

Вінницький національний технічний університет
Міністерство освіти і науки України

Кваліфікаційна наукова
праця на правах рукопису

БОНДАЛЕТОВ КОСТЯНТИН ОЛЕГОВИЧ

УДК 004.89+556.5

ДИСЕРТАЦІЯ
ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ВИДОБУВАННЯ ТА ВЕРИФІКАЦІЇ
ДАНИХ ПРО ВЗАЄМОПОВ'ЯЗАНІ ВОДНІ ОБ'ЄКТИ З ПРОСТОРОВОЮ
ТА ЧАСОВОЮ ПРИВ'ЯЗКОЮ

Спеціальність 126 – «Інформаційні системи та технології»

Галузь знань 12 – «Інформаційні технології»

Подається на здобуття ступеня доктора філософії

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

_____ К. О. Бондалетов

Науковий керівник: Мокін Віталій Борисович,
доктор технічних наук, професор

Вінниця – 2026

АНОТАЦІЯ

Бондалетов К.О. Інформаційна технологія видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою – кваліфікаційна робота на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 126 «Інформаційні системи та технології». – Вінницький національний технічний університет, Вінниця, 2026.

Метою дослідження є підвищення точності автоматизованої геоприв'язки текстів новин із соцмереж та ЗМІ про конкретні локальні об'єкти до описів, що характеризують масиви вод заданої річки, шляхом створення інтелектуальної інформаційної технології видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою.

Найважливіші наукові та практичні результати зводяться у таких положеннях:

1. Проаналізовано предметну область і встановлено доцільність зосередження уваги на стані масивів поверхневих вод України, невизначеність щодо більшості з яких ускладнює перевірку та оптимізацію ефективності програм заходів планів управління річковими басейнами, затверджених Кабінетом Міністрів України. Відзначено, що варто опрацювати різні види числової і текстової інформації про цей стан з гетерогенних джерел, зокрема з соцмереж, ЗМІ, звітів тощо. Досліджено існуючі методи та підходи до видобування та верифікації текстів з просторів інтернету. Визначено, що більшість із наявних сервісів не розглядають словосполучення, через що пошук потрібної інформації про водне джерело займає багато часу та не завжди відповідає дійсності. В ході вивчення даної теми розглянуто найбільш популярні засоби оброблення текстів для визначення недостатньо ефективних принципів аналізу, які вимагають удосконалення. Це і стало передумовою створення нової інформаційної технології видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою.

2. Розроблено метод автоматичного добування екологічних знань із прив'язуванням до часу та географічних координат, який видобуває речення з фактами, утворює SPO-триплети, знаходить геоназви з інформацією про період фіксації події згідно заголовків, встановлює відповідний зв'язок між фактами і масивами водних об'єктів через міру Жаккара; підтримує RAG-оновлення знань.

3. Запропоновано алгоритм геоприв'язки новин за зваженою мірою Жаккара, що відшуковує ключові слова для впорядкованого набору водних ресурсів зі ЗМІ, класифікує їх за локальністю (ваги 3/2/1) та обраховує зважену Жаккар-міру посеред публікацій та окремих масивів вод (MB), обираючи найкращі поміж них.

4. Завдяки критерію MLT-ISOF проведено просторово-зважену адаптацію TF-IDF для геоприв'язки текстів до масивів вод. Проведено вибір релевантних груп за рахунок використання інтегрального показника MLT·ISOF.

5. На основі інтелектуального прив'язування до БГСІС запропоновано метод аугментації бази знань про стан масивів вод, який дозволяє поширювати інформацію з-поміж пов'язаних масивів водних об'єктів після правильної геоприв'язки.

6. Описана система нотацій для порівняння підходів до побудови нових агрегованих варіантів через застосування агрегованих методів отримання ключових фраз.

7. Зведено UML-діаграму варіантів застосування технології із відображенням повноважень, котрі надаються користувачеві-експерту та адміністратору, щоб забезпечити правильне повноцінне функціонування системи. Сформовано структурну схему інформаційної технології. Розроблено пілотну версію вебсистеми «WISEST-SBB», яка реалізує основні елементи цієї технології.

Наукова новизна дисертаційної роботи відображена надалі:

1. Вперше розроблено інформаційну технологію видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою, яка поєднує автоматизоване видобування фактів у форматі SPO-триплетів «subject-predicate-object» з ієрархічним успадкуванням просторово-часового

контексту із заголовків документів та інших метаданих, формалізацію річкової мережі у вигляді багатозв'язної геоінформаційної системи іменованих сутностей (БГСІС) та верифікацію числових даних за критерієм топологічної спостережуваності бази знань, що дало змогу підвищити точність прив'язки фактів до конкретних масивів вод до рівня 85-99%.

2. Удосконалено метод формування користувацьких запитів до великої мовної моделі та RAG-архітектури шляхом прив'язки ембеддінгів до просторових об'єктів і часових меж, що, завдяки диференційованим базовим індексам достовірності за типами джерел (0,3 для авторів дописів, що пишуть у соцмережах, 0,7 для публікацій у ЗМІ і 0,95 для офіційних джерел) дозволяє суттєво підвищити достовірність окремої події з 0,3 до 0,99 і надає можливість більш точно знайти релевантну для користувача інформацію з урахуванням встановлених просторово-часових обмежень.

3. Набув подальшого розвитку метод геоприв'язки, який автоматично відбирає релевантні факти у текстах і більш точно зіставляє ці факти з описами масивів вод на основі зваженої міри Жаккара із застосуванням вагових коефіцієнтів та просторово-зваженого критерію MLT-ISOF («Multiplicity-Length Term with Inverse Spatial Object Frequency»), який враховує довжини лінійних проєкцій геоназв на усереднену течію річки. Цей метод дозволяє забезпечити аугментацію текстів про стан масивів вод на основі БГСІС, за рахунок рекурсивного поширення достовірної інформації між зв'язаними масивами вод.

Розроблену інформаційну технологію рекомендовано до впровадження у практичну діяльність Басейнового управління водних ресурсів (БУВР) річки Південний Буг під час супроводу процесу імплементації програми заходів Плану управління річковим басейном (ПУРБ) Південного Бугу для аналізу даних моніторингу стану та для виявлення проблемних ситуацій, які потребують вирішення під час подальшого оновлення цього ПУРБ (акт впровадження від БУВР р. Південний Буг від 27 квітня 2026 року). Результати дисертаційного дослідження також впроваджено у навчальний процес Вінницького національного технічного університету під час викладання дисципліни «Інтернет речей та

інтелектуальний аналіз даних» для аспірантів, які навчаються за освітньо-науковою програмою «Інформаційні системи та технології» рівня «доктор філософії» спеціальності 126 Інформаційні системи та технології та дисципліни «Машинне навчання та інтелектуальний аналіз даних» для студентів, які навчаються за освітньо-професійною програмою «Інформаційні технології аналізу даних та зображень» рівня «магістр» спеціальності F6 Інформаційні системи і технології у низці лекцій та лабораторних робіт (акт впровадження від ВНТУ від 27 квітня 2026 року).

Варто зауважити, що розроблена інформаційна технологія видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою за певних змін можна використати для інших предметних областей, де є аналіз даних.

Ключові слова: штучний інтелект, оброблення природномовного тексту, NLP, Text Ranking, BERT Model, багатозв'язні геоінформаційні системи, масив вод, великі мовні моделі, аугментація текстів, Python, інтелектуальна технологія, інформаційна технологія, контроль якості вод, аналітична вебсистема, геолокаційні технології

ABSTRACT

Bondalietov K.O. Information technology for extracting and verifying data on interconnected water bodies with spatial and temporal reference – qualification work in the form of a manuscript.

Dissertation for the degree of Doctor of Philosophy in specialty 126 "Information Systems and Technologies". – Vinnytsia National Technical University, Vinnytsia, 2026.

The aim of the research is to increase the accuracy of automated georeferencing of news texts from social networks and the media about specific local objects to descriptions characterizing the water bodies of a given river by creating an intelligent information technology for extracting and verifying data about interconnected water bodies with spatial and temporal reference.

The most important scientific and practical results are summarized in the following provisions:

1. The subject area was analyzed and the feasibility of focusing on the state of surface water bodies of Ukraine was established, the uncertainty regarding most of which complicates the verification and optimization of the effectiveness of the programs of measures of river basin management plans approved by the Cabinet of Ministers of Ukraine. It was noted that it is worth processing various types of numerical and textual information about this state from heterogeneous sources, in particular from social networks, media, reports, etc. Existing methods and approaches to extracting and verifying texts from the Internet were studied. It was determined that most of the available services do not consider word combinations, which is why the search for the necessary information about a water source takes a lot of time and does not always correspond to reality. In the course of studying this topic, the most popular text processing tools were considered to identify insufficiently effective principles of analysis that require improvement. This became the prerequisite for the creation of a new information technology for extracting and verifying data on interconnected water bodies with spatial and temporal reference.

2. A method for automatic extraction of environmental knowledge with binding to time and geographic coordinates has been developed, which extracts sentences with facts, forms SPO triplets, finds geonames with information about the period of event fixation according to headlines, establishes the appropriate connection between facts and water bodies through the Jaccard measure; supports RAG knowledge updating.

3. An algorithm for georeferencing news using the weighted Jaccard measure has been proposed, which finds keywords for an ordered set of water resources from the media, classifies them by locality (weights 3/2/1) and calculates the weighted Jaccard measure among publications and individual water bodies (MB), selecting the best among them.

4. Thanks to the MLT-ISOF criterion, a spatially-weighted adaptation of TF-IDF for georeferencing texts to water bodies has been carried out. Relevant groups have been selected using the MLT-ISOF integral indicator.

5. Based on the intellectual binding to the BGSIS, a method of augmenting the knowledge base about the state of water bodies is proposed, which allows distributing information from related water source bodies after correct georeferencing.

6. A notation system is described for comparing approaches to building new aggregated options through the use of aggregated methods for obtaining key phrases.

7. A UML diagram of the technology application options is compiled, reflecting the powers granted to the expert user and administrator to ensure the correct full-fledged functioning of the system. A structural diagram of information technology has been formed. A pilot version of the web system "WISEST-SBB" has been developed, which implements the main elements of this technology.

The scientific novelty of the dissertation work is reflected in the following:

1. For the first time, an information technology for extracting and verifying data on interconnected water bodies with spatial and temporal binding was developed, which combines automated fact extraction in the format of SPO triplets «subject-predicate-object» with hierarchical inheritance of the spatio-temporal context from document headers and other metadata, formalization of the river network in the form of a multi-connected named entity geoinformation system (BGSIS) and verification of numerical

data according to the criterion of topological observability of the knowledge base, which made it possible to increase the accuracy of linking facts to specific water bodies to the level of 85-99%.

2. The method of forming user queries to the large language model and RAG architecture has been improved by binding embeddings to spatial objects and time limits, which, thanks to differentiated basic reliability indices by type of source (0.3 for authors of posts writing on social networks, 0.7 for publications in the media and 0.95 for official sources), allows to significantly increase the reliability of a single event from 0.3 to 0.99 and provides the opportunity to more accurately find information relevant to the user, taking into account the established spatial and temporal restrictions.

3. The georeferencing method has been further developed, which automatically selects relevant facts in texts and more accurately compares these facts with descriptions of water bodies based on the weighted Jaccard measure using weighting coefficients and the spatially weighted MLT-ISOF criterion («Multiplicity-Length Term with Inverse Spatial Object Frequency»), which takes into account the lengths of linear projections of geonames onto the average river flow. This method allows for the augmentation of texts about the state of water bodies based on the BGSIS, due to the recursive distribution of reliable information between related water bodies.

The developed information technology is recommended for implementation in the practical activities of the Southern Bug River Basin Water Resources Management (SWBWRM) during support for the implementation of the program of measures of the Southern Bug River Basin Management Plan (RBMP) for analyzing condition monitoring data and identifying problem situations that require resolution during further updating of this RBP (implementation act from the Southern Bug River Basin Water Resources Management (SWBWRM) dated April 27, 2026). The results of the dissertation research were also implemented in the educational process of Vinnytsia National Technical University during the teaching of the discipline "Internet of Things and Intelligent Data Analysis" for postgraduate students studying in the educational and scientific program "Information Systems and Technologies" at the level of "Doctor of Philosophy" in specialty 126 Information Systems and Technologies and the discipline

"Machine Learning and Intelligent Data Analysis" for students studying in the educational and professional program "Information Technologies for Data and Image Analysis" at the level of "Master" in specialty F6 Information Systems and Technologies in a number of lectures and laboratory works (act of implementation from VNTU dated April 27, 2026).

It is worth noting that the developed information technology for extracting and verifying data on interconnected water bodies with spatial and temporal reference, with certain changes, can be used for other subject areas where data analysis is available.

Keywords: artificial intelligence, natural language text processing, NLP, Text Ranking, BERT Model, multi-connected geographic information systems, water body, large language models, text augmentation, Python, intelligent technology, information technology, water quality control, analytical web system, geolocation technologies

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Information measuring systems with mobile devices for identification of air pollution parameters caused by transport / Vitalii B. Mokin ; Georgii V. Goriachev ; Dmytro Y. Dziuniak ; Konstantin O. Bondaletov ; Serhii O. Zhukov ; Mariusz Duk ; Saltanat Sailarbek. Proc. SPIE 10031, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2016, 1003128 (September 28, 2016), 8 pages; doi:10.1117/12.2249202 (Scopus).

2. Vitalii B. Mokin. Method for Analyzing and Optimizing the Topologic Observability of Cognitive Maps of Complex Spatially Distributed Systems / Vitalii Mokin, Ilona Varchuk, Kostiantyn Bondaletov, Olena Slobodianiuk. CONFERENCE PROCEEDINGS 2017 IEEE First Ukraine Conference on ELECTRICAL AND COMPUTER ENGINEERING (UKRCON) May 29 – June 2, 2017 Kyiv. – P.: 865-869. – doi 10.1109/UKRCON.2017.8100371. (Scopus)

3. Bondalietov, K., Mokin, V. (2023). Notation System for Comparing and Synthesis of Intelligent Key Phrase Extraction Methods for Ontological Models in Information Systems. In: Dovgyi, S., Trofymchuk, O., Ustimenko, V., Globa, L. (eds) Information and Communication Technologies and Sustainable Development. ICT&SD 2022. Lecture Notes in Networks and Systems, vol 809. Springer, Cham. https://doi.org/10.1007/978-3-031-46880-3_11 ISSN 2367-3370 (Print); ISSN 2367-3389 (Online). (Scopus)

4. Мокін В. Б., Бондалетов К. О., Крижановський Є. М., і Караваєв В. О. Метод аугментації текстів про стан масивів вод на основі інтелектуальної прив'язки до багатозв'язних геоінформаційних систем іменованих сутностей, *Вісник Вінницького політехнічного інституту*, вип. 3, с. 55–65, Черв. 2023. URL: <https://doi.org/10.31649/1997-9266-2023-168-3-55-65> (Index Copernicus, Наукове фахове видання України, категорія «Б» зі спеціальності 126)

5. Мокін В. Б., Бондалетов К. О., Штельмах І. М., Слободянюк О. В. Автоматичне видобування знань з екологічних звітів з прив'язкою до часу та до просторових координат масивів вод, *Вісник Вінницького політехнічного*

інституту, вип. 3, с. 101–110, Черв. 2025, <https://doi.org/10.31649/1997-9266-2025-180-3-101-110> (Index Copernicus, Наукове фахове видання України, категорія «Б» зі спеціальності 126)

6. К. О. Бондалетов і В. Б. Мокін, «Інтелектуальна технологія видобування та верифікації числових та текстових даних у багатозв'язних багатостадійних геоінформаційних системах», *Вісник Вінницького політехнічного інституту*, вип. 5, с. 51–60, Жовт. 2025. Режим доступу: <https://visnyk.vntu.edu.ua/index.php/visnyk/article/view/3335> (Наукове фахове видання України, категорія «Б» зі спеціальності 126)

7. Мокін В. Б. Універсальна інформаційно-вимірювальна система оперативного екологічного моніторингу з використанням мобільних пристроїв / В. Б. Мокін, К. О. Бондалетов, Г. В. Горячев, Д. Ю. Дзюняк. *Вісник Вінницького політехнічного інституту*. — Вінниця. — 2015. — № 5 (122). — С. 116-122. <https://visnyk.vntu.edu.ua/index.php/visnyk/article/view/1836> (Index Copernicus, Наукове фахове видання України, категорія «Б» зі спеціальності 126)

8. Метод і технологія моніторингу стану атмосферного повітря за допомогою універсальної інформаційно-вимірювальної системи з використанням мобільних пристроїв / В. Б. Мокін, Д. Ю. Дзюняк, К. О. Бондалетов, В. В. Олійник. *Наукові праці Вінницького національного технічного університету* [Електронний ресурс]. — 2015. — № 4. — Режим доступу: <http://praci.vntu.edu.ua/index.php/praci/article/view/456/454> (Наукове фахове видання України, категорія «Б» зі спеціальності 126)

9. Бондалетов К. О. Практична реалізація сучасної мобільної системи веб-моніторингу та експрес-аналізу стану повітряного середовища / К. О. Бондалетов, В. Б. Мокін, Г. В. Горячев, Д. Ю. Дзюняк. XLIV Регіональна науково-технічна конференція професорсько-викладацького складу, співробітників та студентів університету з участю працівників науково-дослідних організацій та інженерно-технічних працівників підприємств м. Вінниці та області, електронне наукове видання матеріалів конференції, м. Вінниця, 11—13 березня 2015. — Режим доступу: <http://conf.vntu.edu.ua/allvntu/2015/ineek/txt/bondaletov.pdf>

10. Бондалетов К. О. Мобільна аналітична комп'ютерна система для оперативного моніторингу стану атмосферного повітря міста / К. О. Бондалетов, Д. Ю. Дзюняк, В. Б. Мокін. Молодь в технічних науках: дослідження, проблеми, перспективи : матеріали Міжнародної науково-практичної Інтернет-конференції, 23-26 квітня 2015 року / ВНТУ. – Вінниця, 2015. – С. 76-77.

11. Мокін В. Б. Інформаційно-вимірювальні системи для ідентифікації параметрів моделі забруднення атмосферного повітря автотранспортом з використанням мобільних пристроїв та фотометричних, у т.ч. лазерних, пристроїв і систем / В. Б. Мокін, Г. В. Горячев, Д. Ю. Дзюняк, К. О. Бондалетов. Зб. тез доп. III Міжнародної конференції “Вимірювання, контроль та діагностика в технічних системах (ВКДТС-2015)», м. Вінниця, 27-29 жовтня 2015 р. – Вінниця : ВНТУ, 2015. — С. 130-132.

12. Мокін В. Б. Технологія оперативного екологічного моніторингу стану повітряного середовища та пересувних джерел його забруднення / В. Б. Мокін, Д. Ю. Дзюняк, К. О. Бондалетов. XIV Міжнародна науково-практична конференція: Сучасні інформаційні технології управління екологічною безпекою, природокористуванням, заходами в надзвичайних ситуаціях: 5-9 жовтня 2015 р. – К., 2015. – С. 126-130.

13. Мокін В. Б. Практична реалізація мобільної аналітичної комп'ютерної системи моніторингу стану атмосферного повітря з підсистемою веб-аналізу та виведення даних на геопортал / В. Б. Мокін, Д. Ю. Дзюняк, К. О. Бондалетов, Г. В. Горячев, В. В. Боцян. V-й Всеукраїнський з'їзд екологів з міжнародною участю (Екологія/Ecology-2015), 23-26 вересня, 2015. Збірник наукових праць. – Вінниця: ТОВ «Нілан-ЛТД», 2015. – С. 106.

14. Бондалетов К. О., Мокін В.Б., Крижановський Є.М., Караваєв В.О. Розроблення інтелектуальної веб-системи з інформацією про екологічні проблеми та природоохоронні заходи у масивах вод басейну р. Південний Буг WISEST-SBB / Актуальні проблеми та перспективи розвитку фундаментальних, прикладних, загальнотехнічних та безпекових наук : Матеріали IV всеукр. науково-практ.

конф. пам'яті акад. Акад. наук вищ. освіти, проф. А.В. Касперського, м. Київ, 27 черв. 2024 р. С. 6–12. URL: <https://kztdop.ipf.npu.edu.ua/?view=article&id=244>

15. Мокін В.Б., Бондалетов К.О. Інтелектуальні методи видобування ключових словосполучень із тексту для побудови онтологічних моделей інформаційно-пошукових систем. С. 197-199. – Режим доступу: https://itgip.org/wp-content/uploads/2022/11/tezy_70.pdf. Інформаційно-комунікаційні технології та сталий розвиток: колективна монографія за матеріалами XXI Міжнародної науково-практичної конференції (14-16 листопада 2022 р.). Київ, 2022. ТОВ «Видавництво «стон», 2022, 242 с. URL: <https://itgip.org/kolektivna-monografiya-za-materialam>

16. Бондалетов К. О., Мокін В. Б. Інтелектуальна автоматизація геоприв'язки повідомлень з соцмереж до масивів вод за допомогою зваженої JACCARD-міри. ВІТКП ВНТУ. Факультет інтелектуальних інформаційних технологій та автоматизації (2025): вебресурс. (дата звернення 13 червня 2025: <https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2025/paper/view/23298/19275> .

17. Бондалетов К. О., Мокін В. Б., Григорчук М. В., Джура С. В. та ін. Побудова датасету для тренування інтелектуальних моделей веб-системи з інформацією про екологічні проблеми та заходи у масивах вод басейну р. Південний Буг WISEST-SBB. Матеріали ЛІІ Науково-технічної конференції факультету інтелектуальних інформаційних технологій та автоматизації Вінницького національного технічного університету, Вінниця, 21 – 23 червня 2023 р. – Електрон. текст. дані. – 2023. – Режим доступу: <https://d.conf.vntu.edu.ua/index.php/all-fksa/all-fksa-2023/paper/view/18958>

18. Мокін В. Б., Бондалетов К. О. Новий підхід до автоматизації оптимального управління водогосподарським балансом басейну річки на основі когнітивних карт. Х Міжнародна науково-практична конференція «Інтернет-Освіта-Наука (ІОН-2016)», Вінниця, Україна, 11–14 жовт. 2016. Вінниця : ВНТУ, 2016. С. 101–103. ISBN 978-966-641-646-2. URL: <https://ies.vntu.edu.ua/reports/program/WORK-IES-2016.pdf> (дата звернення: 08.07.2025).

19. Бондалєтов К. О., Мокін В. Б. Розроблення просторово-зваженого критерію геоприв'язки природномовних текстів до масивів вод. Молодь в науці: дослідження, проблеми, перспективи (МН-2025): матеріали Всеукраїнської науково-практичної інтернет-конференції (Вінниця, 15 жовтня 2024 року – 15 червня 2025 року). — Вінниця : ВНТУ, 2025. — С. 1597-1599.

20. Бондалєтов К. О.; Мокін В. Б. Автоматизація геолокації текстової інформації для контролю ефективності заходів з управління річковим басейном: Контроль і управління в складних системах (КУСС-2024), Вінниця, Україна, 16–17 жовтня 2024. с. 99–101.

21. Підвищення точності моделювання забруднення атмосферного повітря міста з використанням мобільних інформаційно-вимірювальних систем [Електронний ресурс] / В. Б. Мокін, Г. В. Горячев, Д. Ю. Дзюняк, К. О. Бондалєтов. Матеріали XLV Науково-технічної конференції Вінницького національного технічного університету, Вінниця, 23-24 березня 2016 р. - Електрон. текст. дані. - 2016. - Режим доступу : <http://conferences.vntu.edu.ua/index.php/all-ebmd/all-ebmd-2016/paper/view/465>.

22. Бондалєтов Костянтин, Мокін Віталій. "Архітектура каскадних парсерів гетерогених джерел даних для автоматизованої побудови датасетів для інтелектуальних систем моніторингу" Науково-технічної конференції факультету інтелектуальних інформаційних технологій та автоматизації Вінницького національного технічного університету, Вінниця, 24-27 березня 2026 р. - Електрон. текст. дані. - 2026. - Режим доступу <https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2026/paper/view/28287/23206>

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ	18
ВСТУП.....	19
РОЗДІЛ 1 АНАЛІЗ ПРОБЛЕМ ВИДОБУВАННЯ ТА ВЕРИФІКАЦІЇ ДАНИХ З РІЗНИХ ДЖЕРЕЛ.....	27
1.1 Аналіз предметної області	27
1.2 Огляд існуючих методів та підходів до видобування текстових даних.....	31
1.3 Проблема верифікації текстових даних.....	47
1.4 Постановка задач	48
Висновки до розділу 1	50
РОЗДІЛ 2 РОЗРОБЛЕННЯ МЕТОДІВ ВИДОБУВАННЯ ТА ВЕРИФІКАЦІЇ ДАНИХ ІЗ ПРОСТОРОВОЮ ТА ЧАСОВОЮ ПРИВ'ЯЗКОЮ	52
2.1 Формалізація вимог до даних та математична постановка задачі	52
2.2 Розроблення методу просторової прив'язки текстів.....	57
2.3 Удосконалення підходу до часової прив'язки текстів	66
2.4 Розроблення методу аугментації даних про водні об'єкти	68
2.5 Розроблення методу видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою	87
Висновки до розділу 2	92
РОЗДІЛ 3 РОЗРОБЛЕННЯ АЛГОРИТМІЧНИХ ОСНОВ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ВИДОБУВАННЯ ТА ВЕРИФІКАЦІЇ ДАНИХ ПРО ВЗАЄМОПОВ'ЯЗАНІ ВОДНІ ОБ'ЄКТИ З ПРОСТОРОВОЮ ТА ЧАСОВОЮ ПРИВ'ЯЗКОЮ	94
3.1 Побудова UML-діаграм варіантів використання технології.....	94

3.2 Розроблення структури інформаційної технології з використанням створеного методу прогнозування взаємозв'язків водних об'єктів з просторовою прив'язкою	96
3.3 Інфраструктура інформаційної технології	108
Висновки до розділу 3	110
РОЗДІЛ 4 ПРИКЛАДНЕ ЗАСТОСУВАННЯ РОЗРОБЛЕНОЇ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ВИДОБУВАННЯ ТА ВЕРИФІКАЦІЇ ДАНИХ ПРО ВЗАЄМОПОВ'ЯЗАНІ ВОДНІ ОБ'ЄКТИ З ПРОСТОРОВОЮ ТА ЧАСОВОЮ ПРИВ'ЯЗКОЮ	112
4.1 Застосування розробленої системи для аналізу даних про водні ресурси....	112
4.2 Застосування розробленої технології до аналізу стану водних ресурсів Вінницької області у басейні річки Південний Буг	115
4.3 Впровадження результатів дисертації у басейнове управління водних ресурсів.....	119
4.4 Впровадження розробленої технології у навчальний процес	121
Висновки до розділу 4.....	122
ВИСНОВКИ	124
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	126
ДОДАТКИ	142
ДОДАТОК А СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ.....	143
ДОДАТОК Б АКТ ВПРОВАДЖЕННЯ У БУВР ПІВДЕННОГО БУГУ	148
ДОДАТОК В АКТ ВПРОВАДЖЕННЯ У НАВЧАЛЬНИЙ ПРОЦЕС	149
ДОДАТОК Г ПРИКЛАДИ ФАКТІВ ПРО СТАН МАСИВІВ ВОД.....	150
ДОДАТОК Д АЛГОРИТМ РОБОТИ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ВИДОБУВАННЯ ТА ВЕРИФІКАЦІЇ ДАНИХ ПРО ВЗАЄМОПОВ'ЯЗАНІ ВОДНІ ОБ'ЄКТИ З ПРОСТОРОВОЮ ТА ЧАСОВОЮ ПРИВ'ЯЗКОЮ.....	151

ДОДАТОК Е UML-ДІАГРАМА ВАРІАНТІВ ВИКОРИСТАННЯ ТЕХНОЛОГІЇ ВИДОБУВАННЯ ТА ВЕРИФІКАЦІЇ ДАНИХ ПРО ВЗАЄМОПОВ'ЯЗАНІ ВОДНІ ОБ'ЄКТИ З ПРОСТОРОВОЮ ТА ЧАСОВОЮ ПРИВ'ЯЗКОЮ	152
ДОДАТОК Ж СТРУКТУРА ІНТЕЛЕКТУАЛЬНОЇ ТЕХНОЛОГІЇ З АВТОМАТИЗОВАНИМ ЗБИРАННЯМ ДАНИХ	153

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

- БГСІС – багатозв’язна геоінформаційна система іменованих сутностей
- БД – база даних
- БЗ – база знань
- БУВР – басейнове управління водних ресурсів
- ВГБ – водогосподарський баланс
- ГІС – геоінформаційна система
- ЗМІ – засоби масової інформації
- МВ – масив вод
- ПУРБ – План управління річковим басейном
- BERT – Bidirectional Encoder Representations from Transformers
- DPSIR – Driver-Pressure-State-Impact-Response Framework
- IDF – Inverse Document Frequency
- LLM (Large language model) – велика мовна модель
- MLT-ISOF – Multiplicity-Length Term with Inverse Spatial Object Frequency
- NER – Named Entity Recognition
- NLP (Natural language Processing) – оброблення природної мови
- RAG (Retrieval Augmented Generation) – генерація з доповненням через пошук
- SPO – subject–predicate–object
- TF – Term Frequency
- WISEST-SBB – Water Information SystEm with Spatial and Temporal references for the Southern Bug Basin

ВСТУП

Обґрунтування вибору теми дослідження. Для підтримки імплементації водної директиви 2000/60/ЄС і забезпечення достатньої якості та стану водних ресурсів необхідно покращення методів екологічного моніторингу. Традиційні рішення для моніторингу обмежені наявною кількістю постів спостереження. Наприклад, на 1091 масив вод р. Південний Буг розташовано біля 50 постів моніторингу, тобто не більше 5% масивів вод системно моніторяться. За даними інформаційного бюлетеня щодо планів управління річковими басейнами України на 2025-2030 рр., підготовленого за фінансової підтримки ЄС в рамках програми «EU4Environment», на території України виділено 5325 поверхневих масивів вод (МВ). Серед них 95% не мають необхідної кількості достовірної інформації для визначення екологічного стану вод. З інших 5% МВ, для яких інформація є достатньою, лише 57% відповідають вимогам ЄС, тобто досягають екологічного стану «не нижче доброго», а решта: 28% у задовільному стані, 12% у поганому та 2% у дуже поганому стані. Нестача інформації стосовно стану МВ є загальнодержавною проблемою. Описана ситуація обумовлює нагальну потребу в інтелектуальних інформаційних технологіях, що здатні підвищити рівень достовірності й повноти інформації про стан масивів вод за рахунок видобування фактів з альтернативних джерел, з обов'язковою прив'язкою до простору та часу.

Зв'язок роботи з науковими програмами, планами, темами, грантами. Тематика дисертаційного дослідження відповідає пріоритетному напрямку розвитку науки і техніки «Інформаційні та комунікаційні технології» та «Рациональне природокористування», визначеним Законом України «Про пріоритетні напрями розвитку науки і техніки» (№2623-III від 11.07.2001 р. зі змінами), а також пріоритетному тематичному напрямку «Інтелектуальні інформаційні та інформаційно-аналітичні технології. Інтегровані системи баз даних та знань. Національні інформаційні ресурси», затверджені Постановою Кабінету Міністрів України від №942 від 07.09.2011 р. Результати дисертаційного дослідження відповідають Водній стратегії України на період до 2050 року,

схвалений розпорядженням Кабінету Міністрів України від 09.12.2022 р. №1134-р; Концепції розвитку штучного інтелекту в Україні, схвалений розпорядженням Кабінету Міністрів України від 02.12.2020 р. №1556-р, в частині застосування технологій машинного навчання, оброблення природної мови та геоінформаційних систем для розв’язання задач у сфері екології та природоохоронної діяльності; і Цілям сталого розвитку ООН, зокрема Цілі 6 «Чиста вода та належні санітарні умови».

Дисертаційна робота виконувалась у межах науково-дослідної роботи (НДР) кафедри САІТ 28-К1 «Технології і системи моніторингу та автоматизації обробки даних у динамічних багатозв’язних системах природничого, технічного та навчального характеру» на замовлення МОН України протягом 2016 р., НДР кафедри САІТ 28-К7 «Інформаційні технології та системи моніторингу, моделювання та системного інтелектуального аналізу й оптимізації даних у складних об’єктах» протягом 2022-2023 рр. на замовлення МОН України та — у межах НДР кафедри САІТ 28-К8 «Системний аналіз та створення інформаційних систем й інтелектуальних технологій процесів у складних об’єктах» протягом 2024-2025 рр. на замовлення МОН України.

Частина результатів дисертаційної роботи про створення інтелектуальної веб-системи з інформацією про екологічні проблеми та природоохоронні заходи була розглянуто на басейновій раді Південного Бугу.

Мета і завдання дослідження. Метою дослідження є підвищення точності автоматизованої геоприв’язки текстів новин із соцмереж та ЗМІ про конкретні локальні об’єкти до описів, що характеризують масиви вод заданої річки, шляхом створення інтелектуальної інформаційної технології видобування та верифікації даних про взаємопов’язані водні об’єкти з просторовою та часовою прив’язкою.

У результаті проведеного аналізу сформульовано наступні задачі дослідження задля досягнення поставленої мети:

1. Провести огляд відомих методів та інформаційних технологій видобування та верифікації даних про взаємопов’язані водні об’єкти з просторовою та часовою прив’язкою.

2. Розробити метод верифікації числових даних для оброблення довільних корпусів текстів, що дасть можливість оцінити повноту та достовірність отриманої інформації про стан масивів вод.

3. Удосконалити метод формування користувацьких запитів до інтелектуальної системи для агрегації інформації про стан масивів вод з урахуванням просторово-часової прив'язки.

4. Удосконалити метод автоматизованої геоприв'язки довільних природномовних корпусів текстів про стан поверхневих вод до конкретних масивів вод басейну річки.

5. Розробити та випробувати на реальних прикладах інтелектуальну інформаційну технологію видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою, яка забезпечить підвищення точності і швидкості автоматизованої геоприв'язки текстів про конкретні водні об'єкти.

Об'єкт дослідження – процес видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою.

Предмет дослідження – інформаційна технологія видобування та верифікації знань з природомовних україномовних корпусів текстів, верифікація даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою.

Методи дослідження. Основними методами дослідження обрані: аналіз предметної області, огляд існуючих методів і алгоритмів видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою, побудова математичних моделей, застосування комп'ютерного моделювання та проєктування інтелектуальної технології видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою.

Наукова новизна отриманих результатів.

1. Вперше розроблено інформаційну технологію видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою, яка поєднує автоматизоване видобування фактів у форматі SPO-

триплетів «subject-predicate-object» з ієрархічним успадкуванням просторово-часового контексту із заголовків документа та інших метаданих, формалізацію річкової мережі у вигляді багатозв'язної геоінформаційної системи іменованих сутностей (БГСІС) та верифікацію числових даних за критерієм топологічної спостережуваності бази знань, що дало змогу підвищити точність прив'язки фактів до конкретних масивів вод до рівня 85-99%.

2. Удосконалено метод формування користувацьких запитів до великої мовної моделі та RAG-архітектури шляхом прив'язки ембеддінгів до просторових об'єктів і часових меж, що, завдяки диференційованим базовим індексам достовірності за типами джерел (0,3 для авторів дописів, що пишуть у соцмережах, 0,7 для публікацій у ЗМІ і 0,95 для офіційних джерел) дозволяє суттєво підвищити достовірність окремої події з 0,3 до 0,99 і надає можливість більш точно знайти релевантну для користувача інформацію з урахуванням встановлених просторово-часових обмежень.

3. Набув подальшого розвитку метод геоприв'язки, який автоматично відбирає релевантні факти у текстах і більш точно зіставляє ці факти з описами масивів вод на основі зваженої міри Жаккара із застосуванням вагових коефіцієнтів та просторово-зваженого критерію MLT-ISOF («Multiplicity-Length Term with Inverse Spatial Object Frequency»), який враховує довжини лінійних проєкцій геоназв на усереднену течію річки. Цей метод дозволяє забезпечити аугментацію текстів про стан масивів вод на основі БГСІС, за рахунок рекурсивного поширення достовірної інформації між зв'язаними масивами вод.

Практичне значення одержаних результатів. Розроблену інформаційну технологію рекомендовано до впровадження у практичну діяльність Басейнового управління водних ресурсів (БУВР) річки Південний Буг під час супроводу процесу імплементації програми заходів Плану управління річковим басейном (ПУРБ) Південного Бугу для аналізу даних моніторингу стану та для виявлення проблемних ситуацій, які потребують вирішення під час подальшого оновлення цього ПУРБ (акт впровадження від БУВР р. Південний Буг від 27 квітня 2026 року наведено у Додатку Б). Результати дисертаційного дослідження також

впроваджено у навчальний процес Вінницького національного технічного університету під час викладання дисципліни «Інтернет речей та інтелектуальний аналіз даних» для аспірантів, які навчаються за освітньо-науковою програмою «Інформаційні системи та технології» рівня «доктор філософії» спеціальності 126 Інформаційні системи та технології та дисципліни «Машинне навчання та інтелектуальний аналіз даних» для студентів, які навчаються за освітньо-професійною програмою «Інформаційні технології аналізу даних та зображень» рівня «магістр» спеціальності F6 Інформаційні системи і технології у низці лекцій та лабораторних робіт (акт впровадження від ВНТУ від 27 квітня 2026 року наведено у Додатку В).

Особистий внесок здобувача. Усі теоретичні результати, що виносяться на захист, отримані здобувачем особисто.

[1] – розроблено логіку для синхронізації датчиків із мобільним пристроєм та реалізовано алгоритм передачі даних на сервер в реальному часі з метою ідентифікації параметрів забруднення; [2] – розроблено алгоритм аналізу та оптимізації топологічної спостережуваності когнітивних карт складних просторово-розподілених систем; [3] – розроблено нову систему нотації та графічного представлення основних етапів методів видобування ключових словосполучень. Запропоновано концепцію нових методів на основі розробленої нотації; [4] – розроблено метод автоматичної аугментації текстів про стан МВ. Формалізовано річкову мережу у вигляді багатозв'язної геоінформаційної системи іменованих сутностей (БГСІС) та удосконалено алгоритм прив'язування даних до іменованих сутностей; [5] – запропоновано метод автоматичного видобування екологічних фактів з текстів у форматі SPO-триплетів для автоматизації оновлення бази знань. Формалізовано контекстно-семантичний критерій оптимальності, який одночасно максимізує повноту та ймовірність зв'язків і мінімізує кількість хибнопозитивних зв'язків; [6] – запропоновано архітектуру програмного інтерфейсу додатка для збирання даних екологічного моніторингу та розроблено модуль для інтеграції GPS із результатами аналізу повітря; [7] – запропоновано і реалізовано інформаційну систему з використанням

мобільних пристроїв для проведення моніторингу повітря та ймовірних джерел забрудження; [8] – запропоновано метод верифікації даних у текстах з використанням критерію топологічної спостережуваності. Удосконалено метод формування користувацьких запитів до великої мовної моделі та RAG-архітектури шляхом прив'язки ембеддингів до просторових об'єктів і часових меж; [9] – розроблено веб підсистему результатів моніторингу та клієнтську частину для візуалізації динаміки забруднення на картах Google Maps; [10] – розроблено структуру бази даних для зберігання результатів експрес-аналізу та запропоновано алгоритми організації мобільної аналітичної системи; [11] – здійснено програмно-апаратне конфігурування Arduino для підсистеми мобільно вимірювальних лабораторій; [12] – розроблено серверне програмне забезпечення для автоматизованого поєднання даних моніторингу із числовими даними; [13] – розроблено інтерфейс взаємодії з геопорталом для автоматичного оновлення шарів просторових даних за результатами мобільного моніторингу; [14] – розроблено базовий інтерфейс інтелектуальної вебсистеми WISEST-SBB з інформацією про екологічні проблеми та природоохоронні заходи у масивах вод р. Південний Буг з інтегрованою ГІС-мапою; [15] – проведено дослідження ефективності алгоритмів MaxSum та MMR для україномовних текстів екологічного спрямування. Обґрунтовано доцільність використання BERT; [16] – запропоновано інтелектуальний алгоритм для обчислення зваженої Jaccard-міри та обчислення ступеня збігу класифікованих ключових слів з текстами новин; [17] – розроблено шаблон датасету та процедуру наповнення системи розміченими фактами про стан водних об'єктів; [18] – запропоновано підхід до автоматизації розрахунків та оброблення вхідних даних водогосподарського балансу на основі ГІС басейну річки; [19] – розроблено просторово-зважений критерій MLT-ISOF з метою адаптації методу TF-IDF до задач геоприв'язки з урахуванням протяжності лінійних проєкцій геоназв на усереднену течію річки; [20] – запропоновано технологію автоматизованого збирання та оброблення текстової інформації із забезпеченням її геолокації щодо масивів вод; [21] – розроблено ПЗ для синхронізації збирання та оцифрування даних із датчиків у складі мобільної IBC.

Проведено експериментальні вимірювання; [22] – запропоновано класифікацію гетерогенних джерел даних з реалізацією спеціалізованих каскадних парсерів, розроблено та реалізовано архітектуру підсистеми каскадних парсерів для WISEST-SBB.

Апробація матеріалів дисертації. Усі описані наукові результати та здобутки, які є частиною дисертаційної роботи оприлюднювались на 15 науково-технічних і науково-практичних конференціях, збірка матеріалів однієї з яких є наукометричній базі Scopus:

- Науково-технічної конференції факультету інтелектуальних інформаційних технологій та автоматизації Вінницького національного технічного університету, Вінниця, 24-27 березня 2026 р.
- LIV Всеукраїнська науково-технічна конференція факультету інтелектуальних інформаційних технологій та автоматизації (2025)
- Актуальні проблеми та перспективи розвитку фундаментальних, прикладних, загальнотехнічних та безпекових наук : Матеріали IV всеукр. науково-практ. конф. пам'яті акад. Акад. наук вищ. освіти, проф. А.В. Касперського, м. Київ, 27 черв. 2024 р.
- Контроль і управління в складних системах (КУСС-2024), Вінниця, Україна, 16–17 жовтня 2024.
- Молодь в науці: дослідження, проблеми, перспективи (МН-2025): матеріали Всеукраїнської науково-практичної інтернет-конференції (Вінниця, 15 жовтня 2024 року
- Матеріали LII Науково-технічної конференції факультету інтелектуальних інформаційних технологій та автоматизації Вінницького національного технічного університету, Вінниця, 21 – 23 червня 2023 р.
- XXI Міжнародної науково-практичної конференції (14-16 листопада 2022 р.). Київ, 2022.
- CONFERENCE PROCEEDINGS 2017 IEEE First Ukraine Conference on ELECTRICAL AND COMPUTER ENGINEERING (UKRCON) May 29 – June 2, 2017 Kyiv.

- Матеріали XLV Науково-технічної конференції Вінницького національного технічного університету, Вінниця, 23-24 березня 2016 р
- X Міжнародна науково-практична конференція «Інтернет-Освіта-Наука (ІОН-2016)», Вінниця, Україна, 11–14 жовт. 2016. Вінниця: ВНТУ
- V-й Всеукраїнський з'їзд екологів з міжнародною участю (Екологія/Ecology-2015), 23-26 вересня, 2015.
- XIV Міжнародна науково-практична конференція: Сучасні інформаційні технології управління екологічною безпекою, природокористуванням, заходами в надзвичайних ситуаціях: 5-9 жовтня 2015 р.
- III Міжнародної конференції “Вимірювання, контроль та діагностика в технічних системах (ВКДТС-2015)», м. Вінниця, 27-29 жовтня 2015 р
- Молодь в технічних науках: дослідження, проблеми, перспективи : матеріали Міжнародної науково-практичної Інтернет-конференції, 23-26 квітня 2015 року / ВНТУ. – Вінниця, 2015

XLIV Регіональна науково-технічна конференція професорсько-викладацького складу, співробітників та студентів університету з участю працівників науково-дослідних організацій та інженерно-технічних працівників підприємств м. Вінниці та області, електронне наукове видання матеріалів конференції, м. Вінниця, 11—13 березня 2015.

Публікації. Основні положення дисертаційної роботи опубліковані у 22 наукових працях, у тому числі 7 наукових публікацій у фахових виданнях України (з яких 2 – у виданнях, що індексовані у Scopus) та 15 апробацій на науково-технічних і науково-практичних конференціях з публікацією матеріалів доповідей (з яких 1 – у збірці конференції IEEE, яка проіндексована у Scopus).

Структура та обсяг дисертації. Дисертаційна робота складається зі вступу, чотирьох розділів, висновків, списку використаних джерел та додатків. Повний обсяг дисертації становить 153 сторінки, у тому числі: 106 сторінок основного тексту, 42 рисунки, 2 таблиці, список використаних джерел із 105 найменувань, кількість додатків – 7.

РОЗДІЛ 1

АНАЛІЗ ПРОБЛЕМ ВИДОБУВАННЯ ТА ВЕРИФІКАЦІЇ ДАНИХ З РІЗНИХ ДЖЕРЕЛ

1.1 Аналіз предметної області

Відповідно до вимог Водної рамкової директиви ЄС та Водного кодексу України [28] постає необхідність розв’язання задачі збирання, впорядкування й оцінювання результативності управлінських заходів, спрямованих на покращення або стабілізацію стану кожного окремого масиву вод у межах річкових басейнів.

Наразі спостерігається істотна невідповідність між наявною мережею пунктів моніторингу якості поверхневих вод у річках України та кількістю масивів вод (МВ), які підлягають контролю [29]. Зокрема, басейн річки Південний Буг поділено на 1091 масив вод, тоді як функціонує лише близько 50 регулярних постів державної системи моніторингу стану вод. Водночас, слід зазначити, що ці пости моніторингу не охоплюють повного спектра потенційних показників якості води. Як наслідок, виникає дефіцит достовірних емпіричних даних щодо екологічного стану більшості МВ, що ускладнює формування ефективних планів управління річковими басейнами.

Таким чином, окреслюється актуальна науково-прикладна задача, що полягає у підвищенні достовірності та точності інформації про стан МВ шляхом інтеграції додаткових джерел даних. До таких джерел, поряд зі звітністю басейнових управлінь водних ресурсів, доцільно віднести матеріали засобів масової інформації, повідомлення очевидців у соціальних мережах тощо. Запропоновано застосування методів аналізу та видобування даних (data mining) з українськомовних природних текстів із подальшою їх формалізованою просторово-часовою прив’язкою до відповідних компонентів багатозв’язної геоінформаційної системи іменованих сутностей.

Для отримання інформації про стан поверхневих вод найчастіше використовуються дані у вигляді тексту. До них можна віднести різноманітні звіти, репортажі з новин, публікації на тематичних інтернет-порталах. Отримана інформація може стосуватися погіршення якості води чи середовища навколо, яке може впливати на водойми і створювати забруднення чи мор риби. Через великий об'єм та темпи зростання самостійне оброблення людиною подібної кількості інформації не є можливим. Та варто взяти до уваги, що потрібно також перевіряти достовірність та точність отриманих фактів. На це може бути витрачено багато часу та ресурсів. У такому випадку виникає актуальність створення та удосконалення існуючих додатків для спостереження за джерелами води, контролю їх якості, своєчасного реагування на негативні зміни. Також оцифрування даних допоможе проконтролювати зміни з плином часу (це можуть бути і солоність води, температура, прозорість, розчинений кисень, наявність пестицидів, нафтопродуктів чи пластику – усе, що фахівці можуть використати для ефективного реагування на проблеми підтримки якості поверхневих вод).

Виходячи з цього, утворюється необхідність розроблення методу автоматичного збирання інформації на просторах інтернету, оброблення звітів, що стосуються масивів вод, та прив'язування важливих фактів до часових інтервалів. Реалізація описаного методу дозволить різноманітним державним, некомерційним та комерційним організаціям, які працюють над моніторингом чи збереженням чистого довкілля, обробляти великі об'єми даних про минулі та актуальні події, за короткий термін на довгий час. Для дослідження і розроблення такого рішення необхідно покладатися на законодавство України та регламенти Європейського Союзу [28-34].

Важливо ефективно формалізувати мережу масивів вод як геопросторовий складений об'єкт.

У роботі [10] було введено термін багатозв'язних геоінформаційних систем (БГС), якими є багатозв'язні просторово-розподілені системи, в яких усі дані та зв'язки між ними можна формалізувати [86] у вигляді геоінформаційних систем, тобто це — системи, основні елементи та об'єкти яких мають просторову

прив'язку. Пропонується серед цих систем виділяти клас, де кожен елемент БГС є іменованою сутністю (скорочено: БГСІС).

В загальному випадку кожен елемент БГСІС може бути складеним об'єктом, який може складатися з декількох точкових, лінійних та площинних об'єктів, не всі з яких є іменованими.

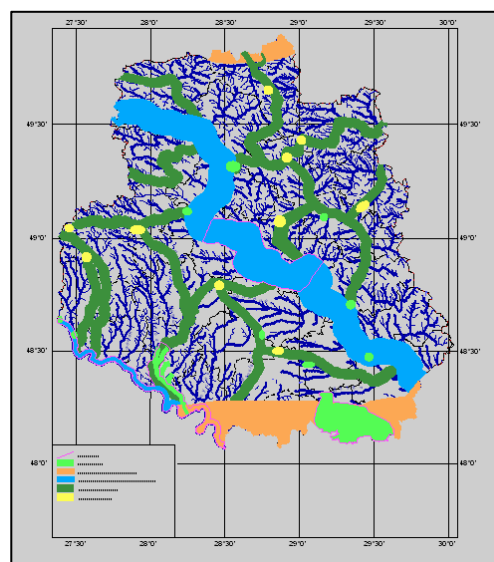
Наприклад, якщо БГС містить елемент, який не є іменованою сутністю, наприклад, точку, яка позначає місце впадіння однієї річки в іншу, тоді він вважається піделементом складеного елемента, яким є одна з цих річок, наприклад, річка-притока. І тоді БГС як сукупність лінійних і точкових об'єктів перетворюється на сукупність складених об'єктів БГСІС, усі з яких є іменованими сутностями, тобто процес трансформування БГС у БГСІС полягає у прив'язуванні усіх об'єктів, які не є іменованими сутностями БГС, до іменованих сутностей, або в їх відкиданні.

Для кожної БГСІС мають виконуватися дві умови: по-перше, кожна іменована сутність географічно та/чи семантично має бути пов'язаною хоча б з однією іншою іменованою сутністю цієї ж БГСІС, а по-друге, не повинно бути жодного об'єкта, який не є іменованою сутністю.

Прикладами БГСІС є гідрографічна мережа річок і водойм (ГМР), екологічна мережа (ЕМ), до якої відносяться прибережна зона річок і водойм, мережа адміністративних утворень (МАУ) площинного типу (області, райони, територіальні громади, населені пункти) тощо, мережа водогосподарських ділянок річкового басейну (МВГД) (рис. 1.1) [4, рис. 1].



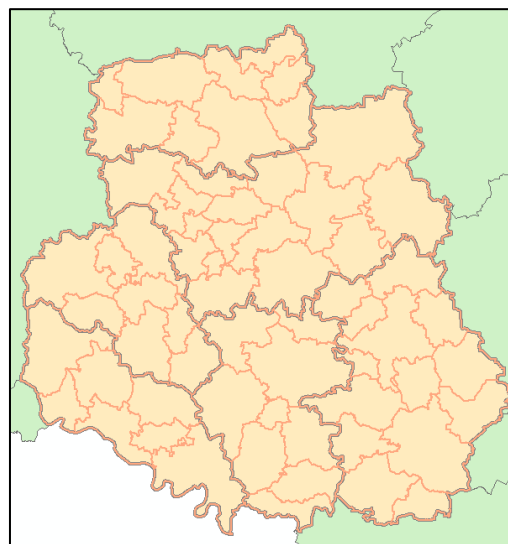
а)



б)



в)



г)

Рисунок 1.1 – Приклади багатозв’язних геоінформаційних систем іменованих сутностей (БГСІС): а) мережа річок і водойм Вінницької області у басейні р. Південний Буг; б) екологічна мережа Вінницької області; в) мережа водогосподарських ділянок басейну р. Південний Буг (з основними річками та адміністративними межами станом на 2017 р.; г) мережа районів та об’єднаних територіальних громад Вінницької області станом на 2023 р.[4, рис. 1]

Південний Буг; б) екологічна мережа Вінницької області; в) мережа водогосподарських ділянок басейну р. Південний Буг (з основними річками та адміністративними межами станом на 2017 р.; г) мережа районів та об’єднаних територіальних громад Вінницької області станом на 2023 р.[4, рис. 1]

1.2 Огляд існуючих методів та підходів до видобування текстових даних

Було розглянуто ряд систем інформаційного пошуку, які використовують онтологічні моделі [31, 35, 36], але при цьому майже ніколи не охоплюють словосполучення для оброблення інформації. Вони зазвичай розглядають лише окремі слова, що унеможлиблює пошук релевантних новин. Саме через використання статичних підходів до аналізу даних пошук повільний та не завжди видає необхідні дані, видаючи нерелевантні документи, особливо, якщо результати опубліковані різними мовами. До методів видобування ключових слів та фраз належать: [37]TF-IDF, RAKE, YAKE[38], KP-Miner, KEA, Single Rank, Text Rank, Expand Rank, Topic Rank, Multipartite Rank, Topical Page Rank та Positional Rank [39].

На відміну від названих методів, є й такі, що можуть шукати інформацію за заданим словосполученням. Це покращує точність пошуку, але при цьому є потреба в удосконаленні цих систем. Одним із рішень могло б бути створення великомасштабного корпусу текстів, який можна надалі використовувати для аналізу алгоритмів і генерації шаблонів аналізу для конкретних текстів.

На основі відомої моделі BERT [40-42] реалізовано популярний пакет KeyBERT [42-44]. KeyBERT має декілька методів для роботи з ключовими висловами. Перший з них визначає набір висловів-кандидатів. Далі цей набір аналізується, і з нього обираються лише найрелевантніші результати відповідно до критерію косинусної подібності з текстом документів. Наступний метод під назвою MaxSum проводить оптимізацію за допомогою максимізації загальної відстані між обраними парами виразів. Так збільшується релевантність відносно документа, але при цьому відбувається зменшення подібності за змістом [45].

Третій метод називається MMR і йому властивий принцип максимально-граничної релевантності, який зменшує надлишковість та збільшує різноманіття обраних результатів [45-47].

Названі три методи створені і автоматизовані у бібліотеці KeyBERT для Python.

Для того, щоб зрозуміти, наскільки ці методи ефективні, проведено дослідження матеріалів в мережі Інтернет, що стосуються результатів спостережень стану та якості водних ресурсів річки Південний Буг, результати наведено на рисунку 1.2 [15]. Приклад реалізації опубліковано в [6].

	sim
keywords	
річки південний	0.173172
південний буг	0.168476
гідрографічна мережа	0.156652
чорного моря	0.107028
малі річки	0.093931

Рисунок 1.2 – Перші п'ять найкращих ключових словосполучень із двох слів, отримані методом MMR з усередненням функції cosine similarity на усі речення, видобуті зі сторінки вебсайту БУВР Південного Бугу про гідрографічну мережу (<https://buvrpb.davr.gov.ua/vodni-resursy/hidrohrafichna-merezha>) [48]

Обидва методи – MaxSum та MMR – показали релевантні ключові слова, але при цьому MMR видав різноманітніші результати. Важко сказати, чи це є перевагою, адже не завжди різноманітність може покращити загальну роботу служб і організацій, які працюють з водою та водними ресурсами. Важливо ще звернути увагу на те, що будь-який текст має пройти якісний попередній аналіз: вилучення цифр, скорочень та шумових слів, які не несуть жодного смислового контексту для пошукових систем. Наступним моментом є проведення додаткової фільтрації для видалення дублікатів серед ключових словосполучень у прямому чи перемішаному порядку.

З огляду на аналіз трьох популярних методів KeyBERT, було сформульовано базові положення (етапи), які необхідно формалізувати:

1. Отримання тексту і його токенізація [3]. Для цього потрібно розбити текст на слова, слоги, символи, скорочення, посилання, тощо. Ці важливі елементи тексту називають токенами. Вони можуть розділятися за допомогою пробілів або знаків пунктуації. Цей етап є одним із найважливіших для багатьох методів роботи з текстовими даними. Саме тут текст перетворюється на структуровані частини, які надалі будуть використані під час автоматизованого оброблення. Якщо відображати це на діаграмі, то процес токенізації позначається через T, якщо це – формальний опис, і у вигляді кіл, якщо це – графіка. Названі кола поширені в діаграмах активності. Пропонується використовувати прямокутники з текстом, що є звичайними для блок-схем.

2. Попереднє оброблення тексту та його підготовка до процесу видалення ключових слів. Під час оброблення завжди перетворюється вхідний текст на множину токенів, які будуть оброблятися. Для кожної системи цей етап може дещо відрізнятися. Зазвичай, цей етап містить видалення зайвих символів чи літер, виправлення помилок (як це робить будь-який перекладач), лематизацію або стеммінг. За бажанням розробників можуть бути додані процеси розбиття публікації на частини, фіксовані кількості слів чи символів. Таким чином, аналіз відбувається для кожного фрагмента окремо, порівнюючи їх надалі, що може видати кращий результат, ніж аналіз цілого тексту. Multipartite Rank може аналізувати увесь текст одразу. Також варто звернути увагу на те, що BERT і подібні моделі можуть потребувати спеціального формату: задання фіксованої довжини для токенів, що може скоротити речення, відсутність збігів, позначок пробілів чи кінця речення. Цей етап можна зобразити у вигляді [PT] для формального опису чи графічно у вигляді прямокутника.

3. Опишемо процес фільтрації «стоп-слів» [49, 50]. Зазвичай, такі слова називають шумовими. Тобто вони не несуть жодного смислового ефекту і відкидаються пошуковими системами (і, або, який, що тощо). Коли ці слова видаляються, текст містить лише вагомі слова і вирази перед початком аналізу. Але цей процес потрібно проводити з обережністю, бо видалення таких слів може зменшити «природність» тексту і точність отриманих результатів може бути

значно меншою. Таке явище зустрічається, якщо застосувати моделі на кшталт BERT. У такому разі можуть замінити автоматизоване оброблення статичним задля збереження смислу. Зобразимо це через [SF] як фільтрацію стоп-слів для формалізованої схеми.

4. Визначення позиції слів у реченнях. Для українських текстів цей етап не грає сильної ролі, а для англійських це дуже важливо. Наприклад, відомо, що в англійській мові структура завжди повинна виглядати таким чином – суб’єкт, що робить, з чим чи ким (об’єкт), де і коли. Об’єкт і суб’єкт можуть мати однакове написання, але суб’єкт завжди стоїть спочатку, тому позиція конкретного слова може відіграти роль під час аналізу тексту. Позначимо цей процес [KP], оскільки англійською це скорочення від словосполучення Keyword Position.

5. Застосування мовних моделей для оброблення і перетворення текстів на їх ембедінги [51, 52]. Це формування числових векторів, які зображують слова чи вирази, зберігаючи семантичні зв’язки, які наявні в тексті. Тобто, для цього етапу (рефакторингу) важливо встановлювати зв’язок між текстовими елементами для збереження специфіки мови [53]. Позначимо це через [BERT] як для формалізованого опису, так і для графічного позначення прямокутника.

6. Визначення сукупності потенційних ключових слів і виразів. Для цього необхідно провести ранжування, вирахування частоти вживання кожного слова чи виразу, встановлення подібності за семантикою тощо. Саме за ключовими словами і виразами пошукові системи будуть знаходити необхідну інформацію краще, пропускаючи невагомі статті. Щоб зібрати ключові слова і фрази, необхідно зрозуміти, що є унікального в тексті, яка тема, головні ідеї, за якими користувач може знайти цю інформацію. Ці комбінації слів можуть подаватися у вигляді n-грам. Якщо присвоєно номери k_1 і k_2 для комбінацій, то запис буде таким: $n\text{-gram}(k_1, k_2)$. Етап підбору ключових слів позначимо через літеру [S]. Цей крок є одним із найвагоміших в алгоритмі. В діаграмі активності UML це буде зображено у вигляді дії (прямокутник із заокругленими кутами). Цей елемент можна перейменувати, якщо потрібно точніше написати про зв’язок між елементом і назвою алгоритму. Розглянемо приклад використання методу

(бібліотеки) KeyBERT. Вибір ключових слів і виразів проводять за рахунок дослідження синтезованих включень моделі BERT, коли відбувається уточнення.

Для цього можуть бути залучені наступні алгоритми:

- $[S(\text{MaxSum}(d))]$ для пошуку максимальної відстані між комбінаціями слів чи виразів. Необхідно обрати ймовірні ключові слова і вирази, які є мінімально подібними і забезпечують максимальне різноманіття;

- $[S(\text{MMR}(d))]$ для мінімізації надлишковості та максимізації різноманітності ключів. Це і буде максимальна релевантність, яку ще називають граничною.

Цікаво, що не кожен метод використовує цей етап як відокремлений, тобто аналіз проводиться для усього документу одразу.

Аналіз. Це базовий крок, який може використовувати різні методи, алгоритми, підходи та метрики, які розробник вирішує застосовувати для реалізації проєкту. На діаграмі UML це показано у вигляді прямокутника, разом з операціями уточнення, але при цьому лінії в кілька разів товщі для відображення узагальнюючого класу. Або це можна відобразити подвійними сторонами. Тип аналізу визначається методами, які використовуються. Якщо методи базові, то тоді знадобиться проводити ряд налаштувань для кожної задачі, що може призвести до збільшення енергозатратності та додаткових витрат. Можна на схемі всередині блоку написати назву алгоритму чи принципу, який є базовим. Зображення – аналогічне попереднім.

Розраховуємо частоти, підраховуючи кількість згадувань кожних ключових слів чи висловлювань в тексті, що розглядається, поділяючи цю кількість згадувань на повну кількість слів. Підрахунок частот пропонується записати як $[[C]]$, що є скороченням англійського слова «Count», що означає «рахувати». Зображення на схемі буде передане прямокутником з жирними лініями.

Наступним кроком є обчислення частоти терміну TF [54] і оберненої частоти документу IDF, яка є оберненою до першого. Тобто потрібен індикатор [55] TF-IDF [54]. Для визначення того, наскільки часто зустрічається якесь слово

в файлі, треба перемножити між собою показники TF і IDF. На схемі буде використана назва індикатора для позначення процесу.

Косинусна подібність [[COS]] також важлива для кожного ключового слова чи виразу [56]. Вона визначає подібність вищевказаних елементів тексту до самого тексту в цілому, вимірюючи кут між векторами вкладень за шкалою від 0 до 1, де чим вища цифра, тим більша відповідність.

Щоб матеріал був більш видимим для пошукових систем, з'являється потреба у ранжуванні ключових елементів у вигляді графіка зв'язків. Одним із методів, який може забезпечити цей процес, є [[TextRank]]. Кожен вагомий елемент тексту розглядається у вигляді вузла зі зв'язками з іншими вузлами. Якщо деякі слова частіше згадуються разом, то між ними встановлюється зв'язок з більшою вагою. Схоже, працює метод PageRank, але він націлений на оцінку важливості вебсторінок.

Агрегація отриманих результатів моделі в одне ціле.

Сучасна практика оброблення даних демонструє необхідність агрегування результатів декількох алгоритмів. Тобто використання лише одного методу не може бути ефективним для будь-якого типу даних. У такому випадку використовуються такі типи з'єднань:

- 1) послідовне з'єднання результатів у вигляді конвеєра;
- 2) комбінування зі зведенням отриманих даних до середньоарифметичних (злиття);
- 3) зважене зведення даних до середньоарифметичних (stacking);
- 4) ансамблеве (ensemble), що передбачає використання результатів роботи одного алгоритму для результатів іншого.

Зазвичай, такі діаграми відображають в бажаній послідовності чи як комбінацію, яку задає розробник. Злиття зображуватиметься у вигляді елемента асоціації для діаграми активності UML (вертикальна жирна лінія до вхідних елементів і з якої виходить лише один вихід). Для стекування — позначення як для нейронів, але як еліпс (*U* для формалізованої діаграми як показчик зв'язку),

пропорційний кількості входів. Якщо для результатів не потрібна агрегація, тоді цей етап пропускається.

Останній пункт алгоритму – це отримання ключових слів і виразів з тих даних, які були отримані в попередніх пунктах. Обов'язково на цьому етапі провести ранжування цих ключів у порядку зменшення показників. І, відповідно, перші показники будуть мати найкращі значення, якщо вони будуть перевищувати заданий поріг. Тобто розробник має вказати, яка мінімальна кількість разів для появи кожного слова чи виразу, щоб вважати це ключем, який варто брати до уваги. Формально цей етап буде відображатися у вигляді [KE], що позначатиме видалення ключових слів. Графічне відображення описується як вписане в коло ще одне коло чорного кольору. Тобто це буде виглядати як типове завершення алгоритму на будь-якій схемі, включаючи UML-діаграму типу позначення стану діяльності.

Для зв'язку блоків необхідно використати стрілки, щоб візуально показати, звідки і куди рухається сигнал.

Якщо певний пункт алгоритму буде повторюватися (мається на увазі використання кілька разів), то краще показати це у вигляді літери L формально – L ([Операція]), що є скороченням слова loop (петля). Це позначення часто зустрічається, якщо є зациклення. Графічно це буде показано безпосередньо циклом.

Відобразимо усі пункти запропонованого алгоритму в таблиці 1.1.

Таблиця 1.1 – Система умовних позначень для кожного пункту алгоритму видобування ключів в тексті

Но- мер п/п	Назва пункту алгоритму або методу	Формалізоване позначення	Графічне позначення
1	Tokenization	[T]	
2	Preprocessing by the certain «Algorithm»	-	Algorithm
2.1	Preprocessing Text	[PT]	PT
2.2	Stop-words Filtering	[SF]	SF

Но- мер п/п	Назва пункту алгоритму або методу	Формалізоване позначення	Графічне позначення
2.3	Keywords Position	[KP]	KP
2.4	BERT-model processing	[BERT]	BERT
3	Keywords Selection	[S]	S
3.1	Keywords Selection with MaxSum	[S(MaxSum(d))]	S(MaxSum(d))
3.2	Keywords Selection with MMR	[S(MMR(d))]	S(MMR(d))
4	Keywords Analysis	[[Method or Metrics]]	Method or Metrics
4.1	Counting Frequency	[[C]]	C
4.2	TF-IDF metrics	[[TF-IDF]]	TF-IDF
4.3	Cosine Similarity metrics	[[COS]]	COS
4.4	TextRank method	[[TextRank]]	TextRank
5	Aggregation results	-	-
5.1.	Merging results	-	
5.2	Stacking results	[U]	
6	Keywords Extraction	[KE]	
-	Loop or Recursion of the “Operation”	L([Operation])	Operation

Для вирішення задач видобування ключових слів дуже часто використовується метод «Мішок слів», який англійською називається «Bag of Words». Принцип роботи полягає у тому, що збирається уся сукупність

потенційних ключів і проводиться ранжування за кількістю вживань в тексті. Щоб підвищити вагомість знайдених ключів, треба видалити стоп-слова, пунктуаційні знаки і символи, які не несуть значущості. Такий крок є досить важливим, тому що саме на цьому етапі відбувається пришвидшення оброблення текстів і пошуку релевантних результатів. Передбачається, що класичний варіант цього методу працює зі словами, але його можливо адаптувати до оброблення n-грам. Наприклад, застосовуючи Python для написання коду разом з методом CountVectorizer та пакетом sklearn.feature_extraction.text. У такому разі можна формувати ключові вирази із двох чи трьох слів: CountVectorizer (ngram_range=(2,3)).

Алгоритм збирання і опрацювання ключових виразів:

1. Токенізація NLP: розбивається текст на менші частини.
2. Видалення символів, що не несуть жодного значення, а також – пунктуаційних знаків, а також переписування слова в нижньому регістрі.
3. Видалення стоп-слів – шумових слів, які ігноруються пошуковими системами.
4. Формування можливих ключових виразів певної довжини, які включатимуть можливі ключові слова.
5. Обрахунок частоти, що відповідає за ранг – скільки разів зустрічається кожне слово чи вираз в обраному тексті.
6. Проведення ранжування результатів у порядку зменшення.

Формалізуємо сказане на рисунку 1.3:

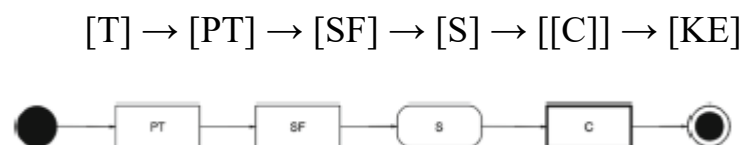


Рисунок 1.3 – Графічна нотація для методу «Bag of Words»

Наступний метод дуже схожий на перші етапи методу «Bag of Words» – «TF-IDF».

Метод «TF-IDF» заснований на використанні показників «Term Frequency» (TF) та «Inverse Document Frequency» (IDF). Перші його етапи ідентичні методу «Bag of Words», але замість простого підрахунку кількості згадувань слів у тексті обчислюється описаний вище показник «TF-IDF», який враховує ще й рідкість та унікальність кожного ключового слова чи фрази.

Алгоритм застосування:

1. Здійснити токенизацію: розділити текст на слова.
2. Передобробити текст: видалити зайві символи, знаки пунктуації та привести слова до нижнього регістру.
3. Видалити стоп-слова: видалити загальновживані та неважливі слова.
4. З потенційних ключових слів сформувати потенційні ключові фрази із заданою кількістю, наприклад, 2 та/чи 3.
5. Підрахувати вагу кожного слова «TF-IDF».
6. Ранжувати слова за зменшенням ваги «TF-IDF».

Формалізований опис (на рис 1.4):

$$[T] \rightarrow [PT] \rightarrow [SF] \rightarrow [S] \rightarrow [[TF-IDF]] \rightarrow [KE]$$



Рисунок 1.4 – Графічна нотація для методу «TF-IDF»

Інші методи, основані на різних статистичних підрахунках частоти появи потенційних ключових слів і фраз, наприклад RAKE, формалізуються аналогічно.

TextRank — це алгоритм оброблення природномовного тексту на основі графової семантичної моделі, де слова або фрази є вузлами, а їхні взаємозв'язки — ребрами (див. опис методу [[TextRank]] вище).

Алгоритм застосування (рис. 1.4):

1. Здійснити токенизацію: розділити текст на слова.

2. Передобробити текст: видалити зайві символи, знаки пунктуації та привести слова до нижнього регістру.
3. Видалити стоп-слова: видалити загальновживані та неважливі слова.
4. З потенційних ключових слів сформувати потенційні ключові фрази із заданою кількістю, наприклад, 2 та/чи 3.
5. Сформувати граф, у вузлах якого будуть усі потенційні ключові слова та фрази й усі інші слова. Проаналізувати усі речення тексту та ітераційно підрахувати ваги усіх ребер графа.
6. Ранжувати слова за зменшенням оцінки TextRank (або, як її частіше називають, PageRank).

Формалізований опис (на рис 1.5):

$$[T] \rightarrow [PT] \rightarrow [SF] \rightarrow [S] \rightarrow [[\text{TextRank}]] \rightarrow [KE]$$


Рисунок 1.5 – Графічна нотація для методу TextRank

PositionRank відрізняється від методу TextRank тим, що враховує позицію слів у реченні. Вузлам графової моделі, які відповідають словам, призначаються вагові коефіцієнти, що враховують позицію цих слів. А в англійській мові, як відомо, у більшості випадків на початку йдуть більш важливі слова [57].

Алгоритм застосування:

1. Здійснити токенизацію: розділити текст на слова.
2. Передобробити текст: видалити зайві символи, знаки пунктуації та привести слова до нижнього регістру.
3. Видалити стоп-слова: видалити загальновживані та неважливі слова.
4. З потенційних ключових слів сформувати потенційні ключові фрази із заданою кількістю, наприклад, 2 та/чи 3.
5. Визначити позицію кожного ключового слова та фрази у реченні.
6. Сформувати граф, у вузлах якого будуть усі потенційні ключові слова та фрази й усі інші слова. Проаналізувати усі речення тексту і ітераційно

підрахувати ваги усіх ребер графа, врахувавши позицію у реченні слів, розташованих у вузлах графа.

7. Ранжувати слова за зменшенням оцінки TextRank (PageRank).

Формалізований опис (на рис. 1.6.):

$[T] \rightarrow [PT] \rightarrow [SF] \rightarrow [Position] \rightarrow [S] \rightarrow [[TextRank]] \rightarrow [KE]$.



Рисунок 1.6 – Графічна нотація для методу PositionRank

Інші методи, основані на графових моделях та оцінці PageRank, наприклад, MultipartiteRank, формалізуються аналогічно, але граф може бути більш складним і може бути більше операцій на етапі передоброблення та визначення додаткової інформації про потенційні ключові слова.

Бібліотека KeyBERT дозволяє видобувати ключові слова однією командою з багатьма параметрами, де вказується яку слід використати мовну модель (англомовну чи багатомовну чи ін.), скільки слів має бути у словосполученнях (параметри n-gram [58]), який використовувати алгоритм відбору слів ([S(MaxSum)], [S(MMR)]) чи не використовувати, а, натомість – брати усі слова та наступним кроком застосовувати до них метрику [[COS]]) [59, 60].

Алгоритм застосування:

1. Здійснити токенизацію: розділити текст на слова.
2. Передобробити текст: до токенів додати спеціалізовані для BERT позначення, видалити зайві символи, знаки пунктуації та привести слова до нижнього регістру.
3. Визначити стоп-слова, щоб потім їх не відбирати як потенційні на етапі визначення оптимальних ключових слів.
4. Застосувати BERT-модель та перетворити усі слова на числові вектори-ембедінги.
5. На основі аналізу числових векторів-ембедінгів сформувати потенційні ключові фрази із заданою кількістю, наприклад, 2 та/чи 3,

прослідкувавши, щоб серед них не було стоп-слів. На цьому етапі можливі 3 варіанти:

6. Відбираються усі слова і фрази, які задовольняють зазначеним вимогам.
7. Слова і фрази відбираються за алгоритмом MaxSum.
8. Слова і фрази відбираються за алгоритмом MMR.
9. Обчислити косинусну подібність та відібрати найкращі ключові слова і фрази із текстом.

Формалізований опис:

- варіант 1:

$$[T] \rightarrow [PT] \rightarrow [SF] \rightarrow [BERT] \rightarrow [S] \rightarrow [[COS]] \rightarrow [KE]$$

- варіант 2:

$$[T] \rightarrow [PT] \rightarrow [SF] \rightarrow [BERT] \rightarrow [S(\text{MaxSum}(d))] \rightarrow [[COS]] \rightarrow [KE]$$

- варіант 3:

$$[T] \rightarrow [PT] \rightarrow [SF] \rightarrow [BERT] \rightarrow [S(\text{MMR}(d))] \rightarrow [[COS]] \rightarrow [KE]$$

Графічна нотація методу KeyBERT наведена на рис. 1.7.



Рисунок 1.7 – Графічна нотація для методу KeyBERT

Запропонований формалізований опис дозволяє легко здійснювати поєднання, трансформацію та гібридизацію цих та інших методів.

Наприклад, команда KeyBERT повертає тільки список оптимальних слів та фраз, тому, щоб покращити візуалізацію, можна ще обчислити частоту появи цих слів і фраз у тексті – звичайну метрику $[[C]]$ або $[[TF-IDF]]$ чи ін.:

$$[T] \rightarrow [PT] \rightarrow [SF] \rightarrow [BERT] \rightarrow [S] \rightarrow [[COS]] \rightarrow [[C]] \rightarrow [KE]$$

або

$$[T] \rightarrow [PT] \rightarrow [SF] \rightarrow [BERT] \rightarrow [S(MMR(d))] \rightarrow [[COS]] \rightarrow [[TF-IDF]] \rightarrow [KE].$$

Таке агрегування відповідає типу «ensembling», коли ключові етапи одного методу застосовуються після іншого, але без дублювання однакових функцій: зокрема, це - «ensembling» методів «KeyBERT»-«Bag of Words» та «KeyBERT з MMR»-«TF-IDF», відповідно.

Інший варіант – це описані вище методи агрегування merging, stacking, pipeline та ensembling. Наприклад, на рис. 1.6 наведено варіант «stacking» методу «TF-IDF» та усі варіації методів «KeyBERT».

Retrieval-Augmented Generation (RAG — генерація з доповненням через пошук) — це метод, який допомагає підсилити великі мовні моделі, які називають LLM [61]. Підсилення відбувається за рахунок того, що ці ж мовні моделі отримують доступ до баз даних в реальному часі. Відмінністю RAG від звичайних LLM є те, що другі генерують результати лише на даних, які були отримані під час тренування, а RAG може працювати з динамічними даними, векторними базами даних [62] або з API пошуком. Тому RAG може надати більш точні, релевантні і обґрунтовані відповіді.

Опишемо, як приблизно може працювати RAG у поєднанні з AI:

1. Користувач-експерт задає запит (наприклад, «Де безпечно купатись влітку»).
2. Будується вектор, складовою якого є отриманий запит, після чого порівнюється з уже наявними векторами.
3. У відповідь надходять k-релевантні документи.
4. Отримані документи надходять мовним моделям у вигляді текстових даних через доповнення промту.
5. Мовні моделі генерують відповідь користувачу на базі отриманих даних.

Тому такий підхід є актуальним для формулювання і реалізації принципів оброблення даних про водні масиви України за допомогою інформаційної технології добування і верифікації даних про водні об'єкти з часовою і просторовою прив'язкою [63].

Важливою частиною роботи є векторні бази даних, тож розглянемо інформацію про них.

Кількість інформації і її різноманітність постійно зростають, тому добувати і обробляти дані стає все складніше. Виникає проблема для стандартних баз даних, які зазвичай використовувалися для подібних задач. Перетворювати інформацію у вектори і маніпулювати цими даними стає простіше. Саме вказані бази даних розроблялися для роботи з високорівневою інформацією, що дозволить якісно обробляти зібрані інформаційні джерела про водні ресурси України у режимі реального часу із прив'язкою до геолокації.

До такої інформації відносяться не тільки дані про стан вод, а й інформація про складові водогосподарського балансу (ВГБ) B для водогосподарських ділянок (ВГД) басейну річки [65]:

$$B = W_{\text{вх}} + W_{\text{біч}} + W_{\text{пзв}} + W_{\text{зв}} + W_{\text{дот}} \pm \Delta V \pm \Delta W_{\text{л}} - W_{\text{вип}} - W_{\text{ф}} - W_{\text{з}} - W_{\text{пер}} - W_{\text{вкр}} - W_{\text{кп}}, \quad (1.1)$$

де $W_{\text{вх}}$ – об'єм стоку, що надходить за розрахунковий період з вище розташованих ВГД;

$W_{\text{біч}}$ – об'єм стоку, що формується на розрахунковій ВГД (бічна приточність);

$W_{\text{пзв}}$ – об'єм водозабору із підземних водних об'єктів, який здійснюється згідно чинного законодавства;

$W_{\text{зв}}$ – об'єм зворотних вод на розрахунковій ВГД;

$W_{\text{дот}}$ – дотаційний об'єм води на ВГД (зовнішні та внутрішньо басейнові перекидання);

$\pm \Delta V$ – спрацювання (+), наповнення (–) ставків та водосховищ;

$\pm\Delta W_{\text{л}}$ – втрати води при осіданні льоду на берегах при зимовому спрацюванні і/або повернення води в результаті танення льоду весною;

$W_{\text{вип}}$ – втрати на додаткове випаровування та льодоутворення з водосховищ;

$W_{\text{ф}}$ – фільтраційні втрати з водосховищ;

$W_{\text{з}}$ – зменшення стоку річки, викликане забором гідравлічно-зв'язаних з нею підземних вод;

$W_{\text{пер}}$ – перекидання частини стоку за межі розрахункової ВГД;

$W_{\text{вкр}}$ – забір поверхневих вод;

$W_{\text{кп}}$ – вимоги до стоку в замикаючому створі ВГД (комплексний попуск).

Приклад частини таких даних подано на рис. 1.8.

Режим роботи водосховищ комплексного призначення
на весняно-літній період 2016 року

№ п/п	Найменування водосховища	Проектні характеристики		Фактичні характеристики		Об'єм регулю- вання при пониженні рівня води на 0,5м, млн.м ³	Розра- хункові санітарні витрати, м ³ /с
		відмітка НПР, м БС	об'єм води, млн.м ³	рівень води, м	об'єм води, млн.м ³		
1	2	3	4	5	6	7	8
р.Рось							
1	Косівське	177,00	9,62	177,04	9,77	1,890	0,35
2	Володарське	167,50	3,44	167,58	3,53	0,600	0,80
3	Шарбавівське	164,00	1,50	163,75	0,05	1,100	0,05

Рисунок 1.8 – Фрагмент рішення Міжвідомчої комісії по узгодженню режимів роботи водосховищ і управлінню водними ресурсами у басейні річки Рось за 25.03.2016 р.

У роботі [18] наведена когнітивна карта зі складовими ВГБ для типової ВГД (рис. 1.9).

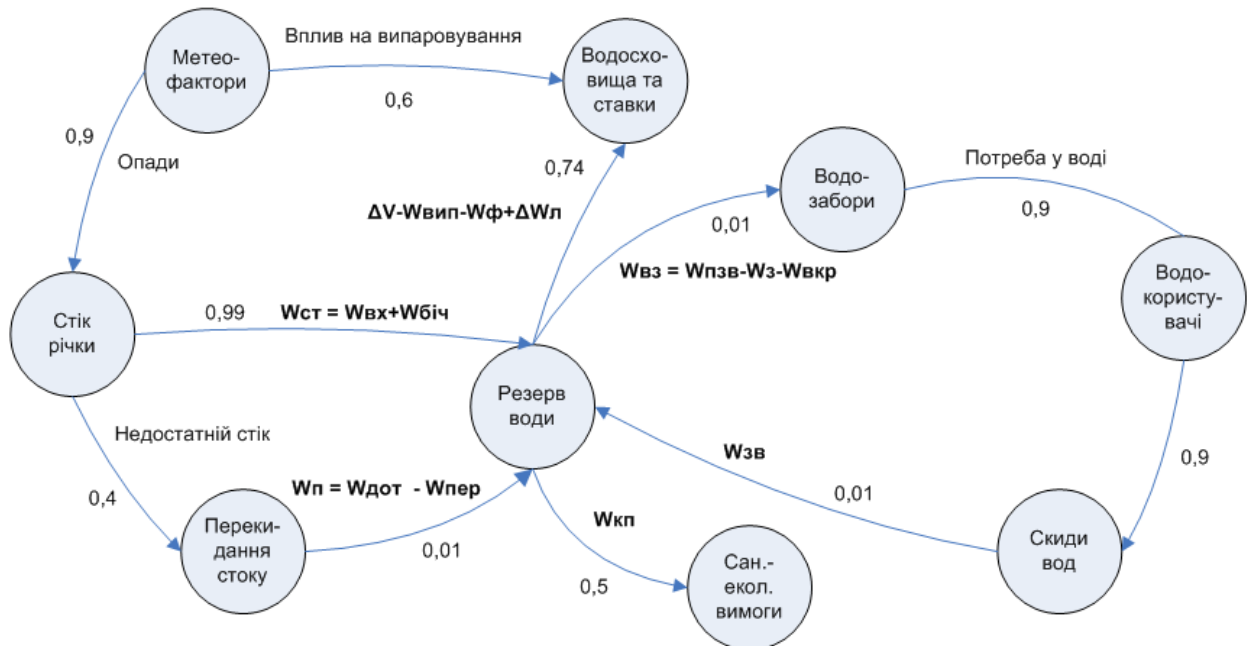


Рисунок 1.9 – Когнітивна карта водогосподарського балансу басейну річки

1.3 Проблема верифікації текстових даних

Для збирання значних обсягів даних про стан складних багатозв'язних багатостадійних геоінформаційних систем доцільно застосовувати прямі вимірювання. Водночас для досягнення високої точності створення їх цифрових двійників перевагу слід надавати не ручним або напівавтоматизованим підходам із залученням людини, а – мобільним інформаційно-вимірювальним комплексам чи рішенням на основі Інтернету речей (IoT) [1, 2, 7, 8, 21, 67, 68]. Проте IoT-системи потребують регулярної верифікації за допомогою альтернативних джерел даних [69].

Сучасні підходи до вебскрапінгу та інтелектуального вилучення числових показників із текстів є достатньо розвиненими, однак для багатостадійних геоінформаційних систем ключовою проблемою їх практичного застосування залишається коректна геоприв'язка до відповідних елементів системи. Інакше кажучи, необхідно чітко ідентифікувати, про який саме об'єкт (наприклад, ділянку річки, трубопровід чи регіон) йдеться в повідомленні. Крім того, через динамічний характер таких систем критично важливо забезпечити точну часову прив'язку даних.

Як зазначено в [5], сучасні сховища знань нового покоління, зокрема Wikidata та YAGO4, передбачають збереження часових і просторових атрибутів. Однак, автоматизоване вилучення таких характеристик безпосередньо з текстів наразі залишається недостатньо дослідженим напрямом [70]. Процес ідентифікації географічних назв (топонімів) у текстах та їх зіставлення з координатами відомий як геопарсинг. Його реалізація ускладнюється різноманітністю мовних конструкцій, неоднозначністю топонімів, контекстуальною залежністю та іншими факторами [53].

Застосування великих мовних моделей (LLM) сприяє встановленню зв'язків між сутностями у структурованих текстових даних [71], що полегшує інтеграцію інформації про різні компоненти геоінформаційних систем. Водночас такі моделі можуть генерувати неточні або вигадані відомості (так звані «галюцинації»). Підхід із доповненим пошуком (RAG), який передбачає попереднє знаходження релевантних фактів у базах даних із подальшим їх використанням для аналізу, також має обмеження — насамперед у частині пошуку з урахуванням часових і просторових характеристик, а також складної структури текстових джерел [72, 73].

Отже, можна виокремити такі ключові проблеми:

1. Забезпечення достовірності числових даних, отриманих інформаційно-вимірювальними системами;
2. Автоматизована просторово-часова прив'язка даних до конкретних елементів багатостадійних геоінформаційних систем.

1.4 Постановка задач

Наявні дані про водні масиви, екологічні проблеми [74], різні види забруднень, скидів та їхні наслідки (наприклад, випадки масової загибелі риби з невстановлених причин) характеризуються значною різноманітністю та фрагментарністю.

Основним завданням є інтеграція цих даних у єдину базу знань. Проте простий пошук за ключовими словами, наприклад, «масова загибель риби», забезпечує лише загальні результати без урахування просторової локалізації подій. У контексті створення інтелектуальної системи важливо забезпечити просторово-часову орієнтацію, зокрема шляхом автоматичного визначення релевантних триплетів: «ДЕ? – ЩО? – КОЛИ?». Це дозволить побудувати систему та інтерфейс, здатні до автоматичного оновлення даних із прив'язкою до часу та конкретної географічної локації.

Гіпотеза: формалізація екологічних даних у вигляді фактів SPO-триплетів з просторовою та часовою прив'язкою забезпечує значне підвищення точності геоприв'язки у порівнянні з видобуванням ключових фраз, без структурування.

Проблема: офіційні звіти часто узагальнені та не деталізують результати на рівні масивів вод.

Для розв'язання вищевказаної проблеми пропонується розв'язати такі задачі:

1. Здійснити аналіз існуючих методів та інформаційних технологій видобування та верифікації даних з україномовних текстів про взаємопов'язані водні об'єкти, включаючи методи видобування ключових слів та словосполучень, методів автоматизованої геоприв'язки до географічних об'єктів, методів верифікації числових даних

2. Розробити метод верифікації числових даних у україномовних текстах про водні об'єкти, для оцінювання повноти та достовірності одержаної інформації.

3. Розробити та апробувати підхід на основі Retrieval-Augmented Generation (RAG) для удосконалення методу формування користувацьких запитів до LLM, з прив'язкою ембеддінгів до просторових об'єктів та часових меж. З метою підвищення достовірності кожної окремої події та точності пошуку інформації з урахуванням просторово-часових обмежень.

4. Удосконалити метод автоматизації геоприв'язки україномовних природномовних текстів про стан поверхневих вод до масивів вод басейну річки.

5. Розробити та апробувати інтелектуальну інформаційну технологію видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою.

Висновки до розділу 1

У першому розділі проаналізовано предметну область та встановлено, що в річках України кількість масивів вод, які моніторять належним чином, незначна. У річці Південний Буг регулярний моніторинг проводиться для, приблизно, 50 масивів вод із 1091. Згідно з інформаційною довідкою[80] щодо плану управління річковими басейнами України на 2025-2030 рр., в Україні екологічний стан визначається для 5325 масивів поверхневих вод, з яких лише про 5% достатньо інформації для визначення оцінки стану. Таким чином, виявлена проблема моніторингу стану масивів вод є актуальною для всіх басейнів річок України. Також варто зазначити, що наявні пости регулярного моніторингу мають обмежений перелік показників моніторингу. Отже, виникає проблема у використанні гетерогенних джерел даних, таких як публікації у ЗМІ, соціальних мережах та ін) та автоматизації оброблення великої кількості україномовних природномовних текстів.

Проведений аналіз існуючих методів та підходів до видобування даних з корпусів текстів показав, що класичні (Bag of Words, TF-IDF), графові (PositionRank, TextRank) та нейромережеві (KeyBERT з MaxSum, MMR) методи мають фундаментальні обмеження для описаної задачі. Наведені методи працюють з словами на рівні лексичних одиниць, не враховують ні контекст ні ієрархію слів та документів та фактичні розміри геоназв. Проведений аналіз з порівнянням MMR та MaxSum на текстах БУВР р. Південний Буг показав, що для цих методів важливо здійснювати якісне передоброблення тексту і вони не враховують особливостей екологічних текстів, що призводить до втрати точності, а також ці методи показують менш корисні результати за умов топонімичної полісемії.

Проведений аналіз проблеми верифікації текстових даних показав головні недоліки існуючих інструментів, а саме: існуючі геопарсери оптимізовані для англomовних текстів, LLM часто «галюцинують», що не допустимо у задачі оброблення числових даних про стан довкілля та класичні реалізації RAG-архітектури не надають можливості явно застосувати фільтрувати за просторово-часовими можливостями, що підвищує хибнопозитивні та хибнонегативні результати запитів. Отже, виникає потреба не тільки забезпечити точність прив'язки конкретних текстів до конкретних масивів вод, але і критерію повноти бази знань.

Виходячи із аналізу (п. 1.4) сформульовано вихідні передумови та наступні задачі: відомі методи не можуть одночасно забезпечити точну геоприв'язку до конкретних масивів вод, часову прив'язку, верифікацію числових даних та регулярного поповнення україномовних фактів щодо водних об'єктів. Таким чином, обґрунтована необхідність розробки інформаційної технології видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою. Поставлено основні задачі, які слід розв'язати для усунення чи мінімізації зазначеної проблеми.

Результати, наведені у цьому розділі, опубліковані у працях [3,15,18] здобувача.

РОЗДІЛ 2

РОЗРОБЛЕННЯ МЕТОДІВ ВИДОБУВАННЯ ТА ВЕРИФІКАЦІЇ ДАНИХ ІЗ ПРОСТОРОВОЮ ТА ЧАСОВОЮ ПРИВ'ЯЗКОЮ

2.1 Формалізація вимог до даних та математична постановка задачі

Введемо систему позначень, яка описані в [5]:

- S – множина речень з початкового корпусу C документів, які аналізуються;
- S_n – речення з нових документів, що надходять у режимі потокового оновлення (RAG);
- $S_{all} = S \cup S_n$ – повний корпус текстів;
- $W = \{w_1, \dots, w_m\}$ – множина масивів вод, які розуміються як їх водозбірні зони, до території яких прив'язуються географічні назви – це, скоріше, можна назвати «просторовий контекст» масивів вод;
- $G = \{g_1, \dots, g_k\}$ – географічні назви (словосполучення);
- $T = \{t_1, \dots, t_l\}$ – допустимі часові інтервали;
- N – абстрактні об'єкти (напр. «вода»), які можуть бути елементами SPO-триплетів, в яких фігурують географічні назви;
- P – множина предикатів SPO-триплетів;
- E – множина речень про «екологічні факти» (стан вод або факти забруднення вод, яке впливає на цей стан);
- E_f – підмножина речень з множини E , які вдасться точно прив'язати до інтервалу часу і координат хоча б одного масиву вод.

Множинні зв'язки:

- $G(w) \subseteq G$ – географічні назви, пов'язані з w з множини W ;
- $G(s) \subseteq G$ – назви в реченні s з множини S_{all} ;

- $T(s) \subseteq T$ – час, прив'язаний до s з множини S_{all} ;
- $R_{Gw} \subseteq G \times W$ – зв'язки між назвами й масивами вод;
- $R_{Ew} \subseteq E_f \times W$ – зв'язки між фактами й масивами вод.

Вхідною інформацією методу є така:

- корпус C українськомовних документів, який містить багато інформації, у т. ч. дані про екологічні факти та згадку про час (наприклад рік чи діапазон дат), за який наведені ці дані, і дані про те, яких річок чи населених пунктів біля річок стосується інформація;

- Ψ – таблиця з кодами, назвами, координатами масивів вод як географічних площинних об'єктів, координатами їх водозбірних зон та списки пов'язаних з цими зонами географічних назв об'єктів.

Метою є пошук C речень з множини S про стан чи забруднення вод річки та їх прив'язка у часі (по суті, з множини E) та до масивів вод Ψ , щоб потім можна було узагальнити інформацію про кожен масив вод за заданий час.

Основні етапи методу, що пропонується, полягають в наступному:

1) розбити вхідний текст (кожен документ корпусу текстів C) на речення і відібрати такі з них, де йдеться про забруднення або стан вод річки, що утворить множину речень з екологічними фактами E ;

2) на основі таблиці Ψ побудувати множину усіх можливих кодів масивів вод W та відповідних їм усіх можливих географічних назв G (об'єднати усі можливі варіанти для кожного елемента з W) як ключових словосполучень чи слів;

3) з використанням LLM у C знайти усі географічні назви з множини G (з урахуванням відмінків, різних закінчень) – усі знайдені назви сформують множину G_f , тобто – назви, які реально є у наведених фактах у текстах C ;

4) з використанням LLM пов'язати кожне речення з E з кожною географічною назвою із G_f (географічна назва може бути заголовком, а текст про стан річки йти далі в тексті в наступних реченнях) шляхом формалізації зв'язків у

вигляді згаданих вище SPO-триплетів "subject–predicate–object", де "subject" та/або "object" – це елемент множини G_f , а "predicate" формується на основі речення з E ;

5) з використанням LLM у текстах C знайти прив'язку до різних часових інтервалів (рік, квартал, сезон року, місяць, дата, діапазон дат тощо) T та спробувати прив'язати за їх допомогою кожне речення з E у часі (весь текст може бути звітом за 2019 рік, тоді це означає, що всі речення будуть за 2019 р., але можуть бути вказані й конкретні дати; важливо, що LLM слід "зрозуміти" текст - дуже часто прив'язка до часу має місце не у самому реченні - а в попередньому, на іншій сторінці, або навіть - на титульній сторінці, а може бути й у заголовках різного рівня, до яких відноситься задане речення);

6) на основі множини E формується множина E_f , яка містить тільки такі речення, які вдалось прив'язати до хоча б однієї географічної назви із G_f та хоча б одного інтервалу часу із T ;

7) на основі відповідності G і W встановити відповідність E_f, T, W тобто визначити які речення з E_f із прив'язкою у часі до елементів T відповідають яким масивам вод W ;

8) прибрати дублікати рядків у таблиці W, E_f, T та об'єднати E_f для однакових пар значень W, T .

Для пояснення головної ідеї методу формалізуємо контекстуально-семантичний критерій повноти та коректності знань про задані просторові об'єкти. Для цього спочатку уточнимо одну з його складових. Нехай:

- $Pr(s \rightarrow G(w)) \in [0, 1]$ – модельна ймовірність того, що речення s стосується просторового контексту масиву вод w ;
- $Pr(s \rightarrow t) \in [0, 1]$ – модельна ймовірність того, що речення s належить часовому інтервалу t ;
- $FP(B)$ – оцінка кількості хибнопозитивних зв'язків у множині B .

Тоді повна ймовірність істинності зв'язку буде розраховуватися за формулою (2.1):

$$Pr(s \rightarrow w, t) = Pr(s \rightarrow G(w)) \times Pr(s \rightarrow t). \quad (2.1)$$

А тоді як критерій оптимальності пропонується знайти за формулою (2.2):

$$L(B) = \sum_{(s, w, t) \in B} Pr(s \rightarrow w, t) - \lambda \times FP(B) \rightarrow \max, \quad (2.2)$$

де:

$Pr\left[\left(e_i \rightarrow w_k, t_l\right)\right]$ – ймовірність того, що зв'язок є семантично та

контекстуально правильним;

$FP(B)$ – кількість хибнопозитивних (англ. «False Positive») зв'язків;

$\lambda \in R, \lambda > 0$ – коефіцієнт штрафу за хибну прив'язку знань.

Головна ідея методу полягає в тому, що він забезпечує максимізацію критерію, одночасно вирішуючи 3 задачі:

1) забезпечити пошук максимально великої кількості варіантів зв'язків між елементами множин G, W та словами з речень із множини S ;

2) обмежити кількість хибнопозитивних зв'язків між реченнями з множин S та G, T шляхом відбору тільки таких SPO-триплетів, які пов'язують кожне речення хоча б з одним триплетом та хоча б один елементом з множини G і хоча б одним – з T ;

3) обмежити кількість хибнопозитивних зв'язків між елементами множин S, G, T та W , щоб уникнути ситуацій, коли новина чітко про конкретний w (де розташований Київський міст у Вінниці через р. Південний Буг у Вінницькій області), а її хибно прив'яжуть до інших значень w (м. Хмельницький на р. Південний Буг або м. Хмільник на р. Південний Буг у Вінницькій області), в яких теж є такі самі географічні назви $g_1 g_2$ із G - для цього використано міру Жаккара для порівняння списку підмножин із G у реченні S та підмножин з G у масиві вод w [16].

Для цього важливо поєднати та припасувати відомі методи видобування і формалізації знань, побудови SPO-триплетів з використанням великих мовних моделей та RAG-режим оновлення векторної бази даних фактів.

Було створено датасет для тренування інтелектуальних моделей для класифікації інформації про екологічні проблеми та заходи у масивах вод на прикладі басейну р. Південний Буг. Для розв'язання поставленої задачі було знайдено 20 фактів про стан масивів вод, наведені у Додатку Г, і на них відпрацьовано, по-перше, шаблон для їх систематизації, а по-друге, інтелектуальні засоби для їх оброблення з використанням отриманого раніше досвіду.

Було прийнято рішення робити шаблон на базі MS Excel (рис. 2.1) [17, рис.1].

	C	D	E	
	link	sour	sour	geonames_text
режжя річки Південний Буг на	https://www.fakty.co	Пост	ФБ	Південний Буг, Староміський міст, Н
ні. Незабаром – почнуть	https://vsim.ua	Пост	ФБ	Південний Буг, водосховище, Хмель
кисню. Дізнавайтеся яку причину				
профілактики хвороб повідомили		Веб-		
. Відтак, стверджують, що кисню	https://vsim.ua/Podii/ekologi-pripustili-chomu-v-pivdennomu-buzi-gine-riba-11638925.html	сторінка		
2/дм3), можемо констатувати, що		нови		
управлінні.		н		
ми). В результаті таких ловів,		Хме		
		льні		
		ччин		
	https://tsn.ua/u	и	сайт	Південний Буг, м. Хмельницький
	https://fakty.co	Веб-	сайт	Південний Буг, м. Первомайськ, Мик
лісцеві жителі та науковці проти,	https://vitatv.co	Веб-	сайт	Південний Буг, Миколаївщина, Олек
	https://vitatv.co	Віта	сайт	Південний Буг, м. Вінниця, Старе міс
	https://vitatv.co	Віта	сайт	Південний Буг, Лаврівка, Стрижавка
	https://vitatv.co	Віта	сайт	Південний Буг, Хмельницький район,
	https://vitatv.co	Віта	сайт	Південний Буг, Хмельник, Старий зам
еберня". Спеціалісти відділу	https://vitatv.co	Віта	сайт	Вінницька реберня, набережна Роше
годом приступлять до робіт	https://vitatv.co	Віта	сайт	Південний Буг, Вінницяводоканал
	https://vitatv.co	Віта	сайт	Південний Буг, м. Вінниця
живлення ситуації.	https://vitatv.co	Віта	сайт	Південний Буг, м. Ладижин, кафе "Кс
	https://vitatv.co	Віта	сайт	Південний Буг, Хмельникводоканал
	https://vitatv.co	Віта	сайт	Південний Буг, Степашки, Гайсинськ
	https://vsim.ua	Нови	сайт	Південний Буг, м. Хмельницький, мік
Буг на пляжах міста в місцях	https://www.fakty.co	Голо	ФБ	Південний Буг, м. Вінниця, пляж "Го

Рисунок 2.1 – Структура шаблону з розміткою датасету [17, рис.1]

Як видно на рисунку 2.1, шаблон передбачає інформацію про сам текст, його джерело, прив'язку у просторі (до карти) та у часі і класифікацію за рядом ознак. Ще можливими є примітки, коли розробник датасету не впевнений в деяких своїх оцінках цих даних.

Проведемо тестування того, як інтелектуальні засоби зможуть здійснити геоприв'язку з використанням бібліотеки SpaCy [75] та класифікацію знайденої інформації з використанням мовних моделей BERT та різних моделей машинного навчання для класифікації даних (рис. 2.2) [17, рис.2] [76].

```
Score = 0.333 for WB code UA_M5.4_0002 and maybe: UA_M5.4_0002 UA_M5.4_0003 UA_M5.4_0004 UA_M
5.4_0011 UA_M5.4_0013 UA_M5.4_0019 UA_M5.4_0028 UA_M5.4_0969 UA_M5.4_0970
['Південний Буг', 'Олександрівське водосховище'] 0.16666666666666663
['Південний Буг', 'Олександрівське водосховище'] 0.21428571428571422
['Південний Буг', 'Олександрівське водосховище'] 0.11111111111111109
['Південний Буг', 'Олександрівське водосховище'] 0.09999999999999996
['Південний Буг', 'Олександрівське водосховище'] 0.03448275862068965
['Південний Буг', 'Олександрівське водосховище'] 0.0
['Південний Буг', 'Олександрівське водосховище'] 0.046874999999999986
['Південний Буг', 'Олександрівське водосховище'] 0.0
['Південний Буг', 'Олександрівське водосховище'] 0.06730769230769229
['Південний Буг', 'Олександрівське водосховище'] 0.08535533905932736
['Південний Буг', 'Олександрівське водосховище'] 0.07142857142857141
```

Рисунок 2.2 – Приклад роботи NLP-програми на Python з автоматичним прив'язуванням інформації до масивів вод [17, рис.2]

Аналіз показав, що потрібно удосконалювати інтелектуальні засоби, враховуючи певні особливості української мови та певну стилістику подання новин, щоб підвищити точність автоматичної геоприв'язки та класифікації новин та іншої інформації.

2.2 Розроблення методу просторової прив'язки текстів

Відповідно до поставленої задачі щодо розроблення методу просторової прив'язки текстів пропонується наступний алгоритм.

Будемо вважати, що вхідний документ з корпусу текстів C вже перетворено на текст із множиною речень S та заголовків H , яким відповідають ці речення (2.3):

$$S = \{s_1, s_2, \dots, s_n\}, H = \{h_1, h_2, \dots, h_k\}. \quad (2.3)$$

Важливо, що кожне речення відповідає як заголовку, в якому воно розміщене $h = p(s)$, так і усім заголовкам вищого рівня, які охоплюють цей заголовок (2.4):

$$H(s) = \{\pi(s)\} \cup \text{Ancestors}(\pi(s)). \quad (2.4)$$

Введемо множину масивів вод з офіційного реєстру (2.5):

$$W = \{w_1, w_2, \dots, w_m\}, \quad (2.5)$$

де кожен об'єкт w_i має унікальний ідентифікатор та назву.

Функція у формлі $G(w_i)$ у формулі (2.6), що повертає множину географічних ключових термінів, асоційованих із об'єктом w_i :

$$G: W \rightarrow 2^U, \quad (2.6)$$

де U – універсум усіх географічних назв (областей, річок, міст, тощо), а $G(w)$ – підмножина U , яка відповідає об'єкту w . Наприклад, $G(UA - M5.4 - 0011)$ {"Вінницька область"} {"Південний Буг"} {"Хмільник"}.

Визначимо функцію виділення географічних термінів в тексті речення. Нехай $X(s)$ – множина географічних назв, знайдених безпосередньо в тексті речення S . Крім того, визначимо розширену множину географічних назв у реченні з урахуванням контексту заголовків, з урахуванням їх ієрархії (2.7):

$$G(s) = X(s) \cup \bigcup_{h \in H(s)} X(h), \quad (2.7)$$

де $X(h)$ – множина географічних назв, присутніх у заголовку розділу h .

Таким чином, $G(s)$ містить усі географічні назви, які стосуються факту із речення s .

Виділяємо речення, що містять екологічні факти (2.8):

$$E = \{s \in S_{all}; IsEcological(s) = 1\}. \quad (2.8)$$

Знайдемо максимальну кількість SPO-триплетів, де "subject" та/або "object" – елементи з множини G , а "predicate" сформуємо на основі речень з E . При цьому, відбираються тільки такі триплети, які містять прив'язку у часі.

Витягуємо триплети (з максимально великої множини варіантів потенційних зв'язків між елементами) з формул (2.9) і (2.10):

$$T_{SPO} \subseteq (G \cup N) \times P \times (G \cup N), ExtractSPO: E \rightarrow 2^{T_{SPO}}, \quad (2.9)$$

$$\forall s \in E, T(s) = ExtractSPO(s). \quad (2.10)$$

Цей крок впливає на $\sum Pr(s \rightarrow w, t)$ – чим більше правильних триплетів, тим вища сума ймовірностей у критерії.

Методика. Для реалізації запропонованого критерію встановимо наступні вхідні дані [19]:

$M = \{M_i\}$ – множина масивів вод із довжинами лінійних осей L_i ;

$G = \{G_i\}$ – множина географічних назв із довжинами проєкцій $L_{g(i,j)}$ на лінійній осі відповідних M_i ;

$L_{g(i,j)}$ відображає, яку частину масиву вод M_i охоплює геоназва, та чи охоплює довшу відстань, ніж має лінійна вісь заданого масиву вод;

$T = \{T_k\}$ – множина довільних текстів, що містять слова з G .

Для адаптації TF-IDF вводяться дві складові – Multiplicity-Length Term (2.11), де замість частоти терміну у документі використовується просторова кратність довжини топоніма до лінійної осі масиву вод, а замість кількості

документів Inverse Spatial Object Frequency представляє кількість просторових об'єктів, на яких зустрічається топонім, що враховує просторові характеристики.

Для кожної пари «масив – геоназва» (M_i, G_j) , визначаємо MLT через r , використовуючи формулу (2.11):

$$MLT_{i,j} = \begin{cases} r \times \frac{L_i}{L_{g(i,j)}}, \text{ якщо } L_{g(i,j)} > L_i, \\ r, \text{ якщо } L_{g(i,j)} \leq L_i, \end{cases} \quad (2.11)$$

де:

r – кількість входжень G_j у тексті;

L_i – це – довжина лінійної осі масиву M_i ;

$L_{g(i,j)}$ – довжина проєкції G_j на лінійну вісь масиву M_i .

Отже, за умови, що проєкція топоніму більш протяжна, ніж лінійна вісь масиву вод, тоді її вага множиться на коефіцієнт співвідношення довжин, інакше вагою є кількість згадувань геоназви.

ISOF – це просторовий аналог класичного IDF, який відображає унікальність G_j [19]:

$$ISOF_j = \log \frac{N}{n_j}, \quad (2.12)$$

де N – загальна кількість масивів вод у річці; n_j – к-ть масивів, до яких належить географічна назва G_j .

Таким чином, поширені топоніми, що присутні у багатьох масивах (наприклад, «водосховище»), мають меншу вагу, а вищі значення $ISOF_j$ вказують на рідкісніші топоніми.

Відобразимо залежність ISOF від кількості масивів вод, до яких належить геоназва, на рисунку 2.3.

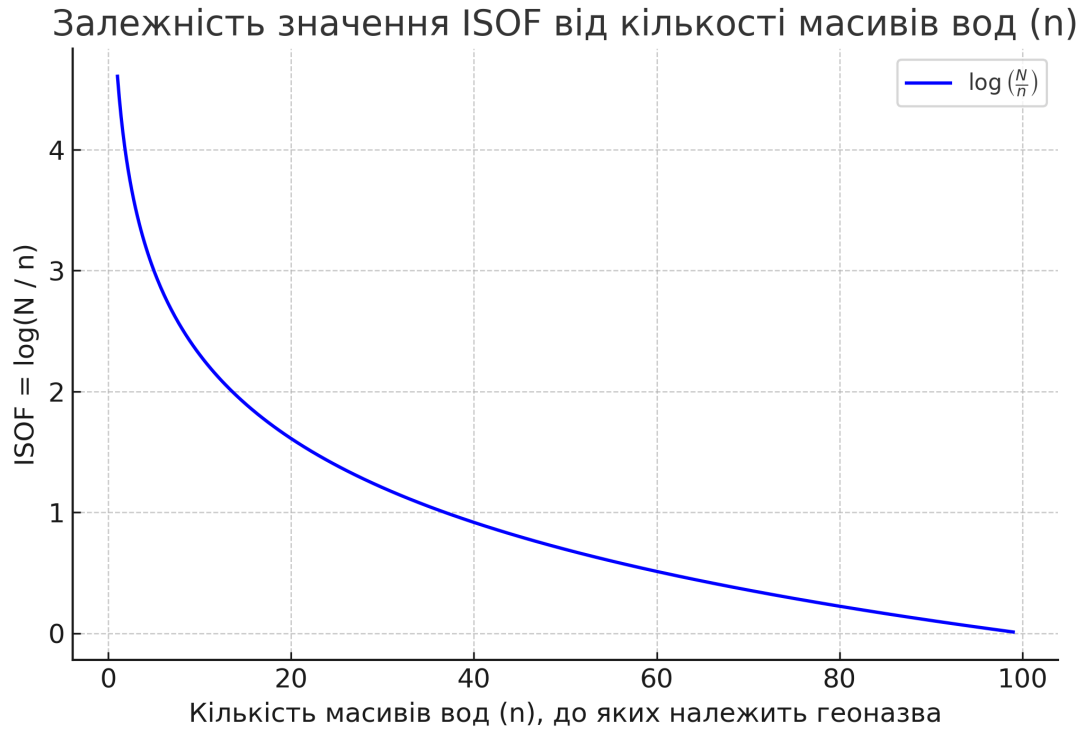


Рисунок 2.3 – Залежність *ISOF* від кількості МВ, до яких належить геоназва

Визначимо зв'язок між текстом та масивом вод, обчисливши інтегральний показник, що і є аналогом TF-IDF для просторової задачі:

$$MLT - ISOF_{i,j} = MLT_{i,j} \times ISOF_j . \quad (2.13)$$

Отже, алгоритм застосування виглядатиме наступним чином:

1. $G^T = \{G_j | G_j \in T\}, G_j \subseteq G,$
2. $MLT - ISOF_{i,j} = \begin{cases} MLT_{i,j} \times ISOF_j, & G_j \subseteq M_i, \\ 0, & G_j \notin M_i, \end{cases}$
3. $S_i = \sum_{j \in G^T} MLT - ISOF_{i,j},$
4. $M^* = \{M_i | S_i = \max(S_1, S_2, \dots, S_n)\},$

де M^* - множина масивів вод з максимальною оцінкою релевантності, до яких прив'язується текст новин.

Приклад наведено на рис. 2.4. Застосуємо декілька відомих і запропонуємо нові методи на прикладі датасету [77] (рис. 2.4).

```
0 In the Southern Bug basin wastewater is discharged from 341 organized source...
1 This amount is distributed as follows: urban and rural housing wastewater tr...
2 Despite the similar volumes of discharged wastewater major part of pollutant...
3 They bring 84% of organic pollution 86% of phosphate ions and 84% of mineral...
4 The input of the industry is between 7–21% and agriculture has the lowest im...
...
95 Chemical composition of water bodies shall comply with requirements of envir...
96 According to the Directive provisions water bodies can be classified as char...
97 Besides in the field of data processing during the assessment of water statu...
98 Use of these values was offered by the Monitoring Expert Group of the Intern...
99 To determine confidence levels for assessment by general physical and chemic...
Name: text, Length: 100, dtype: object
```

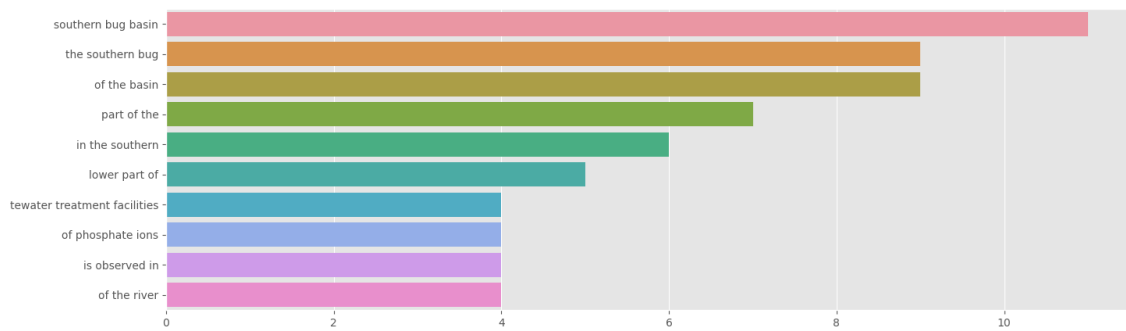


Рисунок 2.6 – Десять найчастіших 3-grams датасету

Оскільки уже використовується *stacking*, можна використати ще *merging* і порівняти отримані результати, приклад такого порівняння наведено на рисунку 2.7.

Одним із найважливіших етапів є формування ключових слів, за якими тексти будуть краще розпізнаватися пошуковими системами. Зобразимо двадцять найважливіших для розглянутої теми на рисунку 2.8.

	keywords	cos_sim	maxsum_sim	mmr_div_0.2_sim	mmr_div_0.7_sim	tf-idf_sim	merging	stacking
0	absolute elevations vary	0.8188	0.8188	0.8188	0.8188	0.488482	0.752736	0.795398
1	thunderstorms begin	0.8397	0.0000	0.8397	0.8397	0.953500	0.694520	0.724694
2	ladyzhyn water storage	0.7406	0.7406	0.7406	0.7406	0.307168	0.653914	0.696541
3	autumn period	0.7138	0.7138	0.7138	0.7138	0.426622	0.656364	0.693532
4	following oblasts vinnytsya	0.7018	0.7018	0.7018	0.7018	0.356229	0.632686	0.671125
5	snow cover height	0.8887	0.0000	0.8887	0.8887	0.364762	0.606172	0.657438
6	dnieper bug estuary	0.8444	0.0000	0.8444	0.8444	0.477817	0.602203	0.646977
7	biggest discharge	0.6469	0.6469	0.6469	0.6469	0.426622	0.602844	0.635329
8	bug estuary	0.7771	0.7771	0.7771	0.0000	0.953500	0.656960	0.628355
9	average snow	0.6468	0.6468	0.6468	0.6468	0.364762	0.590392	0.624725
10	inorganic nitrogen	0.7477	0.5473	0.7477	0.7477	0.000000	0.558080	0.610419
11	southern bug basin	0.7672	0.4710	0.7672	0.7672	0.000000	0.554520	0.608224
12	pollutants containing nitrogen	0.5710	0.5710	0.5710	0.5710	0.518346	0.560469	0.584889
13	vegetation covers	0.5863	0.5863	0.5863	0.5863	0.403158	0.549672	0.578618
14	human extracts	0.5418	0.5418	0.5418	0.5418	0.550342	0.543508	0.564924
15	elevations vary	0.7797	0.7797	0.7797	0.0000	0.488482	0.565516	0.550862
16	bug basin wastewater	0.7015	0.0000	0.7015	0.7015	0.360496	0.492999	0.531289
17	nitrogen ions dominates	0.7127	0.0000	0.7127	0.7127	0.275171	0.482654	0.524288
18	phosphorus belongs	0.6552	0.0000	0.6552	0.6552	0.454352	0.483990	0.516224
19	water bodies eutrofication	0.6531	0.0000	0.6531	0.6531	0.454352	0.482730	0.514817

Рисунок 2.7 – Двадцять найважливіших ключових слів для пошукових систем, використовуючи *merging* [3]

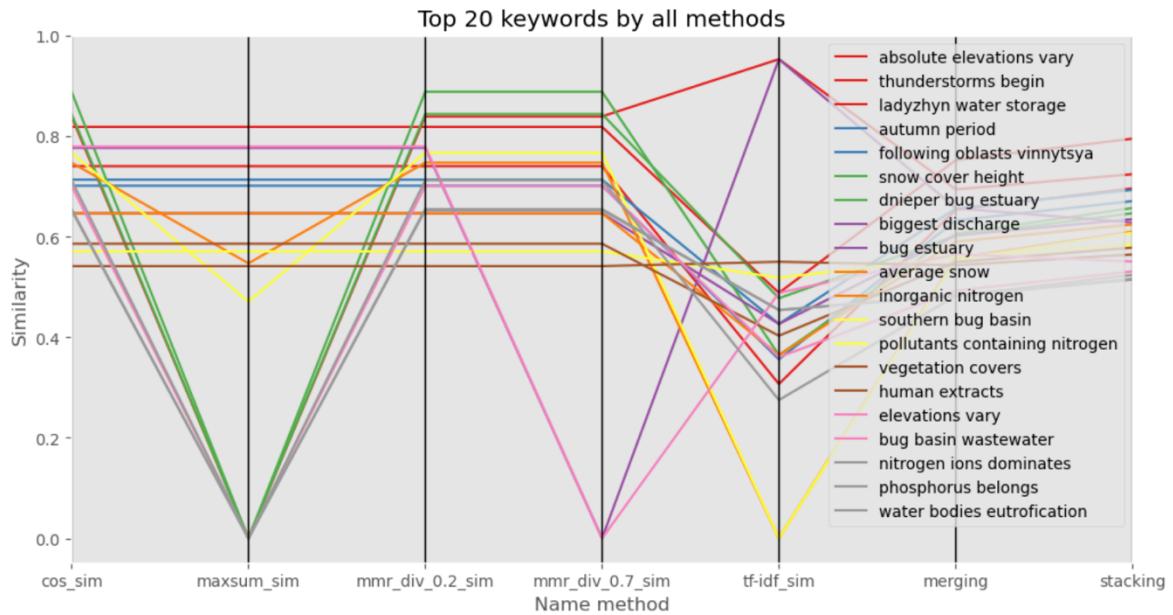


Рисунок 2.8 – Двадцять найважливіших ключових слів
за різними методами [3]

Отже, кожен метод є послідовністю етапів.

Більшість розглянутих методів починають свою роботу однаково, але є деякі розбіжності, які варто ретельно проаналізувати і порівняти. Основні відмінності – на останніх етапах: для статистичних методів (Bag of Words, TF-IDF та RAKE) це — визначення частоти слів словосполучень; для методів TextRank, PositionRank, MultipartiteRank тощо рішення ґрунтується на побудові графа.

У методів на основі BERT спочатку текст перетворюється на систему числових ембедінгів, а потім за одним з алгоритмів з цієї системи відбираються оптимальні ключові слова і фрази.

Проаналізуємо найпоширеніші методи, які розглядаються вище, а саме – особливості та недоліки, які можуть вплинути на точність отриманих результатів оброблення даних про водні масиви та їхні басейни на території України.

У таблиці 2.1 наведено основні особливості та недоліки охарактеризованих вище відомих типових методів видобування ключових слів і фраз із заданого природномовного тексту [78].

Таблиця 2.1. Основні особливості та недоліки існуючих поширених методів видобування ключових слів і фраз із заданого природномовного тексту

Метод	Тип методу	Особливості методу	Недоліки методу
Bag of Words	Статистичний	Представляє текст як множину слів чи фраз, має простий алгоритм, не вимагає спеціальних лінгвістичних знань та розуміння предметної області, порівняно швидкий і простий	Не враховує контекст слів чи фраз, їх порядок у реченні, граматичну структуру та семантичні зв'язки, що призводить до найменш ефективних та релевантних результатів, у порівнянні з іншими методами
TF-IDF	Статистичний	Комплексно враховує і частоту появи слів чи фраз у тексті, і їх рідкість та унікальність.	Ті самі недоліки, але за релевантністю результатів, як правило, вища за метод Bag of Words.
TextRank	Графовий	Використовує графову семантичну модель, яка враховує семантичну близькість заданих слів у тексті в цілому.	Не враховує контексту кожного слова чи фрази у кожному реченні.
PositionRank	Графовий	Застосовує алгоритм TextRank, враховуючи позицію слів і фраз у реченнях для більш точного визначення їх важливості.	Дає кращі результати, ніж TextRank, але також не враховує контекст кожного речення.
KeyBERT	Нейромережевий	Вимірює кут між векторами BERT-ембеддінгів [79] для порівняння подібності між словами і фразами у реченнях.	Може давати надто схожі варіанти оптимальних ключових слів.
KeyBERT з алгоритмом Max Sum Distance	Нейромережевий	Визначає ключові слова на основі міри різноманітності, максимізуючи суму відстаней між різними ключовими словами.	Передбачає різноманітність ключових слів, але, все одно, вони можуть бути або надто подібними, або надто різноманітними і, при цьому, не найбільш релевантними.
KeyBERT з алгоритмом Maximal Marginal Relevance	Нейромережевий	Балансує релевантність та різноманітність ключових слів для вибору оптимальних ключових слів і фраз.	Залежно від параметра «diversity» може давати надто різноманітні і, при цьому, не найбільш релевантні варіанти.

2.3 Удосконалення підходу до часової прив'язки текстів

Визначимо функцію, що задає часовий контекст речення $t(s)$:

$$t(s) \begin{cases} TimeExplicit(s), & \text{if } TimeExplicit(s) \neq 0, \\ Time(h), & \exists h \in H(s) : Time(h) \neq 0, \\ TimeMeta(D), & \text{otherwise.} \end{cases} \quad (2.14)$$

Функція $TimeExplicit(s)$ визначає часовий інтервал зі змісту речення (наприклад, «2019»). Якщо ж така функція не знаходить значення, тоді шукаємо серед заголовків $H(s)$. Якщо серед усіх рівнів заголовків немає вказівки часу, тоді $TimeMeta(D)$ – часовий інтервал, пов'язаний з цілим документом (наприклад, рік публікації звіту). У практичних реалізаціях можна вдосконалити цю функцію, щоб вона враховувала різні формати дат і обирала найбільш специфічний час (якщо є рік і місяць – брати місяць, якщо є день – брати його). Для поставленої задачі достатньо хоча б прив'язки фактів до конкретного року.

Прив'язка у часі може бути й у метаданих документа A (з англ. «About»), тоді, за наявності метаданих $At(s)$ з опису документа (наприклад, рік звіту), формула спрощується до:

$$T(s) = \begin{cases} TimeExplicit(s), & \text{if specified explicitly,} \\ At(s), & \text{if specified metadata,} \\ \emptyset, & \text{otherwise.} \end{cases} \quad (2.15)$$

Цей крок підвищує точність $Pr(s \rightarrow t)$, тим самим збільшуючи $L(B)$.

Наступним кроком є формування множини E_f за формулою (2.16), яка містить тільки такі речення, які вдалось прив'язати до хоча б однієї географічної назви із G та хоча б до одного періоду часу із T :

$$E_f = \{s \in E \mid T(s) \neq \emptyset \wedge T(s) \cap P \neq \emptyset\}. \quad (2.16)$$

Це обмеження зменшує $FP(B)$, залишаючи лише речення з чіткими предикатами й часовими прив'язками.

Зробимо прив'язку R_w географічних назв як ключових словосполучень у реченні s до подібних слів в описі масиву вод з використанням Жаккар-міри за допомогою формул (2.17), (2.18):

$$R_w = \{(s, w) \in E_f \times W \mid Jaccard(s, w) \geq \theta\}, \theta \in [0, 1], \quad (2.17)$$

$$Jaccard(s, w) = \frac{|G(s) \cap G(w)|}{|G(s) \cup G(w)|}. \quad (2.18)$$

Цей крок безпосередньо зменшує $FP(B)$, фільтруючи хибні просторові прив'язки.

Побудуємо триплети «Факт–Масив вод–Інтервал часу» за формулою (2.19).

$$B \subseteq E_f \times W \times T. \quad (2.19)$$

Важливо, що зв'язки включаються до B лише, якщо $T(s) \in t$ та $(s, w) \in R_w$.

Проведемо видалення дублікатів та об'єднання фактів щодо однакових масивів вод за однаковий інтервал часу, для чого зведемо однакові (w, t) в одне значення та об'єднаємо всі відповідні речення s у поле $Text$.

Формуємо фінальне відношення R_Σ за допомогою нижче описаної формули (2.20):

$$R_\Sigma \subseteq W \times T \times Text, \text{ де } Text(w, t) = \bigcup_{\substack{s \in E_f \\ (s, w) \in R_w \\ (s, t) \in R_T}} s. \quad (2.20)$$

Отже, на першому етапі аналізується та обробляється вхідний корпус документів, які містять екологічну інформацію. Важливо, що документи мають певне форматування: заголовки, виділення, відступи та порожні рядки. За допомогою спеціального парсера алгоритм виконує структурний аналіз документу. Виявляються заголовки різних рівнів, короткі ізольовані рядки, та логічно відокремлені блоки тексту. У результаті цього етапу документи перетворюються на багаторівневий JSON-об'єкт.

Текст розбивається на окремі речення, які передаються на фільтрацію. За допомогою LLM-аналізу обираються лише ті речення, які стосуються стану або забруднення вод – формуються множина E . Нові тексти S_n обробляються шляхом перетворення на ембедінги $Emb(s_i) \in R^d$ та індексації за допомогою бібліотеки FAISS. Завдяки цьому та за рахунок семантичного порівняння з новими фрагментами забезпечується швидкий семантичний пошук фактів у майбутньому.

За допомогою LLM в тексті знаходяться всі географічні назви з довідника G . Алгоритм визначає, чи є у реченні пряма вказівка на часовий інтервал. Якщо прив'язка відсутня – використовується дата із заголовка або метаданих документу. Всі прив'язки формалізуються у вигляді множини T .

Створюються дві ключові бінарні матриці – матриця «речення-географічна назва» EG та матриця «речення-час» ET . Враховується ієрархічне наслідування від заголовків. Для кожної знайденої географічної назви визначається, до яких масивів вод W вона належить. Враховується, що одна назва може відповідати кільком масивам і один масив може бути пов'язаний із кількома назвами. Результат очищується від дублювання та подається у структурованому вигляді для подальшого аналізу.

2.4 Розроблення методу аугментації даних про водні об'єкти

У статті [1] була розглянута проблема збирання, систематизації та класифікації інформації про стан масивів вод басейнів річок, яка є необхідною для розроблення, застосування та оцінювання ефективності управлінських заходів,

спрямованих на поліпшення чи стабілізацію доброго екологічного стану вод, відповідно до Водної Рамкової Директиви ЄС та Водного Кодексу України [28].

На виконання цієї директиви та законодавства України розроблені плани управління річковими басейнами, після чого слід з'ясувати стан вод в усіх масивах вод і для кожного з них проаналізувати ефективність плану заходів та обґрунтувати можливі зміни [4, 80].

А для цього потрібно вміти класифікувати результати спостережень про стан цих масивів вод, що можливо з використанням інтелектуальної моделі розробленої за допомогою типових методів машинного навчання. Створення такої моделі потребує якісного датасету, однак, як було зазначено у п.1.1, кількість доступних даних спостережень є недостатньо, оскільки постами спостереження охоплено приблизно 5% масивів. Тому цінним є збільшення розмірів тренувального датасету для створення такої інтелектуальної моделі. Для вирішення таких задач можна використовувати методи аугментації.

Отже, має місце задача аугментації (нарощування кількості) вже розмічених вручну україномовних природномовних текстів про стан масивів вод (МВ).

Зазвичай, задачу аугментації україномовних текстів на Python розв'язують з використанням бібліотек NLTK (Natural Language Toolkit), nlpaug, TextAugment, OpenNMT-py, FastText, GPT, BERT, XLNet та ін. з використанням заміни слів на синоніми чи антоніми, перетворення регістру, вставлення слів, зміни порядку слів, заміни слів з помилками або додавання "шуму" (помилки, помилкові символи, дублювання тощо), заміни іменованих сутностей, синтезу нових речень чи анутовання (саммаризації) тексту тощо [9]. Однак, такі підходи лише урізноманітнюють текст, але не додають нової інформації про кожний масив вод. Крім того, їх слід застосувати з обережністю, наприклад, заміна іменованих сутностей чи заміна слів на антоніми може призвести до отримання фейкової інформації [85].

Для вирішення такої проблеми варто використати той факт, що запропонована у [4] технологія дозволяє здійснювати геоприв'язку інформації до іменованих сутностей, зокрема до масивів вод, до населених пунктів, до

адміністративних областей чи районів країни, для яких, у свою чергу, можуть бути відомими певні закономірності в додаткових дослідженнях про географічні та екологічні особливості цих регіонів.

Розробимо метод автоматичної аугментації природномовних україномовних текстів про стан масивів вод заданого району річкового басейну із забезпеченням верифікації нових даних щодо дослідження іменованих сутностей у цих текстах.

Для розв’язання поставленої задачі для масивів вод як елементів БГСІС, описаних у підрозд. 1.1, пропонуються такі етапи:

1. Створити БГС іменованих сутностей (БГСІС) та формалізувати у них зв’язки між різними іменованими сутностями.
2. Пов’язати кожен іменовану сутність цих БГСІС з кожним масивом вод заданого району річкового басейну.
3. Природномовні україномовні тексти про стан вод прив’язувати до хоча б одного елемента БГСІС. Якщо текст прив’язати не вдається, він відкидається.
4. Здійснювати аугментацію, використовуючи закономірності БГСІС — адаптування текстів про одні масиви вод — до текстів про інші.

Наприклад для гідрографічної мережі річок і водойм (БГСІС ГМР) прикладом застосування такої концепції є наступні дії:

1. Створюється направлена геометрична мережа річок і водойм як БГС з урахуванням басейнового принципу, з урахуванням напрямку руху води від витoku до гирла. Трансформується у БГСІС ГМР [87].
2. Кожна іменована сутність БГСІС ГМР пов’язується з кожним масивом вод, що для цього виду БГСІС є доволі простим, оскільки усі масиви вод створені саме як ділянки ГМР.
3. Щодо прив’язування текстів до елементів БГСІС ГМР, то, наприклад, текст типу «річка рівнинного типу з повільною течією» можна прив’язувати одразу до цілої річки, яку потім, через п.2, можна прив’язати й до усіх масивів вод на цій річці.

4. Використовуючи басейновий принцип, інформацію про антропогенне забруднення масивів вод вище по течії можна, за певних умов, автоматично поширити (аугментувати) на масиви вод, розташовані нижче по течії.

Аналогічно можна діяти і щодо екологічної мережі БГСІС ЕМ. Наприклад, національні природні парки «Кармелюкове Поділля» та «Бузький Гард» розташовані у водозбірній зоні декількох масивів вод, а тоді деякі закономірності одних із них, за певних умов, можна поширювати і на інші.

Багато закономірностей відомі одразу на всю адміністративну область, чи район, чи територіальну громаду, а тоді, у першому наближенні, їх можна поширювати на усі масиви вод цих регіонів.

Аналогічно можна поширювати проблеми межені (маловоддя) чи паводків масивів вод, розташованих вище по течії водогосподарської ділянки (ВГД), на масиви вод, розташовані нижче по течії (межі ВГД, як правило, встановлюються у місцях розташування станцій державної системи моніторингу витрат води, де є максимально достовірна інформація про складові водогосподарського балансу) [88].

Звичайно, такий спосіб аугментації не гарантує повної достовірності фактів. Наприклад, на річці можуть бути масиви вод з більшою глибиною, порогами (відповідно, з різним ступенем аерації), швидкою течією, антропогенними впливами, зарегульованістю тощо, через що їх і поділяли на різні масиви вод, а отже, не усі такі факти можна поширювати автоматично. А тому на першому етапі такі підходи з аугментації потребують експертної перевірки, чи дійсно отримана аугментована інформація є коректною. У разі виявлення невідповідності, слід вносити корективи в алгоритм аугментації та формалізацію закономірностей в БГСІС.

Для віднесення інформації до БГСІС слід передбачити базову ручну розмітку даних, на якій натренувати моделі-класифікатори [11]. Враховуючи структуру планів управління річковим басейном та наявну інформацію про антропогенні джерела забруднення, природно-заповідні об'єкти тощо, дану

технологію варто застосовувати до такого виду класів, по яких інформацію важко зібрати точно по усіх масивах вод басейнів великих річок:

1. «стан» (англ. «status») — всі факти, які характеризують екологічний та хімічний стан вод, її якість, рівні і витрати води, гідроморфологічні показники водних об'єктів у заданому МВ в минулому, в цей момент і прогнози щодо майбутнього стану — як правило, прив'язуються до БГСІС ГМР, іноді і до БГСІС МАУ (часто програми моніторингу вод складаються по областях);

2. «проблеми» («issues») — всі факти, які вже призвели чи можуть призвести до погіршення стану заданого МВ (антропогенні та природні фактори), у т. ч. можливі ризики, точкові та дифузійні джерела забруднень води антропогенного і природного походження, надмірна зарегульованість водних об'єктів, ризик частих паводків чи межені тощо — як правило, прив'язуються до БГСІС ГМР, але можуть прив'язуватись і до БГСІС МВГД, а облік зворотних і стічних вод ведеться по областях, тому актуальною є прив'язка і до БГСІС МАУ;

3. «заходи» («measures») — будь-які природоохоронні заходи, системи очищення вод, спрямовані на поліпшення стану заданого МВ — позитивний і негативний досвід, проекти планів заходів ПУРБ тощо — можуть прив'язуватись до БГСІС МАУ (як правило, ресурси виділяють за адміністративним принципом) або до БГСІС ГМР чи БГСІС МВГД (в самих адміністративних регіонах ресурси виділяють й за басейновим принципом теж);

4. «баланс» («balance») — будь-яка інформація про водогосподарський баланс масиву вод та проблеми, пов'язані з ним (повені, паводки, межень, опади, пересихання колодязів тощо) — цю інформацію можна прив'язувати до БГСІС МВГД.

Важливо відзначити, що до класу «стан» пропонується відносити лише конкретні кількісні оцінки, бажано з розмірністю, які можна виміряти чи оцінити за певними методиками, наприклад, значення концентрацій забруднювальних речовин у воді [32], визначені фахівцями. А якісні оцінки чи загальні факти на кшталт наслідків у вигляді загибелі риби, цвітіння води тощо пропонується відносити до «проблеми», а не до класу «стан», тоді моделі штучного інтелекту

легше буде їх класифікувати. Водночас, пропонується, у першому наближенні, не подрібнювати клас «проблеми» на підкласи (вплив на стан «*impact*», наслідки поганого стану «*pressure*» для довкілля та людей в термінах відомої в ЄС DPSIR-моделі [12]), оскільки, по-перше, моделі важко бути їх розрізнити (експертам під час ручної розмітки коротких речень теж іноді важко зрозуміти що це – причина чи наслідок?), по-друге, даних може бути замало для надійної класифікації, а по-третє, для поставленої задачі не має особливого значення причинно-наслідковий характер проблеми, головним є – наявність такої проблеми і постановка задачі необхідності її розв'язання.

Для формалізації моделі БГСІС G у загальному вигляді запропонуємо нотацію (2.21), основану на нотації направлених зважених графів:

$$G = \{R\{V, P\}, S(V), W(V, P)\}, \quad (2.21)$$

де $R\{V, P\}$ — графова модель, яка складається із двох множин — множини V вершин (вузлів), які є просторовими об'єктами та, одночасно — іменованими сутностями і самі, у свою чергу, можуть бути складеними об'єктами з точкових, лінійних та площинних об'єктів, та із множини P зв'язків (ребер), які формалізують наявність певного відношення між двома і більше вершинами;

(V) — матриця суміжності для множини V вершин (цей параметр потрібний, якщо граф є направленим);

$W(V, P)$ — матриця ваг між множиною V вершин та множиною P ребер (цей параметр потрібен, якщо у графі слід врахувати зменшення сили зв'язку з відстанню).

Модель (2.21) є узагальненою і може враховувати і множину усіх об'єктів БГСІС (вершини), і зв'язки чи відношення між ними (ребра), і напрями та сили впливу цих зв'язків. Наприклад, для БГСІС ГМР-модель матиме саме такий вигляд, оскільки цей граф є і направленим (у напрямку течії води), і зваженим, оскільки забруднення з верхніх по течії ділянок вод впливає на нижні, затухаючи з відстанню, завдяки самоочисним процесам річки. А модель БГСІС МАУ для

адміністративних районів може бути більш спрощеною, оскільки не буде враховувати ні напряму, ні зміни сили впливу між ними, якщо не буде враховуватися відносно розташування цих районів на певних річках.

З використанням ГІС-технологій [89] по попередній формулі (2.21) нескладно знайти усі варіанти відповідності масивів вод, як лінійних (ділянки річок) та площинних об'єктів (водозбірна площа кожного масиву), із множини M та усіх іменованих сутностей із множини V , використовуючи формулу (2.22):

$$\{m\} = \Phi(\{v\}, G), \{v\} = \Psi(\{m\}, G), \{m\} \subseteq M, \{v\} \subseteq V, \quad (2.22)$$

де $\Phi(\{v\}, G)$ — усі правила та відношення, які встановлюються моделлю БГСІС G з (2.21) і дозволяють для заданої підмножини кодів іменованих сутностей $\{v\}$ із множини V визначити підмножину кодів $\{m\}$ масивів вод із множини M ;

$\Psi(\{m\}, G)$ — усі правила та відношення, які встановлюються моделлю БГСІС G з (2.21) і дозволяють для заданої підмножини кодів $\{m\}$ масивів вод із множини M знайти підмножину $\{v\}$ кодів іменованих сутностей із множини V .

У багатьох випадках згадані підмножини можуть бути й одним елементом, але в загальному випадку, наприклад, в одній адміністративній області може бути декілька масивів вод, а у водозбірній площі одного масиву вод можуть бути розташовані багато населених пунктів чи інших іменованих сутностей.

Наприклад, оверлейний ГІС-аналіз за моделлю БГСІС дозволяє легко визначити, в яких адміністративних областях чи районах розташований заданий масив вод і навпаки, які масиви вод розташовані у заданій області чи районі.

Імпортувавши назви іменованих сутностей (вони можуть бути основні та альтернативні як для річок, можуть бути різними мовами або містити попередні назви до перейменування як для населених пунктів та ін.) з атрибутивних даних іменованих сутностей можна сформувану множину усіх назв іменованих сутностей по їх кодах у моделі (2.23):

$$N = \{A(v, G)\}, v \in V, \quad (2.23)$$

де $A(v, G)$ — функція видобування списку усіх можливих назв (можливо, й багатослівних) з атрибутів іменованої сутності v по її коду у БГСІС G .

Але на практиці більш цінною є функція $\chi(*)$ знаходження кодів іменованих сутностей $\{v\}$ по їх назвах $\{n\}$ [4]:

$$\{v\} = \chi(\{n\}, G), \{n\}N. \quad (2.24)$$

Для автоматичного геоприв'язування природномовних україномовних текстів про стан водних об'єктів до іменованих сутностей БГСІС пропонується скористатись алгоритмом, згідно якого геоприв'язування текстів здійснюється у 2 етапи:

Етап 1. Готуються необхідні довідники (списки назв іменованих сутностей):

1) формується максимально велика множина U назв іменованих сутностей (об'єкти на карті, довідники річок, дані земельного кадастру, дані державної статистичної звітності, інформація із різних звітів та довідників тощо), про частину з яких може й не бути відомо просторові координати; з цієї множини виділяється підмножина U_1 з відносно більш великими за площею чи довжиною сутностями (область, район, річка), про які точно відома відповідність масивам вод, інші — відносяться до множини U_2 ;

2) формуються множини X_1 та X_2 з усіма словоформами назв іменованих сутностей із обох підмножин U_1 та U_2 , відповідно.

Етап 2. Застосовується ітеративний алгоритм, який на кожному кроці $i = 0, 1, \dots$, одночасно здійснює і геоприв'язування тексту T_i до масиву(ів) вод, і уточнення моделі G_i з (2.21):

1) у тексті T_i знаходяться назви X_1^T іменованих сутностей із множини X_1 , потім по (2.24) встановлюються їх коди $\{v\}$, а потім по першому із співвідношень (2.22) визначається до яких кодів масивів вод Y_{Ti} їх можна прив'язати;

2) у тому ж тексті T_i знаходяться назви іменованих сутностей із множини U_2 , за (4) визначаються їх коди V_2^T , а тоді щодо них висувається гіпотеза про те,

що вони теж відносяться до масивів вод Y_{Ti} , визначених на попередньому кроці, і це нове правило додається в G_{i+1} та співвідношення (2.22)—(2.24).

Пропонується удосконалити цю технологію, замінивши на другому кроці етапу 1 утворення словоформ на лематизацію (з утворенням множин X_1^* та X_2^* відповідно) і додавши лематизацію до усіх слів текстів T_i . Практичні випробування довели більшу ефективність і надійність результатів з використанням такого удосконалення [90].

Крім того, кроки етапу 1 пропонується доповнити таким чином: по-перше, на карті [91] (2.9) знаходяться об'єкти, які відповідають усім елементам множини U , а якщо їх точне місцезнаходження — поки невідоме, тоді задаються довільні координати (наприклад, в центрі мапи чи в одному з її кутів) і там формується точковий об'єкт певного шару, наприклад «Об'єкти».

Коли на етапі 2 стає відомою прив'язка до певного масиву вод, тоді відповідний об'єкт можна перемістити в точку, яка означає кінцевий створ масиву вод, до тих пір, поки не буде уточнено його справжнє місцезнаходження (варіант розташування в геометричному центрі масиву вод не є кращим, оскільки деякі річки так мандрують, що такий центр може відноситись до іншого масиву вод). По-друге, на першому кроці пропонується відносити до підмножини U_1 усі точно геоприв'язані об'єкти, а не тільки великі.

З урахуванням наведеної вище формалізації запропоноване удосконалення алгоритму можна записати у вигляді формули (2.25):

$$\begin{aligned} X_1^* &= L(U_1), X_2^* = L(U_2), U_1 \cup U_2 = U, \\ X_1^T &= \Omega(L(T_i), X_1^*), \\ V_2^T &= \chi(\Omega(L(T_i), X_2^*), G_i), \\ G_{i+1} &= \Lambda_i(V_2^T, Y_{Ti}, G_i), i=0,1,..., \end{aligned} \quad (2.25)$$

де $L(x)$ — функція лематизації усіх слів тексту x ; функція, які видобуває лематизовані іменовані сутності із множини u у тексті (в загальному випадку, допускаються багатомовні іменовані сутності зі збереженням їх комбінацій, тобто

словосполучення «Вінницький область» (лематизація сутності із назвою «Вінницька область») буде зберігатись саме у такому вигляді, а не як два окремих послідовних слова «Вінницький» і «область»); $\Lambda_i(V_2^T, Y_{Ti}, G_i)$ — функція, яка на i -му кроці додає до графу G_i з (2.21) відповідність ідентифікованих по іменованих сутностях з множини U_1 кодів масивів вод Y_{Ti} , та знайдених у тексті T_i кодів V_2^T іменованих сутностях, утворюючи оновлений граф G_{i+1} .

У загальному випадку, множини U_1, U_2 теж можуть зазнавати оновлення шляхом перенесення вже геоприв'язаних сутностей U_2 з до U_1 .

Найскладнішим елементом цієї технології є функція $\Phi(*)$ з (2.22), оскільки реалізація інших функцій (функції $\chi(*)$ знаходження кодів іменованих сутностей по їх назвах, функції лематизації $L(*)$, функції $\Omega(*)$ пошуку лематизованих іменованих сутностей серед лематизованих слів заданого тексту) є доволі очевидною.

На рис. 2.9а наведено приклад таблиці бази даних, яка може бути згенерована по моделі G_i для усіх комбінацій кодів масивів вод та кодів іменованих сутностей для реалізації функції $\Phi(*)$, а на рис. 2.9б — для цих же кодів іменованих функцій та їх назв для реалізації функції $\chi(*)$.

Приклади фрагментів звичайних таблиць бази даних для формалізації правил наведено на рисунку 2.9 [4, рис. 2].

code_SWB	code_NE
UA_M5.4_0186	1
UA_M5.4_0186	2
UA_M5.4_0186	3
UA_M5.4_0186	4

а)

code_NE	nameua_new
1	с. Березина
2	с. Зарванці
3	с. Якушинці
4	м. Вінниця

б)

Рисунок 2.9 – Приклад фрагментів звичайних таблиць бази даних для формалізації правил для формування вибірок даних для реалізації функцій моделі (2.25): а) для функції $\Phi(*)$; б) для функції $\chi(*)$ [4, рис. 2]

Проте, така формалізація може призводити до зовеликої кількості хибних спрацювань.

Наприклад, текст «У Вінницькій області сталося забруднення р. Південний Буг нітратами біля водозабору м. Вінниці» може призвести до того, що це речення буде прив'язано до усіх масивів вод Вінницької області, р. Південний Буг та м. Вінниці, хоча, насправді, йдеться явно про один масив вод UA_M5.4_0013, розташований на р. Південний Буг біля водозабору м. Вінниця.

Для точнішого прив'язування релевантних даних до масивів вод рекомендується змінити модель даних (2.21), (2.22) і зберігати не всі відповідності за принципом «всі до всіх», а, натомість, для кожного масиву вод за ГІС синтезувати перелік назв у вигляді списку усіх іменованих сутностей, зокрема з множини U_1 , тобто — тих, геоприв'язка яких до цього масиву вод точно відома.

На рисунку 2.10 наведемо приклад списку назв іменованих сутностей, про які зазначено вище [4, рис. 3].

code	geo_names
UA_M5.4_0002	Південний Буг, с. Чорний острів, м. Хмельницький, Хмельницька область, Хмельниччина, Хмельницький район
UA_M5.4_0003	Південний Буг, Хмельницьке водосховище, м. Хмельницький, мікрорайон Гречани, Хмельницька область, Хмельниччина, Хмельницький район
UA_M5.4_0004	Південний Буг, м. Хмельницький, с. Копистін, с. Богданівці, с. Прибузьке, скид Хмельницькводоканалу, Хмельницька область, Хмельниччина, Хмельницький район
UA_M5.4_0011	Південний Буг, м. Хмільник, Сандракське водосховище, с. Березна, с. Крутнів, с. Лелітка, с. Соколова, с. Широка Гребля, с. Володьки, с. Вербівка, Вінницька область, Вінниччина, Хмільницький район, Хмільникводоканал, "Старий замок"
UA_M5.4_0013	Південний Буг, м. Вінниця, с. Стрижавка, с. Гушинці, с. Мизяків, с. Медвідка, с. Лаврівка, водозабір Вінницяводоканалу, Київський міст, Центральний міст, Староміський міст, "Вінницька реберна", пляж "Хімік", пляж "Кумбари", пляж "Спартак", пляж "Гонти", мікрорайон "Старе місто", набережна "Рошен", місце впадіння р. Тяжилівка, Десна, Периорка, Вишня, Згар, Постолова, Вінницька область, Вінниччина, Хмільницький район, Вінницький район, мостовий шляхопровід автодороги "Вінниця – Калинівка"
UA_M5.4_0152	р. Десна, с. Стрижавка, с. Славне, с. Сосонка, Вінницька область, Вінниччина, Вінницький район

Рисунок 2.10 – Приклад списків назв іменованих сутностей декількох масивів вод [4, рис. 3]

Така формалізація і встановлення відповідності назв іменованих сутностей та кодів масивів вод, дозволить замінити функції (2.22)—(2.24) однією функцією (2.26):

$$\{m\} = \Phi(\chi(\{u\}, G), G) = \theta(\{u\}, G), \{m\} \subseteq M, \{u\} \subseteq U, \quad (2.26)$$

де $\theta(\{u\}, G)$ функція буде здійснювати пошук найкращих збігів списків словосполучень, які характеризують кожен масив вод у довіднику як на рис. 2.10, та у заданому списку $\{u\}$ іменованих сутностей із множини U (як правило, для цього використовується косинусна подібність `cosinus_similarity`, яка набуває значень від 1 (повний збіг) до 0 (повний незбіг), також можливі від’ємні значення, де -1 означає протилежність). А це, у свою чергу, дозволить спростити модель (2.25):

Запишемо разом у формулі (2.27).

$$\begin{aligned} G_{i+1} &= \Lambda_i(V_2^T, Y_{Ti}, G_i), \quad i = 0, 1, \dots, \\ V_2^T &= \chi(\Omega(L(T_i), X_2^*), G_i), \quad Y_{Ti} = \Theta(X_1^T, G), \quad X_1^T = \Omega(L(T_i), X_1^*); \\ X_1^* &= L(U_1); \quad X_2^* = L(U_2); \quad U_1 \cup U_2 = U. \end{aligned} \quad (2.27)$$

У лаконічнішому вигляді нову модель (2.27) можна записати у вигляді формули (2.28):

$$G_{i+1} = \Lambda_i(\chi(\Omega(L(T_i), L(U_2)), G_i), \Theta(\Omega(L(T_i), L(U_1)), G_i), G_i), \quad i = 0, 1, \dots \quad (2.28)$$

Проведено експерименти у програмах на Python для вивчення можливості автоматизації порівняння україномовних текстів за моделлю (2.27) з використанням бібліотеки `SpaCy`. Як видно з рис. 2.10, по-перше, найбільше значення метрики «`cosinus_similarity`» дійсно відповідає найбільш релевантному

варіанту, а по-друге, видно, що лематизація речення підвищує значення цієї метрики майже у 2 рази, що збільшує ймовірність вибору найбільш правильного варіанту [92].

На рисунку 2.11 зображено результат порівняння за метрикою `cosinus_similarity` [4, рис.4].

target	test_sentence	similarity_score
Вінницька область, р. Південний Буг	у вінницький область статися забруднення південний буг	0.376
Полтавська область, р. Південний Буг	у вінницький область статися забруднення південний буг	0.36
Вінницька область, р. Соб	у вінницький область статися забруднення південний буг	0.204
Вінницька область, р. Південний Буг	У Вінницькій області сталось забруднення Південного Бугу	0.2
Полтавська область, р. Південний Буг	У Вінницькій області сталось забруднення Південного Бугу	0.181
Вінницька область, р. Соб	У Вінницькій області сталось забруднення Південного Бугу	0.113

Рисунок 2.11 – Результат порівняння за метрикою `cosinus_similarity` україномовних текстів з використанням методів бібліотеки SpaCy та мовної моделі «uk_core_news_sm» [4, рис.4]

Щоб перевірити, наскільки інформація є достовірною, використаємо статистичний метод [93]. Рівень достовірності буде зростати з кожним підтвердженням якогось факту. Якщо перевірятимемо кількісну оцінку, то буде зростати індекс достовірності. Але тут виникає недолік у роботі – неможливо знати, який прилад було використано для моніторингу стану водних об'єктів. Наприклад, сучасні вимірювальні засоби дають точніші результати. Якщо прилад застарілий (що є типовим при недостатньому фінансуванні) або вийшов з ладу (також не завжди можна швидко це виявити), то результати можуть в десять разів відрізнятись від істини.

Щоб оцінити достовірність даних будь-якого типу, можна пробувати комбінувати числа і текстову інформацію в різних співвідношеннях.

Розглянемо, як зміниться результативність, якщо використати багатозв'язність систем. Оскільки система збиратиме характеристики водних джерел, деякі дані можуть бути невідомими. У такому разі можна вичислити приблизні параметри за рахунок оцінювання подібних фактів, використовуючи математичні та інтелектуальні моделі. Для кращого розуміння експерт може

відобразити зв'язки графіком, в якому вершинами будуть змінні, а ребрами – залежності між ними. Отримані результати не є однозначними, верифікація здійснюється шляхом приблизних оцінок у різних комбінаціях ознак [63]. Як такі характеристики пропонується використати топологічну спостережуваність [22].

Як відомо, топологічна спостережуваність визначається на базі біхроматичного графа [22]. Біхроматичний граф, це – граф, що складається з двох типів вузлів: вузлів змінних $y_i (i = \overline{1, 5})$ та вузлів, що представляють залежності між такими змінними $w_j (j = \overline{1, 4})$. На рисунку 2.12 представлено приклад такого графа та його максимальне паросполучення (ребра, що належать до множини збігів, виділені). На прикладі добре видно, що існує декілька розв'язків задачі вибору максимального паросполучення: наприклад, замість ребра $(y_3 - w_3)$ може бути ребро $(y_4 - w_3)$.

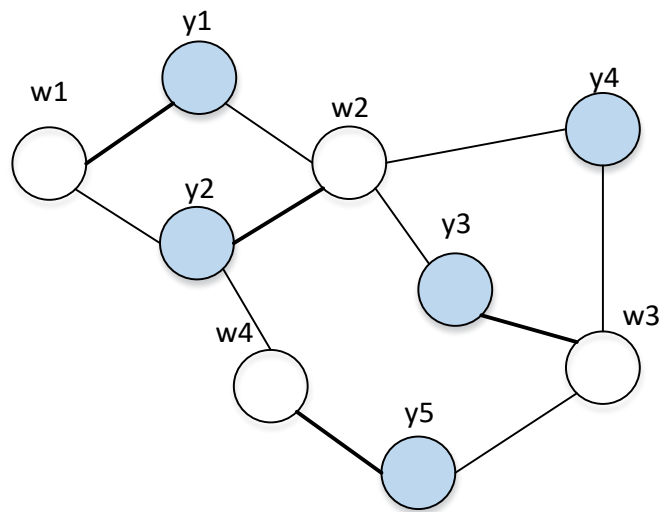


Рисунок 2.12 – Сильні ребра у біхроматичному графі для визначення топологічної спостережуваності складної системи [6, рис. 1] [94]

За умови, що існує максимальне паросполучення, де кожна змінна, що описує залежність, є інцидентною сильному ребру (тобто ребру, яке не належить до зазначеного паросполучення), то система, що описується системою таких залежностей, є повністю топологічно спостережуваною (рис. 2.12) [22].

Пропонується усю доступну числову інформацію максимально пов'язувати між собою з використанням математичних і інтелектуальних моделей, когнітивних карт та інших видів залежностей між змінними (ознаками). По суті, ця модель буде цифровим двійником заданої багатостадійної ГІС. За методологією, викладеною у роботах, будувати граф цієї системи залежностей, по ньому будувати біхроматичний граф, аналізувати його на топологічну спостережуваність. Якщо вона не є повною, додавати ознаки чи залежності. Але при цьому варто оцінювати достовірність цих залежностей, оскільки деякі з них можуть бути доволі наближеними.

Аналогічно варто оцінювати індекс достовірності числових даних, які фігурують у текстовій інформації (у реченнях звітів, новин, повідомлень тощо у різний час у різних місцях).

Для оцінювання індексу достовірності J кожної новини (факту, елементу інформації у вигляді тексту чи числа) пропонується використовувати метод Байєса, спираючись на результати зіставлення апіорних та апостеріорних ймовірностей того, що знайдений факт наближається до істинного [95]:

$$J_i = \frac{j_i \cdot p_{i-1}}{j_i \cdot p_{i-1} + (1 - j_i) \cdot (1 - p_{i-1})}, i = 1, 2, 3, \dots, N, \quad (2.29)$$

де p_i - апіорна достовірність на кроці i , а j_i - індекс достовірності визначеної новини на кроці i . $i = 1, 2, 3, \dots, N$, де N - кількість співставлень новин.

Перевірка імплементації методу. На рис. 2.13 представлено результат покрокового перерахунку індексу достовірності події згідно формули (2.29): свідчення очевидця ($j = 0,3$), дві згадки у ЗМІ ($j = 0,7$) та офіційне джерело ($j = 0,95$).

```
+unverified      j=0.30 → p=0.3000
+media_reference j=0.70 → p=0.5000
+media_reference j=0.70 → p=0.7000
+official        j=0.95 → p=0.9779
OK: final p=0.9779 ∈ (0.97, 0.99)
```

Рисунок 2.13 – Приклад застосування індексу достовірності по одній події з джерел різних типів

На розглянутому прикладі після 4-х ітерацій підсумковим значенням є $p = 0,9779$, що показує вказує на високу вірогідність розглянутої події.

А тоді відкриваються такі можливості:

- якщо оцінка індексу достовірності великого текстового датасету вища за індекс достовірності цифрового двійника, створеного на основі числової інформації, тоді другий можна уточнювати з її використанням;

- якщо індекс достовірності даних та залежностей щодо цифрового двійника має вищу оцінку індексу достовірності, ніж окремі текстові дані, тоді варто корегувати числові значення у тексті;

- якщо співставлення текстових даних з цифровим двійником виявило, що наявних залежностей недостатньо для порівняння, фактично, це означатиме появу нової вершини (змінної) у цифровому двійнику, для якої не вистачатиме залежностей, а отже його спостережуваність вже не є повною, тоді слід провести його удосконалення та знов забезпечити його повноту, після чого співставлення слід повторити.

Цей метод дозволяє максимізувати індекс достовірності та адекватність і цифрового двійника кожного елемента (стадії, ділянки) багатозв'язної багатостадійної ГІС з ГМ, побудованого на основі числових даних, і текстових даних з числовою інформацією про цей елемент. А потім цифровий двійник та текстові дані можна використати для аналізу та узагальнення інформації, для підтримки прийняття рішень щодо покращення стану цього елемента системи та для інших цілей.

Але, наприклад, новина про річку Соб може стосуватися всіх масивів вод, розташованих у її басейні. Однак, практичні випробування алгоритмів, запропонованих у [2] з метрикою `cosine_similarity` для новин, що описують локальні об'єкти, розташовані лише на одному масиві вод, показали, що вони часто працюють неефективно. Варто знайти кращу метрику та спробувати використати потенціал сучасних інформаційних технологій, заснованих на великих мовних моделях [96].

Після проведення дослідження різноманітних метрик виявлено основну проблему – ключові слова нерівнозначні. Для того, щоб результат був максимально точним під час прив'язування текстових даних про об'єкти, пропонується провести класифікацію ключових слів для конкретної локації з вагою $w = 3$ (наявність пляжу, ресторану на березі водойми, ГЕС, греблі), регіону з вагою $w = 2$ (назва міста, району, наявність водосховища) та глобально з вагою $w = 1$ (відношення до національного парку, адміністративної області тощо). Алгоритм роботи пропонується такий:

1. Генерація ключових слів для масивів вод.
2. Провести класифікацію та визначення ваги w згенерованих ключових слів за допомогою сучасних великих мовних моделей (LLM).
3. Отримання ключових слів з новин за рахунок LLM.
4. Застосування зваженої Jaccard-міри та обчислення ступеня збігу класифікованих ключових слів зі словами із новин (2.30):

$$S = \frac{\sum w_i \cdot |K_{\text{новина}} \cap K_{\text{масив}}|}{|K_{\text{новина}} \cup K_{\text{масив}}|}, \quad (2.30)$$

де w_i – вага ключового слова, а K – множина ключових слів у новинах та в описі масивів вод.

5. Пошук масиву води з максимальним показником Jaccard-міри (або усіх зі значенням, більшим за 0).

Для проведення дослідження обрано приклади масивів вод для гіпотетичних новин з об'єктами у масивах вод, розташованими на річці Південний Буг (рис. 2.14) [16, рис. 1].

Новина1. У Хмельницькому нечистоти потрапили в канал, з'єднаний з Південним Бугом: деталі. Витік нечистот з колектора каналізаційно-насосної станції трапився в мікрорайні Гречани, що в Хмельницькому. Стоки потрапили в канал, що сполучається з Південним Бугом. «Дуже велика площа була забруднена»

Новина2. Біля пляжу "Хімік" у Вінниці у Бугу була помічена велика пляма бруду, яка запливла в очерет біля пляжу і там осіла. Екоінспектори досліджують наслідки і можливі причини цього забруднення.

Новина3 Біля села Новоселівка на Ладизинському водосховищі велика кількість риби у серпні загинули і спливла. Екологи роблять аналізи води та намагаються якось боротись з наслідками цього вкрай негативного явища.

Рисунок 2.14 – Гіпотетичні новини про об'єкти у масивах вод, розташованих на річці Південний Буг [16, рис. 1]

Результат геоприв'язки новин до масивів вод за запропонованим алгоритмом з використанням моделі ChatGPT 4o [16, рис. 2] представлено на рисунку 2.15.

Jaccard-Mіра Для 3 Новин

	code	news_1	news_2	news_3
1	UA_M5.4_0002	0.0	0.0	0.0
2	UA_M5.4_0003	7.6923076923076925	0.0	0.0
3	UA_M5.4_0004	0.0	0.0	0.0
4	UA_M5.4_0011	0.0	0.0	0.0
5	UA_M5.4_0013	0.0	3.0303030303030303	0.0
6	UA_M5.4_0152	0.0	0.0	0.0
7	UA_M5.4_0019	0.0	0.0	3.125
8	UA_M5.4_0507	0.0	0.0	0.0
9	UA_M5.4_0028	0.0	0.0	0.0
10	UA_M5.4_0969	0.0	0.0	0.0
11	UA_M5.4_0970	0.0	0.0	0.0

Рисунок 2.15 – Результат геоприв’язки новин до масивів вод за запропонованим алгоритмом з використанням моделі ChatGPT 4o [16, рис. 2]

З рис. 2.15 видно, що точність отриманих результатів (номери ділянок, які відповідають максимальному значенню міри) дорівнює 1,0. Це означає повну відповідність очікуванням і створює перспективи для подальших досліджень.

Під час класифікації ключових слів та словосполучень за ступенем відношення до глобальних, регіональних, локальних доцільно враховувати різні типові ієрархічні структури, передусім, державні реєстри, наприклад, класифікатор КАТОТТГ (Кодифікатор адміністративно-територіальних одиниць та територій територіальних громад), довідник гідрографічної мережі (річки, їх притоки, відтоки, басейни, природне середовище, тощо), а також наявність певних назв у реєстрах назв підприємств, назв національних парків тощо. Також, варто враховувати те, що ця інформація може бути представлена у вигляді електронних таблиць, баз даних або шарів геоінформаційних систем (ГІС) [74].

2.5 Розроблення методу видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою

В основу запропонованого методу покладено принцип Парето. На першому етапі здійснюється відсів інформації, яка однозначно не відповідає умовам запиту. Наприклад, якщо у тексті згадується 2000 рік, а запит користувача стосується подій «за останні п'ять років», такий документ відкидається. Аналогічно, застосовується просторове фільтрування: якщо у повідомленні йдеться про події у мікрорайоні Гречани, а користувач сформулював запит «що трапилося біля Київського мосту?», то ця інформація визнається нерелевантною, тоді як дані, що стосуються р. Південний Буг залишаються у розгляді.

Таким чином, на першому етапі видаляється інформація, яка гарантовано не відповідає часовим або геопросторовим критеріям. Після цього формується підмножина даних, що потенційно релевантні за датою та місцем події.

На наступному етапі пошук виконується із застосуванням підходу Retrieval-Augmented Generation (RAG), що забезпечує більш гнучке та точне отримання знань.

Розглянемо детальніше збирання інформації.

Для практичної реалізації методу верифікації інформації потрібно завчасно організувати конвеєр видобування первинних даних з гетерогенних джерел.

В сучасних дослідженнях соціальних мереж та порталів новин важливу роль відіграють методи збирання та аналізу даних, зокрема вебскрапінг та постоброблення текстових даних. Вебскрапінг є автоматизованим процесом вилучення даних із вебсайтів, який дозволяє ефективно видобувати усі доступні актуальні дані з відомих чи очікуваних джерел інформації.

У рамках задачі використовують спеціалізовані бібліотеки [97]: BeautifulSoup та Newspaper4k, які забезпечують автоматизоване збирання інформації (рис. 2.16) [20, рис.1].

	domain	keyword	title	link
0	fakty.com.ua	Південний Буг	Південний Буг - Факти ICTV	https://fakty.com.ua/ua/tag/pivdennyj-bug/
1	fakty.com.ua	Південний Буг	Південний Буг стрімко всихає: як рятують річку	https://fakty.com.ua/ua/videos/pivdennyj-bug-s...
2	fakty.com.ua	Південний Буг	Одна з найбільших річок України перетворюється...	https://fakty.com.ua/ua/ukraine/suspilstvo/201...
3	fakty.com.ua	Південний Буг	Битва за Південний Буг: чому обміліла річка та...	https://fakty.com.ua/ua/videos/bytva-za-pivden...
4	fakty.com.ua	Південний Буг	На березі Південного Бугу побудували незаконні...	https://fakty.com.ua/ua/videos/___trashed-3/
5	fakty.com.ua	Південний Буг	Була прив'язана мотузка з каменем: на Миколаїв...	https://fakty.com.ua/ua/proisshestvija/2019080...
6	fakty.com.ua	Південний Буг	топ-5 знахідок археологів у Меджибізькій фортеці	https://fakty.com.ua/ua/ukraine/20210815-kubok...
7	fakty.com.ua	Південний Буг	Мер Миколаєва розповів, коли у місті відновлять...	https://fakty.com.ua/ua/ukraine/20220505-mer-m...
8	fakty.com.ua	Південний Буг	Радонове озеро та рафтинг: скільки коштує відп...	https://fakty.com.ua/ua/videos/radonove-ozero-...

Рисунок 2.16 – Приклад результату первинного автоматизованого вебскрапінгу проблем про стан масивів вод басейну р. Південний Буг з обраних джерел інформації [20, рис.1]

Розроблена у [20] методологія дозволяє здійснити аналіз текстової інформації про проблеми чи успіхи реалізації заходів та прив'язати її до конкретних масивів вод, тобто забезпечити геолокацію цієї текстової інформації (рис. 2.17) [20, рис.1]. Прив'язка забезпечується за ключовими словами, які можна видобути у різний спосіб [15].

Для поширення інформації про одні масиви вод на інші, подібні за параметрами (фактично, для текстової «аугментації» даних) було використано бібліотеку SpaCy, що дозволила збільшити обсяг інформації для аналізу та узагальнення.

Узагальнену по багатьох таких текстах інформацію можна використати для контролю ефективності реалізації програми заходів ПУРБ [98] та для пошуку напрямів їх більш ефективної реалізації, з урахуванням досвіду, отриманого щодо інших масивів вод та басейнів.

Similarity of separate word
 Південний Буг 0.7775
 м. Вінниця 0.3543
 с. Стрижавка 0.2033
 с. Гушинці 0.1407
 с. Мизяків 0.1124
 с. Медвідка 0.2695
 с. Лаврівка 0.2273
 водозабір Вінницяводоканалу 0.4462
 Київський міст 0.7352
 Центральний міст 0.7081
 Староміський міст 0.7604
 "Вінницька реберня" 0.3097
 пляж "Хімік" 0.468
 пляж "Кумбари" 0.2082
 пляж "Спартак" 0.4035
 пляж "Гонти" 0.2933
 мікрорайон "Старе місто" 0.535
 набережна "Рошен" 0.4085
 місце впадіння р. Тяжилівка 0.2195
 Десна 0.2957
 Периорка 0.2149
 Вишня 0.3466
 Згар 0.4248
 Постолова 0.2307
 Вінницька область 0.3992
 Вінничина 0.2057
 Хмільницький район 0.6878
 Вінницький район 0.7491
 мостовий шляхопровід автодороги "Вінниця – Калинівка" 0.5817

Рисунок 2.17 – Приклад геолокаційної прив'язки текстових даних до масивів басейну по ключових словах повідомлення [20, рис.2]

Алгоритм роботи удосконаленого методу для технології видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою наведено на рисунку 2.18.

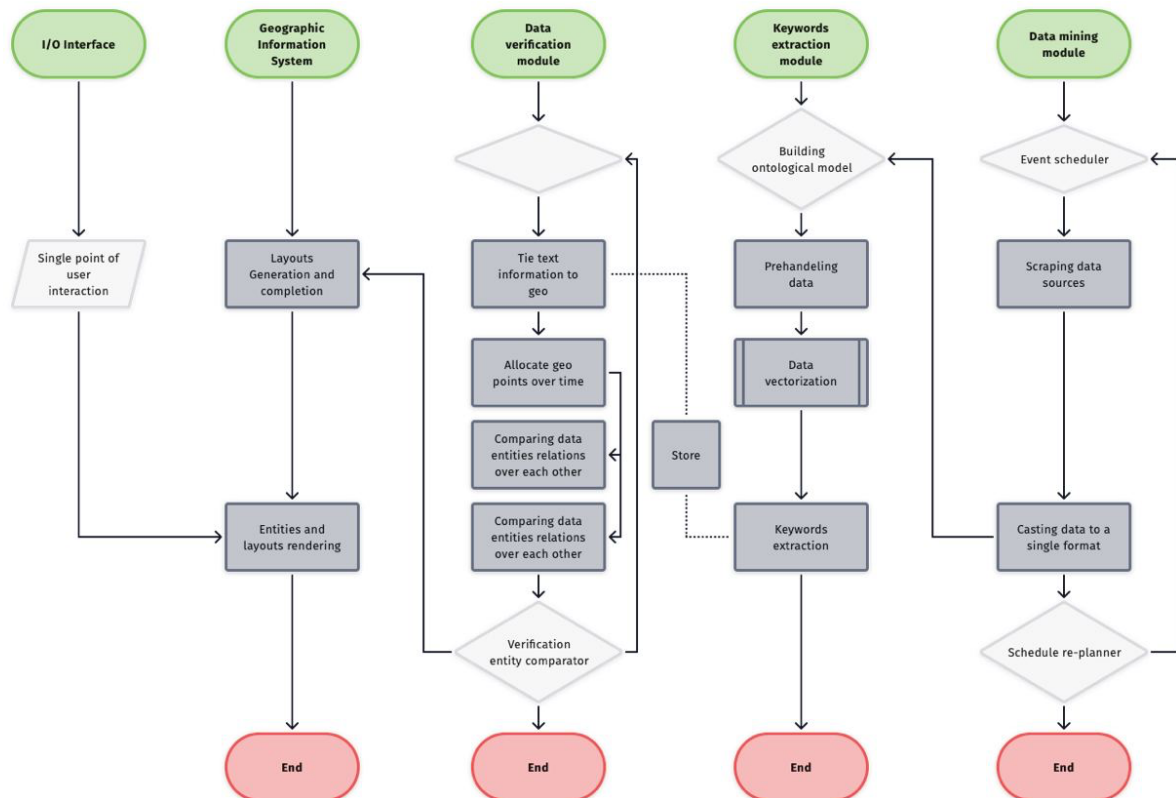


Рисунок 2.18 – Алгоритм роботи інформаційної технології видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою (повний варіант наведено у Додатку Д)

Алгоритм роботи інформаційної технології видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою містить наступні етапи.

1. Збирання вхідних даних — формується корпус неструктурованих екологічних текстів (звіти, новини, поширені в Україні соцмережі) та довідник масивів вод офіційного реєстру;

2. Парсинг документів - документи спершу структурно парсяться: виявляються заголовки різних рівнів і логічні блоки, після чого формується багаторівневий JSON-об'єкт;

3. Попереднє оброблення тексту - далі текст розбивається на речення, з яких LLM відбирає тільки екологічно релевантні (стан/забруднення), а нові речення індексуються ембедінгами та FAISS для семантичного пошуку;

4. Видобування географічних назв - LLM знаходить у реченнях географічні назви (в даному випадку пов'язані з водними ресурсами) з довідника G, а якщо явної дати чи періоду не вказано, час успадковується із заголовків або метаданих зібраних документів;

5. Часова прив'язка - усі часові прив'язки формалізуються в множину T, після чого будуються дві бінарні матриці — "речення–геоназва" (EG) та "речення–час" (ET) із врахуванням ієрархічного спадкування заголовків;

6. Для кожного факту формуються SPO-триплети ("subject–predicate–object"), де subject/object — геоназви з G, а predicate — зміст речення, яке розглядається, і до подальшого аналізу беруться лише триплети з часовою прив'язкою;

7. Фільтрація - формується підмножина речень, яким знайдено і географічну, і часову прив'язку;

8. Прив'язка SPO-триплетів до МВ. Використовуючи відповідність між назвами в G і реєстром масивів вод W, визначається, до яких саме масивів вод належить кожне речення чи триплет;

Метод оптимізує контекстуально-семантичний критерій, який максимізує повноту і ймовірність зв'язків і мінімізує кількість хибнопозитивних. Ключовим методом зіставлення словосполучень з описами МВ є зважена міра Жаккара. Для локальних назв точність міри підвищується за допомогою ваг $w=3$ (локальні), 2 (регіональні), 1 (глобальні), на тестовій виборці досягнуто точність 1,0. Локальність ключових словосполучень доцільно класифікувати за державними реєстрами (наприклад, КАТОТТГ - Кодифікатор адміністративно-територіальних одиниць та територій територіальних громад) та ГІС-шарами. Очікуваним результатом є структурована база знань, у вигляді SPO-триплетів.

Висновки до розділу 2

У другому розділі розроблено та удосконалено комплекс методів видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою.

Формалізовано вимоги до даних та виконано математичну постановку задачі. Введено систему позначень та сформульовано контекстуально-семантичний критерій оптимальності, що одночасно забезпечує пошук максимально великої кількості варіантів зв'язків і мінімізує кількість хибнопозитивних. Завдяки чому було отримано визначення «правильної» геоприв'язки, яке не залежить від конкретного алгоритму, завдяки чому отримали можливість коректно порівнювати різні метрики.

Розроблено метод просторової прив'язки текстів на основі просторово-зваженого критерію MLT-ISOF – адаптацію методу TF-IDF до просторової прив'язки даних. Замість частоти терміну у документі використовується просторова кратність довжини топоніма до лінійної осі масиву вод, а замість кількості документів ISOF представляє кількість просторових об'єктів, на яких зустрічається топонім. Такий підхід знижує вагу для поширених геоназв («річка», «водосховище»).

Удосконалено підхід до часової прив'язки текстів за рахунок аналізу ієрархічної структури контексту. Формалізовано функцію, яка послідовно шукає явну вказівку на час у самому реченні, у заголовках верхнього рівня, у метаданих документа (до прикладу при обробці звіту за 2019 рік автоматично 2019 рік успадковується у якості часового контексту для всіх речень без явно вказаної дати). Створено ключові бінарні матриці «речення-географічна назва» та «речення-час». Корисними вважаються тільки ті речення, що мають і геоназву і часову прив'язку. Завдяки такому підходу знижується кількість хибнопозитивних зв'язків.

Удосконалено метод класифікації україномовних текстових даних. Класичні метрики подібності на україномовних текстах про басейн річки

Південний Буг (на прикладі `cosine_similarity`) не піднімається вище 0,376 для повних коректних збігів текстів та МВ та зберігається проблема полісемії. Запропоновано формалізацію річкової мережі у вигляді багатозв'язної геоінформаційної системи іменованих сутностей, що дозволяє прив'язати всіх об'єктів, які не є іменованими сутностями БГС до іменованих сутностей, або відкинути такі об'єкти. Таким чином текст «у Вінницькій області сталось забруднення р. Південний Буг нітратами біля водозабору м. Вінниця» прив'яжеться не до всіх МВ у Вінницькій області, а до МВ UA_M5.4_0013 (Сабарівське водосховище). Застосовуючи зважену міру Жаккара з вагами $w=3$ (локальні), 2 (регіональні), 1 (глобальні) можливо точно прив'язати новини та МВ. Таким чином отримано рекурсивний алгоритм прив'язки та аугментації даних без синтезу недостовірних фактів.

Результати, наведені у цьому розділі, опубліковані у працях [2, 4, 5, 8, 16, 19] здобувача.

РОЗДІЛ 3

РОЗРОБЛЕННЯ АЛГОРИТМІЧНИХ ОСНОВ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ВИДОБУВАННЯ ТА ВЕРИФІКАЦІЇ ДАНИХ ПРО ВЗАЄМОПОВ'ЯЗАНІ ВОДНІ ОБ'ЄКТИ З ПРОСТОРОВОЮ ТА ЧАСОВОЮ ПРИВ'ЯЗКОЮ

3.1 Побудова UML-діаграм варіантів використання технології

Передбачається, що інформаційною технологією видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою будуть користуватися користувач (експерт) та адміністратор.

Адміністратор повинен керувати користувачами (перевіряти авторизацію, визначати ролі користувачів, адже не кожному буде дозвіл вносити чи змінювати інформацію). Також адміністратор керуватиме шарами ГІС та базовою конфігурацією. При цьому адміністратор і користувач не зобов'язані співпрацювати.

Користувач, у свою чергу, зможе працювати з інформацією і виконувати такі операції:

- 1) переглядати каталог масивів вод (МВ);
- 2) шукати дані з використанням фільтрації;
- 3) переглядати факти про МВ;
- 4) верифікувати дані та аугментації;
- 5) переглядати ГІС-карти;
- 6) займатися скрапінгом і парсингом джерел;
- 7) класифікувати контент за станом, проблемою, заходами для вирішення проблеми тощо;
- 8) виконувати геоприв'язку текстів до масивів вод за допомогою зваженої Jaccard-міри;
- 9) видобувати знання (SPO і RAG);
- 10) аугментувати корпус фактів.

Перелічені функції користувача/експерта та адміністратора на UML-діаграмі варіантів використання технології видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою представлені на рис. 3.1 (у більшому розмірі діаграма подана у Додатку Е).

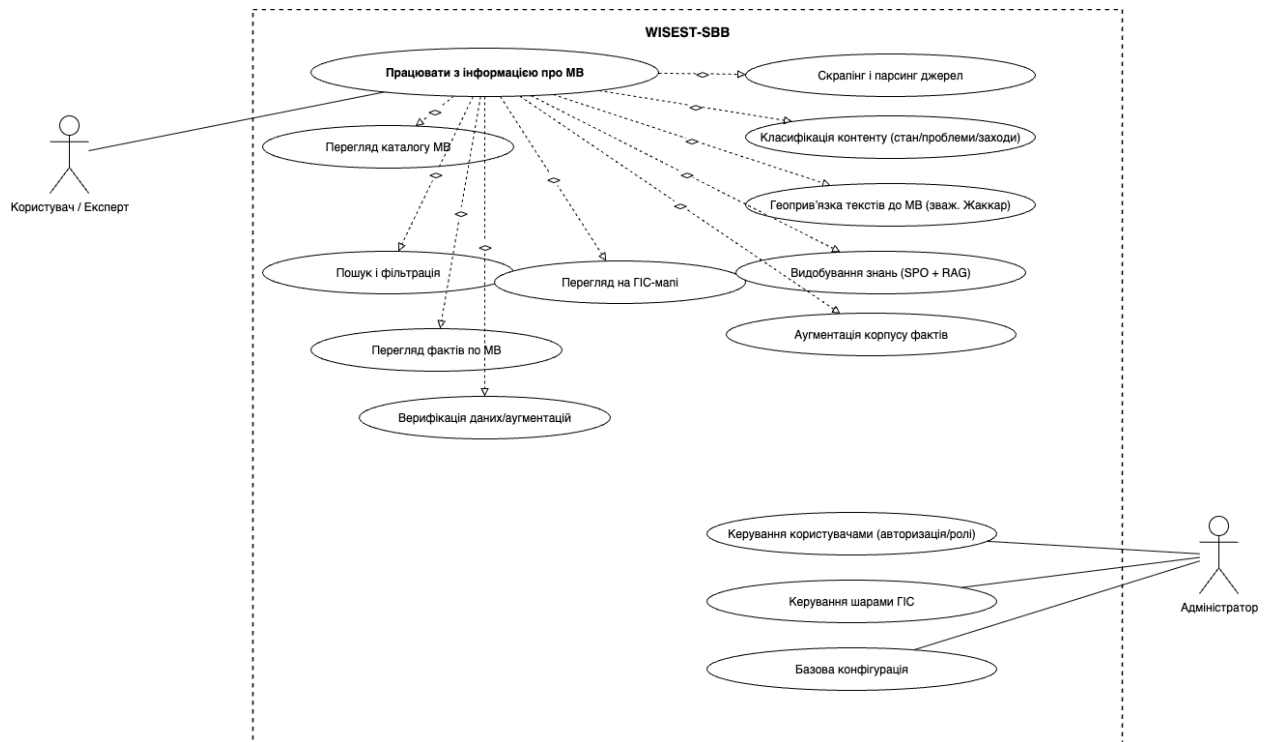


Рисунок 3.1 – UML-діаграма варіантів використання технології видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою

Додатково покажемо структуру системи на рисунку 3.2 та опишемо функціонал.

Через персональний кабінет можна авторизуватися і працювати з функціоналом адміністративної частини.

Модуль «Мапа» дозволить переглядати коди масивів вод.

Модуль «Масив вод» надає можливість шукати інформацію в повному обсязі, переглядати інформацію та зв'язані події, проводити експертну верифікацію.



Рисунок 3.2 – Структура системи на основі технології видобування та верифікації даних про взаємопов’язані водні об’єкти з просторовою та часовою прив’язкою

Як видно з рисунку 3.2, система на основі технології видобування та верифікації даних про взаємопов’язані водні об’єкти з просторовою та часовою прив’язкою враховує найважливіші модулі і взаємодію користувача з ними.

3.2 Розроблення структури інформаційної технології з використанням створеного методу прогнозування взаємозв’язків водних об’єктів з просторовою прив’язкою

Для реалізації усіх відібраних процесів пропонується розробити інтелектуальну технологію видобування та верифікації таким чином, щоб вона базувалася на автоматизованому зборі даних різного типу з будь-яких інтернет-джерел (рис. 3.3) [6, рис.3] (у більшому розмірі цей рисунок подано у Додатку Є).

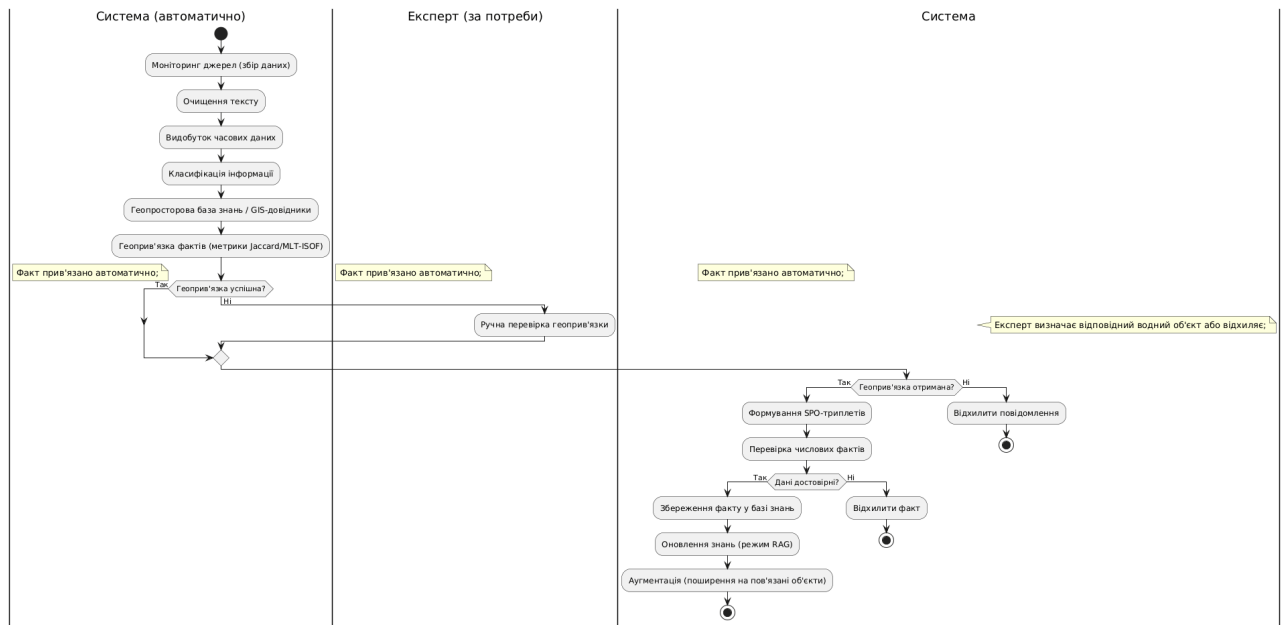


Рисунок 3.3 – Структура інтелектуальної технології з автоматизованим збиранням даних [6, рис. 3]

Опишемо принцип роботи системи з цією технологією. Беруться факти окремо один від одного і встановлюється відповідність відповідної категорії чи класу. Це може бути інформація про стан водного джерела, проблеми, що виникають у воді та навколо нього, які можуть погіршувати стан води. Далі витягуються просторово-часові метадані, які характеризують кожну новину. Ці метадані допоможуть прив'язувати новини до часових рамок і конкретних координат. За базу пропонується взяти ГІС-довідники [74], які використовуються в Україні (через те, що перша версія створюється саме для контролю масивів вод на території України), а також коефіцієнти Жаккара і MLT-ISOF. Щоб зрозуміти, як це працює, розглянемо деталі процесів. Знаходяться подібності серед описів водних джерел, які далі верифікуються та записуються в БЗ, оновлюючи RAG.

Такий принцип роботи точніше прив'язує факти відповідно до геоназв і забезпечує постійне оновлення актуальних даних.

Наведемо для прикладу реалізацію висунутих ідей, аналізуючи зібрану інформацію щодо вод, які відносяться до басейну річки Південного Бугу [30].

1. Завантажуємо відібрану інформацію про кожен водний масив до обраного модуля розроблюваної системи верифікації та обробляємо, створюючи характеристики для водних об'єктів.

2. Зв'язуємо одержані характеристики з об'єктами геоінформаційної системи.

3. Додаємо нові дані з ймовірних ресурсів до системи [89], використовуючи вебскрапінг.

4. Проводимо автоматизовану класифікацію за типом та геоприв'язування до МВ. Якщо не можна провести прив'язку, залучається експерт в області моніторингу, який буде перевіряти кожен етап роботи системи.

5. Оцінюємо індекс достовірності фактів, співставляючи з уже отриманими даними. Якщо є співпадіння, тоді індекс j росте.

6. Отримана інформація прив'язується до відповідного модуля системи, після чого адміністратор чи користувач може отримати до неї доступ.

Найвагомішим плюсом описаної структури з процесами є можливість відстежити зміни в інформації за часом і простором, визначаючи індекс достовірності. Структура першого критерію дозволяє порівняти якомога більше фактів, що підвищує достовірність інформації. І як наслідок цього, більше інформації буде оброблятися за заданий час.

Розроблена система реалізовує інформаційну технологію, яка для кожного МВ знаходить характеристики і прив'язує їх до елементів ГІС. Типова реалізація системи WISEST це авторизована аналітична вебсистема підтримки прийняття рішень. Цільовою аудиторією такої системи є експерти у питаннях моніторингу водних ресурсів та фахівці басейнових управлінь. Для кожного басейну річки розгортається власний віртуальний простір з набором шарів ГІС, довідників та ін.. У робота розглянута конкретна реалізація на прикладі басейну річки Південний Буг (WISEST-SBB), яка описана у п. 3.3. Принципи класифікації інформації описано у п. 2.4. Методи просторої та часової прив'язки у 2.2 та 2.3.

Для підвищення точності автоматичної прив'язки даних до ГІС пропонується класифікувати видобуті ключові слова згідно з рівнями: глобальними, регіональними та локальними, де кожному рівню визначено свої відповідні ваги.

Перед основним обробленням користувацький запит аналізується для визначення часових та просторових обмежень (наприклад, якщо користувач написав «...за останні 5 років...», то вихідні дані не будуть включати старіші дані). За допомогою ембеддінгів документи перетворюються на векторні дані, де кожен вектор містить корисну інформацію щодо конкретного масиву вод та метадані (гео-, часові).

Може додаватися й інша атрибутивна інформація, наприклад, віднесення її до певних класів (для вод – це різні види моніторингу, факти виявлення забруднень, управлінські заходи тощо) [100]. Для виконання пошуку на першому етапі формується набір кандидатів, після чого просторово-часовий фільтр відсіює ті, що не відповідають встановленим користувачем обмеженням.

Технологічно оперування векторною базою даних можна реалізувати за RAG-архітектурою (рис. 3.4) [6, рис. 2].

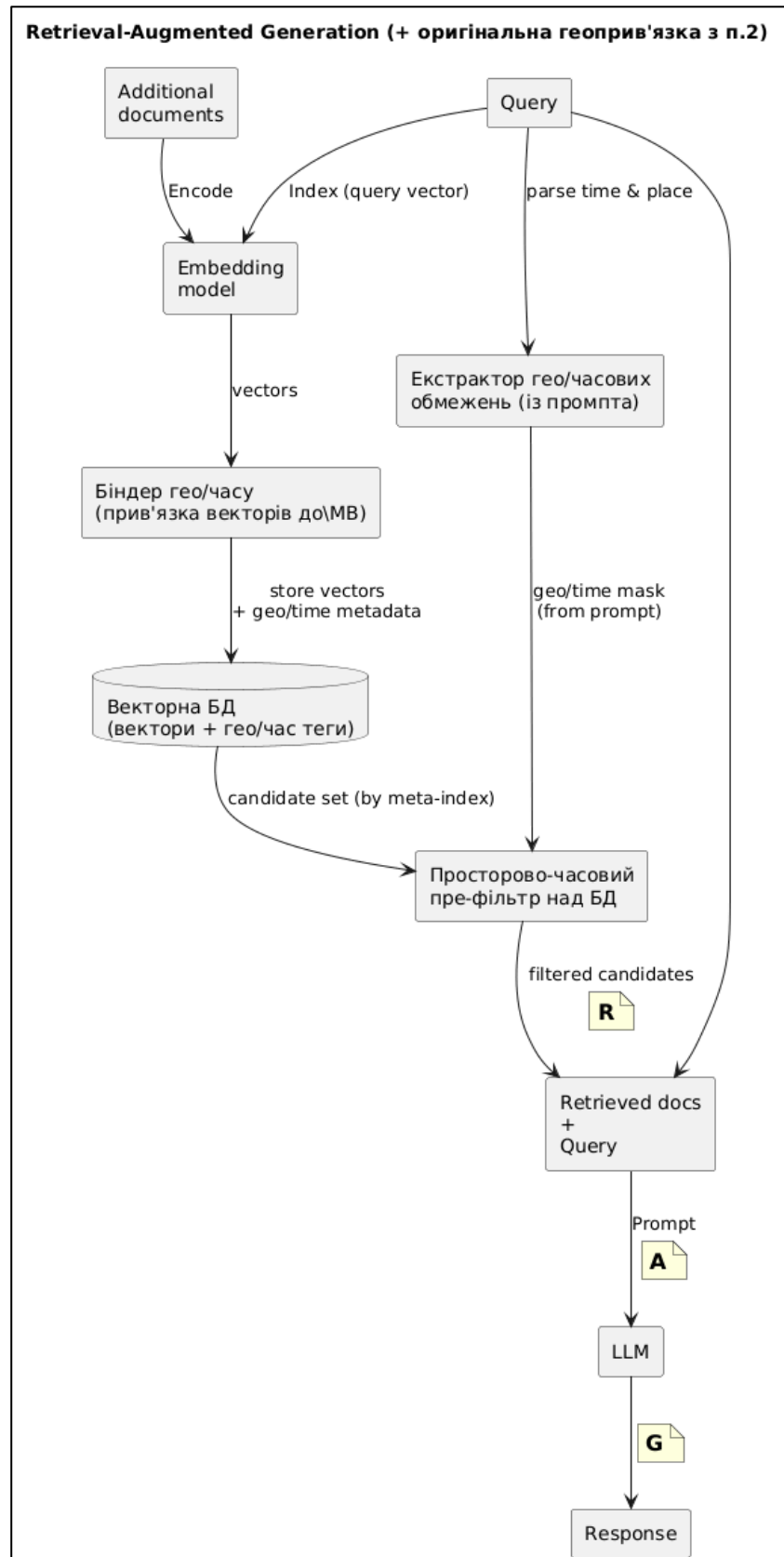


Рисунок 3.4 – Інтеграція Retrieval-Augmented Generation з геоприв'язкою [6, рис. 2]

На наступному етапі RAG-модуль передає релевантні записи до великої мовної моделі, яка формує остаточну відповідь, базуючись на цих записах та на знаннях, отриманих під час навчання. В результаті цього підходу покращується точність результатів.

Для того, щоб підвищити ефективність роботи системи, запропоновано збільшити обсяг датасетів за допомогою джерел гетерогенних даних. Якщо розглядати новини або пости в соцмережах чи звіти громадських організацій, то найкраще використовувати архітектуру каскаду джерел узагальнюючого типу. Для реалізації цієї ідеї неважливо, яким буде формат даних і чи буде API під паролем. Важливим є те, щоб нормалізувати отримані гетерогенні екологічні дані. А отже, буде розроблено каскадну архітектуру парсерів.

Інформація буде класифікуватися на три типи. Для кожного з них буде свій каскад парсерів.

Перший тип буде відповідати за публікації в мережі Інтернет. Принцип побудови підлягає резервуванню. Якщо найпростіший інструмент не може обробити сторінку, то вона передається складнішою. Це допоможе уникнути надлишку обчислень.

Другий тип відповідатиме за статичні дані, які можуть бути представлені звітами. Для цього достатньо використати універсальний API.

Третій тип буде обробляти публічні API провайдерів, які надають інформацію. Для кожного з них важливо розробити індивідуальний парсер.

При цьому для кожного з цих типів будуть застосовані валідація даних і фільтрація.

Покажемо на рисунку 3.5 архітектуру підсистеми каскадних парсерів.

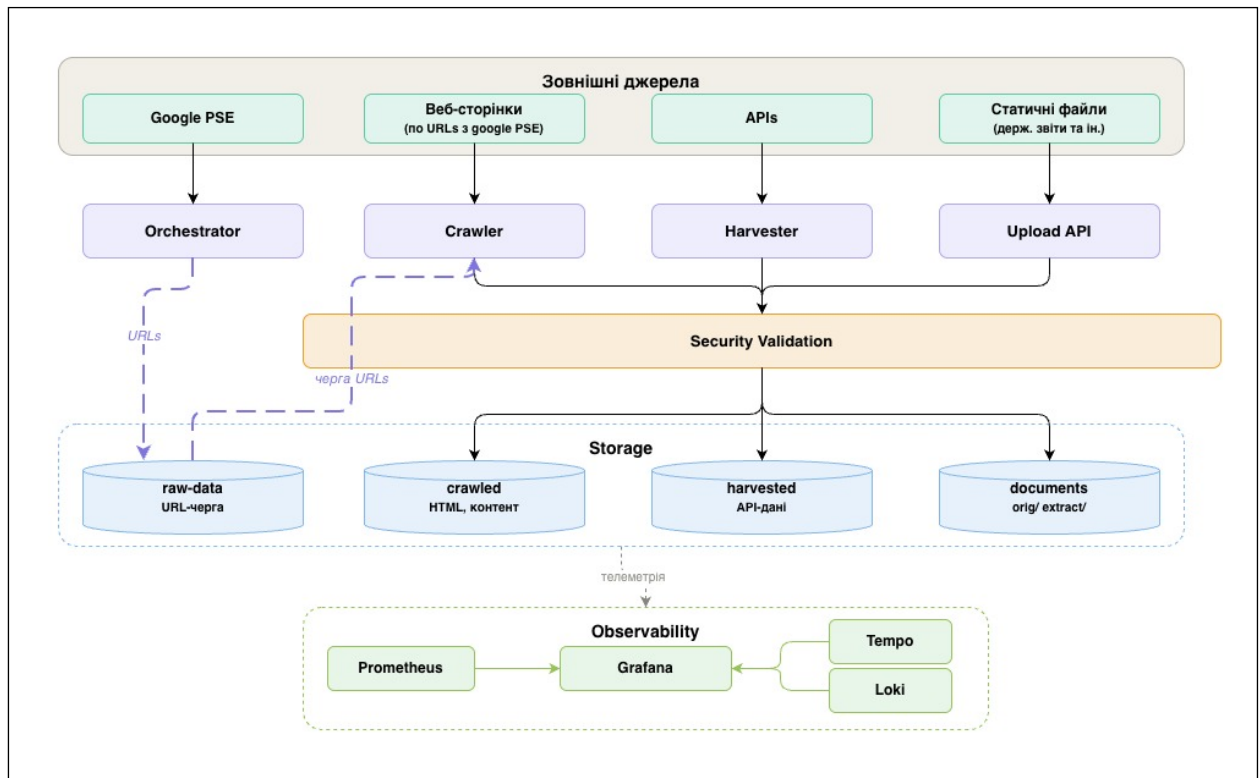


Рисунок 3.5 – Архітектура підсистеми каскадних парсерів

Проаналізуємо отримані дані. Для аналізу роботи системи, що реалізовує інформаційну технологію видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою, взято дані про води басейну річки. Архітектура заснована на модульному моноліті. Це означає, що слабо пов'язані Docker-контейнерами модулі мають спільну архітектуру та мають сховища Redis та minio.

До уваги візьмемо лише ресурси на теренах українського вебпростору.

За перший тип відповідатиме Programmable Search Engine. На вході – два набори ключових слів і виразів (мінус-слова та слова для пошуку).

Ключові слова і фрази обох типів поповнюються через механізм зворотного зв'язку після оброблення видобутого контенту. Посилання зберігаються у вигляді структурованих json-файлів у локальному сховищі minio.

На наступному кроці відбувається певне оброблення та валідація (як-от нормалізація посилань — усунення utm та дедуплікація, таким чином, до наступного етапу потраплять тільки унікальні посилання). Пошук працює

асинхронно, імплементовано пул API-ключів з автоматичною ротацією та обмеженням щоденних лімітів по кожному з ключів.

Для оптимізації видобування контенту з інтернет-джерел рекомендовано застосовувати каскад парсерів: BeautifulSoup для швидкого парсингу статичних сайтів, спеціалізований newspaper4k для сайтів ЗМІ, trafilatura як універсальний інструмент та Playwright у headless режимі для динамічних вебсайтів (як-от SPA-сторінки та ін.). Для кожного домену зберігаються наперед визначені налаштування оптимального способу парсингу.

Для другого типу, а саме — оброблення державних звітів та інших статичних файлів, підготовлено API, що дозволяє завантажувати файли у форматах csv, doc(x), xls(x), pdf, застосовувати необхідні правила оброблення та зберігати у сховищі файлів.

Для конвеєра даних третього типу запроваджена опціональна підтримка інтеграції провайдерів публічних даних через формалізовані контракти даних та нормалізацію до єдиного формату.

З того датасету, який був раніше побудований, було отримано 151 ключове слово, а також 2970 посилань на вебресурси, серед яких 706 дублікатів за ключовими словами, 1134 домени. У результаті подальшої роботи системи було зібрано приблизно 8000 унікальних посилань для приблизно 1200 доменів. Найбільша кількість даних розміщена про водні об'єкти, а найменше — про населені пункти. Якщо будуть нерелевантні результати, вони будуть поповнюватися іншими.

Обрано модель інформаційної технології видобування та верифікації даних про водні масиви, яка базується на семантичному представленні знань щодо взаємопов'язаних водних об'єктів у вигляді SPO-триплетів (subject–predicate–object). В даному випадку врахована просторова та часова прив'язка.

Для моделі, що розглядається, усі параметри [101] зведено до формалізованого вигляду за допомогою використання зваженої міри Жаккара та критерію MLT-ISOF.

Зважена міра Жаккара необхідна задля кількісного оцінювання, яка відображатиме схожість між отриманими даними і наявними в базі даних про водні масиви.

Критерій MLT-ISOF необхідний для адаптації метрики TF-IDF в просторі з доданням ваги, а також інтегральної оцінки об'єктів для відображення релевантності.

Для того, щоб ідентифікувати модель, добуто корпус текстів про водні масиви, який складається з новин та соцмереж, а також інформативні дані про басейни річок.

З цих текстових повідомлень методами NER та KeyBERT вилучено географічні назви й ключові фрази, на основі яких сформовано базу знань у вигляді фактичних SPO-триплетів. З метою підтримання актуальності знань реалізовано механізм RAG-оновлення: нові фрагменти інформації конвертувалися у векторні ембедінги та індексувалися засобами бібліотеки FAISS, що забезпечило швидкий семантичний пошук і динамічне доповнення бази знань перевіреними фактами.

Адекватність моделі було перевірено шляхом порівняння отриманих результатів з фактичними спостереженнями: автоматично визначені моделлю зв'язки між текстами і відповідними водними об'єктами та верифіковані числові дані співставлялися з реальними даними моніторингу і відомими подіями, що підтвердило достовірність та ефективність запропонованої моделі.

Домен для розміщення даної системи знаходиться за адресою <https://wisestr.ai>.

Базовий інтерфейс представлено на рис. 3.6.

Додаток системи, що реалізовує інформаційну технологію видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою, складається з двох частин – стартового вікна та головного меню.

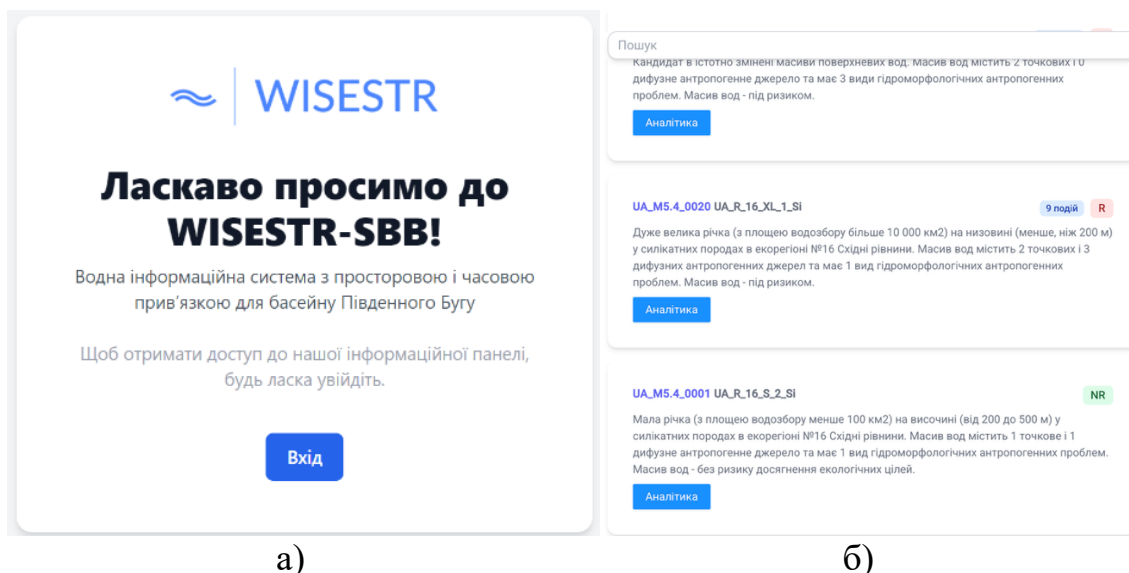


Рисунок 3.6 – Базовий інтерфейс системи «WISESTR-SBB»: а) стартове вікно; б) головне меню

Охарактеризуємо процес додавання інформації до системи. Спочатку опишемо початковий етап створення системи з його структурою і особливостями:

1. Модуль авторизації забезпечує доступ до програми та керування обліковими записами користувачів системи, що реалізовує інформаційну технологію видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою.

2. Головна сторінка представлена у вигляді каталогу з картками, що містять описи кожного з 1091 МВ. Також передбачено пошук за всіма доступними полями МВ.

3. Інтерактивна карта на базі ArcGIS, яка включає векторні шари геоінформаційної системи водних масивів басейну річки Південний Буг. У перспективі планується можливість динамічного підключення потрібних користувачу шарів карти.

Для тестування системи було забезпечено її початкове наповнення. Було вирішено структурувати систему за трьома інформаційними блоками.

Блок 1. Векторні шари геоінформаційної системи водних масивів басейну річки Південний Буг були завантажені до сервісу ArcGIS та інтегровані із системою.

Це дає змогу у зручному форматі відображати розташування кожного водного масиву на карті, а також переглядати його код і опис. Результат показано на рисунку 3.7.

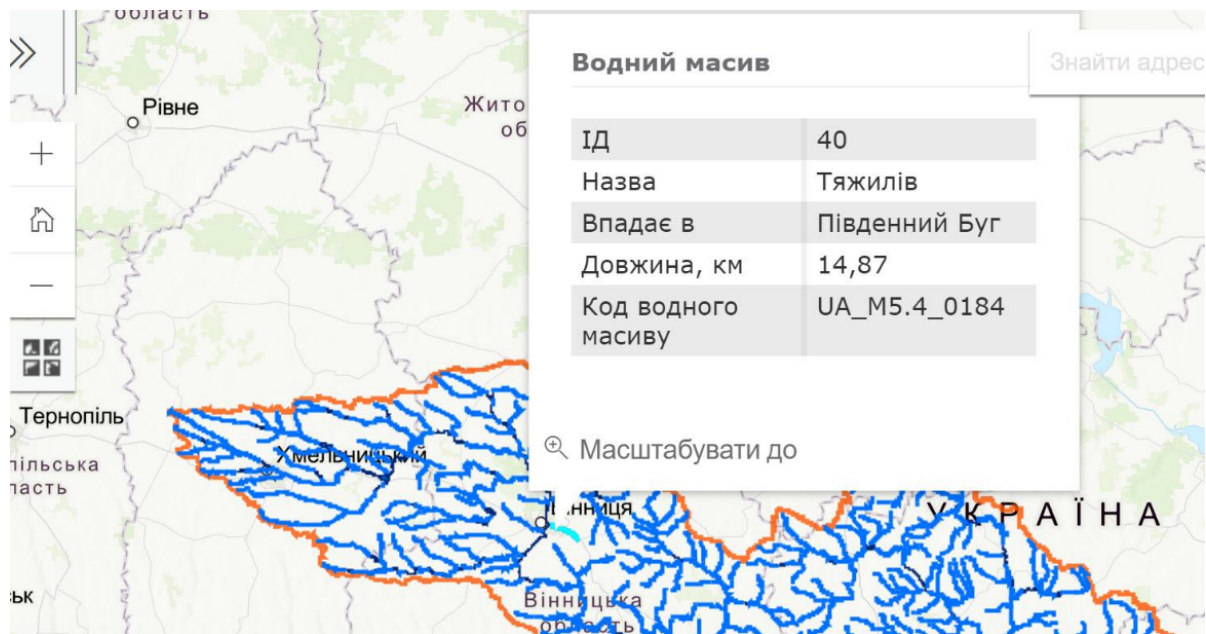


Рисунок 3.7 – Карта з масивами вод басейну р. Південний Буг системи «WISEST-SBB»

Блок 2. Інформацію з проєкту ПУРБ Південного Бугу було програмно оброблено та трансформовано в текстові описи для кожного масиву вод. Результат цього перетворення представлено на рисунку 3.8.

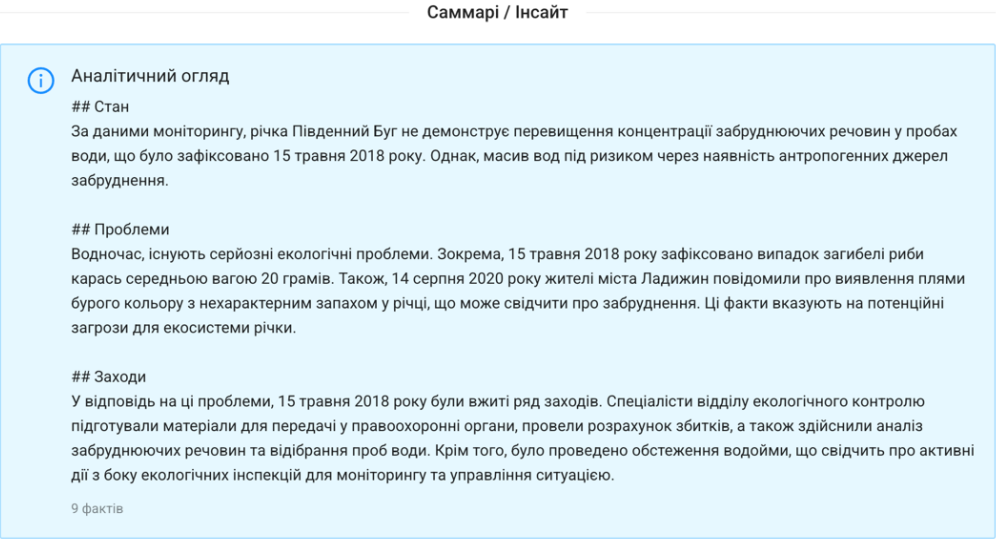


Рисунок 3.8 – Результат перетворення зібраної інформації на текст про кожен МВ

Генерацію текстових речень на основі табличних даних проведено відповідно зазначеного алгоритму:

Показники, що описують стан водних масивів, були оброблені та приведені до формату, який підвищує дружність інтерфейсу, наскільки комфортно і зрозуміло використовувати цю систему, враховуючи морфологічні норми української мови. Це надало можливість детально подати отримані дані у формі, зручній для сприйняття, використовуючи відповідні мовні конструкції залежно від конкретних числових показників.

Після оброблення кожного параметра масиву вод [101] до опису додавався відповідний текст. Результат наведено на рисунку 3.6.

Блок 3. Було відібрано близько 250 релевантних прикладів, після чого їх вручну систематизовано, проведено геоприв'язку та класифікацію за видами інформації.

Надалі їх було інтегровано в систему та реалізовано інтерфейс для візуалізації. Ці приклади застосовуються як тестовий набір для інтелектуальних технологій автоматичної просторово-часової прив'язки та класифікації.

У системі WISEST-SBB реалізовано інтерфейс, що дозволяє одночасно відображати всі три зазначені блоки даних, зокрема на прикладі інформації про тип водного масиву (рис. 3.9).

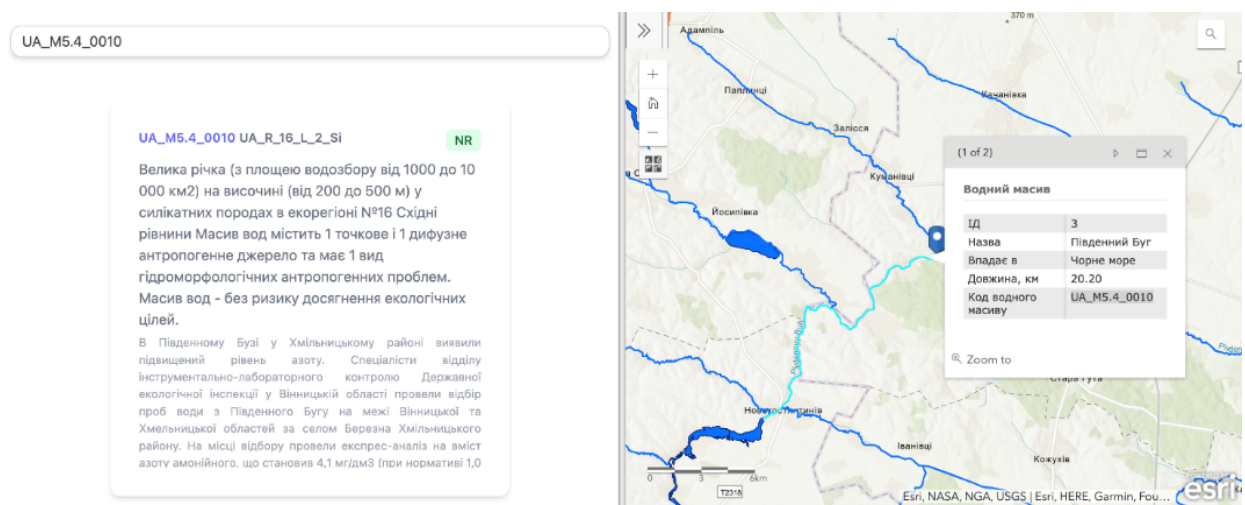


Рисунок 3.9 – Інтерфейс системи «WISEST-SBB» з прикладом інформації про тип масиву вод

3.3 Інфраструктура інформаційної технології

Як було зазначено вище, інфраструктура інформаційної технології базується на архітектурі модульного моноліту та додаткових мікросервісах. Кожен ключовий компонент системи розгортається у власному Docker контейнері. У залежності від поставлених задач, інфраструктурні компоненти системи доцільно розподілити на наступні підкатегорії:

- зберігання даних;
- інформаційна безпека;
- спостережуваність;
- безперервна інтеграція (CI);
- розгортання.

Загальна взаємодія реалізована на базі docker-compose. Сервіси рознесено поміж 4-х окремих Docker-мереж, а саме: публічної (для оброблення вхідних запитів); ізольованої внутрішньої для всіх сервісів, що не потребують доступу до зовнішньої мережі; окремої зовнішньої мережу для сервісів, що потребують власного виходу на зовні та мережу для збирання метрик з метою подальшої аналітики. Для роботи серверу застосовано патерн зворотного проксі (рис. 3.10).

```

Waiting for services (9 total)
  waiting: scraper_redis, scraper_minio,
  ✓ scraper_redis (0s)
  ✓ scraper_minio (0s)
  ✓ promtail (0s)
  ✓ prometheus (0s)
  ✓ tempo (0s)
  ✓ grafana (0s)
  ✓ scraper_worker (7s)
  ✓ loki (20s)
  ✓ caddy (45s)

  ✓ All 9 services ready (45s)

Summary
Keywords: 101 pending / 35 done (25%)
API keys: 0 available / 16 exhausted
Errors: 1 total

```

Рисунок 3.10 – Приклад моніторингу запуску частини мікросервісів

Для зберігання даних використовується декілька сховищ. Для зберігання файлів та передачі великих файлів поміж сервісами системи використовується S3 сумісне сховище MinIO (рис. 3.10). Для координації воркерів, планування задач, дедуплікації, обмеження частоти запитів використовується Redis. PostgreSQL використано як основну реляційну БД, а SQLite на етапі первинного оброблення для зберігання фактів, метаданих тощо.

Кожен завантажений файл аналізується локальним антивірусом ClamAV, специфічні документи перевіряються спеціалізованими інструментами.

За реалізацію спостережуваності відповідає загальноприйнятий Grafana стек, що включає метрики (Prometheus), трейси (Tempo) та логи (Loki). За візуалізацію відповідають окремо побудовані дашборди (статус парсингу, швидкість запитів, статистика збоїв сервісів, та інше). Для встановлення зв'язків поміж різними сервісами реалізовано патерн Distributed tracing.

На рисунку 3.11 представлено скріншот інтерфейсу Grafana системи WISEST-SBB з прикладом логів по сервісам та важливості (info, warning, error).

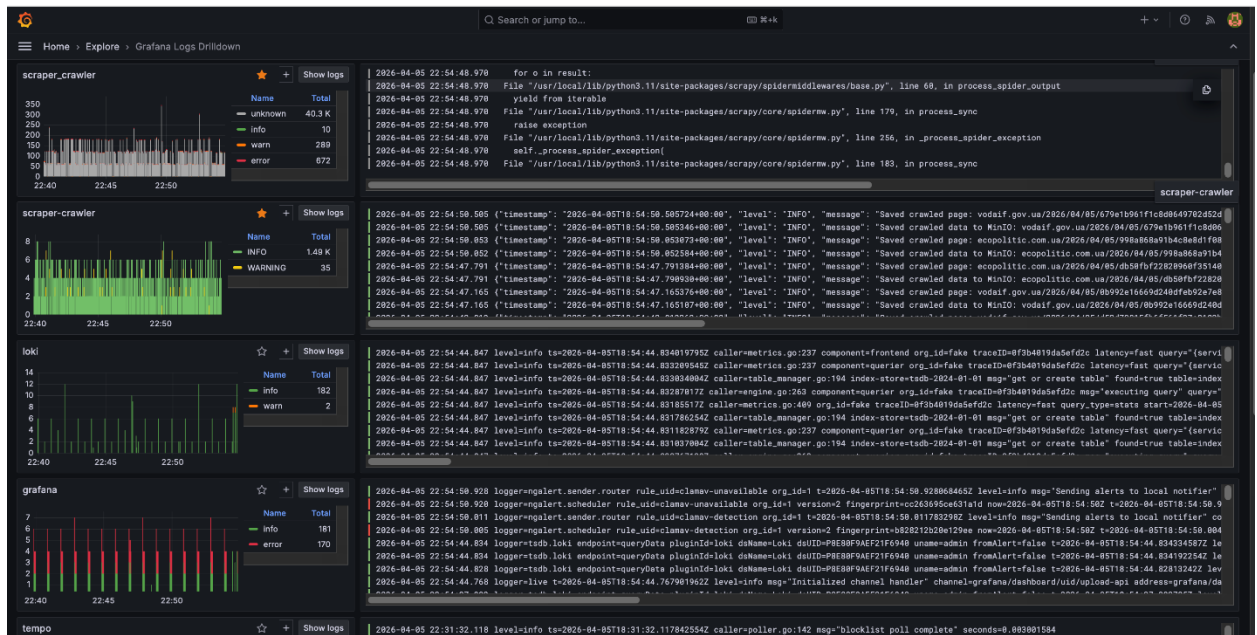


Рисунок 3.11 – інтерфейс Графана для спостереження за роботою системи WISEST-SBB

Для неперервної інтеграції та розгортання використовується Github Actions.

Висновки до розділу 3

Третій розділ присвячений розробленню інформаційної технології видобування та верифікації даних, від UML моделі до розгортання інфраструктури.

UML-діаграма варіантів використання формалізує дві ролі – користувача-експерта та адміністратора та 10 ключових функцій користувача.

Розроблено структуру інтелектуальної технології з автоматизованим збиранням даних з гетерогених джерел та послідовно проводить дані через етапи класифікації згідно визначеним класам (status, issues, measures, balance), видобування просторово-часових метаданих, прив'язки до ГІС-довідників та запису в базу знань. Реалізовано базовий інтерфейс веб системи WISEST-SBB.

Як важливу варто виділити реалізовану архітектуру каскадних парсерів гетерогених джерел 3-х типів (веб ресурси, статичні файли, API-провайдери

даних). На першому проході системи з 151 ключовим словом було отримано також 2970 посилань на вебресурси на 1134 доменів. У результаті подальшої роботи системи було зібрано приблизно 8000 унікальних посилань для приблизно 1200 доменів.

Розроблено інфраструктуру інформаційної технології (на базі docker-compose та Grafana stack).

Результати, наведені у цьому розділі, опубліковані у працях [14, 17, 22] здобувача.

РОЗДІЛ 4

ПРИКЛАДНЕ ЗАСТОСУВАННЯ РОЗРОБЛЕНОЇ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ВИДОБУВАННЯ ТА ВЕРИФІКАЦІЇ ДАНИХ ПРО ВЗАЄМОПОВ'ЯЗАНІ ВОДНІ ОБ'ЄКТИ З ПРОСТОРОВОЮ ТА ЧАСОВОЮ ПРИВ'ЯЗКОЮ

4.1 Застосування розробленої системи для аналізу даних про водні ресурси

Проведено експеримент, метою якого є створення бази знань на основі автоматично структурованих фрагментів тексту з екологічних звітів. Тобто, вирішувалась задача отримання з неструктурованих текстів впорядкованих наборів фактів з метаданими (як-то прив'язка до масиву вод, року) з використанням запропонованого методу, описаного в підрозділ 3.2.

Як вхідні дані було використано фрагмент офіційного звіту про стан водних ресурсів, зокрема розділ IV, підрозділ 4.3, пункт 4.3.1 документа «Річний звіт про діяльність Басейнового управління водних ресурсів річки Південний Буг за 2019 рік» [102]. Цей фрагмент містить узагальнену за результатами моніторингу інформацію про якість поверхневих вод у 4 контрольних пунктах р. Південний Буг у Вінницькій області за 2019 рік.

Другий датасет – це довідник, де зазначені деякі масиви вод басейну річки Південного Бугу: офіційне кодування та назви географічних об'єктів в їх просторовому контексті, що дозволяє однозначно ідентифікувати кожен масив вод.

На рисунку 4.1 наведено приклад цих обох вхідних датасетів [5, рис. 1].

```
=== text_ua ===
```

Узагальнена інформація про стан поверхневих вод Вінницької області
Відповідно до затвердженої програми моніторингу вод, у Вінницькій області, протягом 2019 року, щомісячно здійснювалися спостереження масивів поверхневих вод, забір води з яких здійснюється для задоволення питних і господарсько-побутових потреб населення, на чотирьох пунктах моніторингу, які розташовані в басейні Південного Бугу.
Річка Південний Буг.
Кисневий режим річки Південний Буг задовільний, значення розчиненого кисню знаходяться у межах 4,10-19,10мгО₂/дм³, при нормі не менше 4,0мгО₂/дм³.
Жорсткість води середня, значення середньорічних показників становили 5,52-7,70 мг-екв/дм³(ГДК 7,0 мг-екв/дм³), перевищення у 3 пробах, максимальне у 1,1 рази в питному водозаборі м. Ладижин, м. Хмільник, с. Гушинці питний водозабір, м. Калинівка. ...

```
=== water_bodies ===
```

	code_water_body	list_geo_objects
0	UA_M5.4_0011	Вінницька область, річка Південний Буг, м. Хмільник
1	UA_M5.4_0013	Вінницька область, річка Південний Буг, м. Калинівка, м. Вінниця, с. Гушинці
2	UA_M5.4_0019	Вінницька область, річка Південний Буг, м. Ладижин

Рисунок 4.1 - Вхідні датасети з файлів text_ua.csv (множина *C*) та water_bodies.csv (множини *W* та *G*) [5, рис. 1]

Охарактеризуємо розроблений у п. 2.5. метод, який реалізує автоматичне формування зв'язків за триплетом «факт – масив вод – інтервал часу». Суть методу полягає в поетапному аналізі тексту, контексту та зіставлення даних з множиною *G*. Ключові етапи:

Текстовий фрагмент розбиваємо на окремі речення, проводимо первинний аналіз для розпізнавання речень з потенційно важливими екологічними фактами. До таких речень відносимо ті, що містять метрики або спеціально визначені поняття (наприклад, «концентрація речовин», «перевищення нормативних значень», «показники кисневого режиму» тощо).

Для визначення контексту речень аналізуємо ієрархічну структуру вхідного документа: якщо в реченні відсутня у явному вигляді інформація про географічний об'єкт чи час, то ці дані успадковуються із попередніх речень, розділів або контексту документа загалом (у взятому прикладі 2019 рік згадується у назві документа). Для визначення географічного контексту вважаємо, що текст

відноситься до всієї р. Південний Буг, доки не знайдемо більш специфічних вказівок.

Порівнюємо знайдені в тексті назви у географічному довіднику, порівнюючи підмножини словосполучень в тексті та в довіднику масивів вод на основі міри Жаккара. Наприклад, для «р. Південний Буг, м. Хмільник» алгоритм знаходить частковий збіг у довіднику з офіційною назвою «Вінницька область, річка Південний Буг, місто Хмільник – код UA_M5.4_0011», що дає можливість однозначно прив'язати факт до цього масиву вод.

Формуємо триплети за шаблоном «факт – масив вод – інтервал часу» на підставі аналізу пунктів 1 – 3. Вважаємо, що кожен триплет відображає одну конкретну подію, характеристику чи факт. Після виділення фактів з прив'язкою здійснюємо їх об'єднання про конкретний масив вод у певний інтервал часу в єдиний текст. На прикладі вхідного фрагменту звіту два окремі речення: про перевищення жорсткості води та про кисневий режим було об'єднано в єдиний текст для кожного місця спостереження. Таким чином, отримуємо короткі, але інформативні записи з розрізнених фактів.

У результаті роботи розробленого методу отримано автоматично структурований датасет екологічних фактів R_{Σ} , де кожен запис складається з опису комплексного факту та його взаємозв'язку з масивами вод у часі (див. рис. 4.2) [5, рис. 1].

	code_water_body	time_period	text
0	UA_M5.4_0011	2019	Перевищення середньорічних показників жорсткості мало місце максимально в 1.1 рази (ГДК 7,0 мг-екв/дм3) у м. Хмільник. Кисневий режим річки Південний Буг задовільний, значення розчиненого кисню знаходяться у межах 4,10-19,10мгО2/дм3, при нормі не менше 4,0мгО2/дм3
1	UA_M5.4_0013	2019	Перевищення середньорічних показників жорсткості мало місце максимально в 1.1 рази (ГДК 7,0 мг-екв/дм3) у м. Калинівка. Кисневий режим річки Південний Буг задовільний, значення розчиненого кисню знаходяться у межах 4,10-19,10мгО2/дм3, при нормі не менше 4,0мгО2/дм3
2	UA_M5.4_0013	2019	Перевищення середньорічних показників жорсткості мало місце максимально в 1.1 рази (ГДК 7,0 мг-екв/дм3) у с. Гушчинці. Кисневий режим річки Південний Буг задовільний, значення розчиненого кисню знаходяться у межах 4,10-19,10мгО2/дм3, при нормі не менше 4,0мгО2/дм3
3	UA_M5.4_0019	2019	Перевищення середньорічних показників жорсткості мало місце максимально в 1.1 рази (ГДК 7,0 мг-екв/дм3) у м. Ладизин. Кисневий режим річки Південний Буг задовільний, значення розчиненого кисню знаходяться у межах 4,10-19,10мгО2/дм3, при нормі не менше 4,0мгО2/дм3

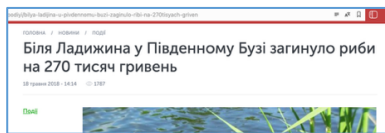
Рисунок 4.2 – Фрагмент описаних вихідних наборів даних [5, рис. 2]

На розглянутому прикладі система сформувала 4 записи, по одному для кожного з 4 пунктів спостереження, успішно прив'язавши кожен екологічний факт до конкретних масивів вод та часових інтервалів.

Таким чином, проведений експеримент підтвердив працездатність методики автоматичного видобування екологічних знань з екологічних звітів із прив'язкою до масивів вод і часу. Вихідна база знань за шаблоном триплетів «факт – масив вод – інтервал часу» узгоджена з експертними даними і зручна для подальшого аналізу. Отже, доведена можливість суттєво автоматизувати агрегацію екологічних звітів, забезпечивши екологам та іншим відповідальним експертам інструментами для швидкого виявлення ключових фактів і тенденцій у великих об'ємах розрізнених текстових документів.

4.2 Застосування розробленої технології до аналізу стану водних ресурсів Вінницької області у басейні річки Південний Буг

Продемонструємо приклад застосування запропонованої інформаційної технології. На рис. 4.3а зображено новину від 18 травня 2018 року, яка опублікована на сайті і пов'язана з мором риби [6, рис. 4]. Розроблена технологія надала можливість здійснити геоприв'язку цієї інформації до масиву вод №UA_M5.4_0019 (Ладизинське водосховище, р. Південний Буг) та, відповідно – до числових даних моніторингу стану цього масиву вод Державним агентством водних ресурсів України на посту біля с. Маньківка (рис. 4.3 б). На рис. 4в наведено графове подання моделі показників з рис. 4б, яке дозволяє прогнозувати їх зміну в часі і враховувати взаємовплив на основі рівнянь евтрофікації, внаслідок якої концентрація розчиненого у воді кисню зменшується, у разі зростання кількості біогенних речовин [103].

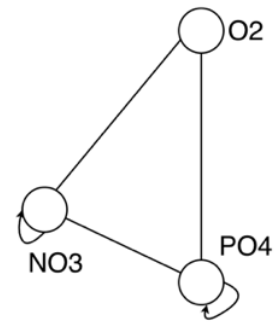


а)

21. Пост: р. Південний Буг, 413 км. с. Маньківка

Значення	Кисень розчинений, мгО ₂ /дм ³	Нітрат-іони, мг/дм ³	Фосфат-іони (поліфосфат и), мг/дм ³
Дата			
17.01.2018	7,950	1,790	0,190
12.02.2018	9,500	2,590	0,210
13.03.2018	7,000	2,300	0,740
17.04.2018	15,200	0,080	0,050
15.05.2018	6,800	1,300	0,540
06.06.2018	11,400	0,000	0,430

б)



в)

Рисунок 4.3 – Результат видобування та геоприв’язки інформації про потенційний мор риби у р. Південний Буг біля м. Ладжин 18.05.2018 р.: а) фрагмент скріншота з текстом новини; б) фрагмент бази даних державного моніторингу; в) графове подання моделі показників з рис. 4б, яке дозволяє прогнозувати їх зміну в часі і враховувати взаємовплив [6, рис. 4]

На рис. 4.4а наведено біхроматичний граф, який відповідає графовій моделі, наведений на рис. 4.3в, побудований за описаною вище технологією. Як видно, з кількості сильних (зелених) ребер, модель є топологічно повністю спостережуваною, тобто модель повністю топологічно спостережувана. Однак, на моделі на рис. 4в відсутня вершина «мор риби».

Пропонується її доповнити інтелектуальною моделлю, яка на основі аналізу даних, наприклад, з рис. 4.3б, буде передбачати мор риби (на рис. 4.3б добре видно, що з квітня по травень концентрація біогенних речовин («NO₃», «PO₄») різко зросла, а кисню – різко впала у декілька разів – такі фактори дозволяють доволі впевнено прогнозувати мор риби. Ця нова модель на біхроматичному графі додана у вигляді червоних ліній (рис. 4.4б).

Оптимізований біхроматичний граф теж є повністю топологічно спостережуваним (4 зелених «сильних» ребра – 4 вершини).

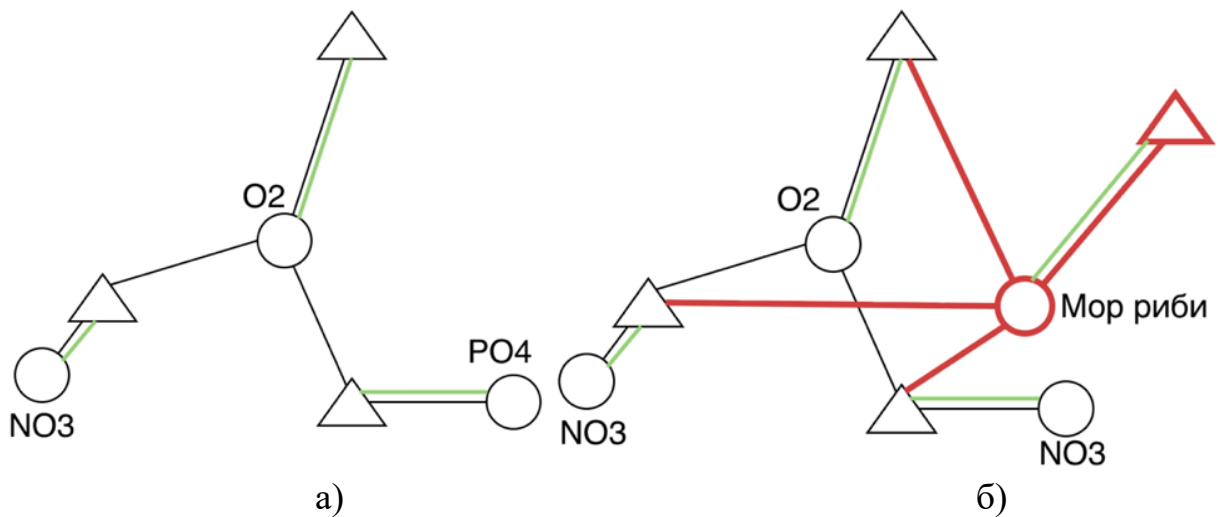


Рисунок 4.4 – Біхроматичні графи для моделей з топологічною спостережуваністю – 1,0: а) відповідає графовій моделі на рис. 4.3в; б) оптимізований (а) під додаткову вершину «мор риби» [6, рис.5]

Але якщо даних для тренування дійсно адекватної інтелектуальної моделі для графа рис. 4.4б буде недостатньо, тоді важливо зіставити різні факти про наявність мору риби з інших джерел, які вдалося прив'язати до цієї ділянки (масиву вод) та до цього ж проміжку часу (рис. 4.5).

Експерти у сфері державного моніторингу вод пропонують брати такі базові індекси достовірності для різних видів джерел інформації: слова жителів-очевидців, блогерів та інших релевантних, але неперифікованих, джерел – 0,3, дані з офіційних сайтів чи звітів з акредитованих лабораторій моніторингу різних держуправлінь чи Держекоінспекції, судів або інших офіційних органів – 0,95, дані зі ЗМІ з посиланням на офіційні джерела, але зі своїми доповненнями – 0,7. Індекси достовірності оцінок за моделями можна оцінювати від 0,1 до 0,9, залежно від обсягу тренувальних даних та досягнутої точності прогнозу на валідаційних даних з крос-валідацією. Візьмемо приклад свідчень очевидця в соціальній мережі від 15 травня 2018 про мор риби біля м. Ладижин (рис. 4.5) [6, рис.6]: без додаткових обробок система оцінює індекс достовірності цієї події як 0,3. Наступним кроком на рис. 4.6 [6, рис.6] наведено приклад знайдених новин про цей же факт, які доповнюють оцінку, отриману за моделлю, для тренування якої було недостатньо даних.

Достовірність події, J: 0.3					
	date	title	link	text	j
0	2018-05-15	-	https://www.facebook.com/plugins/post.php?href...	Це П. Буг, район порохів складів(бакланів) ...	0.3

Рисунок 4.5 – Свідчення очевидця про мор риби на Південному Бузі у масиві вод №UA_M5.4_0019 (біля м. Ладжижина) у травні 2018 року 4 [6, рис.6].

Достовірність події, J: 0.9972375690607735						
	date	title	link	text	j	p
0	2018-05-15	-	https://www.facebook.com/plugins/post.php?href...	Це П. Буг, район порохів складів(бакланів) ...	0.3	0.300000
1	2018-05-18	Біля Ладжижина у Південному Бузі загинуло риби ...	https://vinbazar.com/news/podiyi-bilya-ladijin...	15 травня 2018 року на річці Південний Буг у м...	0.95	0.890625
2	2018-05-16	Екологи відібрали проби води в річці поблизу Л...	https://lada.fm/2018/05/16/ekologi-vdbrali-pro...	Екологи зафіксували загибель риби поблизу Лади...	0.95	0.993578
3	2018-12-06	Якісна вода для містян: ДТЕК Ладжижинська ТЕС п...	https://dtek.com/media-center/news/yakisna-vod...	У прес-релізі ДТЕК зазначено, що Ладжижинська т...	0.7	0.997238

Рисунок 4.6 – Оцінка індексу достовірності новини про мор риби на Південному Бузі у масиві вод №UA_M5.4_0019 (біля м. Ладжижина) у травні 2018 року [6, рис.7]

Отже, з рис. 4.6 видно, що для прогнозу за моделлю або для свідчень очевидців з соцмереж вийшло підвищити індекс достовірності з 0,3 до 0,99, за рахунок прив'язки у часі та просторі різних числових та текстових даних.

На рис. 4.7 представлено фрагмент даних з роботи системи, який ілюструє, що з одного джерела можливо отримати значну кількість екологічних фактів.

```

=== articles ===
{'article_id': 1, 'url_60': 'https://www.facebook.com/permalink.php?story_fbid=pfbid0Lbhw', 'domain': 'Пост на сторінці Держекоінспекції у Вінницькій області'}
{'article_id': 3, 'url_60': 'https://vsim.ua/Podii/ekoaktivisti-hmelnichchini-byut-na-spo', 'domain': 'Пост на сторінці ГО "Річки Хмельниччини"}
{'article_id': 4, 'url_60': 'https://vsim.ua/Podii/ekologi-pripustili-chomu-v-pivdenному-', 'domain': 'Веб-сторінка новин Хмельниччини'}
{'article_id': 5, 'url_60': 'https://tsn.ua/ukrayina/u-mikolayevskiy-oblasti-vidbuvayetsy', 'domain': 'Веб-сторінка служби ТСН'}
{'article_id': 6, 'url_60': 'https://fakty.com.ua/ua/videos/pid-zagrozozyu-ridkisa-flora-', 'domain': 'Веб-сторінка телевізійних новин ФАКТ И'}
{'article_id': 7, 'url_60': 'https://vitatv.com.ua/misto/pivdennoy-buh-na-staromu-misti-u', 'domain': 'BіраТБ'}
{'article_id': 8, 'url_60': 'https://vitatv.com.ua/oblast/u-pivdenному-buzi-bilya-stryzha', 'domain': 'BіраТБ'}
{'article_id': 9, 'url_60': 'https://vitatv.com.ua/oblast/v-pivdenному-buzi-u-khmilnytsko', 'domain': 'BіраТБ'}
{'article_id': 10, 'url_60': 'https://vitatv.com.ua/oblast/u-khmilnyku-v-pivdenному-buzi-z', 'domain': 'BіраТБ'}
{'article_id': 11, 'url_60': 'https://vitatv.com.ua/misto/pivdennoy-buh-u-tsentr-vinnytsi', 'domain': 'BіраТБ'}
{'article_id': 12, 'url_60': 'https://vitatv.com.ua/misto/u-vinnytsi-vidnovly-roboty-z-oc', 'domain': 'BіраТБ'}

=== facts per article ===
{'article_id': 6, 'url': 'https://fakty.com.ua/ua/videos/pid-zagrozozyu-ridki', 'n_facts': 17}
{'article_id': 5, 'url': 'https://tsn.ua/ukrayina/u-mikolayevskiy-oblasti-vi', 'n_facts': 16}
{'article_id': 11, 'url': 'https://vitatv.com.ua/misto/pivdennoy-buh-u-tsentr', 'n_facts': 14}
{'article_id': 4, 'url': 'https://vsim.ua/Podii/ekologi-pripustili-chomu-v-p', 'n_facts': 13}
{'article_id': 3, 'url': 'https://vsim.ua/Podii/ekoaktivisti-hmelnichchini-b', 'n_facts': 12}
{'article_id': 1, 'url': 'https://www.facebook.com/permalink.php?story_fbid=', 'n_facts': 11}
{'article_id': 9, 'url': 'https://vitatv.com.ua/oblast/v-pivdenному-buzi-u-k', 'n_facts': 11}
{'article_id': 8, 'url': 'https://vitatv.com.ua/oblast/u-pivdenному-buzi-bil', 'n_facts': 9}
{'article_id': 7, 'url': 'https://vitatv.com.ua/misto/pivdennoy-buh-na-star', 'n_facts': 8}
{'article_id': 10, 'url': 'https://vitatv.com.ua/oblast/u-khmilnyku-v-pivden', 'n_facts': 5}
{'article_id': 12, 'url': 'https://vitatv.com.ua/misto/u-vinnytsi-vidnovly-r', 'n_facts': 0}

=== NULL-url articles ===
articles with NULL url: 0

```

Рисунок 4.7 – Фрагмент зрізу даних з роботою системи

4.3 Впровадження результатів дисертації у басейнове управління водних ресурсів

Створено аналітичну вебсистему WISEST-SBB («Water Information SystEm with Spatial and Temporal references for the Southern Bug Basin») для підтримки прийняття рішень під час супровіду імплементації Плану управління річковим басейном річки Південний Буг. Рішення розгорнуто за вебадресою <https://wisestr.ai>. Система охоплює 1091 масиви вод р. Південний Буг та включає:

- реєстр карток 1091 МВ з пошуком;
- автоматично згенеровані природномовні описи кожного МВ на основі табличних даних за матеріалами ПУРБ;
- авторизацію користувачів;
- автоматизований парсинг даних із 3-х типів джерел (пошук вебресурсів на базі Google Programmable Search Engine, завантаження файлів у файлів, інтеграцію з API провайдерами даних);
- видобування фактів як SPO-триплетів з просторово-часовою прив'язкою;
- класифікацію видобутих SPO-триплетів за категоріями «Стан», «Заходи», «Проблеми»;
- моніторинг роботи компонентів системи через Grafana stack.

Систему побудовано як модульний моноліт із набором слабозв'язаних мікросервісів у Docker-котейнерах (детальніше про це у п. 3.3).

У квітні 2026 року проведено демонстрацію роботи та тренінг з використання системи у БУВР р. Південний Буг, передано пароль для авторизованого доступу.

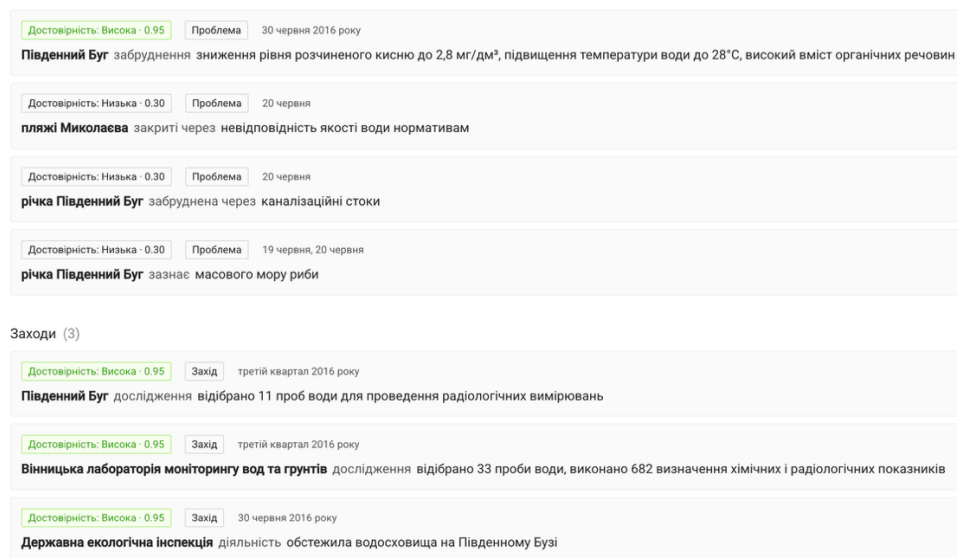


Рисунок 4.8 – Фрагмент класифікованих SPO-триплетів MB

Отримано акт впровадження про те, що результати цієї дисертаційної роботи рекомендовані до впровадження у практичну діяльність Басейнового управління водних ресурсів (БУВР) річки Південний Буг під час супроводу процесу імплементації програми заходів Плану управління річковим басейном (ПУРБ) Південного Бугу для аналізу даних моніторингу стану та для виявлення проблемних ситуацій, які потребують вирішення під час подальшого оновлення цього ПУРБ. Зокрема, цінними є такі отримані результати дослідження:

- метод верифікації числових даних у природномовних текстах з використанням бази знань та експертних оцінок, що дозволяє відібрати найбільш достовірну інформацію;
- метод формування знань та їх прив'язки до просторових об'єктів і часових меж, що надає можливість точніше знайти релевантну для користувача інформацію з урахуванням встановлених просторово-часових обмежень;
- метод геоприв'язки, який автоматично відбирає релевантні факти у текстах і більш точно зіставляє ці факти з описами масивів вод;
- інформаційна технологія, яка поєднує в собі усі ці методи та алгоритми, покращує якість і повноту знань у ГІС та автоматизує оброблення текстів;
- архітектура та тестовий вебінтерфейс аналітичної вебсистеми WISEST-SBB, що забезпечує доступ до узагальненої природномовної та числової

інформації за заданими категоріями за заданий період по кожному з 1091 масивів вод басейну річки Південний Буг з можливістю наповнення системи та її розвитку.

Отримані результати можуть бути поширені і на інші басейни великих річок, розташовані у Вінницькій області, зокрема — Дніпро і Дністер.

4.4 Впровадження розробленої технології у навчальний процес

Результати дослідження впроваджені у навчальний процес Вінницького національного технічного університету під час викладання дисципліни «Інтернет речей та інтелектуальний аналіз даних» для аспірантів, що навчаються за освітньо-науковою програмою «Інформаційні системи та технології» рівня «доктор філософії» спеціальності 126 «Інформаційні системи та технології», та дисципліни «Машинне навчання та інтелектуальний аналіз даних» для студентів, які навчаються за освітньо-професійною програмою «Інформаційні технології аналізу даних та зображень» рівня «магістр» спеціальності F6 «Інформаційні системи і технології» (акт впровадження ВНТУ від 27 квітня 2026 року).

Зокрема, при викладанні цих дисциплін використовуються такі результати досліджень, отримані у дисертаційній роботі здобувача:

1) метод верифікації числових даних у текстах з використанням бази знань, повнота якої оцінюється за допомогою критерію топологічної спостережуваності, що дає можливість оцінити повноту та точність видобутої інформації;

2) метод формування користувацьких запитів до великої мовної моделі та RAG-архітектури шляхом прив'язки ембеддінгів до просторових об'єктів і часових меж, що надає можливість більш точно знайти релевантну для користувача інформацію з урахуванням встановлених просторово-часових обмежень;

3) метод геоприв'язки, який автоматично відбирає релевантні факти у текстах і більш точно зіставляє ці факти з описами масивів вод;

4) інформаційна технологія, яка поєднує в собі усі ці методи та алгоритми, покращує якість і повноту знань у ГІС та автоматизує оброблення текстів, та її

вебреалізація у вигляді авторизованої системи аналітичної вебсистеми WISEST-SBB;

5) методи верифікації числових даних у текстах з використанням баз знань, повнота яких оцінюється за допомогою критерію топологічної спостережуваності;

6) удосконалений метод формування користувацьких запитів до великих мовних моделей на базі RAG-архітектури шляхом просторово-часової прив'язки ембеддінгів.

7) концепція та архітектура вебсистеми WISEST-SBB, що є практичним інструментом для засвоєння технологій автоматизованого скрапінгу та побудови інформаційних пайплайнів даних та ГІС-візуалізацій.

Використання зазначених результатів дозволило підвищити якість навчального процесу із згаданих дисциплін.

Висновки до розділу 4

У четвертому розділі продемонстровано апробацію результатів дослідження методу на фрагменті офіційного звіту про стан водних ресурсів та довіднику, де зазначені деякі масиви вод басейну річки Південного Бугу. Система автоматично сформувала необхідні SPO-триплети «факт-масив вод-інтервал часу», успішно прив'язавши кожний факт до конкретного МВ. Наприклад, для «р. Південний Буг, м. Хмільник» алгоритм знаходить частковий збіг у довіднику з офіційною назвою «Вінницька область, річка Південний Буг, місто Хмільник – код UA_M5.4_0011», що дає можливість однозначно прив'язати факт до цього масиву вод. Отже, система може видобувати екологічні факти з неструктурованих текстів.

Продемонстровано застосування технології на прикладі події «мор риби» у р. Південний Буг біля м. Ладизин у травні 2018 р. Запропонована технологія виконала геоприв'язку новин про подію до конкретного МВ, даних держ моніторингу водних ресурсів на посту поряд з с. Маньківка (концентрація

біогенних речовин «NO₃» та «PO₄» різко зросла, а кисню - різко впала у декілька разів, такі фактори дозволяють доволі впевнено прогнозувати мор риби). Біхроматичний граф моделі повністю топологічно спостережуваний. На основі різноманітних джерел про одну і ту саму подію на тому самому MB індекс достовірності події досяг 0,99. Загалом підсистема збирання даних WISEST-SBB зібрала приблизно 8000 посилань на 1200 унікальних вебдоменів.

Охарактеризовано створену аналітичну вебсистему WISEST-SBB («Water Information SystEm with Spatial and Temporal references for the Southern Bug Basin») для підтримки прийняття рішень під час супроводу імплементації Плану управління річковим басейном річки Південний Буг. Отримано акт впровадження про те, що результати цієї дисертаційної роботи рекомендовані до впровадження у практичну діяльність Басейнового управління водних ресурсів (БУВР) річки Південний Буг під час супроводу процесу імплементації програми заходів Плану управління річковим басейном (ПУРБ) Південного Бугу для аналізу даних моніторингу стану та для виявлення проблемних ситуацій, які потребують вирішення під час подальшого оновлення цього ПУРБ.

Результати дослідження впроваджені у навчальний процес Вінницького національного технічного університету під час викладання дисципліни «Інтернет речей та інтелектуальний аналіз даних» для аспірантів, що навчаються за освітньо-науковою програмою «Інформаційні системи та технології» рівня «доктор філософії» спеціальності 126 «Інформаційні системи та технології», та дисципліни «Машинне навчання та інтелектуальний аналіз даних» для студентів, які навчаються за освітньо-професійною програмою «Інформаційні технології аналізу даних та зображень» рівня «магістр» спеціальності F6 «Інформаційні системи і технології».

Результати, наведені у цьому розділі, опубліковані у працях [4, 6, 14, 17, 20, 22] здобувача.

ВИСНОВКИ

У дисертаційній роботі розв'язано актуальну науково-прикладну задачу підвищення обсягів та достовірності інформації про стан масивів вод річкових басейнів з використанням інформаційної технології автоматизованого видобування та верифікації даних з гетерогених україномовних природомовних джерел.

Вперше розроблено інформаційну технологію видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою, яка поєднує автоматизоване видобування фактів у форматі SPO-триплетів «subject-predicate-object» з ієрархічним успадкуванням просторово-часового контексту із заголовків документа та інших метаданих, формалізацію річкової мережі у вигляді багатозв'язної геоінформаційної системи іменованих сутностей (БГСІС) та верифікацію числових даних за критерієм топологічної спостережуваності бази знань, що дало змогу підвищити точність прив'язки фактів до конкретних масивів вод до рівня 85-99%. Удосконалено метод формування користувацьких запитів до LLM та RAG шляхом прив'язки ембеддінгів до просторових об'єктів і часових метаданих. Цей метод дозволяє достовірніше знайти релевантну інформацію для користувача з урахуванням заданих просторово-часових обмежень.

Набув подальшого розвитку метод автоматизованої геоприв'язки, що автоматично відбирає релевантні факти (триплети) у текстах і більше точно зіставляє ці факти з описами масивів вод. Створено інформаційну технологію, що об'єднала всі наведені методи та алгоритми, автоматизувала оброблення текстів та покращує якість і повноту знань про конкретні МВ.

Отримані результати демонструють, що для подій можливо знайти різноманітні розрізнені підтвердження, завдяки чому значно збільшується індекс достовірності описаних подій (за наявності достатньої кількості інформації про подію – майже до 0,99). Тобто, можливо оцінити достовірність кожної окремої

новини. Модулі системи, що відповідають за автоматизоване збирання даних, зібрали приблизно 8000 посилань з приблизно 1200 вебдоменів.

Підготовлена архітектура каскадних парсерів гетерогенних джерел трьох типів, яка включає в себе унікально налаштовані конвеєри. Також проведено нормалізацію та уніфікацію даних, які використовуються.

Апробація представлена підсистемою спостереження та збирання даних WISEST-SBB. Розроблена технологія автоматизованого збирання та геолокації текстової інформації з альтернативних джерел дозволяє успішно прив'язувати ці дані до конкретних масивів вод на основі ключових слів та характеристик.

Розроблену технологію рекомендовано до впровадження у практичну діяльність Басейнового управління ресурсів річки Південний Буг, для супроводу реалізації Плану управління річковим басейном Південного Бугу та впроваджено у навчальний процес Вінницького національного технічного університету, що підтверджено відповідними актами впровадження.

Це збільшило обсяг даних для аналізу та узагальнення, підвищуючи ефективність контролю реалізації заходів ПУРБ і створюючи основу для поширення досвіду та виявлених проблем між подібними масивами вод.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Information measuring systems with mobile devices for identification of air pollution parameters caused by transport / Vitalii B. Mokin; Georgii V. Goriachev; Dmytro Y. Dziuniak; Konstantin O. Bondaletov; Serhii O. Zhukov; Mariusz Duk ; Saltanat Sailarbek. *Proc. SPIE 10031, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments* 2016, 1003128 (September 28, 2016), 8 pages; doi:10.1117/12.2249202.
2. Vitalii B. Mokin. Method for Analyzing and Optimizing the Topologic Observability of Cognitive Maps of Complex Spatially Distributed Systems / Vitalii Mokin, Ilona Varchuk, Kostiantyn Bondaletov, Olena Slobodianiuk. *CONFERENCE PROCEEDINGS 2017 IEEE First Ukraine Conference on ELECTRICAL AND COMPUTER ENGINEERING (UKRCON)* May 29 – June 2, 2017 Kyiv. – P.: 865-869. – doi 10.1109/UKRCON.2017.8100371.
3. Bondalietov, K., Mokin, V. (2023). Notation System for Comparing and Synthesis of Intelligent Key Phrase Extraction Methods for Ontological Models in Information Systems. In: Dovgyi, S., Trofymchuk, O., Ustimenko, V., Globa, L. (eds) *Information and Communication Technologies and Sustainable Development. ICT&SD 2022. Lecture Notes in Networks and Systems*, vol 809. Springer, Cham. https://doi.org/10.1007/978-3-031-46880-3_11.
4. Мокін В. Б., Бондалєтов К. О., Крижановський Є. М., і Караваєв В. О. Метод аугментації текстів про стан масивів вод на основі інтелектуальної прив'язки до багатозв'язних геоінформаційних систем іменованих сутностей, *Вісник Вінницького політехнічного інституту*, вип. 3, с. 55–65, Черв. 2023. URL: <https://doi.org/10.31649/1997-9266-2023-168-3-55-65>.
5. Мокін В. Б., Бондалєтов К. О., Штельмах І. М., Слободянюк О. В. Автоматичне видобування знань з екологічних звітів з прив'язкою до часу та до просторових координат масивів вод, *Вісник Вінницького політехнічного інституту*, вип. 3, с. 101–110, Черв. 2025, <https://doi.org/10.31649/1997-9266-2025-180-3-101-110>.

6. К. О. Бондалетов і В. Б. Мокін, «Інтелектуальна технологія видобування та верифікації числових та текстових даних у багатозв'язних багатостадійних геоінформаційних системах», *Вісник Вінницького політехнічного інституту*, вип. 5, с. 51–60, Жовт. 2025. Режим доступу: <https://visnyk.vntu.edu.ua/index.php/visnyk/article/view/3335>.

7. Мокін В. Б. Універсальна інформаційно-вимірювальна система оперативного екологічного моніторингу з використанням мобільних пристроїв / В. Б. Мокін, К. О. Бондалетов, Г. В. Горячев, Д. Ю. Дзюняк. *Вісник Вінницького політехнічного інституту*. — Вінниця. — 2015. — № 5 (122). — С. 116-122. <https://visnyk.vntu.edu.ua/index.php/visnyk/article/view/1836>.

8. Метод і технологія моніторингу стану атмосферного повітря за допомогою універсальної інформаційно-вимірювальної системи з використанням мобільних пристроїв / В. Б. Мокін, Д. Ю. Дзюняк, К. О. Бондалетов, В. В. Олійник. *Наукові праці Вінницького національного технічного університету* [Електронний ресурс]. — 2015. — № 4. — Режим доступу: <http://praci.vntu.edu.ua/index.php/praci/article/view/456/454>.

9. Бондалетов К. О. Практична реалізація сучасної мобільної системи веб-моніторингу та експрес-аналізу стану повітряного середовища / К. О. Бондалетов, В. Б. Мокін, Г. В. Горячев, Д. Ю. Дзюняк. XLIV Регіональна науково-технічна конференція професорсько-викладацького складу, співробітників та студентів університету з участю працівників науково-дослідних організацій та інженерно-технічних працівників підприємств м. Вінниці та області, електронне наукове видання матеріалів конференції, м. Вінниця, 11—13 березня 2015. — Режим доступу: <http://conf.vntu.edu.ua/allvntu/2015/ineek/txt/bondaletov.pdf>

10. Бондалетов К. О. Мобільна аналітична комп'ютерна система для оперативного моніторингу стану атмосферного повітря міста / К. О. Бондалетов, Д. Ю. Дзюняк, В. Б. Мокін. Молодь в технічних науках: дослідження, проблеми, перспективи : матеріали Міжнародної науково-практичної Інтернет-конференції, 23-26 квітня 2015 року / ВНТУ. — Вінниця, 2015. — С. 76-77.

11. Мокін В. Б. Інформаційно-вимірювальні системи для ідентифікації параметрів моделі забруднення атмосферного повітря автотранспортом з використанням мобільних пристроїв та фотометричних, у т.ч. лазерних, пристроїв і систем / В. Б. Мокін, Г. В. Горячев, Д. Ю. Дзюняк, К. О. Бондалєтов. Зб. тез доп. III Міжнародної конференції «Вимірювання, контроль та діагностика в технічних системах (ВКДТС-2015)», м. Вінниця, 27-29 жовтня 2015 р. – Вінниця : ВНТУ, 2015. — С. 130-132.

12. Мокін В. Б. Технологія оперативного екологічного моніторингу стану повітряного середовища та пересувних джерел його забруднення / В. Б. Мокін, Д. Ю. Дзюняк, К. О. Бондалєтов. XIV Міжнародна науково-практична конференція: Сучасні інформаційні технології управління екологічною безпекою, природокористуванням, заходами в надзвичайних ситуаціях: 5-9 жовтня 2015 р. – К., 2015. – С. 126-130.

13. Мокін В. Б. Практична реалізація мобільної аналітичної комп'ютерної системи моніторингу стану атмосферного повітря з підсистемою веб-аналізу та виведення даних на геопортал / В. Б. Мокін, Д. Ю. Дзюняк, К. О. Бондалєтов, Г. В. Горячев, В. В. Боцян. V-й Всеукраїнський з'їзд екологів з міжнародною участю (Екологія/Ecology-2015), 23-26 вересня, 2015. Збірник наукових праць. – Вінниця: ТОВ «Нілан-ЛТД», 2015. – С. 106.

14. Бондалєтов К. О., Мокін В.Б., Крижановський Є.М., Караваєв В.О. Розроблення інтелектуальної веб-системи з інформацією про екологічні проблеми та природоохоронні заходи у масивах вод басейну р. Південний Буг WISEST-SBB / Актуальні проблеми та перспективи розвитку фундаментальних, прикладних, загальнотехнічних та безпекових наук : Матеріали IV всеукр. науково-практ. конф. пам'яті акад. Акад. наук вищ. освіти, проф. А.В. Касперського, м. Київ, 27 черв. 2024 р. С. 6–12. URL: <https://kztdop.ipf.npu.edu.ua/?view=article&id=244>

15. Мокін В.Б., Бондалєтов К.О. Інтелектуальні методи видобування ключових словосполучень із тексту для побудови онтологічних моделей інформаційно-пошукових систем. С. 197-199. – Режим доступу: https://itgip.org/wp-content/uploads/2022/11/tezy_70.pdf. Інформаційно-

комунікаційні технології та сталий розвиток: колективна монографія за матеріалами XXI Міжнародної науково-практичної конференції (14-16 листопада 2022 р.). Київ, 2022. ТОВ «Видавництво «стон», 2022, 242. URL: <https://itgip.org/kolektivna-monografiya-za-materialam>

16. Бондалетов К. О., Мокін В. Б. Інтелектуальна автоматизація геоприв'язки повідомлень з соцмереж до масивів вод за допомогою зваженої JACCARD-міри. ВІТКП ВНТУ. Факультет інтелектуальних інформаційних технологій та автоматизації (2025): вебресурс (дата звернення 13 червня 2025): <https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2025/paper/view/23298/19275>.

17. Бондалетов К. О., Мокін В. Б., Григорчук М. В., Джура С. В. та ін. Побудова датасету для тренування інтелектуальних моделей веб-системи з інформацією про екологічні проблеми та заходи у масивах вод басейну р. Південний Буг WISEST-SBB. Матеріали LIІ Науково-технічної конференції факультету інтелектуальних інформаційних технологій та автоматизації Вінницького національного технічного університету, Вінниця, 21 – 23 червня 2023 р. – Електрон. текст. дані. – 2023. – Режим доступу: <https://d.conf.vntu.edu.ua/index.php/all-fksa/all-fksa-2023/paper/view/18958>

18. Мокін В. Б., Бондалетов К. О. Новий підхід до автоматизації оптимального управління водогосподарським балансом басейну річки на основі когнітивних карт. X Міжнародна науково-практична конференція «Інтернет-Освіта-Наука (ІОН-2016)», Вінниця, Україна, 11–14 жовт. 2016. Вінниця: ВНТУ, 2016. С. 101–103. ISBN 978-966-641-646-2. URL: <https://ies.vntu.edu.ua/reports/program/WORK-IES-2016.pdf> (дата звернення: 08.07.2025).

19. Бондалетов К. О., Мокін В. Б. Розроблення просторово-зваженого критерію геоприв'язки природномовних текстів до масивів вод. Молодь в науці: дослідження, проблеми, перспективи (МН-2025): матеріали Всеукраїнської науково-практичної інтернет-конференції (Вінниця, 15 жовтня 2024 року – 15 червня 2025 року). — Вінниця : ВНТУ, 2025. — С. 1597-1599.

20. Бондалетов К. О.; Мокін В. Б. Автоматизація геолокації текстової інформації для контролю ефективності заходів з управління річковим басейном: Контроль і управління в складних системах (КУСС-2024), Вінниця, Україна, 16–17 жовтня 2024. с. 99–101.

21. Підвищення точності моделювання забруднення атмосферного повітря міста з використанням мобільних інформаційно-вимірювальних систем [Електронний ресурс] / В. Б. Мокін, Г. В. Горячев, Д. Ю. Дзюняк, К. О. Бондалетов. Матеріали XLV Науково-технічної конференції Вінницького національного технічного університету, Вінниця, 23-24 березня 2016 р. - Електрон. текст. дані. - 2016. - Режим доступу : <http://conferences.vntu.edu.ua/index.php/all-ebmd/all-ebmd-2016/paper/view/465>.

22. Бондалетов, Костянтин, Мокін, Віталій. "Архітектура каскадних парсерів гетерогених джерел даних для автоматизованої побудови датасетів для інтелектуальних систем моніторингу" Науково-технічної конференції факультету інтелектуальних інформаційних технологій та автоматизації Вінницького національного технічного університету, Вінниця, 24-27 березня 2026 р. - Електрон. текст. дані. - 2026. - Режим доступу <https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2026/paper/view/28287/23206>

23. Berko, A., Chyrun, L., Prokipchuk, O., Chyrun, S., Panasyuk, V., and Luchkevych, M., "Intelligent Analysis of Ukrainian-Language Tweets of Regional Public Opinion Study and Management", in 2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT), 2023, Art. no. 67. doi:10.1109/CSIT61576.2023.10324107.

24. Барановська А. Ю. Підходи та методи застосування технологій штучного інтелекту для розв'язання задач обробки текстової інформації [Електронний ресурс] / А. Ю. Барановська, М. А. Лешок, О. В. Бісікало. Матеріали Всеукраїнської науково-практичної інтернет-конференції «Молодь в науці: дослідження, проблеми, перспективи (МН-2022)», Вінниця, 16-17 червня 2022 р.

- Електрон. текст. дані. – 2022. – Режим доступу: <https://conferences.vntu.edu.ua/index.php/mn/mn2022/paper/view/15365>
25. Кудрик, О.В., Бісікало, О.В. і Здітовецький, Ю.С. 2024. Методи тонкого налаштування штучного інтелекту. *Вісник Вінницького політехнічного інституту*. № 4 (Серп. 2024), 139–146. DOI: <https://doi.org/10.31649/1997-9266-2024-175-4-139-146>.
26. A. Halterman «Mordecai 3: A Neural Geoparser», *arXiv*, 2023, [Electronic resource]. Available: <https://arxiv.org/pdf/2303.13675>
27. Wang, Shengyuan, et al. "Mitigating Geospatial Knowledge Hallucination in Large Language Models: Benchmarking and Dynamic Factuality Aligning." Findings of the Association for Computational Linguistics: EMNLP-2025. 2025.
28. Водний кодекс України : Кодекс України від 06.06.1995 р. № 213/95-ВР : станом на 19 серп. 2022 р. URL: <https://zakon.rada.gov.ua/laws/show/213/95-вр> (дата звернення: 15.04.2025).
29. Про затвердження Порядку розроблення плану управління річковим басейном : Постанова Каб. міністрів України від 18.05.2017 р. № 336. URL: <https://www.kmu.gov.ua/npas/249999756> (дата звернення: 15.04.2025).
30. Водна стратегія України на період до 2050 року. Розпорядженням Кабінету Міністрів України від 9 грудня 2022 р. № 1134-р. [Електронний ресурс]. Режим доступу: <https://zakon.rada.gov.ua/laws/show/1134-2022-%D1%80#Text>.
31. Розпорядження Кабінету Міністрів України від 1 листопада 2024 р. № 1078-р «Про затвердження планів управління річковими басейнами Вісли, Дністра та Південного Бугу на 2025—2030 роки.» <https://zakon.rada.gov.ua/laws/show/1078-2024-p>.
32. Кавац Ю. В., Сергєєва К. Л., Фененко Т. М., Дібрій Д. А. Оцінка стану забруднення водних об'єктів на основі спектральних індексів. Інформаційні технології в металургії та машинобудуванні – ІТММ'2025 : тези доп. Міжнародної наук.-техн. конф. (м. Дніпро, 23-24 березня 2025 р.). Дніпро, 2025, С. 546–550. DOI: 10.34185/1991-7848.itmm.2025.01.097.

33. Europe's state of water 2024: the need for improved water resilience. European Environment Agency (EEA). URL: <https://www.eea.europa.eu/en/analysis/publications/europes-state-of-water-2024>
34. Report from the Commission to the Council and the European Parliament on the implementation of the Water Framework Directive (2000/60/EC) and the Floods Directive (2007/60/EC). Third river basin management plans. Second flood risk management plans : COM(2025) 2 final / European Commission. Brussels, 4.02.2025. 40 p. URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=COM:2025:2:FIN>
35. Стрижак О. Є. Онтологічні інформаційно-аналітичні системи / Олександр Євгенійович Стрижак. *Радіoeлектронні і комп'ютерні системи*. 2014, № 3, С. 71-76.
36. О. Сілагін, Є. Сілагін, В. Денисюк, А. Денисюк, «Розробка онтологічної моделі бази знань «Бібліотека» на базі середовища Protégé», *Інформаційні технології та комп'ютерна інженерія*, вип. 58, вип. 3, с. 12–21, Груд 2023.
37. Інформаційна технологія підтримки функції «запитання-відповідь» на основі образного аналізу фахових текстів : монографія [Електронний ресурс] / А. І. Лісовенко, О. В. Бісікало. – Вінниця : ВНТУ, 2019. – 180 с. ISBN 978-966-641-764-3.
38. LIAAD Yet Another Keyword Extractor (Yake) / Laboratory of Artificial Intelligence and Decision Support. — <https://github.com/LIAAD/yake>
39. Комп'ютерні онтології та їх використання у навчальному процесі. Теорія і практика: монографія / [С. О. Довгий, В. Ю. Велічко, Л. С. Глоба, О. Є. Стрижак та ін.] ; Нац. акад. пед. наук України, Ін-т обдаров. дитини. – Київ: Інститут обдарованої дитини, 2013. – 310 с. URL: <https://lib.iitta.gov.ua/id/eprint/10124>.
40. Batiuk, T., & Dosyn , D. . (2025). Natural language processing of social media text data using BERT and xgboost. *Radio Electronics, Computer Science, Control*, (2), 154–167. <https://doi.org/10.15588/1607-3274-2025-2-14>

41. A. Subakti, H. Murfi, and N. Hariadi, "The performance of BERT as data representation of text clustering," *J. Big Data*, no. 9, 15, 2022. <https://doi.org/10.1186/s40537-022-00564-9>
42. W. Liu, et al., "K-BERT: Enabling Language Representation with Knowledge Graph," in *Proceedings of the AAAI Conference on Artificial Intelligence*, no. 34 (03), 2020, pp. 2901-2908. <https://doi.org/10.1609/aaai.v34i03.5681>
43. M. Grootendorst. Keyword Extraction with BERT. Towards Data Science. — 2022 — <https://towardsdatascience.com/keyword-extraction-with-bert-724efca412ea>
44. М. О. Литвин і Л. М. Олещенко, «Автоматизований підхід до датування англомовного тексту з використанням трансформерних нейронних мереж», *Вісник Вінницького політехнічного інституту*, вип. 3, с. 133–139, Черв. 2025.
45. Grootendorst M. KeyBERT. Github. MaartenGr. 07.02.2025. URL: <https://github.com/MaartenGr/KeyBERT>.
46. Oleh Bisikalo, Alexander Yahimovich. Keyword search based on lexical relationships in the text. — Mauritius : Lap Lambert Academic Publishing, 2019. — 57 p. — ISBN 978-620-0-00314-0.
47. Mao, Yuning, et al. "Multi-document summarization with maximal marginal relevance-guided reinforcement learning." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020. DOI: <https://doi.org/10.48550/arXiv.2010.00117>
48. Гідрографічна мережа. Басейнове управління водних ресурсів річки Південний Буг. Водні ресурси. URL: <https://buvrpb.davr.gov.ua/vodni-resursy/hidrografichna-merezha>.
49. Юхимчук М. С., Горчук Ю. А., Дубовой В. М. Покращення прийняття рішень в управлінні бізнес-процесами за допомогою предикативної аналітики на основі штучного інтелекту. Матеріали XVII Міжнародної конференції «Контроль і управління в складних системах (КУСС-2024)», м. Вінниця, 16-17 жовтня 2024

р. Електрон. текст. дані. 2024. URI:

<https://conferences.vntu.edu.ua/index.php/mccs/mccs2024/paper/viewFile/22082>.

50. Chai CP. Comparison of text preprocessing methods. *Natural Language Engineering*, 2023; 29(3): 509-553. doi:10.1017/S1351324922000213.

51. В. Ковенко, І. Богач, і М. Барабан, «Колаборативна фільтрація на основі каталогу з використанням NLP технології», *Інформаційні технології та комп'ютерна інженерія*, вип. 51, вип. 2, С. 17–22, Жов 2021.

52. Войцеховський В., Кулик Я. Аналіз методів та підходів у використанні технологій штучного інтелекту для обробки текстової інформації на прикладі мовної моделі PaLM. Матеріали Міжнародної науково-практичної інтернет-конференції «Молодь в науці: дослідження, проблеми, перспективи (МН-2024)», Вінниця, Україна, 15–20 черв. 2024. ВНТУ, 2024. С. 995–1001. URL: <https://press.vntu.edu.ua/index.php/vntu/catalog/view/836/1459/2733-1>

53. J. Dagdelen et al., «Structured information extraction from scientific text with large language models», *Nature Commun*, 15, 1418, 2024, [Electronic resource]. Available: <https://doi.org/10.1038/s41467-024-45563-x>.

54. Бісікало О., Урлапова Д., Телега Д. Класифікація жартів за категоріями гумору з використанням методів машинного навчання. Міжнародна науково-практична інтернет-конференція студентів, аспірантів та молодих науковців «Молодь в науці: дослідження, проблеми, перспективи» (МН-2025), Вінниця, Україна, 15 жовт. 2024–черв. 2025. ВНТУ, 2025. URL: <https://ir.lib.vntu.edu.ua/bitstream/handle/123456789/48246/25576.pdf?sequence=3&isAllowed=y>

55. TF-IDF Character N-grams versus Word Embedding-based Models for Fine-grained Event Classification: A Preliminary Study (Piskorski & Jacquet, AESPEN 2020). URL: <https://aclanthology.org/2020.aespen-1.6/>

56. Entropy-extreme concept of data gaps filling in a small-sized collection / V. Kovtun et al. *Egyptian Informatics Journal*, 2025. Vol. 29. P. 100621. URL: <https://doi.org/10.1016/j.eij.2025.100621>

57. Corina Florescu and Cornelia Caragea. 2017. PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1105–1115, Vancouver, Canada. Association for Computational Linguistics.

58. Довгань І. Я., Альфавіцький М. А., Іваніцький Л. Б., Кушнір О. С., Яремків В. В. Розпізнавання текстового плагіату на основі векторної моделі для слів і символічних n-грам. Матер. IX Українсько-Польської наук.-практ. конф. «Електроніка та інформаційні технології» (ЕЛІТ-2017) (Львів–Чинадієво, Україна, 2017 р.). – Львів : Видавн. Львів. ун-ту, 2017. – С. 44–48. URL: https://www.researchgate.net/publication/320290148_ROZPIZNAVANNA_TEKSTOVOGO_PLAGIATU_NA_OSNОВI_VEKTORNOI_MODELI_DLA_SLIV_I_SIMVOLNIH_N-GRAM Recognition of textual plagiarism based on the vector model for words and symbolic n-grams in Ukrainian

59. Vitalii Mokin, Kostyantyn Bondaletov. Kaggle Notebook “NLP for UA : KeyBERT - keywords extractions”, 2023. — <https://www.kaggle.com/code/vbmokin/nlp-for-ua-keybert-keywords-extractions>

60. Dočekal, M., & Smrž, P. (2022). Query-Based Keyphrase Extraction from Long Documents. The International FLAIRS Conference Proceedings, 35. <https://doi.org/10.32473/flairs.v35i.130737>

61. Нич, А., & Праворська, Н. (2026). RAG (retrieval-augmented generation) як нова парадигма корпоративної автоматизації. *Herald of Khmelnytskyi National University. Technical Sciences*, 361(1), 375-382. <https://doi.org/10.31891/2307-5732-2026-361-53>

62. Non-relational approach to developing knowledge bases of expert system prototype. Hnatushenko, V. V.; Hnatushenko, Vik. V.; Dorosh, N. L.; Solodka, N. O.; Liashenko, O. A. *Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu*. Dnipropetrosk Iss. 2, (2022): 112-117. DOI:10.33271/nvngu/2022-2/112.

63. В. Б. Мокін, А. М. Лучко, Давидюк О. М. і Т. Є. Вуж, «Технологія побудови експертної інформаційної веб-системи виявлення та верифікації пріоритетних екологічних проблем у масивах вод басейну річки», *Вісник Вінницького політехнічного інституту*, вип. 1, с. 77–87, Лют. 2021.

64. Гребінь Василь Васильович. Методики гідрографічного та водогосподарського районування території України відповідно до вимог Водної рамкової директиви Європейського Союзу / В. В. Гребінь, В. Б. Мокін, В. А. Сташук, В. К. Хільчевський та ін. – К.: Інтерпрес ЛТД, 2013. – 63 с. (Фінансування за підтримки проекту ЄС).

65. Наукові засади раціонального використання водних ресурсів України за басейновим принципом : монографія / За редакцією В. А. Сташука; [В. А. Сташук, В. Б. Мокін, В. В. Гребінь, О. В. Чунарьов]. — Херсон : Грінь Д. С., 2014. — 320 с.

66. Раціональне використання та відновлення водних ресурсів : монографія / М.О. Клименко, В. Б. Мокін, І. І. Овчаренко, А. М. Прищепа та ін. / За заг. ред. Фещенка В.П. – Житомир: Вид-во ЖДУ ім. І. Франка, 2016. – 250 с.

67. Wójcik Waldemar, Kalizhanova Aliya, Kulyk Yaroslav A., Knysh Bohdan P, Kvyetnyy Roman N, Kulyk Anatoliy I., Sichko Tetiana V, Dumenko Viktoria P., Bezstmertna Oksana V, Adikhanova Saltanat, Zhassandykyzy Maral, Junisbekov Mukhtar, Smailov Nurzhigit, Yussupova Gulbahar. The Method of Time Distribution for Environment Monitoring Using Unmanned Aerial Vehicles According to an Inverse Priority. *Journal of Ecological Engineering*. 2022. С. 179–187. URL: <https://doi.org/10.12911/22998993/153458>.

68. М. С. Юхимчук, В. О. Лесько, В. М. Дубовой, і Ю. Ю. Іванов, «Інтелектуальна система автоматичного керування процесом сушіння зернових культур на основі IoT-технологій», *Наукові праці Вінницького національного технічного університету*, вип. 4, Груд 2025.

69. Kovtun, V.; Zaitseva, E.; Levashenko, V.; Grochla, K.; Kovtun, O. Small Stochastic Data Compactification Concept Justified in the Entropy Basis. *Entropy*, 2023, 25, 1567. <https://doi.org/10.3390/e25121567>

70. Hanwen Zheng et al., «A Comprehensive Survey on Document-Level Information Extraction», In Proceedings of the Workshop on the Future of Event Detection (FuturED), 2024, p. 58–72, USA: Association for Computational Linguistics, [Electronic resource]. Available: <https://aclanthology.org/2024.futured-1.6.pdf>.
71. Haoran Luo et al., «Text2NKG: Fine-Grained N-ary Relation Extraction for N-ary relational Knowledge Graph Construction», *Advances in Neural Information Processing Systems*, 37 (NeurIPS), 2024. [Electronic resource]. Available: <https://doi.org/10.48550/arXiv.2310.05185> (date of access: 06.06.2025)
72. R. Bommasani et al. «On the Opportunities and Risks of Foundation Models», *Computer Science, Machine Learning*, 2021. [Electronic resource]. Available: <https://arxiv.org/abs/2108.07258>.
73. Yunyi Zhang, Ming Zhong, Siru Ouyang, Yizhu Jiao, Sizhe Zhou, Linyi Ding, and Jiawei Han. 2024. Automated Mining of Structured Knowledge from Text in the Era of Large Language Models. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24). Association for Computing Machinery, New York, NY, USA, 6644–6654. <https://doi.org/10.1145/3637528.3671469>
74. Use of geo-information systems for groundwater monitoring / Т. Hrebeniuk та ін. *Енергетика: економіка, технології, екологія*. 2024. № 4. 2024. URL: <https://doi.org/10.20535/1813-5420.4.2024.315602>.
75. Дадиверін В., Бісікало О. Порівняльний аналіз NLP-бібліотек NLTK та spaCy. 2025: Інформаційні технології та комп'ютерне моделювання: матеріали Міжнародної науково-практичної конференції. Івано-Франківськ, 2025. С. 115–117. URL: <https://journal.comp-sc.if.ua/test/index.php/ITCM/article/view/658/636>.
76. Vitalii Mokin, Kaggle Notebook «NLP for UA : BERT CLS & 10 Classifiers», 2022: [Electronic resource]. Available: <https://www.kaggle.com/code/vbmokin/nlp-for-ua-bert-cls-10-classifiers>. Access: 07.06.2025.
77. V. Mokin, D. Pasichniuk, O. Radetskyi, and M. Horash, Kaggle Dataset «NLP: Reports & News Classification. ENG & UKR Automatic Environmental Reports

& News Classification», 2020. [Electronic resource]. Available: <https://www.kaggle.com/vbmokin/nlp-reports-news-classification>.

78. Бісікало О. В., Дадиверін В. В. Доповнення та поглиблення методу пошуку ключових слів. Матеріали Всеукраїнської науково-практичної інтернет-конференції «Молодь в науці: дослідження, проблеми, перспективи (МН-2025)», Вінниця, 15-16 червня 2025 р. Електрон. текст. дані. 2025. URI: <https://conferences.vntu.edu.ua/index.php/mn/mn2025/paper/view/23131>.

79. O. Bisikalo, O. Kovtun and V. Kovtun, "Neural Network Concept of Ukrainian-Language Text Embedding," 2023 13th International Conference on Advanced Computer Information Technologies (ACIT), Wrocław, Poland, 2023, pp. 566-569, doi: 10.1109/ACIT58437.2023.10275511.

80. Плани управління річковими басейнами України 2025–2030 / Blue Rivers екологічний консалтинг ; за фінансової підтримки Європейського Союзу в рамках програми EU4Environment «Водні ресурси та екологічні дані» ; Міністерство захисту довкілля та природних ресурсів України ; Державне агентство водних ресурсів України. — Київ, 2025. — URL: https://eu4waterdata.eu/images/pdf/Summary%20factsheets/ukrainian/Full_Pack_UK_R_compressed_reduced.pdf

81. D. Dessí et al., «CS-KG 2.0: A Large-scale Knowledge Graph of Computer Science», *Scientific Data*, 12, 964, 2025, [Electronic resource]. Available: <https://doi.org/10.1038/s41597-025-05200-8>.

82. Content Monitoring and Big Data Analysis of Smart City Residents Appeals and Complaints Management / A. Y. Berko та ін. 2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT). 2023. URL: <https://doi.org/10.1109/CSIT61576.2023.10324159>.

83. Водна рамкова директива ЄС 2000 / 60 / ЄС. Основні терміни та їх визначення. Київ, 2006. 240 с.

84. Directive 2000/60/EC of the European Parliament and of the Council. EUR-Lex - Access to European Union Law. URL: <https://eur->

lex.europa.eu/resource.html?uri=cellar:5c835afb-2ec6-4577-bdf8-756d3d694eeb.0004.02/DOC_1&format=PDF.

85. Vitalii Mokin, Kaggle Notebook “NLP for WR: Summarizing using BERT, GPT2, XLNET”, 2022: [Electronic resource]. Available: <https://www.kaggle.com/code/vbmokin/nlp-for-wr-summarizing-using-bert-gpt2-xlnet>. Access: 07.06.2025.

86. Геряк, Юрій & Берко, Андрій. (2025). Оцінка семантичної та часової узгодженості розподілених інформаційних ресурсів. *Modeling, Control and Information Technologies*. С. 125-128. DOI 10.31713/MCIT.2025.036.

87. Інформаційні технології автоматизації обробки параметрів геоінформаційних систем з геометричними мережами : монографія / В. Б. Мокін, В. Г. Сторчак, Є. М. Крижановський та ін. – Вінниця: ВНТУ, 2014. – 194 с. ISBN 978-966-641-578-6

88. Аналіз методик автоматизованого картування змін водойм / Кавац О. О., Кавац Ю. В., Сергєєва К. Л., Рой Д. М. *Системні технології*. Том 5 № 160, 2025. URL: <https://doi.org/10.34185/1562-9945-5-160-2025-17>.

89. Zuievskia N., Hrebeniuk T., Zuievskyi Y. Application of GIS for analysis of environmental aspects of mineral deposits. 2023. URL: <https://doi.org/10.33271/crpnmu/75.007>.

90. О. М. Данильчук, В. В. Ковтун, О. Д. Никитенко, Ю. Ю. Нестюк, і В. В. Присяжнюк, «Елементи методології прецизійного фонетичного аналізу фонограм усного мовлення», *Вісник Вінницького політехнічного інституту*, вип. 3, с. 36–51, Черв. 2022.

91. Іванов, Д., Каштан, В., & Гнатушенко, В. (2024). Методологія цифрового картографування виникнення можливих затоплень територій із застосуванням напівкерованого машинного навчання. *Herald of Khmelnytskyi National University. Technical Sciences*, 341(5), 12-19. <https://doi.org/10.31891/2307-5732-2024-341-5-1>

92. Vitalii Mokin, and Kostiantyn Bondaletov, Kaggle Notebook “SpaCy for Ukrainian text similarity”, 2023: [Electronic resource]. Available:

<https://www.kaggle.com/code/bondaletov/spacy-for-ukrainian-text-similarity>. Access: 07.06.2025.

93. П. В. Здебський і А. Ю. Берко, «Перевірка тексту після генерації за допомогою великих мовних моделей для фільтрування неправильних відповідей», *Наукові праці Вінницького національного технічного університету*, вип. 1, Бер 2024. DOI: <https://doi.org/10.31649/2307-5376-2024-1-1-6>

94. Інформаційна технологія аналізу та оптимізації топологічної спостережуваності багатозв'язних геоінформаційних систем: монографія [Електронний ресурс] / В. Б. Мокін, І. В. Варчук, Є. М. Крижановський. – Вінниця : ВНТУ, 2019. – 121 с ISBN 978-966-641-765-0

95. Gelman A., Carlin J. B., Stern H. S., Dunson D. B., Vehtari A., Rubin D. B. Bayesian data analysis (Third edition). 2013; URL: <https://sites.stat.columbia.edu/gelman/book/>

96. Hnatushenko V., Pavlenko I. The use of generative artificial intelligence in software testing. *Системні технології*, Том 2, № 151, 2024. URL: <https://doi.org/10.34185/1562-9945-2-151-2024-10>.

97. Bondaletov K. Kaggle Notebook «Collecting Google Search Data with BeautifulSoup», 2024. – Режим доступу <https://www.kaggle.com/code/bondaletov/collecting-google-search-data-with-beautifulsoup>. Дата звернення 07.08.2024.

98. План управління річковим басейном Південного Бугу на 2025-2030 роки. / Затв. розпорядженням КМУ від 1.11.2024р №1078-р / Державне агентство водних ресурсів України. Київ. URL: https://davr.gov.ua/fls18/PURBPivdenyyBuh7_ukr.pdf

99. P. Zdebskyi, A. Berko and V. Vysotska, "Information Technology for Finding Answers to Questions from Open Web Resources," 2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT), Lviv, Ukraine, 2023, pp. 1-7, doi: 10.1109/CSIT61576.2023.10324087.

100. Супутниковий моніторинг та оцінка ступеня забрудненості водою під впливом військових дій / О. Кавац та ін. *Системні технології*, Том 4 № 159, 2025. URL: <https://doi.org/10.34185/1562-9945-4-159-2025-12>.

101. В. Б. Мокін, «Ідентифікація параметрів малих річок на основі теорії нечітких множин по експертних оцінках та по їх геоінформаційній моделі», Вісник Житомирського державного технологічного університету, с. 133–142, 2004.

102. Річний звіт про діяльність басейнового управління водних ресурсів річки Південний Буг з питань управління водними ресурсами за 2019 рік, Вінниця. Україна: БУВР, 2019.

103. Vinbazar. Біля Ладжижина у Південному Бузі загинуло риби на 270 тисяч гривень, 18 травня 2018, Всі Новини Вінниці. Vinbazar.com. URL: <https://vinbazar.com/news/podiyi/bilya-ladijina-u-pivdenному-buzi-zaginulo-ribi-na-270tisyach-griven>

104. Hnatushenko V., Kashtan, V., Chumychov, D., & Nikulin, S. Comparative analysis of classification methods for high-resolution optical satellite images. *Computer Systems and Information Technologies*, (4), 134–142, 2024. <https://doi.org/10.31891/csit-2024-4-16>

105. Горчук Ю. А., Юхимчук М. С., Дубовой В. М. Розмовний ШІ для автоматизації бізнес-процесів. Матеріали LIV Всеукраїнської науково-технічної конференції підрозділів Вінницького національного технічного університету, Вінниця, 24-27 березня 2025 р. Електрон. текст. дані. 2025. URI: <https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2025/paper/view/23563>.

ДОДАТКИ

ДОДАТОК А

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Information measuring systems with mobile devices for identification of air pollution parameters caused by transport / Vitalii B. Mokin ; Georgii V. Goriachev ; Dmytro Y. Dziuniak ; Konstantin O. Bondaletov ; Serhii O. Zhukov ; Mariusz Duk ; Saltanat Sailarbek. Proc. SPIE 10031, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2016, 1003128 (September 28, 2016), 8 pages; doi:10.1117/12.2249202 (Scopus).

2. Vitalii B. Mokin. Method for Analyzing and Optimizing the Topologic Observability of Cognitive Maps of Complex Spatially Distributed Systems / Vitalii Mokin, Ilona Varchuk, Kostiantyn Bondaletov, Olena Slobodianiuk. CONFERENCE PROCEEDINGS 2017 IEEE First Ukraine Conference on ELECTRICAL AND COMPUTER ENGINEERING (UKRCON) May 29 – June 2, 2017 Kyiv. – P.: 865-869. – doi 10.1109/UKRCON.2017.8100371. (Scopus)

3. Bondalietov, K., Mokin, V. (2023). Notation System for Comparing and Synthesis of Intelligent Key Phrase Extraction Methods for Ontological Models in Information Systems. In: Dovgyi, S., Trofymchuk, O., Ustimenko, V., Globa, L. (eds) Information and Communication Technologies and Sustainable Development. ICT&SD 2022. Lecture Notes in Networks and Systems, vol 809. Springer, Cham. https://doi.org/10.1007/978-3-031-46880-3_11 ISSN 2367-3370 (Print); ISSN 2367-3389 (Online) (Scopus).

4. Мокін В. Б., Бондалєтов К. О., Крижановський Є. М., і Караваєв В. О. Метод аугментації текстів про стан масивів вод на основі інтелектуальної прив'язки до багатозв'язних геоінформаційних систем іменованих сутностей, *Вісник Вінницького політехнічного інституту*, вип. 3, с. 55–65, Черв. 2023. URL: <https://doi.org/10.31649/1997-9266-2023-168-3-55-65> (Index Copernicus, Наукове фахове видання України, категорія «Б» зі спеціальності 126).

5. Мокін В. Б., Бондалєтов К. О., Штельмах І. М., Слободянюк О. В. Автоматичне видобування знань з екологічних звітів з прив'язкою до часу та до

просторових координат масивів вод, *Вісник Вінницького політехнічного інституту*, вип. 3, с. 101–110, Черв. 2025, <https://doi.org/10.31649/1997-9266-2025-180-3-101-110> (Index Copernicus, Наукове фахове видання України, категорія «Б» зі спеціальності 126).

6. К. О. Бондалетов і В. Б. Мокін, «Інтелектуальна технологія видобування та верифікації числових та текстових даних у багатозв'язних багатостадійних геоінформаційних системах», *Вісник Вінницького політехнічного інституту*, вип. 5, с. 51–60, Жовт. 2025. Режим доступу: <https://visnyk.vntu.edu.ua/index.php/visnyk/article/view/3335> (Наукове фахове видання України, категорія «Б» зі спеціальності 126).

7. Мокін В. Б. Універсальна інформаційно-вимірювальна система оперативного екологічного моніторингу з використанням мобільних пристроїв / В. Б. Мокін, К. О. Бондалетов, Г. В. Горячев, Д. Ю. Дзюняк. *Вісник Вінницького політехнічного інституту*. — Вінниця. — 2015. — № 5 (122). — С. 116-122. <https://visnyk.vntu.edu.ua/index.php/visnyk/article/view/1836> (Index Copernicus, Наукове фахове видання України, категорія «Б» зі спеціальності 126).

8. Метод і технологія моніторингу стану атмосферного повітря за допомогою універсальної інформаційно-вимірювальної системи з використанням мобільних пристроїв / В. Б. Мокін, Д. Ю. Дзюняк, К. О. Бондалетов, В. В. Олійник. *Наукові праці Вінницького національного технічного університету* [Електронний ресурс]. — 2015. — № 4. — Режим доступу: <http://praci.vntu.edu.ua/index.php/praci/article/view/456/454> (Наукове фахове видання України, категорія «Б» зі спеціальності 126).

9. Бондалетов К. О. Практична реалізація сучасної мобільної системи веб-моніторингу та експрес-аналізу стану повітряного середовища / К. О. Бондалетов, В. Б. Мокін, Г. В. Горячев, Д. Ю. Дзюняк. XLIV Регіональна науково-технічна конференція професорсько-викладацького складу, співробітників та студентів університету з участю працівників науково-дослідних організацій та інженерно-технічних працівників підприємств м. Вінниці та області, електронне наукове

видання матеріалів конференції, м. Вінниця, 11—13 березня 2015. – Режим доступу: <http://conf.vntu.edu.ua/allvntu/2015/ineek/txt/bondaletov.pdf>

10. Бондалетов К. О. Мобільна аналітична комп'ютерна система для оперативного моніторингу стану атмосферного повітря міста / К. О. Бондалетов, Д. Ю. Дзюняк, В. Б. Мокін. Молодь в технічних науках: дослідження, проблеми, перспективи : матеріали Міжнародної науково-практичної Інтернет-конференції, 23-26 квітня 2015 року / ВНТУ. – Вінниця, 2015. – С. 76-77.

11. Мокін В. Б. Інформаційно-вимірювальні системи для ідентифікації параметрів моделі забруднення атмосферного повітря автотранспортом з використанням мобільних пристроїв та фотометричних, у т.ч. лазерних, пристроїв і систем / В. Б. Мокін, Г. В. Горячев, Д. Ю. Дзюняк, К. О. Бондалетов. Зб. тез доп. III Міжнародної конференції “Вимірювання, контроль та діагностика в технічних системах (ВКДТС-2015)», м. Вінниця, 27-29 жовтня 2015 р. – Вінниця : ВНТУ, 2015. — С. 130-132.

12. Мокін В. Б. Технологія оперативного екологічного моніторингу стану повітряного середовища та пересувних джерел його забруднення / В. Б. Мокін, Д. Ю. Дзюняк, К. О. Бондалетов. XIV Міжнародна науково-практична конференція: Сучасні інформаційні технології управління екологічною безпекою, природокористуванням, заходами в надзвичайних ситуаціях: 5-9 жовтня 2015 р. – К., 2015. – С. 126-130.

13. Мокін В. Б. Практична реалізація мобільної аналітичної комп'ютерної системи моніторингу стану атмосферного повітря з підсистемою веб-аналізу та виведення даних на геопортал / В. Б. Мокін, Д. Ю. Дзюняк, К. О. Бондалетов, Г. В. Горячев, В. В. Боцян. V-й Всеукраїнський з'їзд екологів з міжнародною участю (Екологія/Ecology-2015), 23-26 вересня, 2015. Збірник наукових праць. – Вінниця: ТОВ «Нілан-ЛТД», 2015. – С. 106.

14. Бондалетов К. О., Мокін В.Б., Крижановський Є.М., Караваєв В.О. Розроблення інтелектуальної веб-системи з інформацією про екологічні проблеми та природоохоронні заходи у масивах вод басейну р. Південний Буг WISEST-SBB

/ Актуальні проблеми та перспективи розвитку фундаментальних, прикладних, загальнотехнічних та безпекових наук : Матеріали IV всеукр. науково-практ. конф. пам'яті акад. Акад. наук вищ. освіти, проф. А.В. Касперського, м. Київ, 27 черв. 2024 р. С. 6–12. URL: <https://kztdop.ipf.npu.edu.ua/?view=article&id=244>

15. Мокін В.Б., Бондалетов К.О. Інтелектуальні методи видобування ключових словосполучень із тексту для побудови онтологічних моделей інформаційно-пошукових систем. С. 197-199. – Режим доступу: https://itgip.org/wp-content/uploads/2022/11/tezy_70.pdf. Інформаційно-комунікаційні технології та сталий розвиток: колективна монографія за матеріалами XXI Міжнародної науково-практичної конференції (14-16 листопада 2022 р.). Київ, 2022. ТОВ «Видавництво «стон», 2022, 242. URL: <https://itgip.org/kolektivna-monografiya-za-materialam>

16. Бондалетов К. О., Мокін В. Б. Інтелектуальна автоматизація геоприв'язки повідомлень з соцмереж до масивів вод за допомогою зваженої JACCARD-міри. ВНТКП ВНТУ. Факультет інтелектуальних інформаційних технологій та автоматизації (2025): вебресурс. (дата звернення 13 червня 2025: <https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2025/paper/view/23298/19275> .

17. Бондалетов К. О., Мокін В. Б., Григорчук М. В., Джура С. В. та ін. Побудова датасету для тренування інтелектуальних моделей веб-системи з інформацією про екологічні проблеми та заходи у масивах вод басейну р. Південний Буг WISEST-SBB. Матеріали LI Науково-технічної конференції факультету інтелектуальних інформаційних технологій та автоматизації Вінницького національного технічного університету, Вінниця, 21 – 23 червня 2023 р. – Електрон. текст. дані. – 2023. – Режим доступу: <https://d.conf.vntu.edu.ua/index.php/all-fksa/all-fksa-2023/paper/view/18958>

18. Мокін В. Б., Бондалетов К. О. Новий підхід до автоматизації оптимального управління водогосподарським балансом басейну річки на основі когнітивних карт. X Міжнародна науково-практична конференція «Інтернет-Освіта-Наука (ІОН-2016)», Вінниця, Україна, 11–14 жовт. 2016. Вінниця : ВНТУ, 2016. С. 101–103. ISBN 978-966-641-646-2. URL:

<https://ies.vntu.edu.ua/reports/program/WORK-IES-2016.pdf> (дата звернення: 08.07.2025).

19. Бондалетов К. О., Мокін В. Б. Розроблення просторово-зваженого критерію геоприв'язки природномовних текстів до масивів вод. Молодь в науці: дослідження, проблеми, перспективи (МН-2025) : матеріали Всеукраїнської науково-практичної інтернет-конференції (Вінниця, 15 жовтня 2024 року – 15 червня 2025 року). — Вінниця : ВНТУ, 2025. — С. 1597-1599.

20. Бондалетов К. О.; Мокін В. Б. Автоматизація геолокації текстової інформації для контролю ефективності заходів з управління річковим басейном: Контроль і управління в складних системах (КУСС-2024), Вінниця, Україна, 16–17 жовтня 2024. с. 99–101.

21. Підвищення точності моделювання забруднення атмосферного повітря міста з використанням мобільних інформаційно-вимірювальних систем [Електронний ресурс] / В. Б. Мокін, Г. В. Горячев, Д. Ю. Дзюняк, К. О. Бондалетов. Матеріали XLV Науково-технічної конференції Вінницького національного технічного університету, Вінниця, 23-24 березня 2016 р. - Електрон. текст. дані. - 2016. - Режим доступу : <http://conferences.vntu.edu.ua/index.php/all-ebmd/all-ebmd-2016/paper/view/465>.

22. Бондалетов, Костянтин, Мокін, Віталій. "Архітектура каскадних парсерів гетерогених джерел даних для автоматизованої побудови датасетів для інтелектуальних систем моніторингу" Науково-технічної конференції факультету інтелектуальних інформаційних технологій та автоматизації Вінницького національного технічного університету, Вінниця, 24-27 березня 2026 р. - Електрон. текст. дані. - 2026. - Режим доступу <https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2026/paper/view/28287/23206>

ДОДАТОК Б

АКТ ВПРОВАДЖЕННЯ У БУВР ПІВДЕННОГО БУГУ

ЗАТВЕРДЖУЮ

Начальник Басейнового управління водних
ресурсів річки Південний Буг

Ілля ДЯКОНОВИЧ

«24» 04 2026 р.

АКТ

**про впровадження результатів дисертаційної роботи
«Інформаційна технологія видобування та верифікації даних про взаємопов'язані
водні об'єкти з просторовою та часовою прив'язкою»,
Бондалстова Костянтина Олеговича**

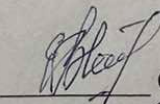
Цей акт підтверджує, що результати дисертаційної роботи на здобуття наукового ступеня доктора філософії Бондалстова К. О. рекомендовані до впровадження у практичну діяльність Басейнового управління водних ресурсів (БУВР) річки Південний Буг під час супроводу процесу імплементації програми заходів Плану управління річковим басейном (ПУРБ) Південного Бугу для аналізу даних моніторингу стану та для виявлення проблемних ситуацій, які потребують вирішення під час подальшого оновлення цього ПУРБ.

Зокрема, цінними є такі отримані результати дослідження Бондалстова К. О.:

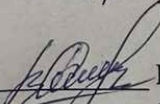
- метод верифікації числових даних у природномовних текстах з використанням бази знань та експертних оцінок, що дозволяє відібрати найбільш достовірну інформацію;
- метод формування знань та їх прив'язки до просторових об'єктів і часових меж, що надає можливість більш точно знайти релевантну для користувача інформацію з урахуванням встановлених просторово-часових обмежень;
- метод геоприв'язки, який автоматично відбирає релевантні факти у текстах і більш точно зіставляє ці факти з описами масивів вод;
- інформаційна технологія, яка поєднує в собі усі ці методи та алгоритми, покращує якість і повноту знань у ГІС та автоматизує обробку текстів;
- архітектура та тестовий веб-інтерфейс аналітичної веб-системи WISEST-SBB (Water Information System with Spatial and Temporal References for the Southern Bug Basin), що забезпечує доступ до узагальненої природномовної та числової інформації за заданими категоріями за заданий період по кожному з 1091 масивів вод басейну річки Південний Буг з можливістю наповнення системи та її розвитку.

Отримані результати можуть бути поширеними і на інші басейни великих річок, розташовані у Вінницькій області, зокрема - Дніпро і Дністер.

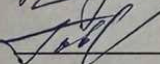
Заступниця начальника Басейнового
управління водних ресурсів
річки Південний Буг

 Олена МИСЛИЦЬКА

Начальник відділу водних
відносин та басейнової взаємодії

 Юлія ДОМЕРАТ

Провідний гідролог

 Юрій ГАВРИКОВ

ДОДАТОК В

АКТ ВПРОВАДЖЕННЯ У НАВЧАЛЬНИЙ ПРОЦЕС



ЗАТВЕРДЖУЮ

Проректор з науково-педагогічної роботи та організації освітнього процесу Вінницького національного технічного університету

доц. Олександр ПЕТРОВ

04 2026 р.

АКТ

про впровадження результатів дисертаційної роботи

Бондалетова Костянтина Олеговича

«Інформаційна технологія видобування та верифікації даних про взаємопов'язані водні об'єкти з просторовою та часовою прив'язкою», представленої до захисту на здобуття наукового ступеня доктора філософії зі спеціальності 126 Інформаційні системи та технології, в навчальному процесі

Комісія Вінницького національного технічного університету у складі:

голова комісії – декан Факультету інтелектуальних інформаційних технологій та автоматизації (ФІТА), к.т.н, доцент Севастьянов В. М. та члени комісії – заступник декана з навчально-методичної роботи ФІТА, ст. викладач кафедри системного аналізу та інформаційних технологій (САІТ) Присяжнюк В. В., заступник завідувача кафедри САІТ, к.т.н., доцент Крижановський Є. М., склала цей акт про підтвердження того, що дійсно у Вінницькому національному технічному університеті під час викладання дисципліни «Інтернет речей та інтелектуальний аналіз даних» для аспірантів, які навчаються за освітньо-науковою програмою «Інформаційні системи та технології» рівня «доктор філософії» спеціальності 126 Інформаційні системи та технології та дисципліни «Машинне навчання та інтелектуальний аналіз даних» для студентів, які навчаються за освітньо-професійною програмою «Інформаційні технології аналізу даних та зображень» рівня «магістр» спеціальності F6 Інформаційні системи і технології впроваджено результати дисертаційної роботи на здобуття доктора філософії аспіранта кафедри САІТ Бондалетова Костянтина Олеговича у низці лекцій та лабораторних робіт.

При викладанні цих дисциплін використовуються такі результати досліджень, отримані у дисертаційній роботі Бондалетова К.О.:

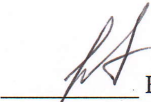
- метод верифікації числових даних у текстах з використанням бази знань та метод формування користувацьких запитів до великої мовної моделі у RAG-архітектурі шляхом прив'язки ембедінгів до просторових об'єктів і часових меж, що надає можливість більш точно знайти релевантну для користувача інформацію з урахуванням встановлених просторово-часових обмежень;

- метод геоприв'язки, який автоматично відбирає релевантні факти у текстах і більш точно зіставляє ці факти з описами масивів вод;

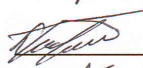
- інформаційна технологія, яка поєднує в собі усі ці методи та алгоритми, покращує якість і повноту знань у ГІС та автоматизує оброблення текстів, а також її веб-реалізація у вигляді авторизованої системи аналітичної веб-системи WISEST-SBB.

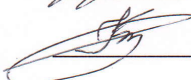
Використання зазначених результатів дозволило підвищити якість навчального процесу із згаданих дисциплін.

Голова комісії

 Володимир СЕВАСТЬЯНОВ

Члени комісії

 Василь ПРИСЯЖНЮК

 Євгеній КРИЖАНОВСЬКИЙ

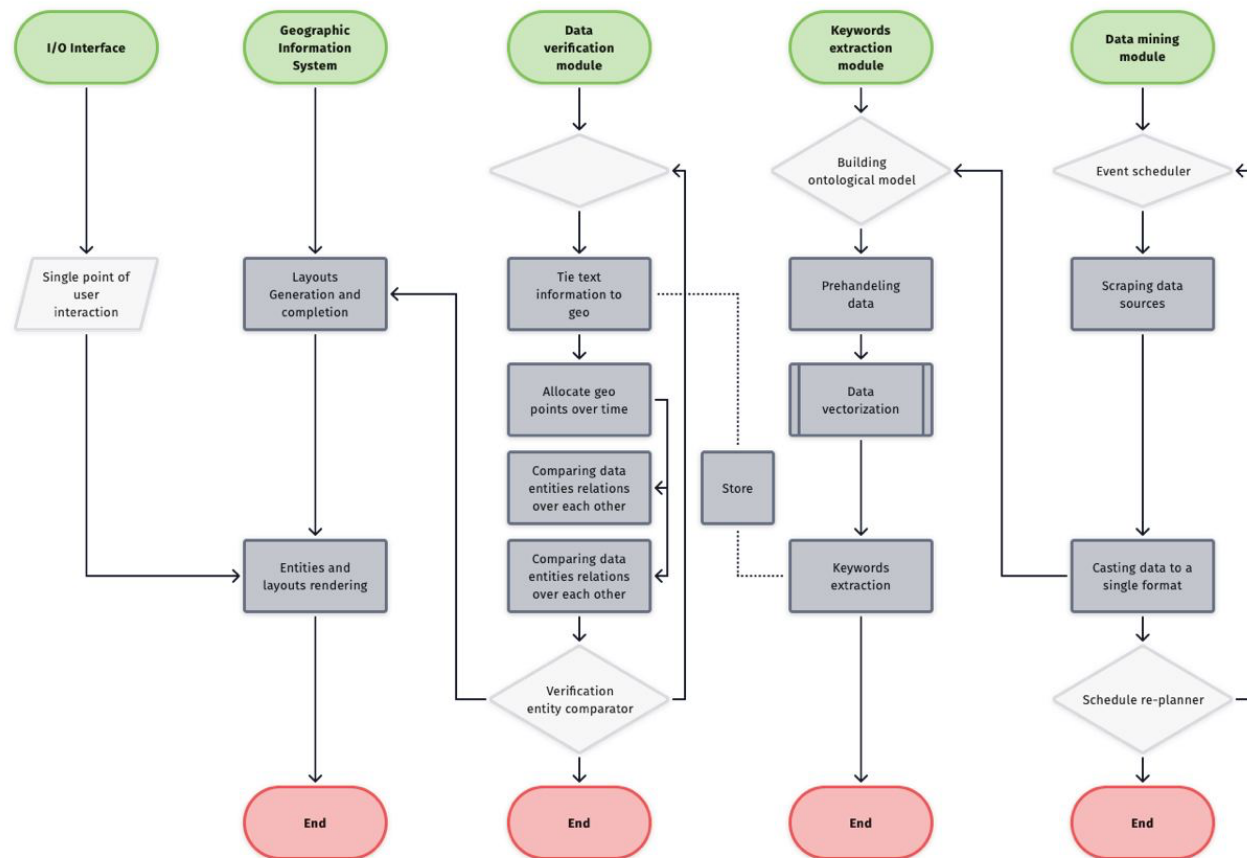
ДОДАТОК Г

ПРИКЛАДИ ФАКТІВ ПРО СТАН МАСИВІВ ВОД

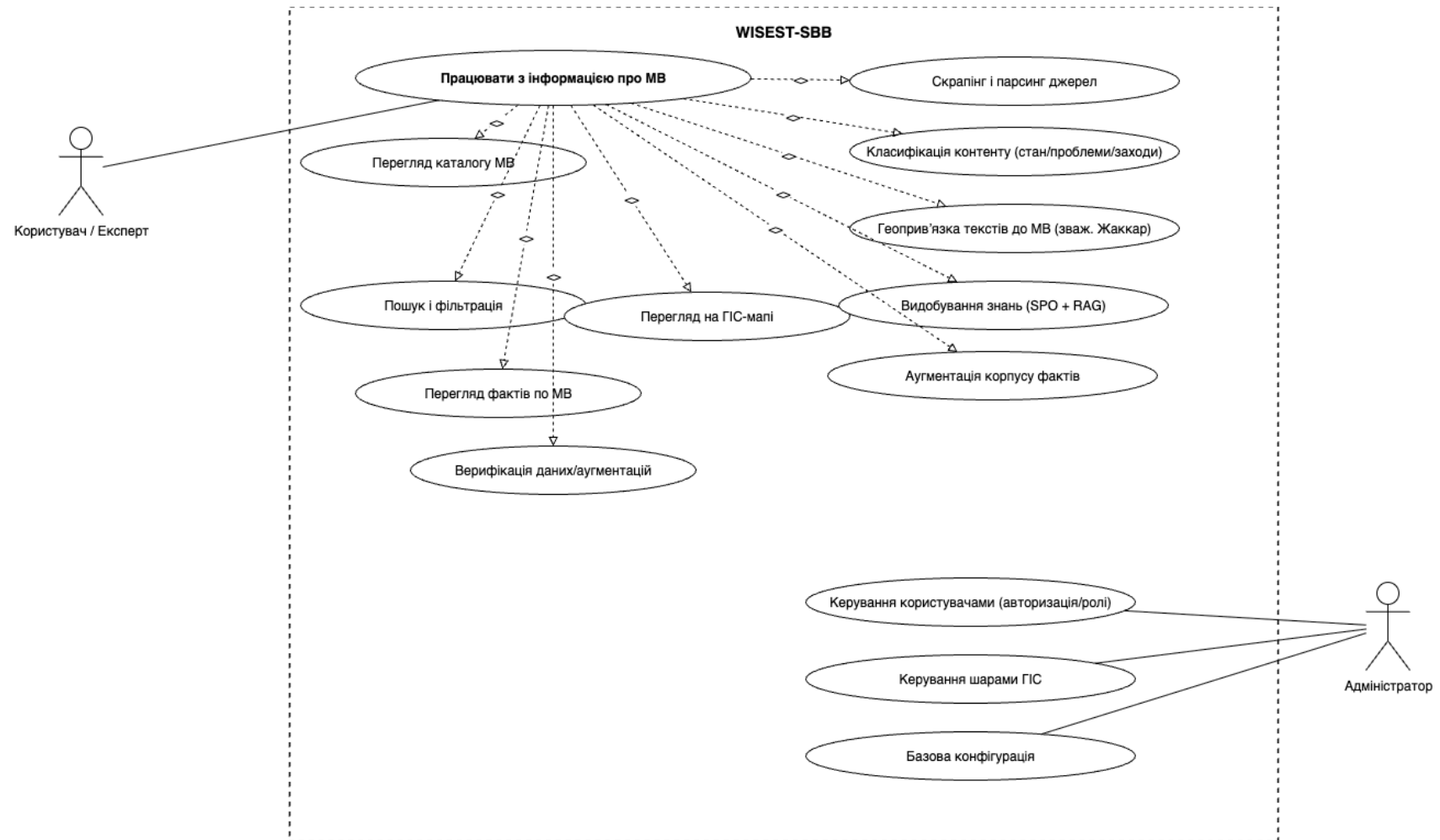
text	link	source_type	geonames_text_ext	geonames_text_nominat_sing	geonames_gps_lat_long	geodata_text	geodata_text_nominat_sing	date_doc	dates_text	dates_text_other	status	issues	measures
Вінницькі екологи п На адресу Держави «11 травня спеціаль За результатами аналіз Для притягнення в Спеціалісти відділу Під час обстеження До Державної еколог Оперативний виїзд «В ході лабораторних Комунальники курорт На річці Південний Буг Як повідомляє Хмельницька міська рада «Зафіксовано загибелі риби» Наразі представники екологічної служби розповідають, що у п'ятницю, 26 листопада, на річці Південний Буг було виявлено мертву рибу. Для покращення екології цього водоему встановлено мережу дозониметрів. Ця риба - природна. Таким чином, за рахунок збільшення кількості водних мешканців та населення водних біологічних видів.	https://vinnitsa.in.ua/	Веб-сторінка "Вінницького екологічного центру"	Південний Буг, м. Вінниця	Південний Буг, м. Вінниця	струмок, який протікає між вулицями Лебединського та Могилянської	5/16/23	11 травня	5/11/23		1	1	0	
«В ході лабораторних досліджень встановлено, що риба загинула через недостатню кількість кисню в воді. Це може бути пов'язано з браком їжі або занадто високими температурами води. Також можливі захворювання риб, спричинені бактеріями, грибами або паразитами. Для подальшого дослідження необхідно взяти проби води та риби для проведення аналізів в лабораторії. Результати будуть надані після завершення всіх необхідних тестів.»	https://today.vnn.ua/vnnytsya-ekologichnyi-centr-pidavizhiv-vinnitsy-11-travnja-spezialni-stadi-za-rezultatami-analizy-178911	Веб-сторінка "Вінницького екологічного центру"	Південний Буг, м. Вінниця	Південний Буг, м. Вінниця	у Вінниці вздовж берега зі сторони району П'ятничан, біля пляжу «Кумбари»	4/16/21	14 квітня	4/14/21		1	1	0	
Комунальники курорту планують провести ремонтні роботи на території курорту. Зокладно буде очищено територію від сміття та опалилих гілок. Також буде виконано ремонт доріжок та огорож. Роботи будуть проводитися протягом наступного тижня. Після закінчення робіт територія буде передана на оренду місцевому підприємству. Надалі планується облаштувати територію для рекреації та відпочинку.	https://glavcom.com.ua/news/178911	Веб-сторінка "ГЛАВКОМ"	Південний Буг, м. Хмельницький	Південний Буг, м. Хмельницький	район «Кармелюкового крісла», яке знаходиться на межі Вінницької та Хмельницької областей	3/1/21	24 лютого 2021 року	2/24/21		1	1	0	
Як повідомляє Хмельницька міська рада, риба загинула через недостатню кількість кисню в воді. Це може бути пов'язано з браком їжі або занадто високими температурами води. Також можливі захворювання риб, спричинені бактеріями, грибами або паразитами. Для подальшого дослідження необхідно взяти проби води та риби для проведення аналізів в лабораторії. Результати будуть надані після завершення всіх необхідних тестів.	https://xaxid.net/	Веб-сторінка "ZAXID.INFO"	Південний Буг, м. Хмельницький	Південний Буг, м. Хмельницький	на правому березі водосховища на річці Південний Буг від вулиці Західної Сторони	7/13/22	12 липня	7/12/22	у вівторок, 12 липня	0	1	0	
Для відновлення екосистеми У п'ятницю, 26 листопада, на річці Південний Буг було виявлено мертву рибу. Для покращення екології цього водоему встановлено мережу дозониметрів. Ця риба - природна. Таким чином, за рахунок збільшення кількості водних мешканців та населення водних біологічних видів.	https://mistaua.com.ua/	Веб-сторінка "Mista UA"	Південний Буг, м. Миколаїв	Південний Буг, м. Миколаїв	у межах міста Миколаєва	11/30/21	26 листопада 2021 року	11/26/21	У п'ятницю, 26 листопада	0	0	1	
Традиційні осінні свята У Державній екологічній службі розповіли, що традиційні осінні свята проводяться щорічно. Вони мають на меті популяризувати екологію та привертати увагу громадян до проблем навколишнього середовища. Традиційні осінні свята проводяться у вигляді екскурсій, лекцій та інших заходів. Вони проводяться протягом осені. Після закінчення свята територія буде передана на оренду місцевому підприємству. Надалі планується облаштувати територію для рекреації та відпочинку.	http://vinbazar.com.ua/	Веб-сторінка "VINBAZAR.COM.UA"	Південний Буг, м. Вінниця	Південний Буг, м. Вінниця	від Старомиського мосту до "Рошену"	11/4/23	25 жовтня	10/25/13		0	0	1	

ДОДАТОК Д

АЛГОРИТМ РОБОТИ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ВИДОБУВАННЯ ТА ВЕРИФІКАЦІЇ ДАНИХ ПРО ВЗАЄМОПОВ'ЯЗАНІ ВОДНІ ОБ'ЄКТИ З ПРОСТОРОВОЮ ТА ЧАСОВОЮ ПРИВ'ЯЗКОЮ



ДОДАТОК Е

**UML-ДІАГРАМА ВАРІАНТІВ ВИКОРИСТАННЯ ТЕХНОЛОГІЇ ВИДОБУВАННЯ ТА ВЕРИФІКАЦІЇ ДАНИХ
ПРО ВЗАЄМОПОВ'ЯЗАНІ ВОДНІ ОБ'ЄКТИ З ПРОСТОРОВОЮ ТА ЧАСОВОЮ ПРИВ'ЯЗКОЮ**

ДОДАТОК Ж

СТРУКТУРА ІНТЕЛЕКТУАЛЬНОЇ ТЕХНОЛОГІЇ З АВТОМАТИЗОВАНИМ ЗБИРАННЯМ ДАНИХ

