Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет інформаційних систем
(повна назва факультету)
Кафедра Кібербезпеки
(повна назва кафедри)

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня

магістр
(назва освітнього ступеня)

на тему: Методи ідентифікації та вилучення географічно пов'язаних об'єктів з даних про атаки на об'єкти критичної інфраструктури України

Виконав(ла): студент(ка) 6 курсу, групи СБм-62
спеціальності 125
кібербезпека
(шифр і назва спеціальності)

(підпис) Чорний П. Р.
(прізвище та ініціали)

Керівник (підпис) Скарга-Бандурова І. С.
(прізвище та ініціали)

Нормоконтроль (підпис) Лечаченко Т. А.
(прізвище та ініціали)

Завідувач кафедри (підпис) Загородна Н. В.
(прізвище та ініціали)

Рецензент (підпис) (прізвище та ініціали)

Тернопіль
2023

Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет комп'ютерно-інформаційних систем і програмної інженерії
(повна назва факультету)
Кафедра Кібербезпеки
(повна назва кафедри)

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Загородна Н. В.
(підпис) (прізвище та ініціали)
« » 2023 р.

З А В Д А Н Н Я
НА КВАЛІФІКАЦІЙНУ РОБОТУ

на здобуття освітнього ступеня магістр
(назва освітнього ступеня)
за спеціальністю 125 Кібербезпека
(шифр і назва спеціальності)
студенту Чорному Павлу Руслановичу
(прізвище, ім'я, по батькові)

1. Тема роботи Методи ідентифікації та вилучення географічно пов'язаних об'єктів з текстових даних про атаки на об'єкти критичної інфраструктури України

Керівник роботи Скарга-Бандурова Інна Сергіївна, доктор технічних наук, професор
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)
Затверджені наказом ректора від «*16*» *листопада* 2023 року № *4/7-1061*
2. Термін подання студентом завершеної роботи 20 грудня 2023 року
3. Вихідні дані до роботи Таблиця з подіями, що стосуються атак на критичну інфраструктуру України
4. Зміст роботи (перелік питань, які потрібно розробити)
РОЗДІЛ 1. ТЕОРЕТИЧНА ОСНОВА
1.1 Методи ідентифікації та вилучення подій з тексту
1.2 Геопросторовий аналіз для побудови зв'язків між подіями в контексті безпекя
РОЗДІЛ 2. ТЕХНОЛОГІЇ РОЗПІЗНАВАННЯ ІМЕНОВАНИХ ОБ'ЄКТІВ ДЛЯ АНАЛІЗУ ТА ВІЗУАЛІЗАЦІЇ ДАНИХ
2.1 Що таке NER?
2.2 Методи навчання, що застосовуються для NER
2.3 Загальна методологія роботи
2.4 Збір даних
2.5 Аналіз даних та візуалізація
РОЗДІЛ 3. ПРАКТИЧНА РЕАЛІЗАЦІЯ
3.1 Попередня обробки даних
3.2 Перетворення об'єктів на географічні одиниці
Висновок
5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень, слайдів)
Слайд 1. Тема роботи, Слайд 2. Вступ, Слайд 3. Мета та задачі дослідження,

Слайд 4. Ідентифікація об’єктів, Слайд 5. Геопросторовий аналіз, Слайд 6. Що таке NER?, Слайд 7. Навчання моделей, Слайд 8. Алгоритм роботи, Слайд 9. Геокодування, Слайд 10. Демонстрація результатів, Слайд 11. Висновки.

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Охорона праці	Осухівська Г.М., к.т.н., доцент		
Безпека в надзвичайних ситуаціях	Клепчик В.М., проректор з адміністративно-господарської роботи та будівництва		

7. Дата видачі завдання 16 листопада 2023 року

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1.	Ознайомлення із завданням кваліфікаційної роботи	16.11.2023	*Виконано*
2.	Пошук статей про атаки на Україну	16.11.2023 – 18.11.2023	*Виконано*
3.	Пошук каналів, що публікують інформацію про атаки на критичну інфраструктуру	19.11.2023 – 21.11.2023	*Виконано*
4.	Пошук наукових статей відповідно згідно теми кваліфікаційної роботи	22.11.2023 – 24.11.2023	*Виконано*
5.	Тестування роботи бібліотеки spaCy	25.11.2023 – 26.11.2023	*Виконано*
6.	Оформлення розділу «Теоретична основа»	27.11.2023 – 30.11.2023	*Виконано*
7.	Оформлення розділу «Технології розпізнавання іменованих об'єктів для аналізу та візуалізації даних»	01.12.2023 – 06.12.2023	*Виконано*
8.	Оформлення розділу «Практична реалізація»	07.12.2023 – 14.12.2023	*Виконано*
9.	Виконання завдання до підрозділу «Охорона праці»	15.12.2023	*Виконано*
10.	Виконання завдання до підрозділу «Безпека в надзвичайних ситуаціях»	16.12.2023	*Виконано*
11.	Оформлення кваліфікаційної роботи	16.12.2023 – 18.12.2023	*Виконано*
12.	Нормоконтроль	22.12.2023	*Виконано*
13.	Перевірка на плагіат	22.12.2023	*Виконано*
14.	Попередній захист кваліфікаційної роботи		
15.	Захист кваліфікаційної роботи	28.12.2023	

Студент _____ *Чорний П.Р.*
(підпис) (прізвище та ініціали)

Керівник роботи _____ *Скарга-Бандурова І. С.*
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Методи ідентифікації та вилучення географічно пов’язаних об’єктів з текстових даних про атаки на об’єкти критичної інфраструктури України // Methods for Identifying and Extracting Geographically Related Objects from Textual Data on Attacks on Critical Infrastructure Facilities in Ukraine // Чорний Павло Русланович // Тернопільський національний технічний університет імені Івана Пулюя, факультет комп’ютерно-інформаційних систем і програмної інженерії, кафедра кібербезпеки, група СБм-62 // Тернопіль, 2023 // С. 72, рис. – 31, табл. – 2, додат. – 1, бібліогр. – 38.

Ключові слова: ІМЕНОВАНІ СУТНОСТІ, ГЕОКОДУВАННЯ, ІДЕНТИФІКАЦІЯ АТАК, КРИТИЧНА ІНФРАСТРУКТУРА, КІБЕРБЕЗПЕКА, NER, ОБРОБКА ПРИРОДНЬОЇ МОВИ, SPACY, PYTHON, NATURAL LANGUAGE PROCESSING, MODEL, GEOCODDING

Кваліфікаційна робота складається із пояснювальної записки та графічної частини (ілюстративний матеріал – слайди).

Об’єм графічної частини кваліфікаційної роботи становить 11 слайдів.

У кваліфікаційній роботі проведено аналіз підходів до ідентифікації та вилучення подій з текстових даних, вперше використано технологію розпізнавання іменованих об'єктів з відкритих джерел для аналізу різних видів атак на критичну інфраструктуру України, із подальшою інтеграцією в системи моніторингу для візуалізацією зазначених об'єктів, виявлення та побудови стратегії реагування на атаки та майбутні загрози.

У ході виконання кваліфікаційної роботи розроблено метод для збору інформації про фізичні та кібератаки з відкритих джерел їх аналіз та виокремлення об’єктів, з можливістю подальшої побудови зв’язків та візуалізації.

ANNOTATION

Methods for Identifying and Extracting Geographically Related Objects from Textual Data on Attacks on Critical Infrastructure Facilities in Ukraine // Chornyi Pavlo Ruslanovych // Ternopil National Technical University named after Ivan Pului, Faculty of Computer Information Systems and Software Engineering, Department of Cyber Security, SBm-62 Group // Ternopil, 2023 // Pages - 72, Figures - 31, Tables - 2, Supplement - 1, Bibliography - 38.

Keywords: NAMED ENTITIES, GEOCODING, ATTACK IDENTIFICATION, CRITICAL INFRASTRUCTURE, CYBER SECURITY, NER, NATURAL LANGUAGE PROCESSING, SPACY, PYTHON, NATURAL LANGUAGE PROCESSING, MODEL, GEOCODING.

The qualification work consists of an explanatory note and a graphic part (illustrative material - slides).

The volume of the graphic part of the qualification work is 11 slides.

The qualification work analyzes approaches to identifying and extracting events from text data. Additionally, for the first time, the technology of recognizing named objects from open sources was used to analyze various types of attacks on the critical infrastructure of Ukraine, with further integration into monitoring systems to visualize these objects, identify and build a strategy for responding to attacks and future threats.

In the course of the qualification work, a method was developed for collecting information about physical and cyber attacks from open sources, analyzing them, and identifying objects, with the possibility of further linking and visualization.

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

NER − Named Entity Recognition

NLP − Natural Language Processing

ELK − Elasticsearch Logstash Kibana

KI − Критична інфраструктура

OKI − Об’єкт критичної інфраструктури

CNN − Convolutional neural networks

RNN − Recurrent neural networks

SML − Supervised machine learning

UML − Unsupervised machine learning

ГІС − Геоінформаційна система

ЗМІСТ

ВСТУП

У сучасному світі безпека інформації, що циркулює в інформаційному середовищі, становить неабияку цінність, що призводить до виникнення багатьох видів кібератак. Такі кібератаки здійснюються на різні організації для отримання конфіденційної та службової інформації, а також націлено на користувачів – для отримання особистої інформації. В умовах сьогодення, кібератаки використовуються як один із видів ведення інформаційної війни, що разом із використанням стандартних методів введення війни, може призводити до значних «ударів» як по економіці держави так і до людських, промислових та інших втрат.

Здійснення атак на критично важливі об’єкти України стала однією із найважливіших проблем безпеки держави. Забезпечення ефективної протидії таким атакам вимагає наявність кваліфікованих кадрів та використання інноваційних методів, що здатні реалізувати ефективну протидію атакам шляхом оперативної їх ідентифікації та ліквідації. Одним із ефективних варіантів вирішення даної проблеми є виявлення географічно пов’язаних об’єктів з текстових та історичних даних, що дає змогу спростити та прискорити аналіз атак у географічному контексті та розробити ефективну стратегію реагування на них, із урахуванням усіх особливостей їх проведення.

Метою даної роботи є аналіз вхідних текстових та історичних даних про атаки на критичну інфраструктуру України та, як наслідок, розробка методів ідентифікації географічно пов’язаних об’єктів атак на об’єкти критичної інфраструктури із подальшим їх застосуванняманаліз вхідних текстових та історичних даних про атаки на критичну інфраструктуру України та, як наслідок, розробка методів ідентифікації географічно пов’язаних об’єктів атак на об’єкти критичної інфраструктури із подальшим їх застосуванням.

Для отримання поставленої мети, необхідно виконати низку наступних **задач**:

- провести аналіз сучасних методів вилучення об’єктів з текстових даних;

- здійснити пошук відкритих джерел, що публікують інформацією про атаки на критичну інфраструктуру України;
- автоматизувати процес отримання повідомлень із найдених джерел;
- автоматизувати процес вилучення об’єктів із повідомлень та процесу геокодування;
- здійснити візуалізацію виявлених об’єктів із використанням геокодів.

Об'єктом дослідження є дані про атаки на об'єкти критичної інфраструктури.

Предметом дослідження – методи, що реалізовують ідентифікацію та вилучення географічно пов'язаних об'єктів з даних про атаки і дають змогу провести ефективний аналіз.

Наукова новизна одержаних результатів кваліфікаційної роботи полягає у наступних положеннях. Вперше використано технологію розпізнавання іменованих об'єктів для даних з відкритих джерел для аналізу кібер та фізичних атак на критичну інфраструктуру України, що дозволяє аналізувати великі обсяги даних та виконувати візуалізацію подій. Набули подальшого розвитку інтегровані рішення систем моніторингу для виявлення та реагування на кібер та фізичні загрози.

Практичне значення одержаних результатів. Розроблено комплексний підхід до ідентифікації та візуаліції загроз для об’єктів критичної інфраструктури України, що дозволяє вчасно виявити загрози та користувачу або оператору центру безпеки швидше прийняти необхідні дії, щоб зменшити вірогідність потрапляння зловмисника у мережу.

Апробація результатів магістерської роботи. Окремі результати проведених досліджень доповідались на XI науково-технічній конференції «Інформаційні моделі, системи та технології», Тернопіль, ТНТУ, 13 – 14 грудня 2023 р.

Публікації. За темою роботи з викладенням її основних результатів опубліковано 1 наукова праця, що являє собою тези в збірнику матеріалів науково-практичних конференцій (див. Додаток А).

1. ТЕОРЕТИЧНА ОСНОВА

1.1 Методи ідентифікації та вилучення подій з тексту

Ідентифікація та вилучення подій з текстових даних є одним із найважливіших етапів аналізу взаємопов’язаних подій, в умовах сучасності. Дані методи активно використовуються в повсякденному житті для отримання коротких відомостей із великих статей [1], реконструкції історичних подій [2], біомедичних дослідженнях [5, 17], екологічних задачах [22, 26], для отримання ключової інформації у великому потоці даних чи для пошуку конкретної інформації за допомогою алгоритмів ідентифікації необхідних джерел [13].

Ідентифікація – це процес виявлення необхідної категорії подій з множини подій, на основі певних ознак. Ідентифікація передбачає визначення унікальних ознак чи параметрів об'єкта з метою його розрізнення від інших об'єктів чи явищ. Цей процес зазвичай включає у себе виявлення ключових характеристик, аналіз властивостей та порівняння отриманих даних з попередньо відомими шаблонами або стандартами. У різних областях ідентифікація може виявлятися високоспеціалізованою, використовуючи унікальні методи та інструментарій, або ж бути універсальною та адаптивною, охоплюючи різноманітні аспекти об'єктів дослідження.

Даний процес відіграє ключову роль у вирішенні ряду завдань, починаючи від забезпечення безпеки та ведення обліку до визначення сутностей в природних мовах чи розпізнавання образів у науці та інших областях. Таким чином, ідентифікація є неодмінною частиною сучасних наукових та технічних розробок, що сприяє розвитку різноманітних сфер знань і вирішенню практичних завдань у різних галузях.

Вилучення подій, в контексті даного дослідження, передбачає систематизоване виділення та ізоляцію ключових подій із текстової інформації чи великих потоків даних з метою подальшого використання, аналізу чи графічного відображення ключових ознак сутності (виступає в якості події).

Такий процес вимагає розробку складних алгоритмів аналізу природної мови та обробку текстової, графічної, відео та аудіо інформації з подальшою автоматизацією, для пошуку необхідних даних та вилучення сутностей, що пов’язані із цими даними.

До методів вилучення подій слід віднести [15]:

- семантичний аналіз;
- глибинне навчання;
- машинне навчання;
- розпізнавання іменованих сутностей.

Семантичний аналіз — це процес розуміння та інтерпретації значень слів та їхніх зв'язків у тексті. Це важливий етап в обробці природної мови (NLP), який дозволяє комп'ютерам розуміти контекст і смислові відносини між словами. Це допомагає системам автоматизованого аналізу тексту витягувати інформацію з документів, електронних повідомлень та інших джерел [25].

Процес аналізу може включати в себе різні лексичні та семантичні зв’язки:

- Гіпоніми - відношення, де одна лексична одиниця (гіпонім) є конкретнішою формою іншої (гіпернім). Наприклад, "апельсин" є гіпонімом "фрукта";
- Мерономія - відношення, де один елемент є складовою частиною або членом іншого. Наприклад, "сегмент" апельсина є частиною цілого;
- Полісемія – зв’язок між значеннями слів або фраз, які, хоча й дещо відрізняються, але мають спільне основне значення. Наприклад: «я читав статтю і я написав статтю»;
- Синоніми - слова, що мають подібне або наближене значення. Для прикладу: "щасливий" і "радісний";
- Антоніми – слова, значення яких є протилежним. Наприклад: "хороший" і "поганий";
- Омоніми - слова, що мають однаковий звук і написання, але різне значення. Наприклад: "ключ" (замок) і "ключ" (джерело).

Семантичний аналіз враховує не лише лексичні зв’язки, але і контекстуальні зв'язки та семантичні ролі слів у конкретному реченні чи тексті. Автоматизований семантичний аналіз використовує алгоритми машинного навчання для навчання комп'ютера розуміти та інтерпретувати семантичні структури.

Семантичний аналіз використовується в різних сферах, таких як розробка чат-ботів, пошукові системи та обробка іншої неструктурованої інформації. Принцип роботи автоматизованого семантичного аналізу включає в себе обробку великої кількості текстових даних, тренування моделей на цих даних та використання навчених моделей для розуміння нових текстів. Такі системи можуть враховувати контекст, синтаксичні структури та семантичні відносиння для точного розуміння тексту. В результаті семантичного аналізу система може визначати не лише слова, але і їхні концептуальні значення, що сприяє ефективнішому взаєморозумінню між машиною і людиною.

Глибинне навчання - це галузь машинного навчання, що використовує штучні нейронні мережі з багатьма шарами для вирішення завдань інтелектуального аналізу даних. Цей підхід здатен автоматично вивчати внутрішні представлення даних за допомогою великої кількості шарів, що дозволяє йому розпізнавати складні залежності та виконувати завдання з вражаючою точністю.

Методи глибинного навчання включають в себе нижчеописані аспекти [7].

Штучні нейронні мережі (НМ), які складаються з великої кількості штучних нейронів, які розподілені в різні шари. Інформація передається від входу до виходу через ці шари за допомогою «ваг» і зважених з'єднань.

Зворотне поширення помилок (Backpropagation) - використовується для корекції ваг в нейронних мережах. Він вимагає порівняння прогнозованих результатів з фактичними та використовує цю інформацію для оновлення ваг, зменшуючи помилку передбачення.

Функції активації, що визначають вихід нейрона на підставі взаємодії з його входами. Популярні функції активації включають сигмоїду, гіперболічний тангенс та Rectified Linear Unit (ReLU).

Функції втрат визначають, як вимірюється відмінність між прогнозованими та фактичними значеннями.

Конволюційні нейронні мережі (CNN) – використовуються для обробки візуальної інформації і використовують фільтри для виявлення різних особливостей та структур в зображеннях.

Рекурентні нейронні мережі (RNN) – необхідні для обробки послідовностей даних (мовлення або часових рядів). Характерною ознакою таких мереж є пам’ять, що дозволяє зберігати інформацію про попередні стани.

Машинне навчання представляє собою метод функціонування штучного інтелекту, що використовує алгоритми для виявлення закономірностей під час аналізу обширних даних та їхнього використання для автоматизації та оптимізації рішень [11]. Такий підхід спрямований на самонавчання, де система намагається оптимізувати свою продуктивність в процесі розв'язання подібних завдань. Загальна мета машинного навчання полягає в автоматизації та оптимізації рішень у різних сферах, розширюючи використання його функціоналу від метеорології до комунікацій. Використання алгоритмів машинного навчання призводить до покращення продуктивності та досягнення бізнес-цілей, а також реалізації низки інноваційних рішень [27].

Метою штучного інтелекту є створення комп'ютерних моделей, які виявляють "розумову поведінку", схожу на людську, за словами Бориса Катца, головного дослідника та керівника групи InfoLab в CSAIL. Це означає створення машин, які можуть визнавати візуальні сцени, розуміти текст на природній мові або виконувати дії в фізичному світі [4].

Одним із способів реалізації штучного інтелекту є машинне навчання. Це було визначено в 1950-х роках піонером штучного інтелекту Артуром Самуелем як "галузь дослідження, яка надає комп'ютерам здатність вчитися без явного програмування".

Є три підкатегорії машинного навчання [8]:

- Навчання з учителем;
- Навчання без учителя;
- Навчання за підсиленням.

Навчання з учителем (Supervised machine learning): Моделі тренуються з використанням маркованих наборів даних, що дозволяє їм навчатися та покращувати свою точність з часом. Наприклад, алгоритм може навчитися розпізнавати моделі автомобілів, на основі вже наданих йому зображень, що були опрацьовані людьми.

Навчання без учителя (Unsupervised machine learning): Програма шукає патерни в немаркованих даних, знаходячи тенденції чи зв'язки, про які людина не задумується. Наприклад, програма може аналізувати мережевий трафік без попередньо заданих правил чи опису нормальної поведінки. Замість того, щоб вивчати конкретні патерни атак, вона шукає відхилення від звичайних зв'язків та тенденцій у мережі. Це може включати в себе різні параметри, такі як обсяг передачі даних, часові інтервали, частоту з'єднань тощо.

Навчання за підсиленням (Reinforcement machine learning) - навчання через випробування та помилки, отримуючи нагороди за правильні дії.

Розпізнавання іменованих сутностей (NER) — це компонент обробки природної мови (NLP), який ідентифікує попередньо визначені категорії об'єктів у тексті. Ці категорії можуть включати, імена осіб, організацій, локації, вирази часу, кількості, грошові значення та відсотки. Фактично, NER - це процес взяття текстового рядка (речення, абзацу чи цілого документа) та ідентифікація та класифікація сутностей, що відносяться до кожної категорії [24].

Першовідкривачеві терміну "NER" ставлено завдання спростити завдання екстракції інформації, спрямоване на аналіз великих обсягів неструктурованого тексту та виділення ключової інформації. Відтоді, NER еволюціонував та розширив свій функціонал завдяки досягненням в галузі методик машинного та глибинного навчання.

Організації, які використовують технологію розпізнавання іменованих сутностей (NER) для вилучення неструктурованих даних, виділяюсь три великі категорії підходів [3]:

- **підходи на основі правил** передбачають створення набору правил для граматики мови. Правила використовуються для ідентифікації сутностей у тексті на основі їх структурних і граматичних особливостей. Ці методи можуть вимагати багато часу та можуть погано узагальнювати невидимі дані;
- **підходи машинного навчання** передбачають навчання моделі машинного навчання, керованої штучним інтелектом, на заданому наборі даних за допомогою алгоритмів, на основі випадкових полів та максимальної ентропії. Техніки можуть варіюватися від традиційних методів машинного навчання (наприклад, дерева рішень і опорні векторні машини) до більш складних підходів глибокого навчання, таких як рекурентні нейронні мережі (RNN) і трансформатори. Ці методи краще узагальнюють невидимі дані, але вони вимагають великої кількості позначених навчальних даних і можуть бути вимагати велику кількість ресурсів для обчислень;
- **гібридні підходи** поєднують методи на основі правил і машинного навчання, щоб використовувати сильні сторони обох. Вони можуть використовувати систему на основі правил для швидкої ідентифікації легких для розпізнавання сутностей і систему машинного навчання для ідентифікації більш складних сутностей. В даному дослідженні для вилучення об’єктів із текстових даних використовується метод розпізнавання іменованих сутностей, що реалізований на базі бібліотеки spaCy.

1.2 Геопросторовий аналіз для побудови зв’язків між подіями в контексті безпеки

Цифрове середовище існування геоінформаційних систем (ГІС) передбачає використання цифрової форми оброблюваних даних. У своєму розмаїтті відображених видів даних, які використовують ГІС, відзначається широкий спектр цілей цих систем. ГІС може знаходити застосування в різних сферах, починаючи міським плануванням, закінчуючи відстеженням кіберзагроз, що призводить до великого розмаїття вихідного матеріалу [6, 14]. Класифікація і інвентаризація цього матеріалу стає викликом, особливо враховуючи різноманіття необхідних даних навіть в рамках одного ГІС-проекту.

Навіть при взятті до уваги того, що вигляд матеріалу може значно відрізнятися в залежності від проекту, фахівці з геоінформаційних систем повинні мати глибоке розуміння доступних джерел даних, їх характеристик і обмежень [19]. Це стає ключовим фактором для ефективної обробки та аналізу інформації, незалежно від конкретного застосування ГІС. Знання про різноманітність джерел даних дозволяє геоінформаційним спеціалістам вибирати найбільш відповідні та надійні джерела для конкретного завдання, сприяючи успішному впровадженню та розвитку проектів в різних галузях.

В галузі кібербезпеки геопросторовий аналіз використовується як інструмент для виявлення зв’язків між кіберінцидентами та кіберзагрозами. Використання таких даних дає змогу відстежувати джерела атак та їх цілі, що в результаті дозволить локалізувати загрози та передбачити появу нових.

За допомогою геопросторового аналізу здійснюється:

- визначення місцезнаходження (локалізація) систем, що критично важливі або вразливі, в порівнянні з іншими об’єктами;
- аналіз розповсюдження кібератак за географічними регіонами для виявлення тенденцій та паттернів проведення атак, це може допомогти в з'ясуванні можливих мотивацій зловмисників;

- визначення зв'язків між кіберінцидентами та їхнім географічним контекстом для виявлення частоти атак у певних географічних областях чи країнах;
- використання геопросторових аналітичних інструментів для моніторингу кіберінцидентів та атак в реальному часі, та виявлення взаємодій між ними на мапі;
- аналіз кібератак на критичну інфраструктуру з точки зору географічного розташування об'єктів, що може вказувати на стратегічне спрямування атак.

В роботі передбачається проведення геопросторового аналізу із використанням бази подій про атаки та кіберінциденти, що відображаються на мапі. Таким чином проведено відстеження та відображення подій на мапі в середовищі ELK. Усі дані, використані в цьому дослідженні, є загальнодоступними, безкоштовними та в цифровому форматі.

1.3 Висновки

Аналіз літератури зосереджено на двох ключових аспектах: методах ідентифікації та вилучення подій з тексту та використання геопросторового аналізу для побудови зв'язків між подіями в контексті безпеки (кібербезпеки).

У підрозділі 1.1 розглянуто основні методи ідентифікації та вилучення подій з текстових джерел. Зокрема, акцент був зроблений на сучасних методологіях обробки природної мови, машинному навчанні та аналізі семантики для вдосконалення процесу розуміння та екстрагування інформації з текстів. Підрозділ 1.2 акцентує увагу на важливості використання геопросторового аналізу для побудови зв'язків між подіями. Такий підхід значно покращує розуміння та відстеження безпекових інцидентів, а також допомагає покращити уяву про природу атак та утворених наслідків, за допомогою прив’язки до території.

Комбінування методів ідентифікації подій з тексту та геопросторового аналізу відкриває нові перспективи для розвитку систем безпеки та моніторингу подій. Ці підходи повинні використовуватися комплексно, створюючи унікальний підхід до аналізу та управління кожною подією в реальному часі.

2 ТЕХНОЛОГІЇ РОЗПІЗНАВАННЯ ІМЕНОВАНИХ ОБ'ЄКТІВ ДЛЯ АНАЛІЗУ ТА ВІЗУАЛІЗАЦІЇ ДАНИХ

Алгоритми NER разом з іншими алгоритмами, які використовуються під час цього проекту, можна знайти в [17] для індивідуальної перевірки. У цьому розділі детально описано обґрунтування та процеси, пов’язані з кожним аспектом NER для візуального та логічного розуміння.

2.1 Що таке NER?

Сучасний інформаційний простір включає в себе постійний обмін інформацією. Така інформація не завжди передається в структурованому вигляді, а іноді – великими неструктурованими масивами, що потребує обробки та фільтрації.

Прагнення отримати інформацію з таких джерел виявляється складним завданням, і саме тут виникає важлива роль обробки природної мови (Natural Language Processing, NLP). NLP є ключовим напрямком в галузі штучного інтелекту, спрямованим на розуміння, інтерпретацію та обробку мови в її природній формі.

Технологія NLP дозволяє отримувати корисну інформацію з потоків даних для подальшої обробки. Природна мова є надзвичайно складною, тому для її обробки використовуються багато методів та алгоритмів, що здатні адаптуватися до різноманітності мовленнєвих конструкцій, семантичних відтінків та контекстуальних ньюансів.

Natural Language Proccessing – це процес, який при аналізі тексту розпізнає іменовані сутності та вилучає інформацію про них. Іменовані сутності - це фрази, які містять імена осіб, організації, місця розташування, час і кількість. Розпізнавання іменованих сутностей (NER) є підзавданням вилучення інформації. NER може обробляти як структуровані, так і неструктуровані тексти шляхом ідентифікації та визначення місцезнаходження сутностей.

Початковим етапом у виконанні завдання з ідентифікації іменованих об'єктів (NER) є процес екстрагування інформації, що визначається виявленням та підготовкою специфічних об'єктів, що узгоджуються з текстами документів (рис. 2.1), абзаців, речень та текстів загалом. Весь цей етап екстрагування включає в себе ряд дій, таких як тегування мовлення, визначення границь речень, застосування правил використання великих літер і розгляд посилань в документах, які виявляються ключовими для подальшого використання та пошуку більш конкретних термінів під час пошуку інформації.

Наступним кроком є пошук потенційних об'єктів для визначення їх вмісту в документі. В процесі цього пошуку враховуються різні форми згадування, такі як імена, посилання на кілька сторінок, а також інформативні псевдоніми веб-сторінок, розглядаючи їх як синоніми. Алгоритм пошуку зберігає деякий баланс між точністю визначення та обчислювальною ефективністю, вибираючи потрібні сутності та зберігаючи лише обмежений набір сутностей, що дозволяє оптимізувати ресурси, що необхідні для розпізнавання та ідентифікації зазначених об'єктів.

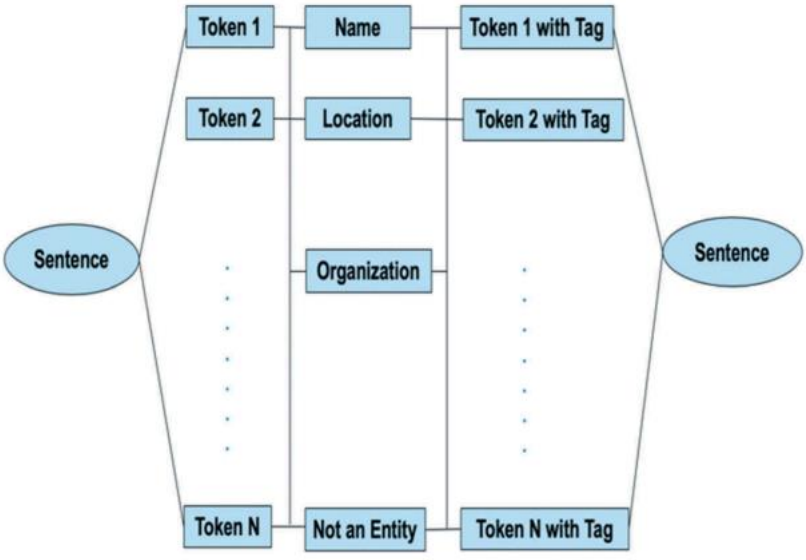

Рисунок 2.1 – Структура NER [21]

В процесі NER все, що позначається власним ім’ям або тегом, наприклад, місцезнаходження, організація чи особа, ідентифікується як об’єкт. Іменовані сутності включають такі речі, як географічне розташування, дата, час або гроші,

крім того, можливе налаштування моделі NER для визначених користувачем іменованих сутностей. Наприклад, у наведеному нижче тексті названі об’єкти позначені наступним чином:

Сьогодні [ЧАС] *директор* [ОСОБА] *Ладижинській ТЕС* [ОРГАНІЗАЦІЯ] *дав розгорнуте інтерв’ю щодо команд диспетчера по перемиканню ліній*.

Це речення містить три сутності імені: одне слово, позначене як ОРГАНІЗАЦІЯ, одне слово, позначене як ЧАС, одне слово, позначене як ОСОБА.

Таким чином, NER — це завдання обробки природної мови (NLP), яке передбачає ідентифікацію та класифікацію іменованих об’єктів (таких як імена людей, організацій, місцеположення, дати тощо) у неструктурованому тексті.

Витягуючи сутності, NER сприяє розумінню контексту інформації. Наприклад, він може відрізнити звичайне згадування назви програмного забезпечення від його згадування в контексті інциденту безпеки. NER допомагає відстежувати події та інциденти, визначаючи дати, час і місця, згадані в твітах. Це важливо для розуміння того, коли і де сталася конкретна атака. NER допомагає ідентифікувати осіб, групи чи організації, причетні до кібератак. Це стосується хакерів, експертів із безпеки або компаній, які постраждали від атаки. Витягуючи сутності, NER полегшує агрегацію інформації. Це допомагає консолідувати деталі про атаку, полегшуючи аналітикам розуміння загроз безпеці та реагування на них.

NER покращує ситуаційну обізнаність шляхом виділення та категоризації об’єктів, які мають відношення до виявлення атак. Це дозволяє аналітикам безпеки швидко зрозуміти ключові елементи інциденту безпеки. NER забезпечує швидкий і ефективний пошук інформації. Аналітики можуть шукати конкретні об’єкти, щоб зібрати інформацію про схожі інциденти або пов’язаних суб’єктів загрози. Ідентифікація ключових об’єктів допомагає сортувати та пріоритезувати сповіщення системи безпеки. Аналітики можуть зосередитися на твітах, що

містять критичні об’єкти, що призводить до більш ефективної відповіді. NER можна інтегрувати в автоматизовані системи обробки великих обсягів твітів. Ця автоматизація пришвидшує аналіз даних соціальних мереж, дозволяючи своєчасно реагувати на нові загрози [23].

Нарешті, витягнуті сутності можна порівняти з існуючими базами даних аналізу загроз. Ця інтеграція допомагає аналітикам співвідносити інформацію з соціальних мереж, наприклад Twitter із відомими профілями загроз або шаблонами атак.

2.2 Методи навчання, що застосовуються для NER

Методи навчання, що застосовуються для NER, включають навчання під контролем, напівконтрольоване навчання і навчання без контролю, кожен з яких використовує різні підходи. Навчання під контролем включає такі методи, як метод опорних векторів (SVM), моделі максимальної ентропії (ME), приховані марківські моделі (Hidden Markov model, HMM), дерева рішень та умовні випадкові поля (CRF) [6].

Для NER доступна величезна кількість бібліотек NLP, розроблених з акцентом на конкретних мовах як області розпізнавання об'єктів. Варті уваги приклади включають Stanford NER, створену за допомогою JAVA, і бібліотек на основі Python: SpaCy та TensorFlow. Ці бібліотеки надають попередньо навчені моделі NER для стандартних об'єктів (рис. 2.1), таких як Person, Organization, Time і Location. Проте, потреби реальних додатків вимагають створення нових моделей NER для виконання унікальних вимог до ідентифікації об'єктів. Оцінюючи різні моделі NER, міркування виходять за рамки простої точності і охоплюють такі фактори, як час прогнозування, розмір моделі і легкість навчання, забезпечуючи таким чином всебічну оцінку їхньої корисності для практичних застосувань.

Отримання геоданих із соціальних мереж зазвичай включають геоаналіз, усунення неоднозначності та геотеги [16].

2.3 Загальна методологія роботи

В контексті даного дослідження розроблено концептуальну схему (рис. 2.2), згідно якої відбувається вилучення сутностей з потоку даних.

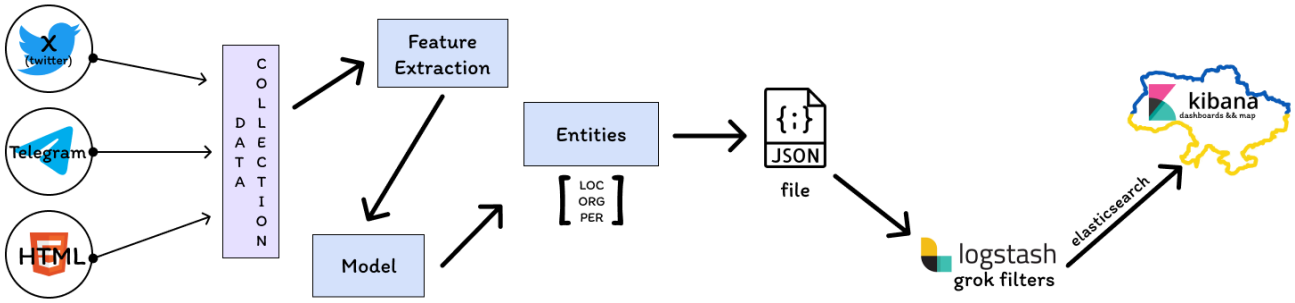

Рисунок 2.2 – Концептуальна схема

Алгоритм роботи, включаючи етапи від збору даних до візуалізації, починається з визначення конкретних об'єктивів аналізу та вибору джерел даних, які відповідають цим цілям. Під час збору даних необхідно враховувати різноманітність джерел, від баз даних та сенсорів до текстових джерел та соціальних мереж. Для забезпечення ефективності та автоматизації процесу збору, використання спеціалізованих інструментів та технологій є ключовим. Після збору даних важливим є їхнє очищення та перевірка на наявність помилок, відсутніх значень чи дублікатів, щоб забезпечити якість вихідного набору.

Далі, на етапі попередньої обробки даних, проводиться їхня структуризація та перетворення для виготовлення їх придатними для аналізу. Це може включати зміни формату, кодування категорій, інтеграцію різних джерел та перевірку якості. Попередня обробка даних також включає створення метаданих, які допомагають зберігати інформацію про походження та характеристики даних. Після цих етапів даних готові для подальшого використання в аналітичних задачах. Наступним етапом є вибір та застосування відповідних методів аналізу, а також візуалізація отриманих результатів для зручності сприйняття та прийняття вирішень.

Як зазначається в [20], інтеграція просторових та часових даних дозволяє виконувати інтелектуальний аналіз тексту з урахуванням географічного контексту. Цю ідею покладено з основку робочого процесу ідентифікації та вилучення географічно пов'язаних об'єктів для аналізу інформації з текстових файлів на основі розпізнавання топонімів і відображено на рис. 2.3.

Схематично, процес класифії топонімів та оцінки відстані представлено на рис. 2.4. На вхід подаються різні типи даних, для прикладу дані із web-сторінки, повідомлення Telegram, пости X, що збережені у різних форматах, для прикладу: html, pdf або txt відповідно.

За допомогою моделей розпізнавання топонімів здійснюється виявлення геоточок (країн, міст, населених пунктів). Дані геоточки за допомогою розроблених алгоритмів конвертуються одразу в регіональний код, або спершу перевіряються на приналежність до конкретного регіону, а тоді вже отримується регіональний код відповідного регіону.

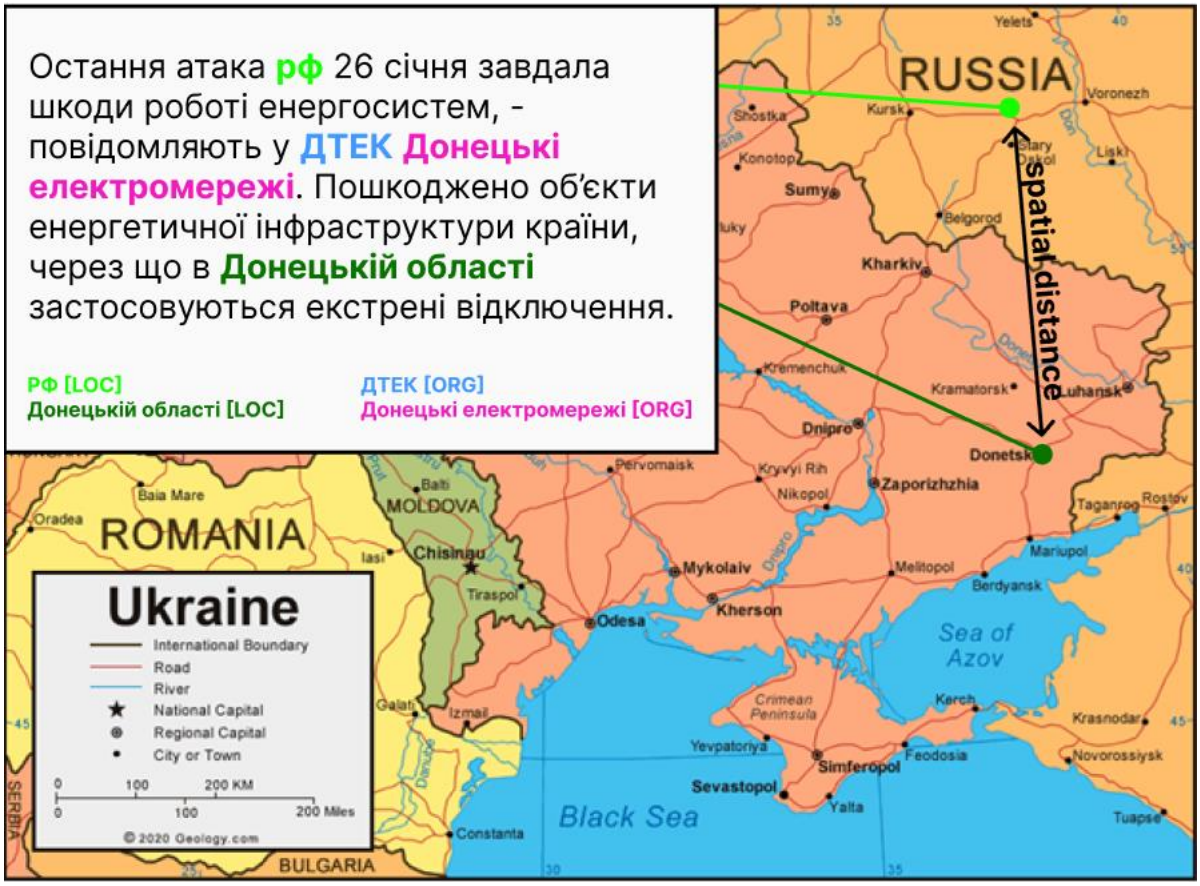

Рисунок 2.3 – Розпізнавання топонімів

Якщо в тексті вказані 2 геоточки, то в результаті здійснюється обрахунок мінімальної відстані між даними точками (регіонами).

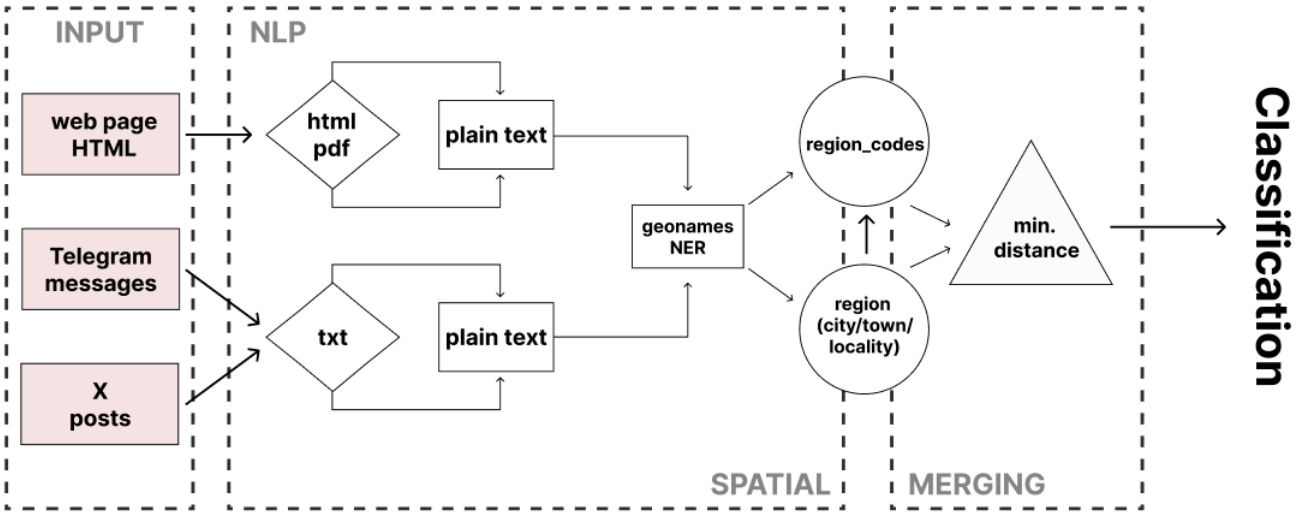

Рисунок 2.4 – Схема на основі розпізнавання топонімів та оцінки відстані

Ще одним можливим варіантом застосування даної схеми є використання координат геоточок. Таким чином, кожний населений пункт відповідатиме певним координатам, за якими здійснюється їхнє відстеження на карті та обраховується відстань між ними чи від точки до конкретної цілі (відповідно до потреб). Для реалізації такого варіанту необхідно розробити алгоритм отримання координат кожного населеного пункту з бази даних, та додатково створити таку базу, якщо її не буде у відкритому доступі, або використовувати API інтерфейси онлайн-сервісів, для отримання координат. Таким чином описано ще один можливий вектор розвитку ідеї, на якій базуватиметься розробка комлексного підходу.

2.4 Збір даних

В сучасному світі практично кожна важлива подія відображена в Інтернеті, люди публікують багато інформації, не завжди підозрюючи яку цінність вона має. Збір даних відбувається з відкритих джерел, таких як: платформа Twitter (X), новинні ресурси, спільноти Telegram, сайти державних структур та ін. з використанням технологій web crawling [10] і web scrapping [12].

Соціальна мережа Twitter, з липня 2023 року - X, пропонує доволі широкий спектр інформації та зручний пошук із можливістю розширеного пошуку за великою кількістю критеріїв, таких як:

- Пошук із використанням усіх слів, що введено в пошуковому рядку;
- Пошук із використанням конкретної фрази;
- Пошук із використанням будь-якого слова із введених в пошуковому рядку;
- Не демонструвати пости, у яких міститиметься введені в пошуку слова;
- Пошук за хештегами;
- Пошук лише за обраною мовою;
- Пошук конкретного профілю чи публікацій даного профілю.

@NetBlocks – онлайн Інтернет обсерваторія, що займається відстеженням підключень до мережі Інтернет. Вони аналізують ситуацію у світі та при цьому використовують аналіз мережевої активності. Таким чином вони надають інформацію про подію (зникнення мережі, «падіння» сервісів, мережеві атаки і т.д.) та причину виникнення цієї події, якщо вдається її встановити. Кожний містить відповідне повідомлення, хештег, з позначкою певної країни чи регіону та графіків, що підтверджують дану подію (рис. 2.5).

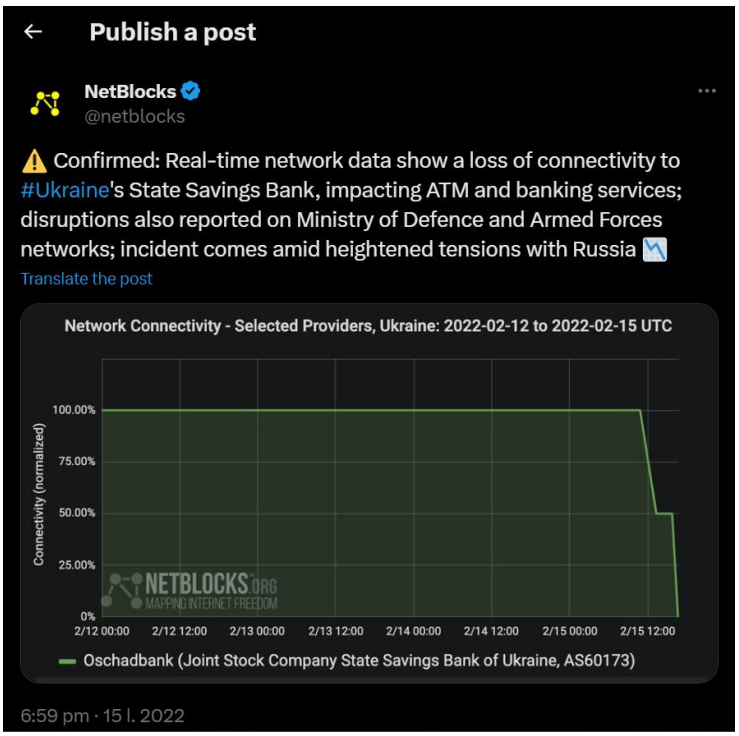

Рисунок 2.5 – Вигляд постів в соціальній мережі Twitter

Енергетика під куполом (@rada_en) - найбільший український телеграм-канал про відновлювальні джерела енергії, актуальні законодавчі ініціативи та енергетичний порядок денний в Україні та світі. Канал публікує інформацію, що містить результат діяльності країни агресора.

Державної служби спеціального зв’язку та захисту інформації (@dsszzi_official) – телеграм канал, в якому публікуються щоденні події широкомасштабної збройної агресії російської федерації проти України.

ДСНС України (@dsns_telegram) – офіційний телеграм канал Державної служби України з надзвичайних ситуацій. Публікації інформативного характеру містять інформацію про: діяльність підрозділів ДСНС, запобігання, ліквідацію та захист від наслідків надзвичайних ситуацій (до яких входять наслідки збройних атак країни агресора), рятувальну справу і т. д.

Оперативний ЗСУ (@operativnoZSU) – новини про ситуацію на фронті та атаки на Україну.

Таблиця 2.1 – Кількість джерел використовуваних в даній роботі

Джерело	**Назва ресурсу Канал / URL**	**Зібрані документи (txt)**	**Помічено як атака**
X	@NetBlocks	1100	-
X	Енергетика під куполом (@rada_en)	2158	
Telegram	Державна служба спеціального зв'язку та захисту інформації (@dsszzi_official)	5282	
Telegram	ДСНС України (@dsns_telegram)	10792	
Telegram	Оперативний ЗСУ (@operativnoZSU)	90265	
HTML	Щоденні аналітичні статті Тома Купера про ситуацію в Україні https://shorturl.at/msCT9		
Декілька джерел	Таблиця з описом інцидентів на об'єктах критичної інфраструктури з відкритих джерел (телеграм, твіттер) https://shorturl.at/hmpBW		

2.5 Аналіз даних та візуалізація

Дане дослідження передбачає використання бібліотеки spaCy, за допомогою якої було реалізовано виділення ключових елементів з потоку даних (таблиці) та текстових файлів. Бібліотека spaCy надає готові моделі для розпізнавання іменованих об'єктів для різних мов, що дозволяє ефективно визначати та виділяти іменовані об'єкти, такі як особи, місця, організації та інші ключові елементи інформації в тексті. Застосування spaCy дозволяє автоматизувати цей процес, що є критичним для обробки великих обсягів текстової інформації.

Процес використання spaCy для розпізнавання іменованих об'єктів включає завантаження відповідної мовної моделі, передачу текстових даних до

цієї моделі та отримання результатів, які містять інформацію про знайдені іменовані об'єкти та їх класифікацію.

Завантаження необхідної мовної моделі здійснюється за допомогою команди

python -m spacy download en_core_web_trf

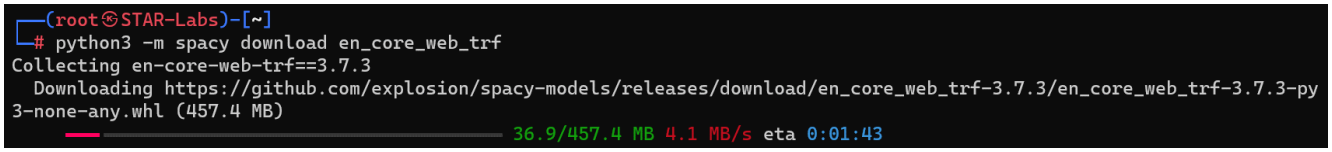

Рисунок 2.6 – Завантаження мовної моделі

Після завантаження моделі (рис. 2.6) слід перевірити, яку інформацію бібліотека може розпізнати з наданих користувачем даних. В даному випадку, вхідними даними для моделі є таблиця з даними про атаки на критичну інфраструктуру України.

Лістинг 2.1 – Візуалізація іменованих сутностей

```
import pandas as pd
import spacy
from spacy import displacy
import argparse

def save_entities_to_file(entities, output_file):
    with open(output_file, 'w') as file:
        for entity in entities:
            file.write(f"{entity[0]}: {entity[1]}\n")

def render_and_save_visualization(texts, displacy_options, output_file):
    combined_text = ' '.join(texts)
    doc = nlp(combined_text)
    html = displacy.render(doc, style="ent", **displacy_options)
    with open(output_file, 'w', encoding='utf-8') as file:
        file.write(html)


parser = argparse.ArgumentParser(description='Analyze and visualize named
entities in a data.')
parser.add_argument('output_file')
parser.add_argument('visualization_file')
args = parser.parse_args()
data = pd.read_excel("table.xlsx")
print(f"All columns: {list(data.columns)}")
nlp = spacy.load("en_core_web_sm")

if data.empty:
    print("No rows in the table")
else:
    all_entities = []
    all_texts = []
    for index, row in data.iterrows():
        print(f"\nAnalysis for row {index + 1}:")
```

Продовження лістингу 2.1

```
            text_entities = []
            for column in data.columns:
                doc = nlp(str(row[column]))
                entities = [(ent.text, ent.label_) for ent in doc.ents]
                print(f"\t- {column}:")
                for entity in entities:
                    print(f"\t\t+ {entity[0]}: {entity[1]}")
                    text_entities.append(entity)

            all_entities.extend(text_entities)
            all_texts.append(' '.join([ent[0] for ent in text_entities]))

        if all_entities:
            save_entities_to_file(all_entities, args.output_file)
            print(f"\nNamed entities saved to {args.output_file}")
        if all_texts:
            render_and_save_visualization(all_texts, {},
    args.visualization_file)
            print(f"\nVisualization saved to {args.visualization_file}")

        print("\nAnalysis complete.")
```

Для запуску програмного коду із лістингу 2.1 необхідно використати інтерпретатор python3 та задати назву файлу для збереження проаналізованих даних та назву файлу візуалізації (за допомогою бібліотеки spaCy), приклад наведено на рисунку 2.7.

```
┌──(root㉿STAR-Labs)-[~/NER]
└─# python3 spacy_analyse_to_file.py analysed_information.txt visualize_information.html
All columns: ['date', 'time', 'target_name', 'Critical_instrastructure_sector ', 'target_location ', 'region_code', 'att
ack_type', 'attack_domain', 'Summary_of_the_attack', 'Result', 'Countermeasures/Comments', 'Reference']

Analysis for row 1:
        - date:
                + 15.02.2022: CARDINAL
        - time:
        - target_name:
                + Ukraine: GPE
                + State Savings Bank: ORG
```

Рисунок 2.7 – Запуск розпізнавання іменованих об’єктів

В результаті виконання коду (лістинг 2.1) отримано файл *analysed_information.txt* в якому міститься ключова інформація про розпізнавані іменовані об’єкти, та *visualize_information.html* – html-сторінка, яка графічно відображає іменовані об’єкти в тексті (рис. 2.8).

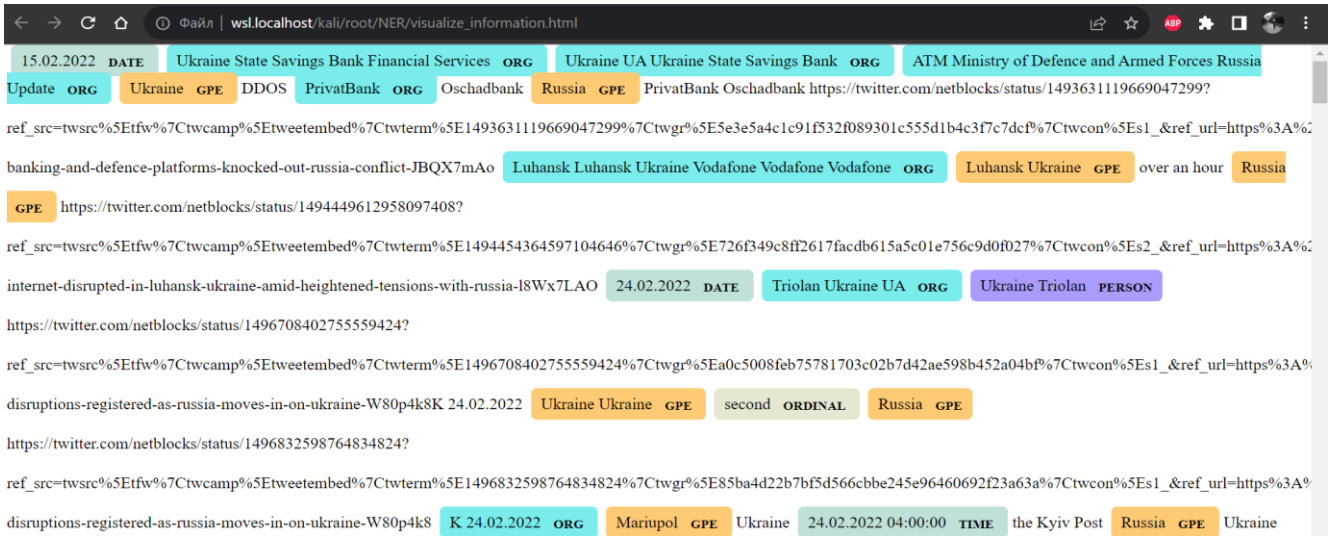

Рисунок 2.8 – Графічне відображення іменованих об’єктів

2.6 Висновки

У даному розділі розглянуто ключові аспекти, пов'язані з розпізнаванням іменованих об'єктів (NER) в контексті аналізу атак на критичну інфраструктуру. Починаючи з визначення самого поняття NER в підрозділі 2.1, проведено огляд різних методів навчання, які використовуються для ефективного розпізнавання і класифікації іменованих об'єктів.

Загальна методологія роботи, що була ретельно розглянута в підрозділі 2.3, включає важливі етапи від збору даних до аналізу та візуалізації. Процес збору даних визначає критичну основу для подальших етапів дослідження, дозволяючи отримати необхідні вхідні дані для навчання та тестування системи NER. Аналіз даних та їх візуалізація, як важливі аспекти дослідження, дозволяють не лише виявити іменовані об'єкти, але й створити зрозумілу графічну репрезентацію взаємозв'язків між ними.

У результаті проведеного дослідження можна зробити висновок, що ефективне використання технологій NER у поєднанні з методами навчання та аналізу даних є ключовим для розпізнавання та класифікації іменованих об'єктів в контексті атак на критичну інфраструктуру. Такий підхід може виявитися корисним для підвищення ефективності систем безпеки та виявлення потенційних загроз для інфраструктури національного рівня.

3. ПРАКТИЧНА РЕАЛІЗАЦІЯ

3.1Попередня обробки даних

Попередня обробка інформації включає в себе аналіз та сортування даних, отриманих із текстових джерел, а саме заповнення відповідних полів таблиці (рис. 3.1) та збереження інформаційних повідомлень у текстові файли. Таблиця містить наступні поля:

- Дата;
- Час (якщо вдається визначити);
- Ціль (об’єкт);
- Тип критичної інфраструктури;
- Ціль (місто, локація);
- Код регіону (згідно ISO 3166-2 [9]);
- Тип атаки;
- Домен атаки (кібер атака, фізична, соціальна, гібридна);
- Короткий опис атаки (взяти зі статті чи постів);
- Результат атаки (якщо відомо);
- Контрзаходи;
- Посилання.

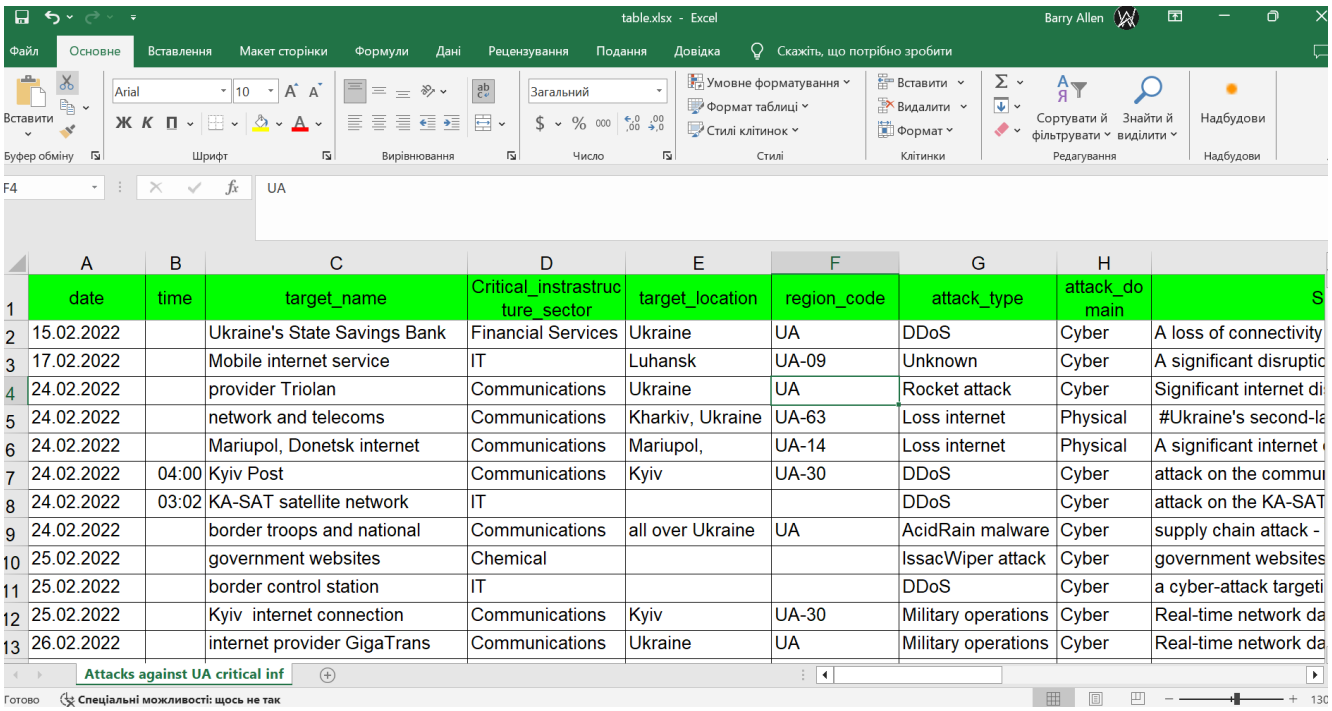

table.xlsx - Excel

	A	B	C	D	E	F	G	H	S
1	date	time	target_name	Critical_instrastructure_sector	target_location	region_code	attack_type	attack_domain	
2	15.02.2022		Ukraine's State Savings Bank	Financial Services	Ukraine	UA	DDoS	Cyber	A loss of connectivity
3	17.02.2022		Mobile internet service	IT	Luhansk	UA-09	Unknown	Cyber	A significant disruptic
4	24.02.2022		provider Triolan	Communications	Ukraine	UA	Rocket attack	Cyber	Significant internet di
5	24.02.2022		network and telecoms	Communications	Kharkiv, Ukraine	UA-63	Loss internet	Physical	#Ukraine's second-la
6	24.02.2022		Mariupol, Donetsk internet	Communications	Mariupol,	UA-14	Loss internet	Physical	A significant internet
7	24.02.2022	04:00	Kyiv Post	Communications	Kyiv	UA-30	DDoS	Cyber	attack on the commu
8	24.02.2022	03:02	KA-SAT satellite network	IT			DDoS	Cyber	attack on the KA-SAT
9	24.02.2022		border troops and national	Communications	all over Ukraine	UA	AcidRain malware	Cyber	supply chain attack -
10	25.02.2022		government websites	Chemical			IssacWiper attack	Cyber	government websites
11	25.02.2022		border control station	IT			DDoS	Cyber	a cyber-attack targeti
12	25.02.2022		Kyiv internet connection	Communications	Kyiv	UA-30	Military operations	Cyber	Real-time network da
13	26.02.2022		internet provider GigaTrans	Communications	Ukraine	UA	Military operations	Cyber	Real-time network da

Attacks against UA critical inf

Рисунок 3.1 – Заповнення таблиці з даними

Збір та аналіз інформації з телеграм каналів здійснено за наступною схемою (рис.3.2):

Рисунок 3.2 – Алгоритм аналізу текстових даних з Telegram

Для того, щоб реалізувати завантаження текстових даних з усіх постів з зазначених каналів Telegram було реалізовано функцію *collect_data()* (рис. 3.3). На вхід функція приймає є значення api_id, api_hash, список каналів та шлях до робочої директорії куди зберігатимуться текстові дані.

```python
# Colelct messages from Telegram channels
def collect_data(api_id, api_hash, channels, directory_path):
    logging.basicConfig(level=logging.DEBUG)
    client = TelegramClient('client', api_id, api_hash)
    async def save_messages(channel):
        async for message in client.iter_messages(channel):
            if message.text:
                folder_name = os.path.join(directory_path, f'{channel[1:]}')
                os.makedirs(folder_name, exist_ok=True)
                file_path = os.path.join(folder_name, f'{channel[1:]}_{message.id}.txt')
                with open(file_path, 'w', encoding='utf-8') as file:
                    file.write(message.text)
                    message_link = f"https://t.me/{channel[1:]}/{message.id}"
                    file.write(f"\n\nMessage Link: {message_link}")
                    file.write(f"\nPublication Date: {message.date}")

    with client:
        for channel in channels:
            client.loop.run_until_complete(save_messages(channel))
            print(f"Messages from {channel} collected.")
```

Рисунок 3.3 – Функція collect_data

Метод *collect_data* реалізує збір текстових повідомлень з каналів Telegram. Спочатку, він налаштовує параметри та ініціалізує клієнт TelegramClient з використанням наданих ідентифікатора та хеш-ключа API. Далі, використовуючи асинхронний цикл, метод ітеративно переглядає повідомлення кожного каналу, перевіряє наявність текстового контенту та зберігає кожне повідомлення у відповідному файлі, разом із додатковою інформацією, такою як посилання на повідомлення та дата публікації.

Даний метод допомагає автоматизувати процес збору інформації повідомлень з Telegram із вказаного каналу.

Наступний метод - *process_files()* виконує обробку текстових файлів, які містять зібрані повідомлення з вказаних каналів Telegram. На початку кожної ітерації по списку каналів, визначається порожній список *results_for_channel* для збереження результатів обробки файлів конкретного каналу.

Усередині циклу по файлам кожного каналу, метод перевіряє, чи файл має розширення ".txt", і якщо так, визначає шлях до файлу та ініціює змінні для збереження посилання на повідомлення та часу публікації. Шлях до файлу формується шляхом конкатенації значення *directory_path* та назви каналу.

Наступним кроком, метод відкриває кожен файл у режимі читання та зчитує дані з нього, з вилученням посилання на пост та дату публікації. Ця інформація передається до функції *extract_information* (рис. 3.5), яка відповідає за подальшу обробку та витягнення додаткової інформації з текстового файлу.

У кінці кожної ітерації отримані дані зберігаються у форматі JSON у файлі з форматом {channel[1:]}_result.json. Цей файл містить список словників, де кожен словник представляє інформацію про конкретний файл, включаючи назву файлу та витягнуті дані (рис. 3.4). Після завершення обробки всіх файлів виводиться відповідне повідомлення, що відображає завершення процесу та розміщення файлу із результатом роботи (JSON файлу).

```
# Iterate all text files in the directory
def process_files(directory_path, channels):
    for channel in channels:
        results_for_channel = []

        channel_directory_path = os.path.join(directory_path, channel[1:])
        for filename in os.listdir(channel_directory_path):
            if filename.endswith(".txt"):
                file_path = os.path.join(channel_directory_path, filename)
                message_time = None
                message_link = None

                # Get the message link and publication date from the text file
                with open(file_path, "r", encoding="utf-8") as file:
                    lines = file.readlines()
                    if lines:
                        last_line = lines[-1]
                        if last_line.startswith("Message Link:"):
                            message_link = last_line.split(":")[1].strip()
                        if last_line.startswith("Publication Date:"):
                            message_time = last_line.split(":")[1].strip()

                print("\nExtracting information from:", file_path)
                extracted_info = extract_information(file_path, message_time, message_link)
                print("Extracted Information:", extracted_info)
                results_for_channel.append({
```

Рисунок 3.4 – Функція process_files

Останній метод **extract_information()** використовує бібліотеку spaCy для обробки текстового файлу, отримуючи інформацію про сутності (entities), такі як іменовані особи, місця та інші ключові елементи тексту. У початковому етапі, метод приймає шлях до текстового файлу (file_path), а також інформацію про час публікації повідомлення (message_time) та посилання на повідомлення (message_link), що відображено на рис. 3.5.

```python
# Load model for Ukrainian language and analyse messages from *.txt files using NLP
nlp = spacy.load("uk_core_news_trf")
def extract_information(file_path, message_time, message_link):
    with open(file_path, "r", encoding="utf-8") as file:
        text = file.read()
    doc = nlp(text)
    # Extract entities using spaCy's NER
    entities = [(ent.text, ent.label_) for ent in doc.ents]
    return {"message_time": message_time, "message_link": message_link, "extracted_info": entities}
```

Рисунок 3.5 – Функція extract_information

Під час виконання методу, вміст текстового файлу зчитується та передається до обробки за допомого моделі spaCy, яка була завантажена зазначеним попередньо завантаженим пакетом української мови ("uk_core_news_trf"). Щоб завантажити україномовний пакет необхідно використати команду:

python3 -m spacy download uk_core_news_trf

Для завантажених з телеграм даних здійснюється аналіз за наступною схемою, що відображена на рис. 3.6.

Інформація з текстових файлів аналізується бібліотекою spaCy за допомогою функції nlp, а результати зберігаються в форматі JSON.

Рисунок 3.6 – Аналіз текстових файлів за допомогою бібліотеки spaCy

За допомогою Named Entity Recognition (NER) spaCy виділяє сутності в тексті, інтерпретуючи їхні типи, такі як особи, локації, події тощо. Отримані сутності представлені у вигляді списку кортежів, де кожен кортеж містить текст сутності та її відповідний тип. Результати обробки (час повідомлення, посилання та витягнуті сутності) повертаються у вигляді словника для подальшого використання в основному методі, який збирає та обробляє дані з різних файлів.

Збереження результатів обробки тексту відбувається у файл із форматом {{channel_name}}_result.json (рис. 3.7). Формат JSON доволі простий у роботі і

дозволяє легко сортувати та отримувати необхідну інформацію і не вимагає значних витрат ресурсів процесора чи оперативної пам’яті.

```
    # Save as JSON
    output_path = f"{channel[1:]}_result.json"
    with open(output_path, "w", encoding="utf-8") as json_file:
        json.dump(results_for_channel, json_file, ensure_ascii=False, indent=4)
    print(f"\nData from {channel} saved to:", output_path)
```

Рисунок 3.7 – Збереження результатів у форматі JSON

В результаті розроблено комплексне рішення для формування бази даних (таблиці) із завантажених файлів та подальшим їх аналізом. В базі відображено наступні дані:

- channel_name – назва каналу; шрифт зробіть Times New Roman
- post_date – дата публікації;
- text_data – текстові дані з повідомлення;
- link – посилання на пост;
- file_path – місце розташування файлу із повідомленням;

Реалізація потрібного функціоналу потребує внесення змін в метод process_file(). Кінцевий варіант відображено на рис. 3.8.

```
def process_files(root_folder, output_excel, regions_data, regions_code_data):
    nlp = spacy.load('uk_core_news_trf')
    df_list = []
    for object_name in os.listdir(root_folder):
        object_path = os.path.join(root_folder, object_name)
        if os.path.isdir(object_path):
            for txt_file in tqdm(os.listdir(object_path), desc=object_name, position=0, leave=True):
                file_path = os.path.join(object_path, txt_file)
                with open(file_path, 'r', encoding='utf-8') as file:
                    text = file.read()

                date_match = re.search(r'\b\d{4}-\d{2}-\d{2}\b', text)
                if date_match:
                    message_link = text.split('\n')[-2].split(': ')[1]
                    text_from_file = text.split(f'Message Link: {message_link}')[0]
                    doc = nlp(text_from_file)
                    ner_results = [ent.text for ent in doc.ents]
                    cities_in_text = []
                    for result in ner_results:
                        cities_in_text.extend(result.split())

                    region_for_city = None
                    for city_in_text in cities_in_text:
                        region_for_city = get_region_for_city(city_in_text, regions_data, nlp)
                        if region_for_city:
                            break
```

Рисунок 3.8 – Модернізація функції process_files

3.2 Перетворення об'єктів на географічні одиниці

Перетворення об’єктів на координати здійснюється за допомогою тієї ж бібліотеки spaCy, в разі розпізнавання міст/країн вона конвертує дані в кодовий формат згідно ISO 3166-2. Стандарт ISO 3166-2 є другою частиною більш широкого стандарту ISO 3166, що був розроблений та опублікований Міжнародною організацією зі стандартизації (ISO). Ця система геокодів призначена для кодування назв основних адміністративно-територіальних одиниць на першому та, в окремих випадках, на другому рівні адміністративно-територіального поділу територій усіх країн та незалежних регіонів [5].

Щоб отримати дані про регіони, а точніше назви населених пунктів, міст та країн, використано бібліотеку spaCy та модель uk_core_news_trf. При аналізі тексту за допомогою бібліотеки spaCy вдалося вилучити іменовані сутності наступних типів що відображені у таблиці 3.1.

Таблиця 3.1 – Типи іменованих сутностей

Тип іменованої сутності	**Опис**
[LOC]	розташування
[ORG]	організація, установа
[PER]	персона

Отримані дані відображено в колонці “NER” на рис. 3.9.

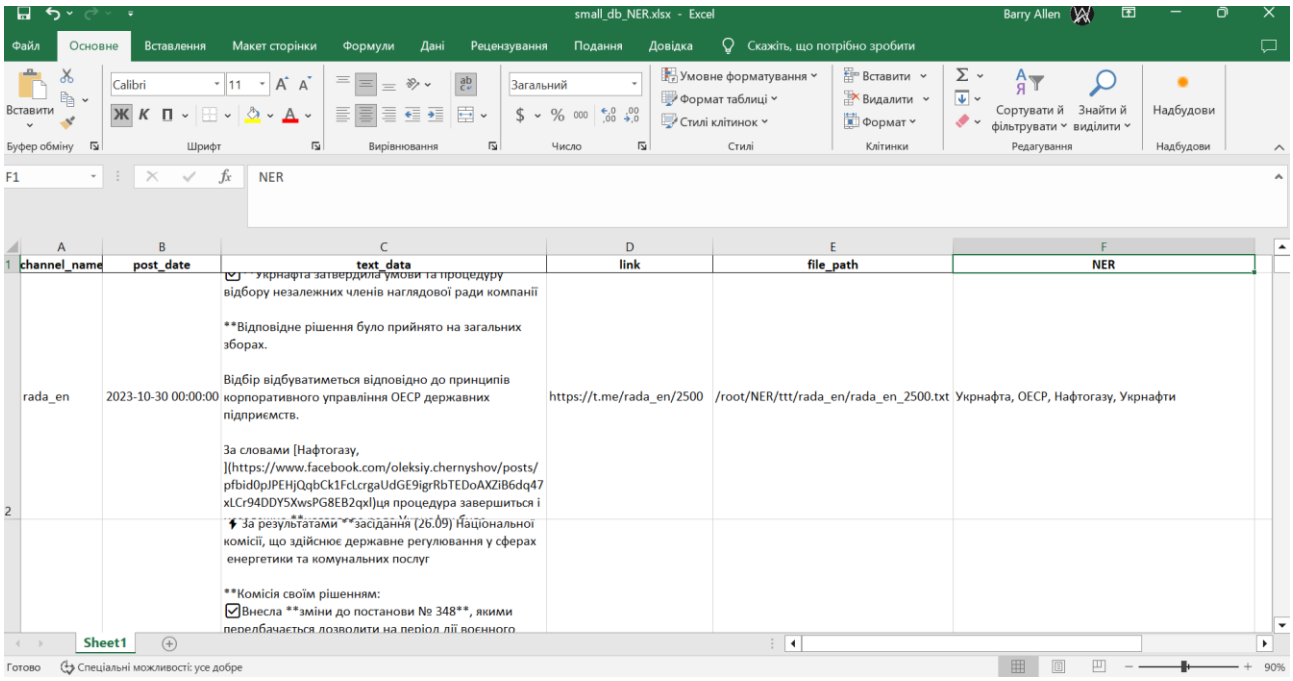

Рисунок 3.9 – Відображення результатів вилучення іменованих сутностей

Наступний крок – визначення регіонального коду для відображення подій на мапі. Щоб отримати регіональний код було створено колекцію *regions_code_ukr.json* із усіма регіонами України та їх регіональним кодом (згідно ISO-3166-2)[5]. Формат даної колекції відповідає формату “ключ:значення”, вміст колекції відображено на рис. 3.10.

```
┌──(root㉿STAR-Labs)-[~/NER/ttt]
└─# cat ../attack/regions_code_ukr.json
{
    "Вінницька область": "UA-05",
    "Волинська область": "UA-07",
    "Дніпропетровська область": "UA-12",
    "Донецька область": "UA-14",
    "Житомирська область": "UA-18",
    "Закарпатська область": "UA-21",
    "Запорізька область": "UA-23",
    "Івано-Франківська область": "UA-26",
    "Київська область": "UA-32",
    "Кіровоградська область": "UA-35",
    "Автономна Республіка Крим": "UA-43",
    "Луганська область": "UA-09",
    "Львівська область": "UA-46",
    "Миколаївська область": "UA-48",
    "Одеська область": "UA-51",
    "Полтавська область": "UA-53",
    "Рівненська область": "UA-56",
    "Сумська область": "UA-59",
    "Тернопільська область": "UA-61",
    "Харківська область": "UA-63",
    "Херсонська область": "UA-65",
    "Хмельницька область": "UA-68",
    "Черкаська область": "UA-71",
    "Чернівецька область": "UA-77",
    "Чернігівська область": "UA-74",
    "Київ" : "UA-30",
    "Севастополь" : "UA-40"
```

Рисунок 3.10 – Регіональний код для кожної області України

Додатково, створено колекцію *regions.json* (рис. 3.11), у якій відображені області (регіони) та міста, які входять до складу даних областей. Для автоматизації визначення приналежності міста до певного регіону та отримання регіонального коду було модифіковано метод **process_files()**.

```
┌──(root㉿STAR-Labs)-[~/NER/ttt]
└─# cat ../attack/regions.json
{
    "Вінницька область": ["Вінниця", "Хмільник", "Могилів-Подільський", "Козятин", "Жмеринка", "Тульчин", "Ладижин", "Гайсин", "Бар"],
    "Волинська область": ["Луцьк", "Ковель", "Нововолинськ", "Камінь-Каширський", "Володимир-Волинський", "Рожище", "Любомль", "Устилуг", "Шацьк"],
    "Дніпропетровська область": ["Дніпро", "Кривий Ріг", "Дніпровськ", "Нікополь", "Павлоград", "Кам'янське", "Жовті Води", "Новомосковськ", "П'ятихатки"],
    "Донецька область": ["Донецьк", "Маріуполь", "Макіївка", "Харцизьк", "Ясинувата", "Слов'янськ", "Краматорськ", "Артемівськ", "Дружківка"],
    "Житомирська область": ["Житомир", "Бердичів", "Коростень", "Новоград-Волинський", "Малин", "Баранівка", "Іршанськ", "Маневичі", "Овруч"],
    "Закарпатська область": ["Ужгород", "Мукачево", "Берегово", "Хуст", "Тячів", "Свалява", "Виноградів", "Міжгір'я", "Іршава"],
    "Запорізька область": ["Запоріжжя", "Мелітополь", "Бердянськ", "Приморськ", "Енергодар", "Токмак", "Пологи", "Василівка", "Молочанськ"],
    "Івано-Франківська область": ["Івано-Франківськ", "Коломия", "Інтернаціональна", "Снятин", "Долина", "Калуш", "Яремче", "Болехів", "Бурштин"],
    "Київська область": ["Бровари", "Буча", "Вишневе", "Васильків", "Обухів", "Ірпінь", "Фастів", "Боярка", "Біла Церква"],
    "Кіровоградська область": ["Кіровоград", "Олександрія", "Світловодськ", "Знам'янка", "Гайворон", "Долинська", "Кропивницький", "Новомиргород", "Маловисків"],
    "Автономна Республіка Крим": ["Сімферополь", "Ялта", "Алушта","Феодосія", "Керч", "Алупка", "Євпаторія", "Армянськ", "Бахчисарай"],
    "Луганська область": ["Луганськ", "Алчевськ", "Сєвєродонецьк", "Лисичанськ", "Краснодон", "Сорокине", "Сватове", "Старобільськ", "Антрацит"],
    "Львівська область": ["Львів", "Дрогобич", "Самбір", "Стрий", "Тернопіль", "Івано-Франківськ", "Трускавець", "Сколе", "Борислав"],
```

Рисунок 3.11 – Приналежність міст до областей

В лістингу 2.1, наведено автоматизацію процесу аналізу текстових повідомлень з *.txt файлів, з подальшим вилученням іменованих сутностей декількох типів, що наведені в таблиці 1, вилучення релевантної інформації з подальшим створенням структурованого DataFrame і експортом результатів у файл Excel. Метод використовує бібліотеку spaCy, зокрема модель 'uk_core_news_trf', для обробки природної мови.

Також, метод сканує усі директорії в робочому каталозі та переглядає кожний *.txt файл знайдений в цих директоріях. Програма вилучає посилання на пост, текстовий вміст поста і виконує розпізнавання іменованих об'єктів (NER), використовуючи моделі spaCy для ідентифікації об'єктів у тексті. Після чого результати NER обробляються для виявлення міст, згаданих у тексті. Для кожного міста метод намагається визначити відповідний регіон із файлу regions.json, порівнюючи його її з попередньо визначеним списком міст,

пов'язаних з регіонами. Якщо місто співпадає, отримується назва регіону і по цій назві із файлу regions_code_ukr.json отримується регіональний код (рис. 3.12).

```
        date_match = re.search(r'\b\d{4}-\d{2}-\d{2}\b', text)
        if date_match:
            message_link = text.split('\n')[-2].split(': ')[1]
            text_from_file = text.split(f'Message Link: {message_link}')[0]
            doc = nlp(text_from_file)
            ner_results = [ent.text for ent in doc.ents]
            cities_in_text = []
            for result in ner_results:
                cities_in_text.extend(result.split())

            region_for_city = None
            for city_in_text in cities_in_text:
                region_for_city = get_region_for_city(city_in_text, regions_data, nlp)
                if region_for_city:
                    break

            region_code_for_city = None
            if region_for_city:
                region_code_for_city = regions_code_data.get(region_for_city)

            data_dict = {
                'channel_name': object_name,
                'post_date': pd.to_datetime(date_match.group()),
                'text_data': text_from_file,
                'link': message_link,
                'file_path': file_path,
                'NER': ', '.join(ner_results),
                'region': region_for_city,
                'region_code': region_code_for_city,
            }
```

Рисунок 3.12 – Автоматизація отримання регіонального коду

Отримані дані (назву каналу, дату публікації, текстові дані, посилання на повідомлення, шлях до файлу, результати NER, ідентифікований регіон і код регіону, збираються в словник для кожного файлу). Ці словники об'єднуються у список, і, на основі цього списку створюється Pandas DataFrame. Із створеного DataFrame дані експортуються в Excel-файл.

По суті, метод *process_files* поєднує методи обробки природної мови, розпізнавання дат та ідентифікацію регіонів для перетворення неструктурованих текстових даних у структурований і придатний для аналізу формат, надаючи цінну інформацію про географічний контекст інформації, що міститься в проаналізованих файлах. Створення бази даних (таблиці), що містить вилучені іменовані сутності, а також населені пункти та їх регіональний код продемонстровано на рис. 3.13.

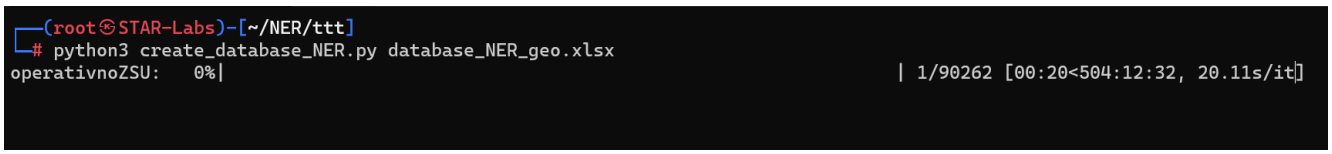

Рисунок 3.13 – Створення бази

Як можемо бачити, процес аналізу файлів та вилучення сутностей доволі вимогливий до ресурсів та вимагає багато часу, адже здійснюється аналіз абсолютно кожного слова. До прикладу, серед наявних повідомлень, їх загальний список складає близько 111 тисяч файлів. Даний процес можливо прирівняти до майнінгу криптовалюти.

3.3 Візуалізація даних

Процес відображення отриманих даних включає в себе розгортання та налаштування відповідної систем. В даній роботі пропонується використання стеку ELK (Elasticsearch Logstash Kibana).

В першу чергу необхідно привести дані із xlsx формату в json, для цього розроблено скрипт *xlsx_to_json.py*, який відображено в лістингу 3.1. Функціонал конвертера доволі простий, він призначений для перетворення даних з таблиці Excel (у форматі XLSX) у файл JSON. Скрипт використовує бібліотеку pandas для зчитування вхідного XLSX-файлу в DataFrame. Після цього він обробляє стовпець "*post_date*" і перетворює його у формат datetime. Потім DataFrame перетворюється на рядок у форматі JSON за допомогою методу 'to_json', вказуючи орієнтацію 'records' і встановлюючи для 'force_ascii' значення false, щоб обробляти символи, що не є ASCII символами. Отриманий рядок JSON розбирається на список записів, і кожен запис записується у вказаний вихідний JSON-файл. Для запуску програми необхідно вказати xlsx файл та назву вихідного json файлу. Приклад використання відображено на рис. 3.14.

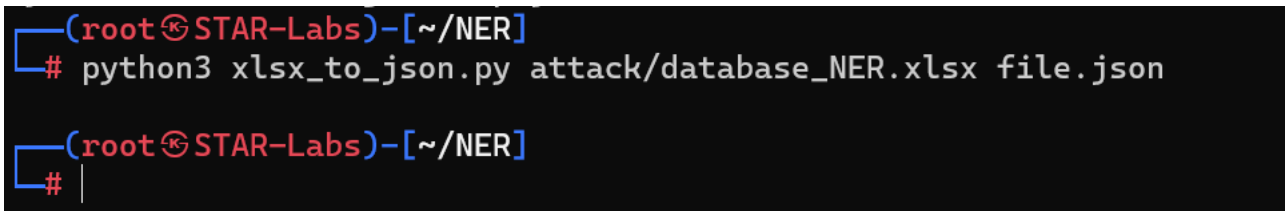

Рисунок 3.14 – Приклад використання конвертера з xlsx в json

Лістинг 3.1 – Конвертер з xlsx в json

```
import pandas as pd
import sys
import json

def convert_xlsx_to_json(xlsx_file, json_file):
    data = pd.read_excel(xlsx_file)
    data['post_date'] = pd.to_datetime(data['post_date'], unit='ms')
    data_json = data.to_json(orient='records', force_ascii=False)
    parsed_data = json.loads(data_json)
    with open(json_file, 'w', encoding='utf-8') as f:
        for record in parsed_data:
            json.dump(record, f, ensure_ascii=False)
            f.write('\n')
if len(sys.argv) != 3:
    print("Usage: python3 xlsx_to_json.py <xlsx_file> <json_file>")
else:
    xlsx_file = sys.argv[1]
    json_file = sys.argv[2]
    convert_xlsx_to_json(xlsx_file, json_file)
```

Щоб надіслати отриманий файл на хост (рис. 3.15), який візуалізуватиме ці дані можна використати утиліту **scp** за допомогою команди

scp file.json barry@192.168.177.130:/home/barry/data.json

Рисунок 3.15 – Передача файлів через scp

Щоб налаштувати отримання даних через Logstash необхідно створити новий *pipeline*, який отримуватиме дані та розбиватиме їх по частинках (як це відображено у вихідній таблиці з даними).

Файл *pipeline* конфігурації повинен міститися в “**/etc/logstash/conf.d/**”. Створено файл *logstash.conf*, вміст якого відображено в лістингу 3.2.

Лістинг 3.2 – Конфігурація logstash.conf

```
input {
  file {
    path => "/home/barry/data.json"
    start_position => "beginning"
    sincedb_path => "/dev/null"
    codec => "json"
  }
}
filter {
  grok {
    match => {
      "message" => '%{DATA:channel_name}, %{NUMBER:post_date:long},
%{QUOTEDSTRING:text_data:json}, %{URI:link}, %{UNIXPATH:file_path},
%{QUOTEDSTRING:NER:json}, %{DATA:region}, %{DATA:region_code}'
    }
  }
}
output {
  elasticsearch {
    hosts => ["192.168.177.130:9200"]
    index => "news_data"
  }
  stdout {
    codec => rubydebug
  }
}
```

Дана конфігурація призначена для обробки файлу json. В якості **input** можна подати файл або вказати порт для прослуховування і надсилати дані на вказаний порт. У секції input вказано вхідний файл, з якого Logstash безперервно зчитує вказаний JSON-файл з початку (start_position => "початок"). Параметр sincedb_path має значення "/dev/null", що вказує на те, що Logstash не повинен відстежувати прочитані позиції в базі даних. Параметр codec має значення "json", що вказує на те, що Logstash повинен інтерпретувати кожен рядок файлу як JSON-документ.

Фільтри використовуються для розбору вхідних JSON-повідомлень використовується фільтр grok. Директива match визначає шаблон grok для вилучення структурованих полів з поля "message". Шаблон включає різні типи даних, такі як DATA, NUMBER, QUOTEDSTRING, URI, UNIXPATH і т.д.

Оброблені (відфільтровані) дані надсилаються в Elasticsearch, що запущений на "192.168.177.130:9200" з індексом "news_data". Другий вивід спрямовує оброблені дані до стандартного виводу (stdout) для налагодження за допомогою кодека rubydebug.

Таким чином, дана конфігурація Logstash дозволяє отримувати файл в форматі JSON, виконувати структурований розбір повідомлень за допомогою фільтрів *grok* і індексувати витягнуті поля в Elasticsearch, а також виводити результати на консоль для дебагу.

Щоб запустити logstash використовуючи створену конфігурацію необхідно використати команду:

```
/usr/share/logstash/bin/logstash -f
/etc/logstash/conf.d/logstash.conf --config.reload.automatic
```

Таким чином Logstash запустить компіляцію усіх конфігураційних файлів і розпочне фільтрацію вхідних даних, що задані в конфігураційному файлі, створення якого описано вище. При обробці кожного повідомлення відображається результат фільтрації (рис. 3.16).

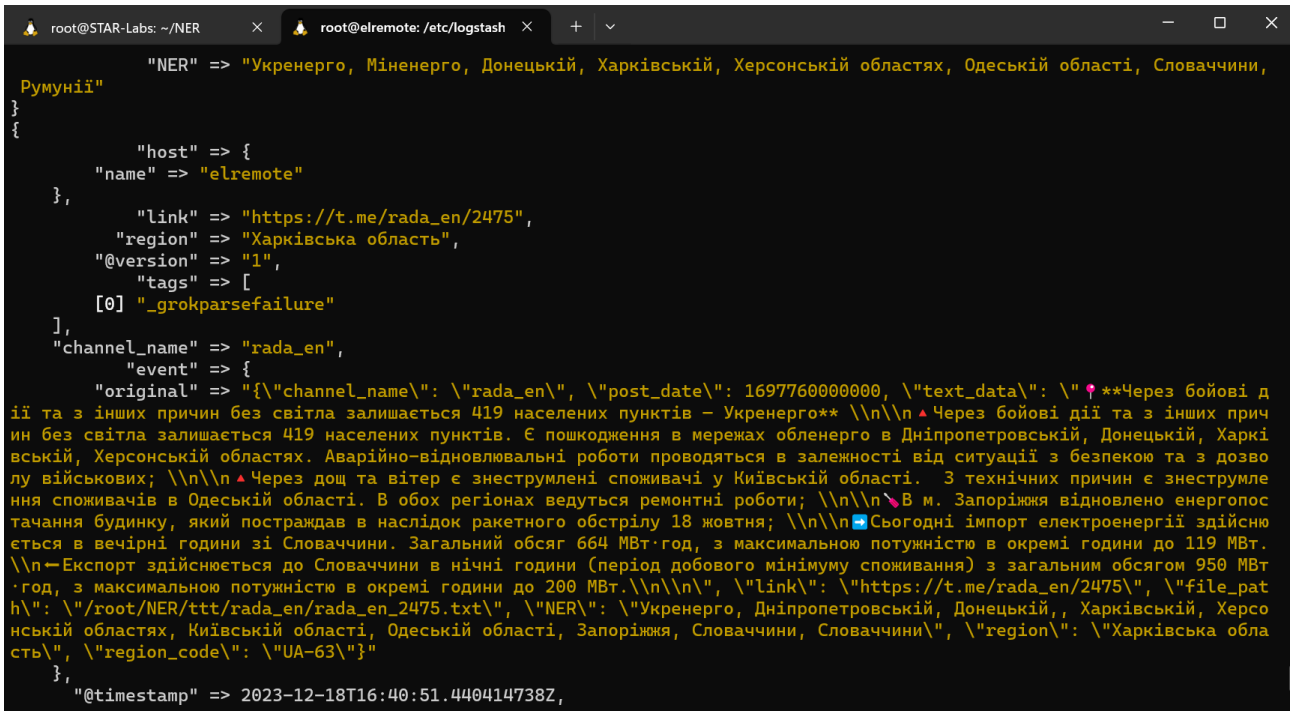

Рисунок 3.16 – Процес обробки даних logstash

Оброблені дані відображення збережені під індексом **news_data**. Щоб переглянути ці дані необхідно перейти на https://192.168.177.130:5601 розділ Discovery (рис. 3.17).

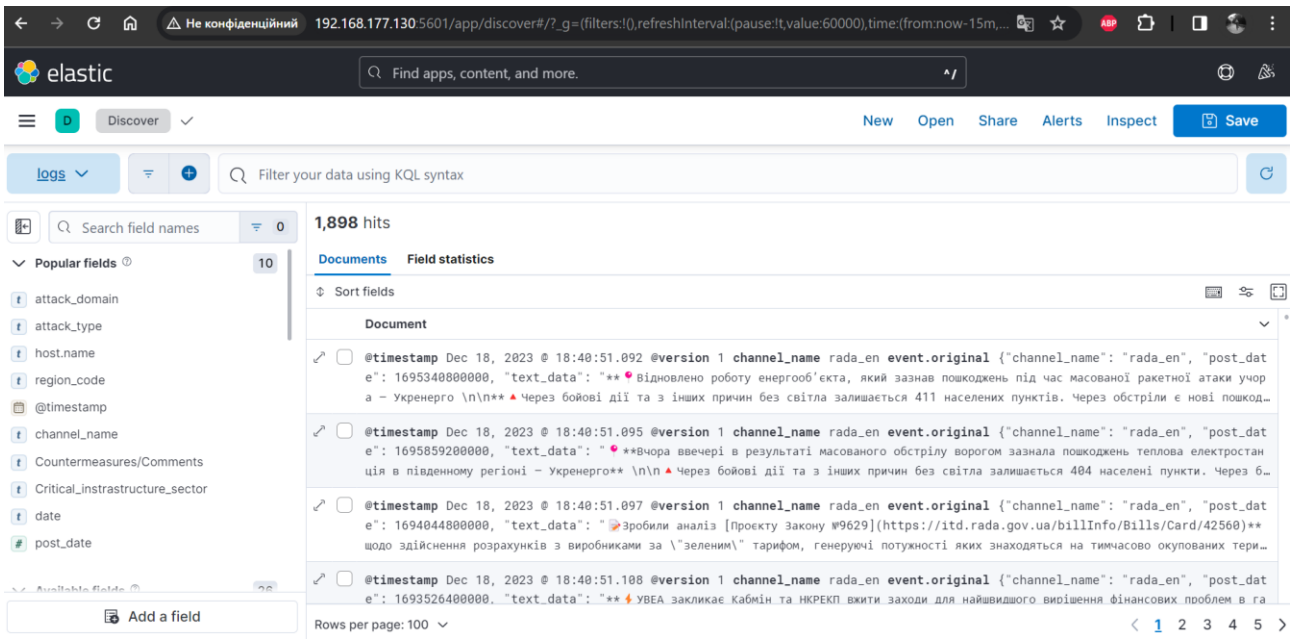

Рисунок 3.17 – Відображення повідомлень на Kibana

Для зручності роботи із даними та легкістю відображення необхідно розробити дашборд. Щоб створити дашборд (інформаційну панель) необхідно перейти на вкладку *Dashboard* та натиснути кнопку *Create dashboard*.

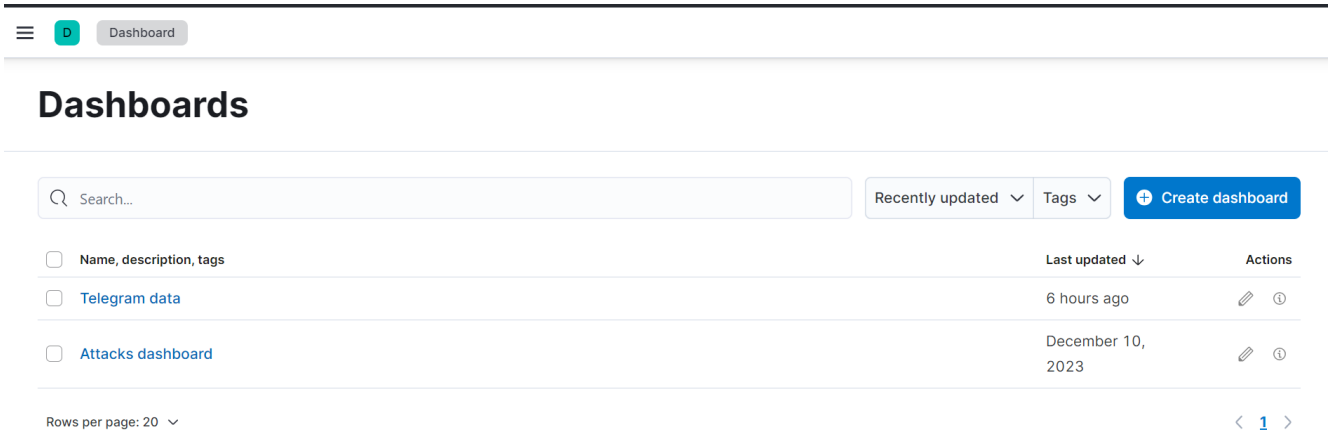

Рисунок 3.18 – Вкладка dashboard

Kibana надає широкий спектр графічних елементів для візуалізації будь-яких типів інформації. Щоб створити панель для відображення типів атак, потрібно натиснути *Create visualization* та обрати необхідний варіант відображення. Типи візуалізацій відображено на рисунку 3.19.

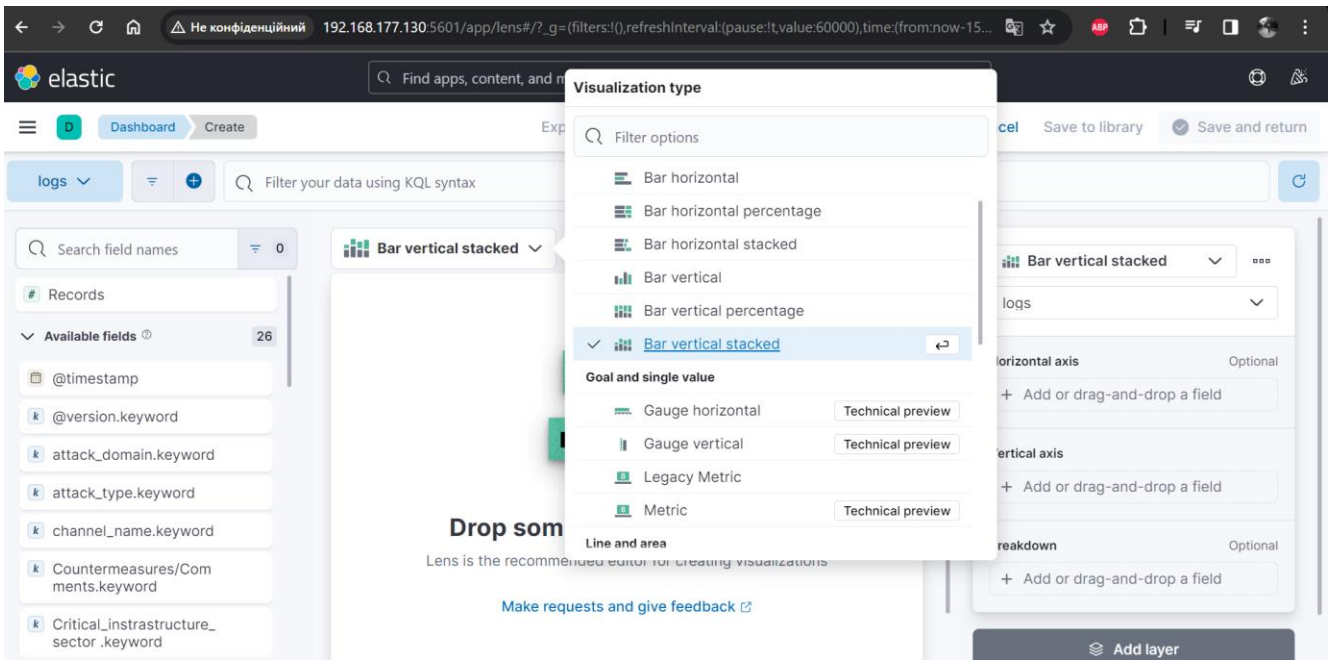

Рисунок 3.19 – Типи візуалізації

Для поставленої цілі необхідно реалізувати відображення у вигляді кругової діаграми, тому обрано тип – *Donut*. У панелі налаштувань (з правої сторони) необхідно вказати Data view, в даному випадку він один – logs. Усі інші налаштування потрібно здійснити так, як зображено на рис. 3.20.

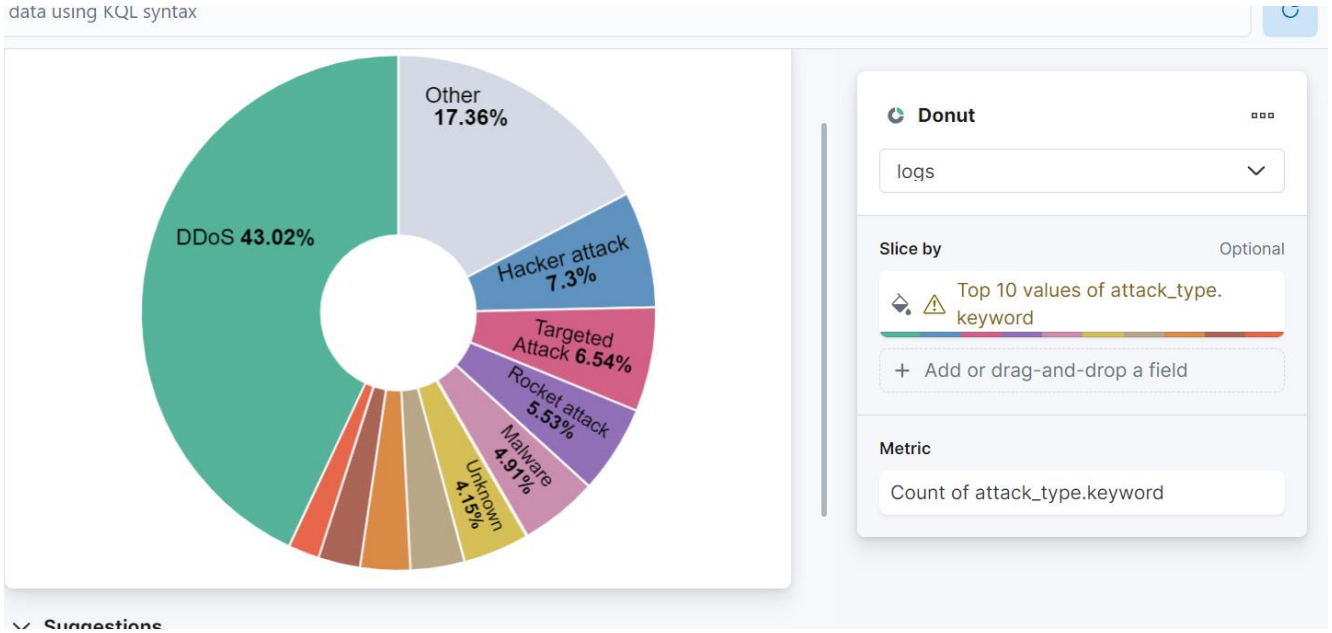

Рисунок 3.20 – Налаштування діаграми

Наступна панель Data sources передбачає відображення усіх каналів, звідки збирається інформація. Візуалізація така ж як і на рис. 3.20, проте тип діаграми – Pie, поле для сортування – channel_name.keyword. Оскільки тестова вибірка даних зібрана з одного каналу, тому на рис. 3.21 відображено лише один канал.

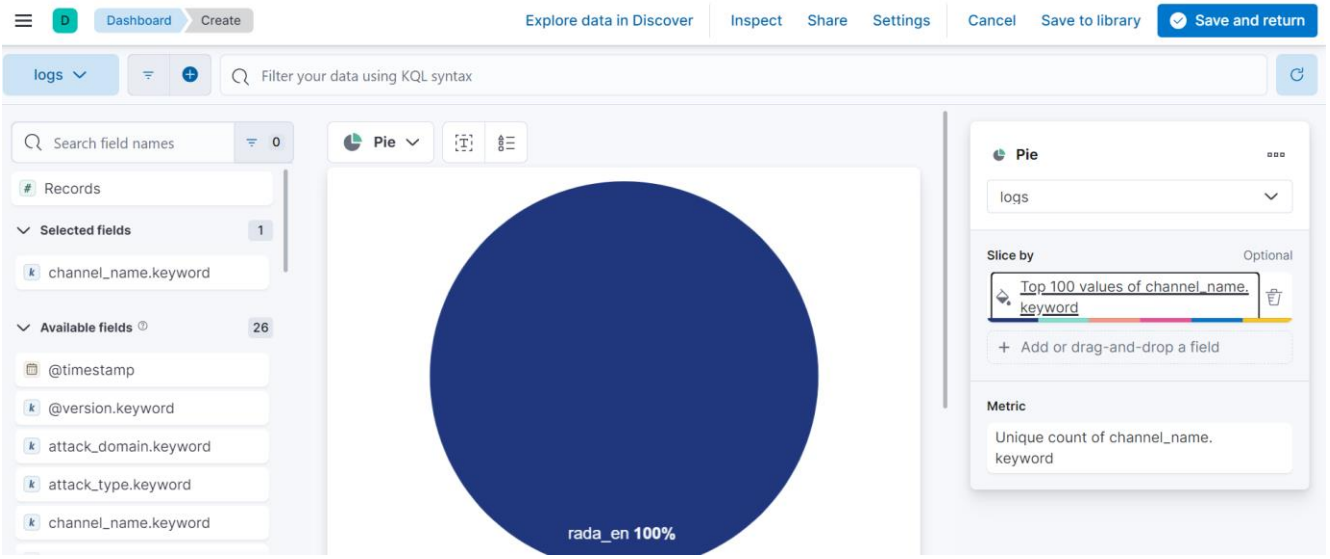

Рисунок 3.21 – Налаштування діаграми Data Sources

Найважливіше що потрібно додати – карта, на якій буде позначено відповідний регіон та кількість пов’язаних із ним подій. На панелі інструментів потрібно обрати *Region map* та вибрати поле, що містить код регіону, в даному випадку це поле region_code.keyword. В розділі *Metric* вказати *Count.* Таким чином, в результаті отримано картографічну візуалізацію на рис. 3.22, на якій відображається кількість подій із прив’язкою до регіону. Це інформація дає змогу зрозуміти динаміку змін в атаках та виявити найбільш “активні” регіони.

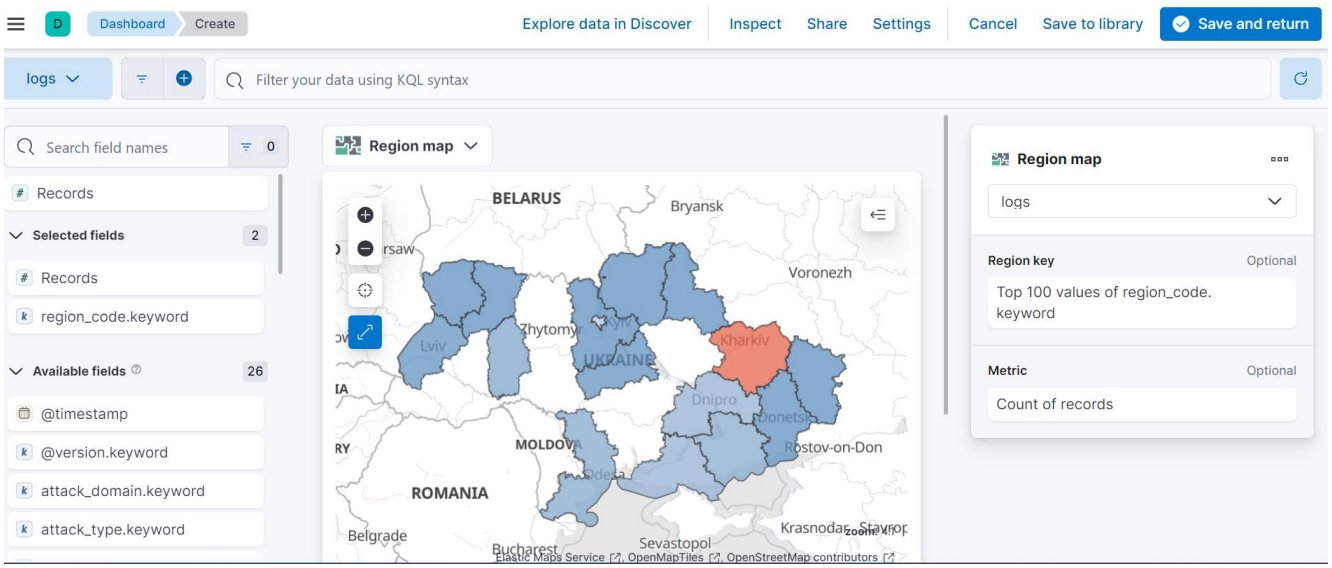

Рисунок 3.22 – Налаштування карти

Таким чином реалізовано та автоматизовано процес збору інформаційних повідомлень та вилучення іменованих сутностей. Завдяки такому процесу вдалося здійснити прив'язку інформаційного повідомлення до певного регіону. При достатньо великому обсязі вхідних даних, візуалізація відіграє значну роль для розуміння глобальної ситуації та виокремлення конкретних подій (за типом, часом, характером, місцем ураження).

3.4 Висновки

В даному розділі описані основні етапи збору та аналізу інформації про атаки на критичну інфраструктуру України із застосуванням моделей розпізнавання іменованих сутностей за допомогою бібліотеки spaCy та мови програмування python.

За допомогою стеку ELK реалізовано візуалізацію інофрмації, створено окремі дашборди відповідно до потреб та створено мапу, на якій відображено проаналізовані події.

4. ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

4.1 Охорона праці

Кваліфікаційна робота на тему «Методи ідентифікації та вилучення географічно пов'язаних об'єктів з даних про атаки на об'єкти критичної інфраструктури України» ставить за мету аналіз вхідних текстових та історичних даних про атаки на критичну інфраструктуру України, а також розробка методів ідентифікації географічно пов’язаних об’єктів атак на об’єкти критичної інфраструктури із подальшим їх застосуванням, що в свою чергу, потребує використання комп’ютерної техніки.

З метою забезпечення ефективної та безпечної праці аналітиків, що працюють над ідентифікацією та вилученням географічно пов’язаних об’єктів із зібраної текстової інформації про атаки на КІ України, важливо організувати безпечні умови праці. Керівник організації несе пряму відповідальність за дотримання нормативно-правових актів у сфері охорони праці, а саме тому при облаштуванні робочих місць працівників необхідно дотримуватись вимог, що зазначені у НПАОП 0.00-7.15-18 "Вимоги щодо безпеки та захисту здоров'я працівників під час роботи з екранними пристроями" [28]:

- при облаштуванні робочого місця працівника, який взаємодіє з екранними пристроями, важливо вибирати обладнання, що не породжує надмірного шуму та не випромінює зайвого тепла. Рівні шуму на місцях праці осіб, які використовують екранні пристрої, повинні відповідати встановленим санітарним нормам виробничого шуму, ультразвуку та інфразвуку, які були схвалені постановою Головного державного санітарного лікаря України від 01 грудня 1999 року № 37;
- робочі місця повинні облаштовуватися так, щоб надати користувачу достатньо робочого простору для зміни свого робочого положення та вільних рухів;

- прилади, які застосовуються повинні відповідати вимогам технічних регламентів (дотримання гранично-допустимих норм випромінювання тощо);
- організація робочого місця аналітика повинна враховувати вимоги ергономіки, антропології та психофізіології, а також специфіку виконуваних робіт;
- освітлення повинно відповідати вимогам, встановленим у Державних санітарних нормах щодо роботи з візуальними дисплейними терміналами електронно-обчислювальних машин (ДСанПІН 3.3.2.007-98).

Приміщення з робочими місцями аналітиків повинні відповідати вимогам протипожежного захисту згідно вимог НАПБ А.01.001.-2014 «Правил пожежної безпеки в Україні» [29], та обладнані системами протипожежної сигналізації, за необхідністю приміщення можуть бути обладнанні системами автоматичного пожежогасіння та системами автоматичної пожежної сигналізації. У допоміжних приміщеннях можна встановлювати теплові пожежні сповіщувачі.

У випадках, коли функціонує більше п'яти комп'ютерів в окремому приміщенні, рекомендовано розглядати можливість встановлення аварійного резервного вимикача для повного відключення електропостачання приміщення, при цьому залишаючи в роботі лише систему освітлення. Ця рекомендація ґрунтується на принципах безпеки та забезпечення ефективного управління електропостачанням у великих офісних приміщеннях, таким чином забезпечується швидке та безпечне вимкнення електроенергії в умовах надзвичайних ситуацій чи аварій, сприяючи виконанню евакуаційних заходів. Введення аварійного резервного вимикача з підтримкою освітлення дозволяє зберігати необхідне освітлення для забезпечення видимості та безпеки протягом процедур евакуації.

Для зберігання обладнання, що забезпечує роботу усіх сервісів та надає доступ до мережі необхідно використовувати серверну кімнату. З метою забезпечення ефективності та надійності функціонування мережевого, кросового

та комунікаційного обладнання в серверних приміщеннях рекомендується розміщення щонайменше однієї серверної стійки висотою 42U для монтажу 19-дюймового устаткування.

Габаритні параметри серверного приміщення визначаються кількістю стійок та іншого обладнання, забезпечуючи просторі проходи перед та за обладнанням не менше 80 сантиметрів. Площа серверного приміщення повинна становити хоча б 20 м².

Забороняється постійне перебування людей в серверних приміщеннях та їх використання як робочих кабінетів. Також заборонено використання серверних приміщень як складських приміщень.

Розташування серверного приміщення рекомендується проводити безпосередньо відділено від зовнішніх стін будівлі та проходів в інші приміщення, уникаючи впливу зовнішніх факторів, також розміщувати серверне обладнання в приміщеннях без вікон та уникати встановлення джерел електромагнітного випромінювання, що може негативно впливати на роботу серверного та телекомунікаційного обладнання.

Проектування конструкції стін, дверей та підлоги приміщення серверної повинно враховувати герметичність, рівність та відсутність елементів, які сприяють осіданню пилу. Матеріали повинні бути антистатичними та не сприяти накопиченню пилу чи електростатичного заряду.

Для таких приміщень обов’язково необхідно встановити:

- систему контролю доступу та постійного відеоспостереження;
- систему безперебійного електроживлення;
- систему протипожежної сигналізації;
- систему пожежогасіння;
- системи заземлення електроустановок відповідно до вимог норм та правил державних нормативів ПУЕ2017, ПТЕЕС, ПБЕЕС;
- систему клімат-контролю, що включає в себе вентиляцію, регулювання температури та вологості;
- структуровану кабельну систему.

Аналіз ситуацій надзвичайного характеру за останні 5–8 років вказує на те, що багато надзвичайних ситуацій виникає на рівні конкретних об'єктів, таких як невеликі підприємства, установи, організації тощо, особливо в галузях виробництва, логістики, торгівлі, освіти, науки, медицини та розважальної індустрії.

Важливість розробки та ефективності впровадження заходів з запобігання та ліквідації надзвичайних ситуацій на підприємствах визначається не лише безпекою працівників, але й можливим великим збитком для підприємства та його відвідувачів.

Відповідно до Кодексу цивільного захисту України [30], незалежно від форми власності, підготовка персоналу на підприємствах до дій у надзвичайних ситуаціях проводиться відповідно до спеціально розробленої схеми заходів захисту населення та територій.

Для якісного захисту від надзвичайних ситуацій на великих і малих підприємствах передбачається планування та виконання заходів щодо захисту працівників та об'єктів господарювання, розробка планів локалізації та ліквідації аварій, утримання сил і засобів для запобігання та ліквідації наслідків надзвичайних ситуацій, а також створення матеріальних резервів.

Важливо враховувати, що наведені заходи мають загальний характер і не повністю охоплюють специфіку роботи конкретного підприємства.

Малі підприємства, зокрема, мають особливості в захисті персоналу та відвідувачів, що обумовлено їхньою розмірністю та функціональністю. Згідно зі статтею 130 Кодексу цивільного захисту України [30], підприємства з чисельністю персоналу 50 осіб і менше повинні розробляти інструкції для дій при загрозі або виникненні надзвичайних ситуацій. Крім того, у сфері промислового виробництва до малих підприємств можуть бути віднесені і ті, де чисельність працівників перевищує 50 осіб, а інструкції для них розроблюються відповідно до рішення територіального органу Держслужби України з надзвичайних ситуацій.

Розроблені інструкції мають відповідати вимогам Кодексу цивільного захисту України [30] та враховувати специфіку конкретного підприємства. Також на малих підприємствах рекомендується розробка Плану евакуації при пожежі або загрозі вибуху, що є особливо важливим для об'єктів з великою кількістю відвідувачів.

4.2 Безпека життєдіяльності

Джерелами факторів, які мають вплив на виробничий процес можуть бути підвищений рівень напруги електричної мережі, замикання якої може відбутися через тіло людини [31]; недостатність природного світла (при порушенні умов праці і вимог до приміщень) [32]; недостатнє освітлення робочої зони [32]. А також нервово-психічна перевантаження (розумове, перенапруження аналізаторів-зорових) і фізичні (статичне – сидіння) [33,34].

Відповідно до НПАОП 0.00-7.15-18 [28], при обслуговуванні ПЕОМ мають місце фізичні і психофізичні небезпечні, а також шкідливі виробничі чинники:

- підвищене значення напруги в електричному ланцюзі, замикання якої може відбутися через тіло людини;
- підвищений рівень статичної електрики;
- підвищений рівень електромагнітних випромінювань;
- підвищена або знижена температура повітря робочої зони;
- підвищений або знижений рух повітря;
- підвищена або знижена вологість повітря;
- відсутність або недостатність природного світла;
- підвищена пульсація світлового потоку;
- недостатня освітленість робочого місця;
- підвищений рівень шуму на робочому місці;
- розумове перенапруження;
- емоційні навантаження;
- монотонність праці.

4.2.1 Пожежна безпека

Приміщення оснащено системою автоматичної пожежної сигналізації, має 1 вогнегасник ВП-5 із зарядом вогнегасної речовини 8-12 кг, відповідно до вимог чинного законодавства України. Проходи до засобів пожежогасіння вільні, не захаращуються та у разі потреби забезпечувати евакуацію всіх людей, які перебувають у приміщенні через один евакуаційний вихід з дверима на шляху евакуації, що відчинятися в напрямку виходу з будівлі від робочого місця. В приміщенні наявна затверджена «План-схема евакуації з кабінету (приміщення)».

Пожежна безпека при застосуванні ЕОМ забезпечується:

– системою запобігання пожежі,

– системою протипожежного захисту,

– організаційно-технічними заходами.

Згідно ДСТУ Б В.1.1-36:2016 «Визначення категорій приміщень, будинків та зовнішніх установок за вибухопожежною та пожежною небезпекою» [35] офісні або побутові приміщення, відноситься до категорії "В" (пожежонебезпечної) та для протипожежного захисту в ньому проектом передбачено устаткування автоматичною пожежною сигналізацією із застосуванням датчиків-сповіщувачів РІД-1 (сповіщувач димовий ізоляційний) в кількості 1 шт., і застосуванням первинних засобів пожежогасіння.

Горючими матеріалами в приміщенні, де розташовані ЕОМ, є:

– поліамід – матеріал корпусу мікросхем, горюча речовина, температура самозаймання 420º С,

– полівінілхлорид – ізоляційний матеріал, горюча речовина, температура запалювання 335º С, температура самозаймання 530º С,

– склотекстоліт ДЦ – матеріал друкарських плат, важкогорючий матеріал, показник горючості 1.7А, не схильний до температурного самозаймання,

– пластикат кабельний №.489 – матеріал ізоляції кабелів, горючий матеріал, показник горючості більше 2.1,

– деревина – будівельний і обробний матеріал, з якого виготовлені меблі, горючий матеріал, показник горючості більше 2.1, температура запалювання 255º С, температура самозаймання 399º С.

Згідно ДСТУ Б В.1.1-36-2016 [8] приміщення відносяться до категорії В (пожежовибухонебезпечним) і згідно правилам побудови електроустановок простір усередині приміщення відноситься до вогненебезпечної зони класу П - IIа (зони, розташовані в приміщеннях, в яких зберігаються тверді горючі речовини).

Потенційними джерелами запалення при роботі ПЕОМ є:

− іскри при замиканні і розмиканні ланцюгів;

− іскри і дуги коротких замикань;

− перегріви від тривалого перевантаження і наявності перехідного опору.

Продуктами згорання, що виділяються на пожежі, є: окис вуглецю; сірчистий газ; окис азоту; синильна кислота; акромін; фосген; хлор і ін. При горінні пластмас, окрім звичних продуктів згорання, виділяються різні продукти термічного розкладання: хлорангідридні кислоти, формальдегіди, хлористий водень, фосген, синильна кислота, аміак, фенол, ацетон, стирол.

Зменшити горюче навантаження не представляється можливим, тому проектом передбачається застосувати наступні способи і їх комбінації для запобігання утворенню(внесення) джерел запалення :

- застосування устаткування, що задовольняє вимогам електростатичної безпеки;
- застосування в конструкції швидкодіючих засобів захисного відключення можливих джерел запалення;
- виключення можливості появи іскрового заряду статичної електрики в горючому середовищі з енергією, рівної і вище мінімальної енергії запалення;
- підтримка температури нагріву поверхні машин, механізмів, устаткування, пристроїв, речовин і матеріалів, які можуть увійти до контакту з пальним середовищем, нижче гранично допустимої,

становить 80% як найменшої температури самозаймання пального.

- заміна небезпечних технологічних операцій більш безпечними;
- ізольоване розташування небезпечних технологічних установок і устаткування;
- зменшення кількості пальних і вибухонебезпечних речовин, що знаходяться у виробничих приміщеннях;
- запобігання можливості утворення пальних сумішей на лінії, вентиляційних системах і ін.;
- механізація, автоматизація та справність(потокова) виробництва;
- суворе дотримання стандартів і точне виконання встановленого технологічного режиму;
- запобігання можливості появи в небезпечних місцях джерел запалення;
- запобігання розповсюдженню пожеж і вибухів;
- використовування устаткування і пристроїв, при роботі яких не виникає джерел запалення;
- виконання вимог сумісного зберігання речовин і матеріалів;
- наявність громовідводу;
- ліквідація можливості самозаймання речовин і матеріалів .
- Для запобігання пожежі в обчислювальних центрах проектом пропонується виконання наступних вимог :
- електроживлення ЕОМ повинно мати автоматичне блокування відключення електроенергії на випадок зупинки системи охолоджування і кондиціонування;
- система вентиляції обчислювальних центрів повинна бути обладнана блокуючими пристроями, що забезпечують її відключення на випадок пожежі;
- робочі місця повинні бути оснащені пожежними щитами, сигналізацією, засобами для сповіщення про пожежну небезпеку (телефонами), медичними аптечками для надання першої медичної

допомоги, розробленим планом евакуації.

Для зниження пожежної небезпеки в приміщеннях використовуються первинні засоби гасіння пожеж, а також система автоматичної пожежної сигналізації, яка дозволяє знайти початкову стадію загоряння, швидко і точно оповістити службу пожежної охорони про час і місце виникнення пожежі.

Відповідно до правил пожежної безпеки для промислових підприємств приміщення категорії В підлягають устаткуванню системами автоматичної пожежної сигналізації. Проектом передбачається застосування датчика типу ІДФ – 1 (димовий фотоелектричний датчик), оскільки специфікою пожеж обчислювальної техніки і радіоапаратури є, в першу чергу, виділення диму, а потім - підвищення температури.

При виникненні пожежі в робочому приміщенні обслуговуючий персонал зобов'язаний негайно вжити заходи по ліквідації пожежі. Для ліквідації пожежі використовують вогнегасники (пінні для повітря ОП-5, ОП-6, ОП-9, вуглекислотні ОУ-5), пісок, пожежний інвентар (сокири, ломи, багри, шерстяну або азбестову ковдри). Як засіб індивідуального захисту проектом передбачається використовування промислового протигаза з маскою, фільтруючої коробки В.

В якості організаційно-технічних заходів рекомендується проводити навчання робочого персоналу правилам пожежної безпеки.

4.2.2 Освітлення

Основним небезпечним чинником при роботі з ЕОМ є небезпека поразки людини електричним струмом, яка посилюється тим, що органи чуття людини не можуть на відстані знайти наявності електричної напруги на устаткуванні.

Проходячи через тіло людини, електричний струм чинить на нього складну дію, що є сукупністю термічної (нагрів тканин і біологічних середовищ), електролітичної (розкладання крові і плазми) і біологічної (роздратування і збудження нервових волокон і інших органів тканин організму) дій.

Тяжкість поразки людини електричним струмом залежить від цілого ряду чинників:

- значення сили струму;
- електричного опору тіла людини і тривалості протікання через нього струму;
- роду і частоти струму;
- індивідуальних властивостей людини і навколишнього середовища.

Розроблений дипломний проект передбачає наступні технічні способи і засоби, що застерігають людину від ураження електричним струмом:

- заземлення електроустановок;
- занулення;
- захисне відключення;
- електричне розділення мережі;
- використовування малої напруги;
- ізоляція частин, що проводять струм;
- огорожа електроустановок.
- занулення зменшує напругу дотику і обмежує години, протягом яких людина, ткнувшись до корпусу, може потрапити під дію напруги.

Збільшення освітленості сприяє поліпшенню працездатності навіть в тих випадках, коли процес праці практично не залежить від зорового сприйняття. При поганому освітленні людина швидко втомлюється, працює менш продуктивно, виникає потенційна небезпека помилкових дій і нещасних випадків.

Освітленість приміщення має велике значення при роботі на ПЕОМ. Вона багато в чому визначається колірною і мережевий обстановкою. Для зменшеного поглинання світла стеля і стіни вище панелей (1,5-1,7м.). Якщо вони не облицьовані звукопоглинальним матеріалом, фарбуються білою водоемульсійною фарбою (коефіцієнт відбиття повинен бути не менше 0,7). Для забарвлення стіни панелей рекомендується віддавати перевагу світлим фарбам.

Природне освітлення, коли робочі місця з ПЕОМ розташовуються в один

ряд по довжині приміщення на відстані 0,8 - 1,0 м від стіни з віконними прорізами, і екрани знаходяться перпендикулярно цієї стіни. Основний потік природного світла при цій повинен бути зліва. Не допускається спрямування основного світлового потоку природного світла праворуч, ззаду і спереду працює на ПЕОМ. Оптимальна відстань очей до екрана відео монітора повинна становити 60-70 см, допустиме не менше 50 см. Розглядати інформацію ближче 50 см не рекомендується.

У приміщенні, де розташовані ЕОМ передбачається природне бічне освітлення, рівень якого відповідає ДБН В.2.5-28:2018 [32]. Джерелом природного освітлення є сонячне світло. Регулярно повинен проводитися контроль освітленості, який підтверджує, що рівень освітленості задовольняє [36] і для даного приміщення в світлий час доби достатньо природного освітлення.

4.2.3 Мікроклімат і вентилювання

Здійснюється провітрювання приміщення, в залежності від погодних умов, тривалість повинна бути не менше 10 хв. Найкращий обмін повітря здійснюється при наскрізному провітрюванні.

Мікроклімат виробничих приміщень визначається діючими на організм людини поєднаннями температури, вологості і швидкості руху повітря, а також температури навколишніх поверхонь. Значне коливання параметрів мікроклімату приводити до порушення систем кровообігу, нервової і потовидільної, що може викликати підвищення або пониження температури тіла, слабкість, запаморочення і навіть непритомність.

Відповідно до ДСН 3.3.6.042-99 [36] встановлюють оптимальну і допустиму температуру, відносну вологість і швидкість руху повітря в робочій зоні . За відсутності надмірного тепла, вологи, шкідливих речовин в приміщенні досить природної вентиляції.

4.3 Дії у надзвичайних ситуаціях

Основною загрозою для життя и здоров’я працівника є небезпека застосування ворожого ракетної атаки. Життєво-важливим є наявність працюючих засобів оповіщення про повітряну тривогу (сигнально-гучномовні пристрої – сирени, дзвони на церквах; офіційні джерела інформації – регіональне радіо, телебачення, застосунок «Повітряна тривога»). Відповідно до Кодексу цивільного захисту [37] та Положення про організацію оповіщення про загрозу виникнення або виникнення надзвичайних ситуацій та зв’язку у сфері цивільного захисту України, затвердженого постановою КМУ від 27.09.2017 р. №733 [38], при оголошенні повітряної тривоги усім працівникам слід негайно призупинення роботу робочого обладнання (комп’ютерного обладнання та допоміжних засобів), забезпечивши збереження інформації, у т.ч. на хмарних сховищах. Забезпечити збереження інших інтелектуальних чи матеріальних цінностей. Вимкнути електроживлення в приміщенні, нагрівальні прилади, закрити вікна, оповістити колег, які поруч та негайно евакуюватися у визначені укриття (бомбосховище). Впевнитися, що всі залишили приміщення. Рухатися вздовж несучих стін, подалі від скляних поверхонь чи вікон. Важливо зберігати спокій, і за можливості власної безпеки, здійснювати супровід іншого персоналу або відвідувачів, особливо з обмеженими можливостями, як найшвидше перейти в безпечну зону. Неухильно стежити за сигналами сповіщення і виконувати вказівки адміністрації підприємства/організації. Під час евакуації не користуватися сходами або ліфтами. При прибутті до сховища провести перекличку для з’ясування наявності всіх евакуйованих.

4.4 Висновки

В даному розділі проведено аналіз потенційних небезпечних та шкідливих виробничих факторів, причин пожеж та дій у надзвичайних ситуаціях. Розглянуті заходи, які дозволяють забезпечити гігієну праці і виробничу санітарію. Приведені рекомендації щодо організації робочого місця, а також

важливу інформацію щодо пожежної та електробезпеки. Була наведені значення яких впливає на умови праці робітника. На підставі аналізу розроблені заходи з техніки безпеки, організації вимог освітлення робочого місця та рекомендації з пожежної профілактики.

ВИСНОВОК

Аналіз методів ідентифікації та вилучення географічно пов’язаних об’єктів з текстових даних про атаки на об’єкти критичної інфраструктури України дає змогу описати процес ідентифікації та процес вилучення об’єктів, із врахуванням їх географічних зв’язків. В даній роботі проведено аналіз наявних методів та засобів, що дозволяють досягнути поставленої мети, розглянуто статті науковців, що досліджували дану тему. В основу даного дослідження покладено методи аналізу природньої мови, що фактично є основою для автоматизованого збору та аналізу інформації з будь яких впорядкованих так і не впорядкованих джерел.

Пропонується використання моделі розпізнавання іменованих сутностей (Named Entity Recognition, NER), що значно спрощує процес ідентифікації об’єктів серед інших та дозволяє пов’язувати об’єкти за допомогою географічних даних. Додатково, розроблено декілька програм, що використовують сутності для пошуку регіональних кодів, за допомогою яких здійснюється відображення усіх об’єктів на мапі.

В процесі виконання роботи автоматизовано збір близько 111 тисячі повідомлень про кібер та фізичні втручання до об’єктів критичної іфраструктури України за період з 22 лютого по 16 грудня з різних джерел та використання моделей NER за допомогою бібліотеки spaCy мовою програмування Python. Слід зазначити, що процес аналізу тексту та вилучення іменованих сутностей вимагає значних потужностей, наявність відеокарти завжди буде в пріоритеті. Для прикладу, щоб проаналізувати більше 100 тисяч повідомлень та вилучити із них сутності, потрібно близько 857 годин (35 днів, 7 годин) при наступних технічних характеристиках (процесор: Intel(R) Xeon(R) Gold, RAM: 32 Gb, Cores: 16).

В подальшому планується розроблення універсального рішення для збору даних із соціальних мереж, веб сторінок та месенджерів та побудови власної моделі, для вилучення більшої кількості видів сутностей ти покращення алгоритму опрацювання виявлених подій.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Supporting knowledge re-use with effective searches of related engineering documents—A comparison of search engine and natural language processing-based algorithms. [Електронний ресурс] / I. Ö.Arnarsson, O. Frost, E. Gustavsson, D. Stenholm // INTERNATIONAL CONFERENCE ON ENGINEERING DESIGN. – 2019. – Режим доступу до ресурсу: https://www.cambridge.org/core/services/aop-cambridge-core/content/view/23692B10CC75D406F048464EDFE2C665/S222043421900266Xa.pdf/supporting-knowledge-re-use-with-effective-searches-of-related-engineering-documents-a-comparison-of-search-engine-and-natural-language-processing-based-algorithms.pdf.
2. Reconstruction of the 1874 Santa Tecla's rainstorm inWestern Catalonia (NE Spain) from flood marks and historical accounts [Електронний ресурс] / J.Balasch, J. Ruiz-Bellet, J. Tuset, J. Martin de Oliva // Natural Hazards and Earth System Science. – 2010. – Режим доступу до ресурсу: https://www.researchgate.net/profile/Jordi-Tuset/publication/235720321_Reconstruction_of_the_1874_Santa_Tecla\'s_rainstorm_in_Western_Catalonia_NE_Spain_from_flood_marks_and_historical_accounts/links/0912f512dde78817ad000000/Reconstruction-of-the-1874-Santa-Teclas-rainstorm-in-Western-Catalonia-NE-Spain-from-flood-marks-and-historical-accounts.pdf?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19.
3. Transformer based named entity recognition for place name extraction from unstructured text [Електронний ресурс] / C.Berragan, A. Singleton, A. Calafiore, J. Morley // International Journal of Geographical Information Science. – 2022. – Режим доступу до ресурсу: https://doi.org/10.1080/13658816.2022.2133125.

4. Brown S. Machine learning, explained. [Електронний ресурс] / S. Brown. – 2021. – Режим доступу до ресурсу: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained.
5. Casero A. A. Named entity recognition and normalization in biomedical literature: A practical case in sars- cov-2 literature [Електронний ресурс] / A. A. Casero. – 2021. – Режим доступу до ресурсу: https://oa.upm.es/67933/1/TFM_ALVARO_ALONSO_CASERO.pdf.
6. Chen H. Geo-referencing place from everyday natural language descriptions [Електронний ресурс] / H. Chen, M. Vasardani, S. Winter. – 2017. – Режим доступу до ресурсу: https://doi.org/10.48550/arXiv.1710.03346.
7. Deep learning-based named entity recognition and knowledge graph construction for geological hazards [Електронний ресурс] / R.Fan, L. Wang, J. Yan, W. Song // ISPRS Int. J. Geo Inf. – 2019. – Режим доступу до ресурсу: https://www.mdpi.com/2220-9964/9/1/15.
8. Hu Y. H. A Supervised Machine Learning Approach to Toponym Disambiguation [Електронний ресурс] / Y. H. Hu, L. Ge // The Geospatial Web. – 2007. – Режим доступу до ресурсу: https://www.geospatialweb.com/chapter-11.html.
9. ISO 3166 — Codes for the representation of names of countries and their subdivisions [Електронний ресурс] – Режим доступу до ресурсу: https://www.iso.org/obp/ui/#iso:code:3166:UA.
10. Dumbacher B. SABLE: Tools for web crawling, web scraping, and text classification [Електронний ресурс] / B. Dumbacher, L. K. Diamond. – 2018. – Режим доступу до ресурсу: https://nces.ed.gov/FCSM/pdf/A_1Dumbacher_2018FCSM.pdf.
11. Spatial Planning Text Information Processing with Use of Machine Learning Methods [Електронний ресурс] / I.Kaczmarek, A. Iwaniak, A. Swietlicka, M. Piwowarczyk // ISPRS Annals of the Photogrammetry. – 2020. – Режим доступу до ресурсу: https://isprs-annals.copernicus.org/articles/VI-4-W2-2020/95/2020/isprs-annals-VI-4-W2-2020-95-2020.pdf.

12. Personalized content extraction and text classification using effective web scraping techniques [Електронний ресурс] / T.Karthikeyan, K. Sekaran, D. Ranjith, K. Vinoth // Int. J. Web Portals. – 2019. – Режим доступу до ресурсу: https://ericbrasiln.github.io/ferramentas_digitais_UNILAB/textos/10.4018.pdf.
13. Geographic Question Answering: Challenges, Uniqueness, Classification, and Future Directions [Електронний ресурс] / G.Mai, K. Janowicz, R. Zhu, L. Cai // AGILE. – 2021. – Режим доступу до ресурсу: https://agile-giss.copernicus.org/articles/2/8/2021/agile-giss-2-8-2021.pdf.
14. Comparing supervised learning algorithms for spatial nominal entity recognition [Електронний ресурс] / A.Medad, M. Gaio, L. Moncla, S. Mustiere // AGILE: GIScience Series. – 2020. – Режим доступу до ресурсу: https://doi.org/10.5194/agile-giss-1-15-2020.
15. Medlock B. W. Investigating Classification for Natural Language Processing Tasks [Електронний ресурс] / B. W. Medlock // University of Cambridge. – 2008. – Режим доступу до ресурсу: https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-721.pdf.
16. Location extraction from social media: Geoparsing, location disambiguation, and geotagging. [Електронний ресурс] / S. E.Middleton, G. Kordopatis-Zilos, S. Papadopoulos, Y. Kompatsiaris // TOIS. – 2018. – Режим доступу до ресурсу: https://doi.org/10.1145/3202662.
17. Chornyi P. Data extraction using NER [Електронний ресурс] / P. Chornyi, I. Skarga-Bandurova. – 2023. – Режим доступу до ресурсу: https://github.com/BarryAllen7/Data_extraction_using_NER.git.
18. Phum A. Toponym detection in the bio-medical domain: A hybrid approach with deep learning [Електронний ресурс] / A. Phum, T. Rcanasinghe, C. Orˇasan // International Conference on Recent Advances in Natural Language. – 2019. – Режим доступу до ресурсу: https://aclanthology.org/R19-1106.pdf.
19. Radford B. J. Regressing Location on Text for Probabilistic Geocoding [Електронний ресурс] / B. J. Radford. – 2021. – Режим доступу до ресурсу: https://arxiv.org/pdf/2107.00080.pdf.

20. Scheele C. Geographic context-aware text mining: Enhance social media message classification for situational awareness by integrating spatial and temporal features. [Електронний ресурс] / C. Scheele, M. Yu, Q. Huang // Int. J. Digit. Earth. – 2021. – Режим доступу до ресурсу: https://www.tandfonline.com/doi/full/10.1080/17538947.2021.1968048.
21. Named Entity Recognition Approaches and Their Comparison for Custom NER Model [Електронний ресурс] / H.Shelar, G. Kaur, N. Heda, P. Agrawal // Science & Technology Libraries. – 2020. – Режим доступу до ресурсу: https://www.researchgate.net/publication/341501760_Named_Entity_Recognition_Approaches_and_Their_Comparison_for_Custom_NER_Model.
22. Sit M. A. Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: A case study of Hurricane Irma [Електронний ресурс] / M. A. Sit, C. Koylu, I. Demir // Int. J. Digit. Earth. – 2019. – Режим доступу до ресурсу: https://www.researchgate.net/publication/347648534_Identifying_disaster-related_tweets_and_their_semantic_spatial_and_temporal_context_using_deep_learning_natural_language_processing_and_spatial_analysis_a_case_study_of_Hurricane_Irma.
23. Wang J. NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages [Електронний ресурс] / J. Wang, Y. Hu, K. Joseph // Trans. GIS. – 2020. – Режим доступу до ресурсу: https://www.researchgate.net/profile/Yingjie-Hu/publication/341378611_NeuroTPR_A_neuro-net_toponym_recognition_model_for_extracting_locations_from_social_media_messages/links/5f49c8db458515a88b82df38/NeuroTPR-A-neuro-net-toponym-recognition-model-for-extracting-locations-from-social-media-messages.pdf?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19.
24. What is named entity recognition? [Електронний ресурс] // IBM – Режим доступу до ресурсу: https://www.ibm.com/topics/named-entity-recognition.

25. Wolff R. Semantic Analysis, Explained [Електронний ресурс] / R. Wolff. – 2020. – Режим доступу до ресурсу: https://monkeylearn.com/blog/semantic-analysis/.
26. Deep learning for real-time social media text classification for situation awareness—Using Hurricanes Sandy, Harvey, and Irma as case studies [Електронний ресурс] / M.Yu, Q. Huang, H. Qin, C. Scheele // Int. J. Digit. Earth. – 2019. – Режим доступу до ресурсу: https://www.tandfonline.com/doi/full/10.1080/17538947.2019.1574316.
27. Prysiagniuk A. How machine learning works and its practical applications. [Електронний ресурс] / A. Prysiagniuk. – 2019. – Режим доступу до ресурсу: https://nachasi.com/tech/2019/01/31/yak-pratsyuye-machine-learning/.
28. НПАОП 0.00-7.15-18 «Вимоги щодо безпеки та захисту здоров'я працівників під час ро-боти з екранними пристроями». Наказ Міністерства соціальної політики України від 14.02.2018 №207 "Про затвердження вимог щодо безпеки та захисту здоров'я працівників під час роботи з екранними пристроями". Зареєстровано в Міністерстві юстиції України 25 квітня 2018 р. за № 508/31960. Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/z0508-18
29. НАПБ А.01.001.-2014 «Правил пожежної безпеки в Україні». Наказ Міністерства внутрішніх справ України від 30.12.2014 № 1417 Про затвердження Правил пожежної безпеки в Україні. Зареєстровано в Міністерстві юстиції України 05 березня 2015 р. за № 252/26697. Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/z0252-15
30. Кодексу цивільного захисту України. Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/5403-17
31. Електробезпека в будівлях і спорудах. Вимоги до захисних заходів від ураження електричним струмом. Наказ від 1 липня 2016 року N 204. Режим доступу до ресурсу: http://epicentre.co.ua/dstu/doc28522.html.

32. ДБН В.2.5-28:2018 «Природне і штучне освітлення». Режим доступу до ресурсу: http://www.minregion.gov.ua/wp-content/uploads/2018/12/V2528-1.pdf.
33. НПАОП 0.00-7.15-18 «Вимоги щодо безпеки та захисту здоров'я працівників під час роботи з екранними пристроями». Зареєстровано в Міністерстві юстиції України 25 квітня 2018 р. за № 508/31960. Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/z0508-18.
34. Державні санітарні правила і норми роботи з візуальними дисплейними терміналами електронно-обчислювальних машин ДСанПІН 3.3.2.007-98. Затверджено Постановою Головного державного санітарного лікаря України 10 грудня 1998 р. N 7. Режим доступу до ресурсу: https://zakon.rada.gov.ua/rada/show/v0007282-98.
35. ДСТУ Б В.1.1-36:2016 «Визначення категорій приміщень, будинків та зовнішніх установок за вибухопожежною та пожежною небезпекою». Наказ від 15.06.2016 №158. Режим доступу до ресурсу: https://zakon.rada.gov.ua/rada/show/v0158858-16.
36. Санітарні норми мікроклімату виробничих приміщень ДСН 3.3.6.042-99. Постанова N 42 від 01.12.99. Режим доступу до ресурсу: https://zakon.rada.gov.ua/rada/show/va042282-99.
37. Кодекс цивільного захисту України. Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/5403-17#Text.
38. Положення про організацію оповіщення про загрозу виникнення або виникнення надзвичайних ситуацій та зв'язку у сфері цивільного захисту, затвердженого постановою КМУ від 27.09.2017 р. №733. Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/733-2017-%D0%BF#Text.

ДОДАТКИ

Додаток А

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ ІМЕНІ ІВАНА ПУЛЮЯ

М А Т Е Р І А Л И

XI НАУКОВО-ТЕХНІЧНОЇ КОНФЕРЕНЦІЇ

«ІНФОРМАЦІЙНІ МОДЕЛІ, СИСТЕМИ ТА ТЕХНОЛОГІЇ»

13-14 грудня 2023 року

ТЕРНОПІЛЬ
2023

Продовження додатку А

УДК 004.56
Чорний П.Р., Скарга-Бандурова І. С., д.т.н., проф.
Тернопільський національний технічний університет ім. Івана Пулюя

МЕТОДИ ІДЕНТИФІКАЦІЇ ТА ВИЛУЧЕННЯ ГЕОГРАФІЧНО ПОВ'ЯЗАНИХ ОБ'ЄКТІВ З ДАНИХ ПРО АТАКИ НА ОБ'ЄКТИ КРИТИЧНОЇ ІНФРАСТРУКТУРИ УКРАЇНИ

P.R. Chornyi, I.S. Skarga-Bandurova, DSc, Prof

METHODS FOR IDENTIFYING AND EXTRACTING GEOGRAPHICALLY RELATED OBJECTS FROM TEXTUAL DATA ON ATTACKS ON CRITICAL INFRASTRUCTURE FACILITIES IN UKRAINE

Кібератаки на об'єкти критичної інфраструктури в Україні стали однією з найактуальніших проблем національної безпеки. Для забезпечення ефективної протидії таким атакам необхідно точно та оперативно ідентифікувати та ліквідувати пов'язані з ними загрози. Виявлення географічно пов'язаних об'єктів з текстових та історичних даних спрощує аналіз та розуміння впливу кібератак у географічному контексті, що сприяє розробці ефективних стратегій реагування на інциденти та забезпечує захист кіберпростору.

Метою даного дослідження є розробка та застосування методів ідентифікації географічно пов'язаних об'єктів та їх взаємозв'язків. Це дозволить виявити та аналізувати закономірності між об'єктами.

Основними етапами роботи є:

1. Збір даних і попередня обробка. На цьому етапі здійснюється збір необхідних даних про об'єкти, їх географічні координати та атрибути, також попередня обробка;
2. Етап використання методів ідентифікації передбачає застосовування різних методів для ідентифікації географічно пов'язаних об'єктів;
3. Геокодування. Процес полягає у перетворенні об'єктів, пов'язаних із розташуванням, у географічні координати за допомогою бібліотек геокодування;
4. Геопросторовий аналіз допомагає виявити розміщення об'єктів у просторі та знаходити закономірності їх розподілу;
5. На етапі візуалізації та аналізу даних географічно пов’язані об’єкти представляються в графічній формі.

На основі текстової інформації з відкритих джерел в роботі планується картографування щоденних подій із використанням технологій для обробки природньої мови та геопросторового аналізу. Для візуалізації обробленої інформації пропонується використання сучасних інструментів для створення графіків та картографічних візуалізацій.

Один із найважливіших технологічних компонентів методології полягає у використанні методів обробки природньої мови для ідентифікації іменованих сутностей, що є ключовим елементом аналізу текстових даних. За допомогою таких методів, іменовані сутності, зокрема імена, локації, дати, організації та інші, можуть бути виділені та класифіковані у текстових даних для подальшого використання в аналізі та визначенні ключових елементів тексту. Для роботи з іменованими сутностями використовуються завдання з бібліотек spaCy, NLTK та Stanford NER, в яких реалізовані відповідні алгоритми та класифікатори для виявлення іменованих сутностей у текстових даних.

В результаті дослідження очікується набуття всебічного розуміння складної природи атак на критичну інфраструктуру України (ККІ). Засновуючись на аналізі історичних даних та закономірностей фізичних та кібератак, проводиться ідентифікація критичних аспектів у захисті об'єктів ККІ і розробка ефективних стратегій захисту.

Продовження додатку А