресурсу: https://moderndiplomacy.eu/2023/05/17/the-role-of-open-source-intelligence-in-the-war-in-ukraine/.
7. OSINT Integration Tool. [Електронний ресурс]. – 2018. – Режим доступу до ресурсу: https://www.osti.gov/biblio/1569427.
8. Ekman I. Ukraine's Information Front – Strategic Communication dur- ing Russia's Full-Scale Invasion of Ukraine / I. Ekman, P. Nilsson., 2023. – 97 с. – (FOI-R—5451—SE).
9. Przetacznik J. Russia's war on Ukraine: Timeline of cyber-attacks / J. Przetacznik, S. Tarpova., 2022. – 7 с. – (PE 733.549).
10. Cyber Dimensions of the Armed Conflict in Ukraine [Електронний ресурс]. – 2023. – Режим доступу до ресурсу: https://cyberconflicts.cyberpeaceinstitute.org/report.
11. Cyber Attacks in Times of Conflict [Електронний ресурс] – Режим доступу до ресурсу: https://cyberconflicts.cyberpeaceinstitute.org/.

12. Cyber-attacks during the Russian invasion of Ukraine [Електронний ресурс] – Режим доступу до ресурсу: https://obr.uk/box/cyber-attacks-during-the-russian-invasion-of-ukraine/.
13. Russia's Cyber Tactics: Lessons Learned 2022 [Електронний ресурс]. – 2023. – Режим доступу до ресурсу: https://cip.gov.ua/en/news/russia-s-cyber-tactics-lessons-learned-in-2022-ssscip-analytical-report-on-the-year-of-russia-s-full-scale-cyberwar-against-ukraine.
14. The Role of OSINT in Russia's Invasion of Ukraine [Електронний ресурс] – Режим доступу до ресурсу: https://flashpoint.io/resources/report/role-of-osint-russia-invasion-of-ukraine/.
15. From Discord to 4chan: The Improbable Journey of a US Intelligence Leak [Електронний ресурс] – Режим доступу до ресурсу: https://www.bellingcat.com/news/2023/04/09/from-discord-to-4chan-the-improbable-journey-of-a-us-defence-leak/.
16. Freeman D. Machine Learning & Security / D. Freeman, C. Chio., 2018. – 385 с. – (ISBN: 9781491979907).
17. How to Scrape Everything from Telegram Using Python [Електронний ресурс] – Режим доступу до ресурсу: https://python.plainenglish.io/how-to-scrape-everything-from-telegram-using-python-f71437194569.
18. Telethon's Documentation [Електронний ресурс] – Режим доступу до ресурсу: https://docs.telethon.dev/en/stable/.
19. Ланде Д. В. ОБРОБЛЕННЯ НАДВЕЛИКИХ МАСИВІВ ДАНИХ / Д. В. Ланде, І. Ю. Субач, А. Я. Гладун. – Київ: ІСЗЗІ КПІ ім. Ігоря Сікорського, 2021. – 167 с. – (ISBN 978-966-2344-83-7).
20. System Properties Comparison Elasticsearch vs. Manticore Search vs. PostgreSQL [Електронний ресурс] – Режим доступу до ресурсу: https://db-engines.com/en/system/Elasticsearch%3BManticore+Search%3BPostgreSQL.
21. What is PostgreSQL? Introduction, Advantages & Disadvantages [Електронний ресурс] – Режим доступу до ресурсу: https://www.guru99.com/introduction-postgresql.html.
22. Manticore Search Introduction [Електронний ресурс] – Режим доступу до ресурсу: https://manual.manticoresearch.com/Introduction#Stream-filtering-made-easy.

23. Analysis of Lucene — Basic Concepts [Електронний ресурс] – Режим доступу до ресурсу: https://alibaba-cloud.medium.com/analysis-of-lucene-basic-concepts-5ff5d8b90a53.

24. Manticore: a faster alternative to Elasticsearch in C++ with a 21-year history [Електронний ресурс] – Режим доступу до ресурсу: https://manticoresearch.com/blog/manticore-alternative-to-elasticsearch/.

25. Importing CSV and Log Data into Elasticsearch with File Data Visualizer [Електронний ресурс] – Режим доступу до ресурсу: https://www.elastic.co/blog/importing-csv-and-log-data-into-elasticsearch-with-file-data-visualizer.

26. A Practical Introduction to Elasticsearch with Kibana [Електронний ресурс] – Режим доступу до ресурсу: https://medium.com/@victorsmelopoa/an-introduction-to-elasticsearch-with-kibana-78071db3704.

27. КОМПЬЮТЕРНАЯ КОНКУРЕНТНАЯ РАЗВЕДКА / А. Г.Додонов, Д. В. Ландэ, В. В. Прищепа, В. Г. Путятин. – Киев: ТОВ «Інжиніринг», 2021. – 354 с. – (ISBN 978-966-2344-79-0).

28. СИСТЕМА КОНТЕНТ-МОНІТОРИНГУ НОВИННИХ ІНТЕРНЕТ-РЕСУРСІВ [Електронний ресурс] – Режим доступу до ресурсу: https://science.lpnu.ua/sites/default/files/journal-paper/2017/jun/3423/0271.pdf.

29. Lande D. Агрегація інформації з питань кібербезпеки як основа навчального курсу «Оброблення надвеликих масивів даних» [Електронний ресурс] / D. Lande, O. Puchkov, I. Subach – Режим доступу до ресурсу: http://dwl.kiev.ua/art/its2021-1/article.pdf.

ДОДАТОК А ЛІСТИНГ ПРОГРАМИ ЗБИРАЧА ДАНИХ

```
import telethon
from telethon.sync import TelegramClient
import pytz
import datetime


api_id = ***
api_hash = '***'
phone_number = '***'

channel_username = 'https://t.me/***'

start_date = datetime.datetime(2023, *, *)

utc = pytz.UTC

with TelegramClient(phone_number, api_id, api_hash) as client:

    channel = client.get_entity(channel_username)
    channel_id = channel.id
    channel_name = channel.title
    output_filename = f"{channel_username.split('/')[-1]}.txt"

    with open(output_filename, 'w', encoding='utf-8') as file:

        post_count = 0

        for message in client.iter_messages(channel):
            if message.date.replace(tzinfo=utc) >= start_date.replace(tzinfo=utc):
                text = message.message

                date_created = message.date.strftime('%Y-%m-%d %H:%M:%S')

                if (
                    message.media
                    and hasattr(message.media, 'document')
                    and hasattr(message.media.document, 'attributes')
                ):
                    filename = ''

                    for attribute in message.media.document.attributes:
```

```
                if isinstance(attribute,
telethon.tl.types.DocumentAttributeFilename):

                    filename = attribute.file_name
                    break

            file.write(f"Channel ID: {channel_id}\n")
            file.write(f"Channel Name: {channel_name}\n")
            file.write(f"Message Text: {text}\n")
            file.write(f"Date Created: {date_created}\n")
            file.write(f"Attached media: {filename}\n")
            file.write("-----------------------------\n")

        else:

            file.write(f"Channel ID: {channel_id}\n")
            file.write(f"Channel Name: {channel_name}\n")
            file.write(f"Message Text: {text}\n")
            file.write(f"Date Created: {date_created}\n")
            file.write("Attached media:\n")  # Empty field for the filename
            file.write("-----------------------------\n")

        post_count += 1

print("Output written to file:", output_filename)
```

ДОДАТОК Б ЛІСТИНГ ПРОГРАМИ ОБРОБКИ ДАНИХ

```
import json
def filter(filename, delete_which_has=['Message Text: None', 'Message Text:
\nDate Created'],
           delimiter='------------------------------'):
    f = open(filename, 'r', encoding='utf-8')
    content = f.read()

    new_txt = []
    for item in content.split(delimiter)[:-1]:
        tmp_l = len(item.split(delete_which_has[0])) - 1
        tmp1_l = len(item.split(delete_which_has[1])) - 1

        if len(item[item.find('Attached media:'):item.find('Attached media:') + 100])
== 16:
            attch_is_empty = 1
        else:
            attch_is_empty = 0

        msg_is_empty = tmp_l + tmp1_l

        if not (attch_is_empty == 1 and msg_is_empty == 1):
            new_txt.append(item)

    js_text = []
    keys = ['Channel ID', 'Channel Name', 'Message Text', 'Date Created', 'Attached
media']
    for i in new_txt:
        ch_id = int(i[i.find('Channel ID:') + 11:i.find('Channel Name')].strip())
        ch_name = i[i.find('Channel Name:') + 13:i.find('Message Text:')].strip()
        msg_text = i[i.find('Message Text:') + 13:i.find('Date Created')].strip()
        date_cr = i[i.find('Date Created:') + 13:i.find('Attached media:')].strip()
        att_med = i[i.find('Attached media:') + 15:].strip()

        values = [ch_id, ch_name, msg_text, date_cr, att_med]
        dict1 = {keys[i]: values[i] for i in range(len(keys))}
        js_text.append(dict1)

    json_string = json.dumps(js_text, indent=2, ensure_ascii=False)
    output_name = filename.replace('.txt', '') + '_out.json'

    with open(output_name, 'w', encoding='utf-8') as f:
        f.write(json_string)
        f.close()
```

```
    print('output file name:', output_name)
    print('input items count:', len(content.split(delimiter)))
    print('output items count:', len(js_text))

filter('*.txt')
```