DOI 10.31891/2307-5732-2024-343-6-23
УДК 004.8

МАЗУРЕЦЬ ОЛЕКСАНДР
Хмельницький національний університет
https://orcid.org/0000-0002-8900-0650
e-mail: exe.chong@gmail.com
ВІТ РОМАН
Хмельницький національний університет
https://orcid.org/0009-0009-6958-4730
e-mail: vit.roman.vit@gmail.com

МЕТОД ВИЯВЛЕННЯ ЦІЛЬОВИХ ОБ'ЄКТІВ ПРЕДМЕТНОЇ ОБЛАСТІ У ТЕКСТОВОМУ КОНТЕНТІ

Розроблено метод виявлення цільових об'єктів предметної області, який використовує алгоритми машинного навчання для адаптивного розпізнавання об'єктів, враховуючи специфіку предметної області, що дозволяє значно скоротити час обробки даних і знизити ризик втрати важливої інформації. Метод виявлення цільових об'єктів предметної області дозволяє перетворювати вхідні дані у вигляді досліджуваного тексту і попередньо обробленого та збалансованого корпусу текстів досліджуваної предметної області в вихідні дані у вигляді сформованої множини цільових об'єктів з досліджуваного тексту, яка є об'єднаною множиною ключових слів знайденими різними методами без повторів та множиною NER що згруповані шляхом лематизації. Запропонований метод виявлення цільових об'єктів предметної області відрізняється від існуючих урахуванням ключових слів та іменникових сутностей предметної області, що дало змогу підвищити точність виявлення цільових об'єктів предметної області внаслідок врахування іменникових сутностей.

Для дослідження ефективності розробленого методу виявлення цільових об'єктів предметної області було сформовано навчальний датасет обсягом 400 текстів українською мовою. Також для валідації запропонованого методу було розроблено програмний застосунок для перетворення текстового контенту файлів із тестової вибірки у множину цільових об'єктів предметної області; створено окреме консольне програмне забезпечення для використання отриманого списку цільових об'єктів для досліджуваних текстів та словників з предметних областей, обраних відповідно до датасету. Виконане дослідження ефективності розробленого методу виявлення цільових об'єктів предметної області виявило, що знайдені за методом цільові об'єкти предметних областей спроможні виконувати подальшу задачу класифікації, демонструючи на метриці Евклідових відстаней групування текстів однієї категорії та збільшення ортогональної їй відстані.

Ключові слова: машинне навчання, NLP, NER, іменовані сутності, цільові об'єкти.

MAZURETS OLEKSANDR, VIT ROMAN
Khmelnytskyi National University

METHOD FOR DETECTING SUBJECT AREA TARGET OBJECTS IN TEXT CONTENT

Method for detecting subject area target objects is proposed, which uses machine learning algorithms for adaptive object recognition, taking into account the specifics of the subject area, which allows to significantly reduce the time of data processing and reduce the risk of losing important information. The method for detecting subject area target objects allows to transform input data in the form of the researched text and a pre-processed and balanced corpus of texts of the researched subject area into output data in the form of a formed set of target objects from the researched text, which is a combined set of keywords found by various methods without repetitions and the set of NER grouped by lemmatization. The proposed method for detecting subject area target objects differs from the existing ones by taking into account keywords and noun entities of the subject area, which made it possible to increase accuracy of detection of detecting subject area target objects as result of taking into account noun entities.

To research the effectiveness of developed method for detecting subject area target objects, a training dataset of 400 texts in the Ukrainian language was formed. Also, for the validation of proposed method, the software was developed to convert the text content of files from the test sample into a set of target objects of the subject area; separate console software was created to use the obtained list of target objects for the studied texts and dictionaries from the subject areas selected according to the dataset. The performed study of the effectiveness of the developed method for detecting subject area target objects revealed that the target objects of the subject areas found by the method are able to perform the further task of classification, demonstrating on the metric of Euclidean distances the grouping of texts of the same category and the increase of the distance orthogonal to it. This determines the beneficial effect and scope of the developed method.

Keywords: machine learning, NLP, NER, named entities, target objects.

Аналіз предметної області

Методи виявлення цільових об'єктів у предметній області є критично важливими для ефективного аналізу та обробки великих обсягів інформації. В умовах зростаючої складності даних, які охоплюють різноманітні предметні області, необхідність розробки та вдосконалення методів автоматизованого виявлення цільових об'єктів стає все більш актуальною [1]. Це особливо важливо в таких сферах, як штучний інтелект, а саме системи обробки природної мови та інформаційний пошук. Відсутність надійних та ефективних методів виявлення цільових об'єктів може призвести до втрати важливої інформації, зниження точності прийняття рішень та збільшення витрат на аналіз даних. Враховуючи швидкий розвиток технологій та постійне зростання обсягів інформації, дослідження методів виявлення цільових об'єктів набуває особливої ваги.

Виявлення цільових об'єктів у заданій предметній області передбачає застосування спеціальних алгоритмів та методів, спрямованих на ідентифікацію та класифікацію елементів, які мають ключове значення

для аналізу конкретної задачі. У роботі цільові об’єкти будуть шукатись у текстових даних, а під терміном «цільові об’єкти» буде матись на увазі сукупність множини ключових слів та множини NER з групуванням шляхом лематизації.

Виявлення цільових об’єктів у системах NLP, зокрема розпізнавання іменованих сутностей, відіграє важливу роль у багатьох завданнях аналізу тексту та обробки інформації. Основна мета NER полягає в ідентифікації і класифікації значущих елементів тексту, таких як імена людей, назви організацій, географічні назви, дати та інші сутності, які мають специфічне значення для конкретного контексту. Це завдання є ключовим для ряду практичних задач, таких як інформаційний пошук, машинний переклад, обробка юридичних документів та аналіз даних у соціальних медіа.

Одним із перспективних напрямків для задачі виявлення цільових об’єктів є використання методів машинного навчання, які дозволяють автоматично адаптуватися до особливостей даних та поліпшувати точність виявлення об'єктів з часом [2].

З проведеного аналізу, запропоновано автоматизувати виявлення цільових об’єктів предметної області з використанням підходів машинного навчання. Автоматизація виявлення цільових об’єктів предметної області сприятиме значному підвищенню ефективності та точності ідентифікації релевантних об’єктів у великих обсягах даних.

Останні публікації

Проблему виявлення цільових об’єктів предметної області варто розглядати у контексті пошуку іменованих сутностей та пошуку ключових слів [2]. Даними задачами широко займаються науковці як по всьому світу, так і в Україні [3, 4].

Модель отримання ключових слів загального призначення, що призначена для роботи з групами документів різних розмірів, доменів і читабельності, а також наявності міток ключових слів запропоновано у [3]. Для отримання кращого вибору ключових слів використано модель логістичної регресії з найменшим стисненням і регуляризацією оператора вибору. Заснована на класифікації структура такого підходу забезпечує вивчення слів, які чітко характеризують цю групу документів у порівнянні з групами порівняння, що підвищує репрезентативність вилучених ключових слів.

Щодо задачі NER, у [4] запропоновано підхід до оптимізації завдання розпізнавання іменованих сутностей шляхом використання попередньо навчених мовних моделей для автоматичного дослідження слів, пов’язаних з віртуальними мітками, що представляють категорії сутностей. Метод передбачає розробку міток через встановлення зв’язків між початковими словами міток і відповідними словами сутності на основі розподілу даних, отриманих з попередньо навченої мовної моделі. Завдяки цьому покращується семантичне представлення слів міток, що в результаті підвищує точність моделі в ідентифікації конкретних сутностей. Крім того, завдання NER переналаштовується у формат text2text, що дозволяє краще використовувати знання мовної моделі та оптимізує процес вилучення інформації.

Метою роботи є створення методу виявлення цільових об’єктів предметної області, який відрізняється від існуючих урахуванням ключових слів та іменникових сутностей предметної області, що дало змогу підвищити точність виявлення цільових об’єктів предметної області внаслідок врахування іменникових сутностей.

Основна частина

Метод виявлення цільових об’єктів предметної області у текстовому контенті призначений для автоматизації процесу ідентифікації ключових елементів у великих масивах текстових даних, спрямований на підвищення точності та ефективності аналізу текстової інформації. Схема та кроки методу наведені на рис. 1.

Цей метод використовує алгоритми машинного навчання для адаптивного розпізнавання об’єктів, враховуючи специфіку предметної області, що дозволяє значно скоротити час обробки даних і знизити ризик упущення важливої інформації. Метод виявлення цільових об’єктів предметної області дозволяє перетворювати вхідні дані у вигляді досліджуваного тексту і попередньо обробленого та збалансованого корпусу текстів досліджуваної предметної області в вихідні дані у вигляді сформованої множини цільових об’єктів з досліджуваного тексту, яка є об’єднаною множиною ключових слів знайденими різними методами без повторів та множиною NER що згруповані шляхом лематизації.

Вхідними даними методу є досліджуваний текст та попередньо оброблений збалансований корпус текстів досліджуваної предметної області.

Першим етапом є підготовка досліджуваного тексту для аналізу, який включає в себе токенізацію, лематизацію та видалення стоп-слів.

Наступним етапом є пошук ключових слів різними методами, такими як TF-IDF, TF, YAKE! та методом дисперсної оцінки. Кожним перерахованим методом відбувається формування множини ключових слів.

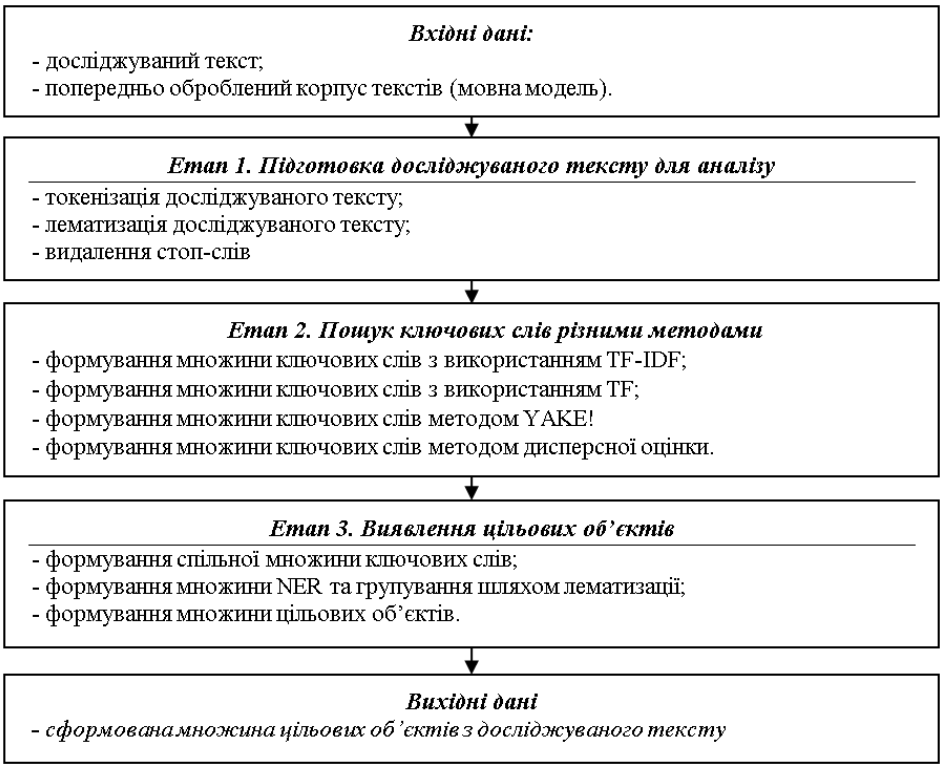

Рис. 1. Етапи роботи методу виявлення цільових об'єктів предметної області

На третьому етапі здійснюється виявлення цільових об'єктів, яке включає в себе декілька кроків. Схематично виявлення цільових об'єктів зображено на рис. 2.

Рис. 2. Етап формування цільових об'єктів

Цільові об'єкти є об'єднаною множиною ключових слів знайденими різними методами без повторів та множиною NER що згруповані шляхом лематизації.

Таким чином працює запропонований метод виявлення цільових об'єктів предметної області, призначений для автоматизації процесу ідентифікації ключових елементів у великих масивах текстових даних.

Дослідження ефективності методу виявлення цільових об'єктів предметної області

Для валідації запропонованого методу для пошуку цільових об'єктів предметної області було розроблено програмний застосунок мовою C# для перетворення текстового контенту файлів із тестової вибірки у множину цільових об'єктів предметної області. Головне вікно розробленого застосунку зображено на рис.3.

Оскільки українська мова повсякденного спілкування значно відрізняється від літературної через велику кількість діалектів, слів-запозичень та слів-покручів, наявні частотні словники не здатні охопити всю множину української мови. Для створення вектора значущих слів українською мовою було вирішено об'єднати кілька частотних словників [5], з відсіканням стоп-слів. Після об'єднання та фільтрації довжина вектора значущих слів склала 1500 елементів. Для цього було використано тексти з двох ортогональних множин, взятих з наступних ресурсів:

«Карпати буд каркас» – набір статей, що містять інформацію про будівництво, новини архітектури та технології. Колекція включає понад 200 текстів, середній обсяг кожного з яких становить близько 500 слів [6].

«Блог садівника» – сайт, що містить понад 200 текстів, кожен з яких має розмір близько 500 слів, тематика сайту – садівництво [7].

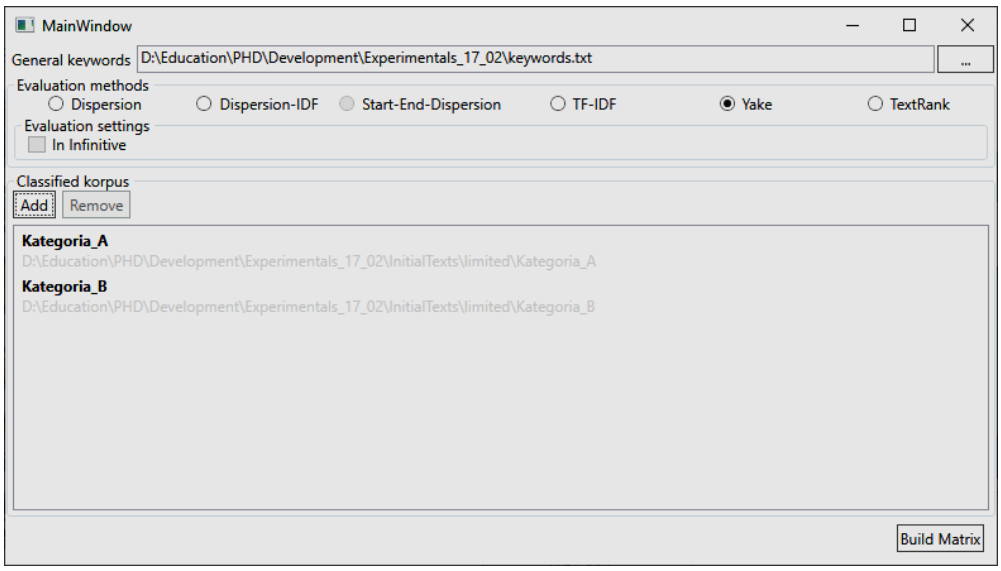

Рис. 3. Експериментальний застосунок для пошуку ключових слів досліджуваними методами

Такий вибір ресурсів обумовлений необхідністю забезпечити достатній обсяг текстів, що мають понад 200 слів, для навчання та перевірки запропонованого підходу.

Результати експерименту з дослідження ефективності запропонованого методу

Для дослідження ефективності запропонованого підходу було створено окреме консольне програмне забезпечення мовою Python, яке передбачає використання отриманого списку цільових об’єктів для досліджуваних текстів, та словників для окреслених тем «Карпати буд каркас» та «Блог садівника». Відповідно, знайдені цільові об’єкти були переведені у векторне представлення розміром 1500 (як розмір словника) методом One-Hot Encoding. Надалі було перевірено Евклідові відстань між текстами одного спрямування (5 текстів категорії «Карпати буд каркас» та 5 текстів «Блог садівника»), а також були обраховані Евклідові відстані між векторами протилежних категорій. Дані експерименту наведено в таблиці 1.

Таблиця 1

Евклідові відстані між текстами одного спрямування

	Текст 1	Текст 2	Текст 3	Текст 4	Текст 5	Текст 6	Текст 7	Текст 8	Текст 9	Текст 10
Текст 1	0	10.3	11.2	9.75	14.7	25.7	23.4	28.6	29.6	24.7
Текст 2	10.3	0	15.7	17.1	16.4	30.21	24.5	26.3	23.34	26.5
Текст 3	11.2	15.7	0	9.4	8.89	27.6	24.9	23.8	25.7	27.1
Текст 4	9.75	17.1	9.4	0	5.47	32.4	30.7	26.1	27.6	23.6
Текст 5	14.7	16.4	8.89	5.47	0	19.4	23.45	26.12	28.4	24.7
Текст 6	25.7	30.21	27.6	32.4	19.4	0	9.78	6.99	9.1	14.3
Текст 7	23.4	24.5	24.9	30.7	23.45	9.78	0	11.9	12.45	7.98
Текст 8	28.6	26.3	23.8	26.1	26.12	6.99	11.9	0	6.33	8.91
Текст 9	29.6	23.34	25.7	27.6	28.4	9.1	12.45	6.33	0	13.5
Текст 10	24.7	26.5	27.1	23.6	24.7	14.3	7.98	8.91	13.5	0

Результати отримані з таблиці 1 проілюстровані на графіку рис.4. Матриця відстаней таблиці 1 та рис.4 демонструють чітке розділення текстів на дві основні групи з різним змістом [8]. Перша група текстів (1–5), що належать категорії «Карпати буд каркас» має тісніші зв'язки між собою, аналогічно як друга група (6–10) також має менші внутрішні відстані (категорія «Блог садівника»), але водночас має великі відстані до текстів з першої групи, що свідчить про те, що ці групи належать до різних тематик. Тексти всередині кожної групи мають невеликі відстані, що свідчить про їхню тематичну схожість.

Таким чином, напрямками подальшої роботи з розглядуваної проблеми є розширення кількості категорій та експерименти із іншими метриками оцінки знайдених цільових об’єктів, у порівнянні їх із відомими великими мовними моделями.

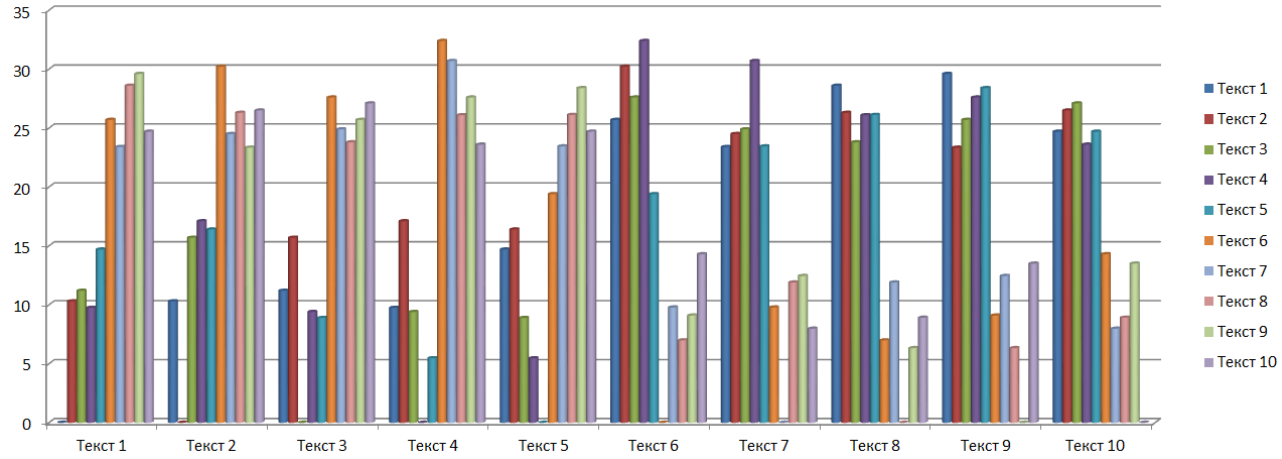

Рис. 4. Евклідові відстані між тестовими текстами двох категорій

Висновки

У статті розглянуто поточний стан наукового напряму виявлення цільових об'єктів предметної області у контексті пошуку іменованих сутностей та пошуку ключових слів, та на основі опрацьованого матеріалу запропоновано власний метод виявлення цільових об'єктів предметної області. Цей метод використовує алгоритми машинного навчання для адаптивного розпізнавання об'єктів, враховуючи специфіку предметної області, що дозволяє значно скоротити час обробки даних і знизити ризик втрати важливої інформації. Розроблений метод виявлення цільових об'єктів предметної області призначений для автоматизації процесу ідентифікації ключових елементів у великих масивах текстових даних, й спрямований на підвищення точності та ефективності аналізу текстової інформації.

Метод виявлення цільових об'єктів предметної області дозволяє перетворювати вхідні дані у вигляді досліджуваного тексту і попередньо обробленого та збалансованого корпусу текстів досліджуваної предметної області в вихідні дані у вигляді сформованої множини цільових об'єктів з досліджуваного тексту, яка є об'єднаною множиною ключових слів знайденими різними методами без повторів та множиною NER що згруповані шляхом лематизації. Запропонований метод виявлення цільових об'єктів відрізняється від існуючих урахуванням ключових слів та іменникових сутностей предметної області, що дало змогу підвищити точність виявлення цільових об'єктів предметної області внаслідок урахування іменникових сутностей.

Для дослідження ефективності розробленого методу виявлення цільових об'єктів предметної області було сформовано навчальний датасет обсягом 400 текстів побутовою українською мовою. Також для валідації запропонованого методу було розроблено програмний застосунок для перетворення текстового контенту файлів із тестової вибірки у множину цільових об'єктів предметної області; створено окреме консольне програмне забезпечення для використання отриманого списку цільових об'єктів для досліджуваних текстів та словників з предметних областей, обраних відповідно до датасету.

Виконане дослідження ефективності розробленого методу виявило, що знайдені за методом цільові об'єкти предметних областей спроможні виконувати подальшу задачу класифікації, демонструючи на метриці Евклідових відстаней групування текстів однієї категорії та збільшення відстані ортогональної їй. Це визначає корисний ефект та область застосування розробленого методу.

Подальші дослідження будуть спрямовані на розширення кількості категорій та експерименти із іншими метриками оцінки знайдених цільових об'єктів, у порівнянні їх із відомими великими мовними моделями, на кшталт GPT, Gemini тощо.

Література

1. Алгоритм виявлення аб'юзивного вмісту в україномовному аудіоконтенті для імплементації в об'єктно-орієтовану інформаційну систему / [М. О. Молчанова, О. В. Мазурець, О. В. Собко та ін.]. // Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. – 2024. – №1 (331). – С. 101–106.

2. Practical Approach for Detection by Deep Learning of Target Objects of Subject Area Based on Semantic Connectivity Indicators in Audio Database / O.Mazurets, O. Sobko, R. Vit, V. Pasternak. // Proceedings of XXIV International Scientific and Practical Conference «Modern Scientific Challenges are the Driving Force of the Development of Scientific Research». – 2024. – С. 91–96.

3. Shin H. General-use unsupervised keyword extraction model for keyword analysis / H. Shin, J. Lee, S. Cho. // Expert Systems with Applications. – 2023. – №233. – С. 120889.

4. Chen X. Named Entity Recognition via Unified Information Extraction Framework / X. Chen, Z. Zhang, X. Lu. // 4th International Conference on Computer Communication and Artificial Intelligence. – 2024. – С. 308–313.

5. An Approach Based on the Visualization Model for the Ukrainian Web Content Classification / [V.

Slobodzian, M. Molchanova, O. Kovalchuk та ін.]. // 2022 12th International Conference on Advanced Computer Information Technologies. – 2022. – С. 400–405.

6. Набір статей «Карпати буд каркас» [Електронний ресурс]. – 2024. – Режим доступу до ресурсу: https://karpatybud.com.ua/statti/.

7. Набір статей «Блог садівника» [Електронний ресурс]. – 2024. – Режим доступу до ресурсу: https://agro-market.net/ua/news/.

8. Text Data Vectorization Model of Ukrainian-Language Internet Communication Content / [V. Slobodzian, O. Kovalchuk, M. Molchanova та ін.]. // CEUR Workshop Proceedings. – 2022. – №3171. – С. 561–571.

References

1. Alhorytm vyiavlennia abiuzyvnoho vmistu v ukrainomovnomu audiokontenti dlia implementatsii v obiektno-oriietovanu informatsiinu systemu / [M. O. Molchanova, O. V. Mazurets, O. V. Sobko ta in.]. // Naukovyi zhurnal «Visnyk Khmelnytskoho natsionalnoho universytetu» seriia: Tekhnichni nauky. – 2024. – №1 (331). – S. 101–106.

2. Practical Approach for Detection by Deep Learning of Target Objects of Subject Area Based on Semantic Connectivity Indicators in Audio Database / O.Mazurets, O. Sobko, R. Vit, V. Pasternak. // Proceedings of XXIV International Scientific and Practical Conference «Modern Scientific Challenges are the Driving Force of the Development of Scientific Research». – 2024. – S. 91–96.

3. Shin H. General-use unsupervised keyword extraction model for keyword analysis / H. Shin, J. Lee, S. Cho. // Expert Systems with Applications. – 2023. – №233. – S. 120889.

4. Chen X. Named Entity Recognition via Unified Information Extraction Framework / X. Chen, Z. Zhang, X. Lu. // 4th International Conference on Computer Communication and Artificial Intelligence. – 2024. – S. 308–313.

5. An Approach Based on the Visualization Model for the Ukrainian Web Content Classification / [V. Slobodzian, M. Molchanova, O. Kovalchuk ta in.]. // 2022 12th International Conference on Advanced Computer Information Technologies. – 2022. – S. 400–405.

6. Nabir statei «Karpaty bud karkas» [Elektronnyi resurs]. – 2024. – Rezhym dostupu do resursu: https://karpatybud.com.ua/statti/.

7. Nabir statei «Bloh sadivnyka» [Elektronnyi resurs]. – 2024. – Rezhym dostupu do resursu: https://agro-market.net/ua/news/.

8. Text Data Vectorization Model of Ukrainian-Language Internet Communication Content / [V. Slobodzian, O. Kovalchuk, M. Molchanova ta in.]. // CEUR Workshop Proceedings. – 2022. – №3171. – S. 561–571.