https://doi.org/10.31649/1997-9266-2023-168-3-55-65

УДК 004.9+556

В. Б. Мокін[1]
К. О. Бондалєтов[1]
Є. М. Крижановський[1]
В. О. Караваєв[1]

МЕТОД АУГМЕНТАЦІЇ ТЕКСТІВ ПРО СТАН МАСИВІВ ВОД НА ОСНОВІ ІНТЕЛЕКТУАЛЬНОЇ ПРИВ'ЯЗКИ ДО БАГАТОЗВ'ЯЗНИХ ГЕОІНФОРМАЦІЙНИХ СИСТЕМ ІМЕНОВАНИХ СУТНОСТЕЙ

[1]Вінницький національний технічний університет

Досліджено аугментацію україномовних текстів про стан масивів поверхневих вод басейну річки для тренування інтелектуальних моделей машинного навчання, які повинні автоматично розмічати ці тексти, тобто прив'язувати у просторі й часі та здійснювати їхню класифікацію.

Охарактеризовано стан створення авторами статті системи «Водна інформаційна система з просторовою і часовою прив'язкою для басейну Південного Бугу» («WISEST Southern Bug Basin» — «WISEST-SBB»), яка наповнюється розміченими даними про стан масивів вод басейну річки з використанням технологій та алгоритмів, розробленими авторами раніше. Зазначено, що досвід показав недостатність інформації для тренування інтелектуальних моделей машинного навчання, призначених для автоматизації її розмітки. Проведено аналіз сучасних методів аугментації текстової інформації, які можна застосувати до україномовних текстів, та відмічено їхні недоліки, передусім — високу ймовірність синтезу недостовірної інформації.

Запропоновано здійснювати аугментацію даних про масиви вод річкової мережі з урахуванням поширення достовірної інформації про одні масиви вод на інші, розташовані вище чи нижче за течією, або зв'язані з ними в інший спосіб. Для формалізації та автоматизації цього процесу запропоновано нову формалізацію річкової мережі у вигляді багатозв'язної геоінформаційної системи іменованих сутностей (БГСІС), яка передбачає виділення серед усіх об'єктів саме іменованих сутностей, а потім встановлення просторових зв'язків між ними. Охарактеризовано приклади БГСІС у вигляді гідрографічної чи екологічної мережі, мережі адміністративних утворень тощо. Удосконалено раніше запропонований авторами рекурсивний алгоритм прив'язування даних про масиви вод до іменованих сутностей БГСІС та розроблено його формалізований опис. Після прив'язування текстів до масивів вод запропоновано здійснювати їхню аугментацію з подальшою верифікацією результатів в напівавтоматизований спосіб, який, згодом, теж можна зробити автоматизованішим.

Охарактеризовано результати апробації запропонованого методу, алгоритму та підходів у системі «WISEST-SBB», які довели їхню ефективність.

Результати роботи можуть бути поширені й на інші типи БГСІС — як на басейни інших річок, так і на системи іншого характеру.

Ключові слова: аугментація текстів, іменовані сутності, просторові дані, багатозв'язні геоінформаційні системи, аналітична веб-система, інтелектуальна технологія, оброблення природномовного тексту.

Постановка задачі та вихідні передумови

У статті [1], за участі авторів цього дослідження, розглянуто проблему 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 [2] та Водного Кодексу України [3]. На виконання цієї директиви та законодавства України розробляються плани управління річковими басейнами, в яких слід з'ясувати стан вод в усіх маси-

вах вод та для кожного з них розробити і затвердити відповідний план заходів, виконання якого потім слід регулярно моніторити [4]. Зокрема, у тій статті запропонована інтелектуальна технологія автоматизованої геоприв'язки екологічної текстової природномовної інформації за допомогою технології розпізнавання іменованих сутностей (англ.: «Name Entity Recognition», скорочено — NER) та технологій опрацювання природної мови (англ. «Natural Language Processing», скорочено — NLP), які дозволяють всю інформацію автоматизовано прив'язувати до масивів вод, яких стосується ця інформація. Наведено алгоритм і приклад застосування цієї технології для басейну річки Південний Буг. Основні елементи цієї технології та алгоритми реалізовані в прототипі системи «Водна інформаційна система з просторовою і часовою прив'язкою для басейну Південного Бугу» (англ.: Water Information System with Spatial and Temporal References for the Southern Bug Basin, скорочено: WISEST-SBB) [5], [6]. Розпочато наповнення системи, яке показало наявність нової проблеми — для повної автоматизації технології з її інтелектуальними алгоритмами класифікації потрібна значна кількість розмічених даних, якої виявилось замало. Крім того, заплановане узагальнення усіх знань про кожний масив вод в більшості випадків буде мати замало інформації для цього. Наприклад, у басейні річки Південного Бугу виділено 1091 масивів вод (тут і надалі йдеться тільки про масиви поверхневих, а не підземних, вод), а постів спостереження за якістю вод — за різними програмами не більше сотні і пунктів регулярних спостережень — лише 50 та тільки за хімічних станом [5]. Отже, постала 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» (друкарські помилки, помилкові символи, дублювання тощо), заміни іменованих сутностей, синтезу нових речень чи узагальнення (анотування) тексту тощо [7]—[9]. Однак, такі підходи лише урізноманітнюють текст, але не додають нової інформації про кожний масив вод. До того ж їх слід застосовувати з обережністю, наприклад, заміна іменованих сутностей чи заміна слів на антоніми може призвести до отримання фейкової інформації. Для вирішення такої проблеми варто використати той факт, що запропонована у статті [1] 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.

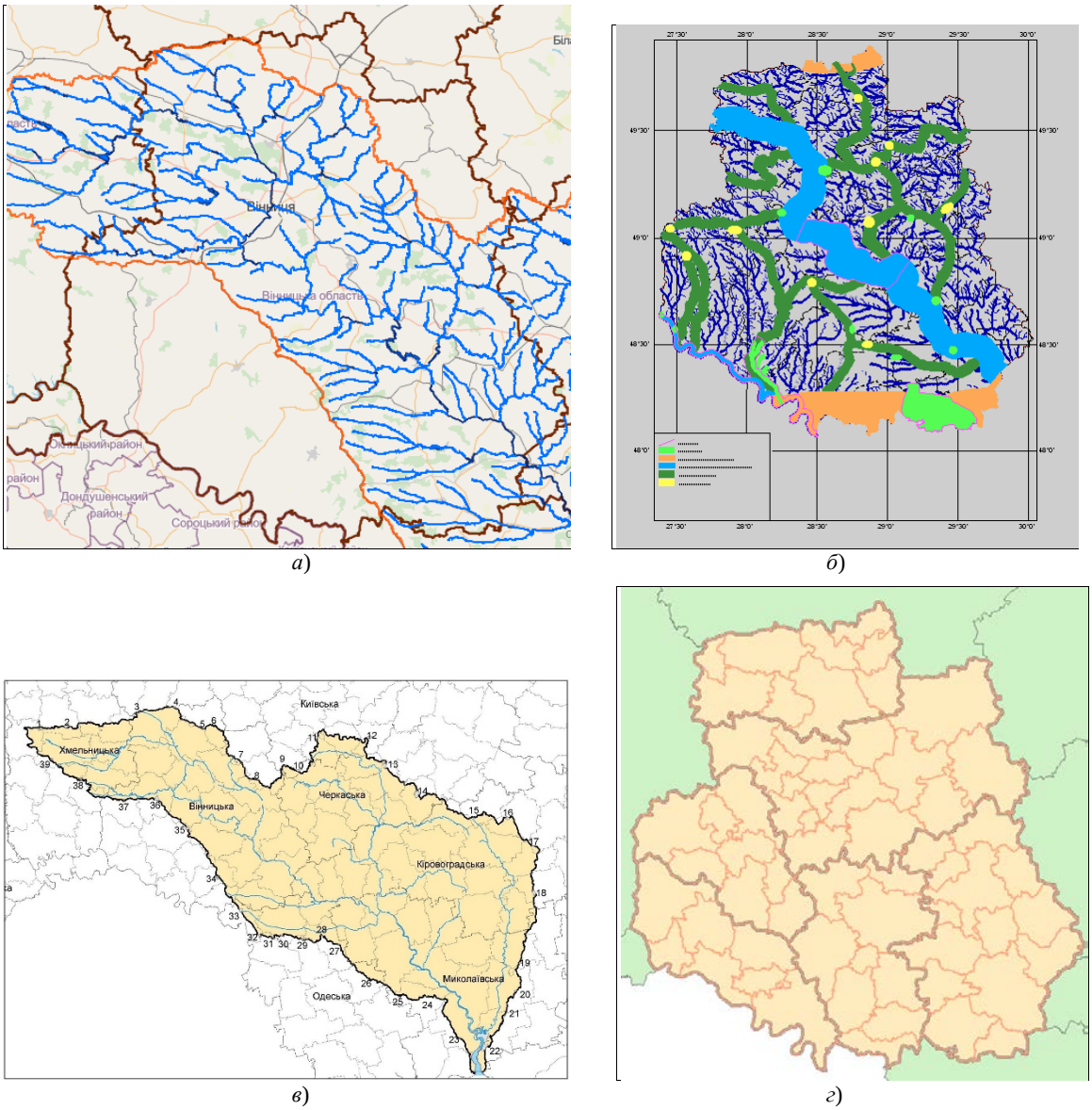
Метою статті є розроблення методу автоматичної аугментації природномовних україномовних текстів про стан масивів вод заданого району річкового басейну із забезпеченням верифікації нових даних по дослідженнях про іменовані сутності у цих текстах.

Основні поняття та формалізація постановки задачі

У роботі [10], за участі частини авторів цієї статті, введено термін багатозв'язних геоінформаційних систем (БГС), якими є багатозв'язні просторово-розподілені системи, де всі дані та зв'язки між ними можна формалізувати 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 складатись 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 — точка, яка позначає місце де одна річка впадає в іншу, тоді він вважається піделементом складеного елементу, 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 всіх об'єктів, які не є іменованими сутностями БГС, до іменованих сутностей, або в їхньому відкиданні. Для кожної БГСІС має виконуватись дві умови: по-перше, кожна іменована сутність географічно та/чи семантична має бути пов'язаною хоча б з однією іншою іменованою сутністю цієї ж БГСІС, а по-друге, не повинно бути жодного об'єкту, 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) пло-

щинного типу (області, райони, тергромади, населені пункти) тощо, мережа водогосподарських ділянок річкового басейну (МВГД) (рис. 1).



а) *б*) *в*) *г*)

Рис. 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.

Для розв'язання поставленої задачі пропонуються такі етапи:

1. Створити БГС іменованих сутностей (БГСІС) та формалізувати у них зв'язки між різними іменованими сутностями.

2. Пов'язати кожну 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.

Для гідрографічної мережі річок і водойм (БГСІС ГМР) прикладом застосування такої концепції є такі дії:

1. Створюється направлена геометрична мережа річок і водойм як БГС з урахуванням басейнового принципу, з урахуванням напрямку руху води від витоку до гирла. Трансформується у БГСІС ГМР.

2. Кожна іменована сутність БГСІС ГМР пов'язується з кожним масивом вод, що для цього ви-

ду 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 за течією можна, за певних умов, автоматично поширити (аугментувати) на масиви вод, розташованих нижче за течією.

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 тергромаду, а тоді, у першому наближенні, їх можна поширювати на усі масиви вод цих регіонів.

Аналогічно можна поширювати проблеми межені (маловоддя чи паводки) масивів вод, розташованих вище за течією водогосподарської ділянки (ВГД), на масиви вод, розташовані нижче за течією (межі ВГД, зазвичай, встановлюються у місцях розташування станцій державної системи моніторингу витрат води, де є максимально достовірна інформація про складові водогосподарського балансу).

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 поділили 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 [1], [5], [11]. Враховуючи структуру планів управління річковим басейном [4] та наявну інформацію про антропогенні джерела забруднення, природно-заповідні об'єкти тощо, таку технологію варто застосовувати до такого виду класів, для яких інформацію важко зібрати точно по всіх 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, на сьогодні і прогнози щодо майбутнього стану — зазвичай, 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 — зазвичай, 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, проєкти планів заходів ПУРБ тощо — можуть прив'язуватись до БГСІС МАУ (як правило, гроші виділяють за адміністративним принципом) або до БГСІС ГМР чи БГСІС МВГД (в самих адміністративних регіонах гроші 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 відмітити, 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, 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 буде їх розрізнити (екс-

пертам 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.

Можуть бути й інші класи. Наприклад, раніше були проведені дослідження і натреновані моделі для класів "Env_problems", "Pollution", "Treatment", "Climate", "Biomonitoring" [11].

Формалізація та алгоритм застосування методу

Для формалізації моделі БГСІС G у загальному вигляді пропонуємо нотацію, основану на нотації направлених зважених графів [13]:

$$G = \{R\{V,P\}, S(V), W(V,P)\}, \quad (1)$$

де $R\{V,P\}$ — графова модель, яка складається з двох множин — множини V вершин (вузлів), які є просторовими об'єктами та, одночасно — іменованими сутностями, які і собі можуть бути складеними об'єктами з точкових, лінійних та площинних об'єктів, та з множини P зв'язків (ребер), які формалізують наявність певного відношення між двома і більше вершинами; $S(V)$ — 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 потрібний, якщо, у графі слід врахувати зменшення сили зв'язку з відстанню).

Модель (1) є узагальненою і може враховувати і множину усіх об'єктів БГСІС (вершини), і зв'язки чи відношення між ними (ребра), і направлення та силу впливу цих зв'язків. Наприклад, для БГСІС ГМР модель буде мати саме такий вигляд, оскільки цей граф є і направленим (у напрямку течії води), і зваженим, оскільки забруднення з верхніх за течією ділянок вод впливає на нижні, затухаючи з відстанню, завдяки самоочисним процесам річки. А модель БГСІС МАУ для адміністративних районів може бути спрощеною, оскільки не буде враховувати ні напрямок, ні зміну сили впливу між ними, якщо не буде враховуватись відносне розташування цих районів на певних річках.

З використанням ГІС-технологій за (1) нескладно знайти усі варіанти відповідності масивів вод, як лінійних (ділянки річок) та площинних об'єктів (водозбірна площа кожного масиву), із множини M та всіх іменованих сутностей із множини V:

$$\{m\} = \Phi(\{v\}, G);\ \{v\} = \Psi(\{m\}, G),\ \{m\} \subseteq M,\ \{v\} \subseteq V, \quad (2)$$

де $\Phi(\{v\}, G)$ — всі правила та відношення, які встановлюються моделлю БГСІС G з (1) і дозволяють для заданої підмножини кодів іменованих сутностей $\{v\}$ із множини V визначити підмножину кодів $\{m\}$ масивів вод із множини M; $\Psi(\{m\}, G)$ — всі правила та відношення, які встановлюються моделлю БГСІС G з (1) і дозволяють для заданої підмножини кодів $\{m\}$ масивів вод із множини M знайти підмножину $\{v\}$ кодів іменованих сутностей із множини V. В 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 [1].

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 сформувати множину N всіх назв іменованих сутностей по їхніх кодах у моделі (1):

$$N = \{A(v,G)\},\ v \in V, \quad (3)$$

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

Але на практиці ціннішою є функція $\chi(*)$ знаходження кодів іменованих сутностей $\{v\}$ за їхніми назвами $\{n\}$

$$\{v\} = \chi\left(\{n\}, G\right),\ \{n\} \subseteq N. \tag{4}$$

Для автоматичного геоприв'язування природномовних україномовних текстів про стан водних об'єктів до іменованих сутностей БГСІС варто скористатись інтелектуальною технологією, запропонованою у статті [1] для екологічної текстової інформації (ETI), згідно з якою геоприв'язування текстів здійснюється у 2 етапи:

Eman 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 з різних звітів та довідників тощо), про частину з яких можуть бути невідомі просторові координати; з цієї множини виділяється підмножина U_1 з відносно більшими 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 з обох підмножин U_1 та U_2 відповідно.

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

3) у тексті T_i знаходяться назви X_1^T іменованих сутностей із множини X_1, потім, використовуючи (4), встановлюються їхні коди {v}, а потім за першим із співвідношень (2) визначається до яких кодів масивів вод Y_{Ti} їх можна прив'язати;

4) у тому ж тексті T_i знаходяться назви іменованих сутностей із множини U_2, за (4) визначаються їхні 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)—(4).

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 авторів довели вищу ефективність і надійність результатів з використанням такого удосконалення.

Крім того, кроки етапу 1 пропонується доповнити таким чином: по-перше, на карті (1) знаходяться об'єкти, які відповідають усім елементам множини U, а якщо їхнє 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 меандрують, що такий центр може відноситись до іншого масиву вод). По-друге, на першому кроці пропонується відносити до підмножини U_1 всі точно геоприв'язані об'єкти, а не тільки великі.

З урахуванням вищенаведеної формалізації запропоноване удосконалення алгоритму зі статті [1] можна записати таким чином:

$$\begin{gathered}
X_1^* = L\left(U_1\right),\ \ X_2^* = L\left(U_2\right),\ \ U_1 \cup U_2 = U; \\
X_1^T = \Omega\left(L\left(T_i\right), X_1^*\right); \\
Y_{Ti} = \Phi\left(\chi\left(X_1^T, G_i\right),\ G_i\right); \\
V_2^T = \chi\left(Q\left(L\left(T_i\right), X_2^*\right), G_i\right); \\
G_{i+1} = \Lambda_i\left(V_2^T, Y_{Ti}, G_i\right),\ \ i = 0, 1, \ldots,
\end{gathered} \tag{5}$$

де $L(x)$ — функція лематизації усіх слів тексту x; $\Omega(\alpha, \beta)$ — функція, яка знаходить лематизовані

іменовані сутності з множини β у тексті α (в загальному випадку, допускаються багатомовні іменовані сутності зі збереженням їхніх комбінацій, тобто словосполучення «Вінницький область» (лематизація сутності з назвою «Вінницька область») буде зберігатись саме у такому вигляді, а не як два окремих послідовних слова «Вінницький» і «область»); $\Lambda_i\left(V_2^T, Y_{Ti}, G_i\right)$ — функція, яка на i-му кроці додає до графу G_i з (1) відповідність ідентифікованих за іменованими сутностями з множини U_1 кодів масивів вод Y_{Ti} та знайденими у тексті T_i кодами V_2^T іменованих сутностей, утворюючи оновлений граф G_{i+1}.

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

Найскладнішим елементом цієї технології є функція Φ(*) з (2), оскільки реалізація інших функцій (функції χ(*) знаходження кодів іменованих сутностей за їхніми назвами, функцією лематизації L(*), функцією Ω(*) пошуку лематизованих іменованих сутностей серед лематизованих слів заданого тексту) є доволі очевидною. На рис. 2*а* наведено приклад таблиці бази даних, яка може бути згенерована за моделлю G_i для усіх комбінацій кодів масивів вод та кодів іменованих сутностей для реалізації функції Φ(*), а на рис. 2*б* — для цих же кодів іменованих функцій та їхніх назв для реалізації функції χ(*).

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)

Рис. 2. Приклад фрагментів звичайних таблиць бази даних для формалізації правил з формування вибірок даних для реалізації функцій моделі (5): *а* — для функції Φ(*); *б* — для функції χ(*) (ЄМ)

Проте, така формалізація може призводити до завеликої кількості хибних спрацювань. Наприклад, текст «У Вінницькій області сталось 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 даних до масивів вод пропонується змінити модель даних (1)—(2) і зберігати не всі відповідності за принципом «всі до всіх», натомість, для кожного масиву вод за ГІС (1) синтезувати список назв усіх іменованих сутностей, зокрема з множини U_1, тобто — тих, геоприв'язка яких до цього масиву вод точно відома (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. Гущинці, с. Мизяків, с. Медвідка, с. Лаврівка, водозабір Вінницяводоканалу, Київський міст, Центральний міст, Староміський міст, "Вінницька реберня", 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

Рис. 3. Приклад списків назв іменованих сутностей декількох масивів вод, синтезованих за моделлю (1)

Така формалізація і встановлення відповідності назв іменованих сутностей та кодів масивів вод, дозволить замінити функції (2)—(4) однією функцією:

$$\{m\} = \Phi\left(\chi\left(\{u\}, G\right), G\right) = \Theta\left(\{u\}, G\right),\ \{m\} \subseteq M,\ \{u\} \subseteq U, \tag{6}$$

де функція $\Theta\left(\{u\}, G\right)$ буде здійснювати пошук найкращих збігів списків словосполучень, які характеризують кожен масив вод у довіднику як на рис. 3, 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 (5):

$$G_{i+1} = \Lambda_i\left(V_2^T, Y_{Ti}, G_i\right),\ \ i = 0, 1, \ldots, \tag{7}$$

$$V_2^T = \chi\left(\Omega\left(L\left(T_i\right), X_2^*\right), G_i\right),\ Y_{Ti} = \Theta\left(X_1^T, G\right),\ X_1^T = \Omega\left(L\left(T_i\right), X_1^*\right);$$

$$X_1^* = L\left(U_1\right);\ X_2^* = L\left(U_2\right);\ U_1 \cup U_2 = U.$$

У лаконічнішому вигляді нову модель (7) можна записати у вигляді:

$$G_{i+1} = \Lambda_i\left(\chi\left(\Omega\left(L\left(T_i\right), L\left(U_2\right)\right),\ G_i\right), \Theta\left(\Omega\left(L\left(T_i\right), L\left(U_1\right)\right), G_i\right),\ G_i\right),\ \ i = 0,\ 1, \ldots. \tag{8}$$

Авторами у Kaggle перевірено на мові Python можливість автоматизації порівняння україномовних текстів за моделлю (7) з використанням бібліотеки SpaCy [14]. Як видно з рис. 4, по-перше, найбільше значення метрики «cosinus_similarity» дійсно відповідає найрелевантнішому варіанту, а по-друге, видно, що лематизація речення підвищує значення цієї метрики майже у 2 рази, що збільшує ймовірність вибору найправильнішого варіанта.

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

Рис. 4. Результат порівняння за метрикою cosinus_similarity україномовних текстів з використанням методів бібліотеки SpaCy та мовної моделі «uk_core_news_sm» [14]

Наведемо приклад застосування запропонованого методу та моделі (8) на реальній вже створеній веб-системі.

Приклад застосування запропонованого методу

Відповідно до рішення басейнової ради Південного Бугу (протокол № 12 від 7 грудня 2022 р.) авторами створена інтелектуальна веб-система з інформацією про екологічні проблеми, природоохоронні заходи тощо, про масиви вод басейну р. Південний Буг «WISEST Southern Bug Basin» (скорочено — «WISEST-SBB» або «WISESTR-SBB» — англ.: Water Information SystEm with Spatial and Temporal References for the Southern Bug Basin – «Водна інформаційна система з просторовою і часовою прив'язкою для басейну Південного Бугу») за адресою: https://wisestr.ai/ .

Триває її наповнення даними. Базове наповнення здійснюється вручну. Потім планується максимально автоматизувати усі процеси скрапингу і парсингу (автоматичного пошуку текстів у вебресурсах та перетворення їх на таблиці даних), класифікації (на охарактеризовані вище 4 види: «status», «issues», «measures», «balance»), прив'язування у часі та у просторі за запропонованим у цій статті методом на основі моделі (8). Вже є пілотні версії усіх цих технологій, методів та алгоритмів. Частина їх опублікована у вигляді ноутбуків до нашого датасету у платформі Kaggle (див. розділ «Code») [11]. Здійснюється їхнє удосконалення. На рис. 5 подано приклад інтерфейсу системи «WISEST-SBB».

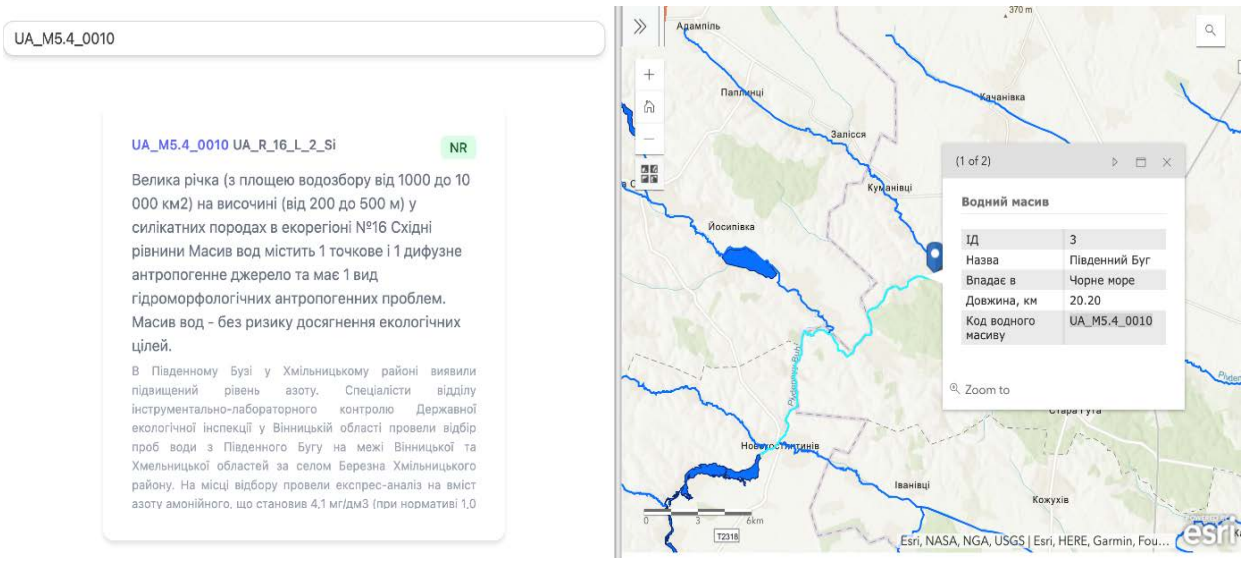


Рис. 5. Приклад інтерфейсу системи «WISEST-SBB», яка використовує запропоновані у статті методи, алгоритми та моделі (https://wisestr.ai/)

Доступ до системи авторизований, але планується його надати усім бажаючим, небайдужим до процесу поліпшення стану вод у басейні р. Південний Буг: усі члени басейнової ради, її робочих органів, представники органів влади, водокористувачів, активісти та науковці, які досліджують проблеми басейну, та ін.

Після прив'язування текстів до масивів вод здійснено її аугментацію, тобто поширення на сусідні масиви вод вище і нижче за течією та в межах тих же областей чи районів, залежно від того, через яку БГСІС вони були прив'язані до масивів вод. Розроблено інтерфейс, який дозволить експертам здійснювати верифікацію результатів аугментації. В подальшому, коли буде накопичено достатньо таких експертних оцінок, планується розробити ще один метод для автоматизації такої верифікації з використанням інтелектуальних моделей машинного навчання.

Отже, запропонований метод довів свою ефективність під час використання для розв'язання складної прикладної задачі.

Висновки

У роботі розглянуто задачу аугментації україномовних текстів про стан масивів поверхневих вод басейну річки для тренування інтелектуальних моделей машинного навчання, які повинні автоматично прив'язувати ці тексти у просторі й часі та здійснювати їхню класифікацію.

Зазначено, що авторами цієї статті створюється система «WISEST-SBB», яка наповнюється розміченими даними про стан масивів вод басейну річки Південний Буг з використанням технологій та алгоритмів, розроблених авторами раніше. Досвід показав недостатність інформації для тренування інтелектуальних моделей машинного навчання, призначених для автоматизації її розмітки (прив'язування у просторі й часі та для класифікації за низкою ознак). Проаналізовано сучасні методи аугментації текстової інформації, які можна застосувати до україномовних текстів, та відмічено їхні недоліки, основним з яких є те, що аугментація без урахування просторової прив'язки даних, з високою ймовірністю може спричиняти синтез недостовірної інформації.

Запропоновано здійснювати аугментацію даних про масиви вод річкової мережі, з урахуванням поширення достовірної інформації про одні масиви вод на інші, розташовані вище чи нижче за течією, або зв'язані з ними в інший спосіб. Для формалізації та автоматизації цього процесу запропоновано нову формалізацію річкової мережі у вигляді багатозв'язної геоінформаційної системи іменованих сутностей (БГСІС). Охарактеризовано раніше запропонований авторами рекурсивний алгоритм прив'язування даних про масиви вод до іменованих сутностей БГСІС та розроблено його формалізований опис. Запропоновано низку удосконалень цього алгоритму, які дозволять спростити модель та пришвидшити її ідентифікацію на практиці, за рахунок використання методів Python-бібліотеки SpaCy для задач оброблення україномовних природномовних текстів. Після прив'язування текстів до масивів вод запропоновано здійснювати їхню аугментацію з подальшою верифікацією результатів в напівавтоматизований спосіб, який, згодом, теж можна зробити автоматизованішим.

Охарактеризовано результати апробації запропонованого методу, алгоритму та підходів у системі «WISEST-SBB», які довели їхню ефективність. Результати роботи можуть бути поширені й на інші типи БГСІС — як на басейни інших річок, так і на системи іншого характеру.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] В. Б. Мокін, М. А. Гораш, Є. М. Крижановський, і Т. Є. Вуж, «Інформаційна інтелектуальна технологія автоматизованої геоприв'язки екологічної текстової природно-мовної інформації,» *Наукові праці ВНТУ*, № 4, 2020. [Електронний ресурс]. Режим доступу: https://praci.vntu.edu.ua/index.php/praci/article/view/624 .

[2] *Directive 2000/60/ec of the European Parliament and of the Council*. EUR-Lex – Access to European Union Law. [Electronic resource]. Available: https://eur-lex.europa.eu/resource.html?uri=cellar:5c835afb-2ec6-4577-bdf8-756d3d694eeb.0004.02/DOC_1&format=PDF . Access: 07.06.2023.

[3] Верховна Рада України, *Водний кодекс України, Кодекс України від 06.06.1995 р. № 213/95-ВР*, станом на 19 серп. 2022 р. [Електронний ресурс]. Режим доступу: https://zakon.rada.gov.ua/laws/show/213/95-вр#Text . Дата звернення: 07.06.2023.

[4] Кабінет міністрів України, *Постанова від 18.05.2017 р. № 336, Про затвердження Порядку розроблення плану управління річковим басейном* [Електронний ресурс]. Режим доступу: https://www.kmu.gov.ua/npas/249999756 . Дата звернення 04.06.2023.

[5] В. Б. Мокін, і К. О. Бондалєтов, *Інтелектуальні методи видобування ключових словосполучень із тексту для побудови онтологічних моделей інформаційно-пошукових систем. Інформаційно-комунікаційні технології та сталий розвиток*, колективна моногр. за матеріалами XXI Міжнародної науково-практичної конференції, Київ, 14-16 листопада 2022 р., С. О. Довгий, Заг. ред. Київ, Україна: ТОВ «Видавництво «стон», 2022, 242 с.

[6] А. І. Лісовенко, і О. В. Бісікало, *Інформаційна технологія підтримки функції «запитання-відповідь» на основі образного аналізу фахових текстів*, моногр. Вінниця, Україна: ВНТУ, 2019, 180 с. ISBN 978-966-641-764-3. [Електронний ресурс]. Режим доступу: https://press.vntu.edu.ua/index.php/vntu/catalog/book/512 .

[7] Vitalii Mokin, "NLP for WR: Summarizing using BERT, GPT2, XLNET," *Kaggle: Your Machine Learning and Data Science Community*. [Electronic resource]. Available: https://www.kaggle.com/code/vbmokin/nlp-for-wr-summarizing-using-bert-gpt2-xlnet . Access: 07.06.2023.

[8] Oleh Bisikalo, and 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 .

[9] A. Fiori, *Trends and Applications of Text Summarization Techniques*. IGI Global, 2019.

[10] В. Б. Мокін, І. В. Варчук, і Є. М. Крижановський, І*нформаційна технологія аналізу та оптимізації топологічної спостережуваності багатозв'язних геоінформаційних систем*: моногр., Вінниця, Україна: ВНТУ, 2019, 121 с.

[11] Vitalii Mokin, "NLP for UA : BERT CLS & 10 Classifiers," *Kaggle: Your Machine Learning and Data Science Community*. [Electronic resource]. Available: https://www.kaggle.com/code/vbmokin/nlp-for-ua-bert-cls-10-classifiers. Access: 07.06.2023.

[12] "Environmental indicators: typology and overview," *European Environment Agency*. [Electronic resource]. Available: https://www.eea.europa.eu/publications/TEC25 .

[13] В. М. Дубовой, Р. Н. Квєтний, О. І. Михальов, і А. В. Усов, *Моделювання та оптимізація систем*, підруч. Вінниця, Україна: ПП «ТД«Едельвейс», 2017, 804 с.

[14] Vitalii Mokin, and Kostiantyn Bondaletov, "SpaCy for Ukrainian text similarity," *Kaggle: Your Machine Learning and Data Science Community*. [Electronic resource]. Available: https://www.kaggle.com/code/bondaletov/spacy-for-ukrainian-text-similarity . Access: 07.06.2023.

Рекомендована до друку кафедрою системного аналізу та інформаційних технологій ВНТУ

Стаття надійшла до редакції 28.08.2022

Мокін Віталій Борисович — д-р техн. наук, професор, завідувач кафедри системного аналізу та інформаційних технологій; e-mail: vbmokin@vntu.edu.ua ;

Бондалєтов Костянтин Олегович — аспірант кафедри системного аналізу та інформаційних технологій; e-mail: bondaletov.k@gmail.com ;

Крижановський Євгеній Миколайович — канд. техн. наук, доцент кафедри системного аналізу та інформаційних технологій; e-mail: kruzhan@gmail.com ;

Караваєв Вадим Олександрович — студент факультету інтелектуальних інформаційних технологій та автоматизації, e-mail: karavaevvadim1999@gmail.com .

Вінницький національний технічний університет, Вінниця

V. B. Mokin[1]
K. O. Bondalietov[1]
Ye. M. Kryzhanovskyi[1]
V. O. Karavaiev[1]

Method of Augmentation of Texts About the State of Water Bodies on the Base of Intellectual Referencing to Multi-Related Geoinformation Systems of Named Entities

[1]Vinnytsia National Technical University

The article is dedicated to the augmentation of Ukrainian-language texts about the state of surface water bodies in a river basin for the training of machine learning models that should automatically annotate these texts, i. e. referencing in space and time and performing their classification.

The authors describe the progress made in creating the "Water Information System with Spatial and Temporal Referencing for the Southern Bug Basin" ("WISEST-SBB"), which is being populated with annotated data on the state of water bodies in the river basin using technologies and algorithms developed by the authors earlier. It is noted that the experience has shown a lack of information for training machine learning models intended for automating its annotation. An analysis of modern methods of text data augmentation applicable to Ukrainian texts has been conducted, highlighting their drawbacks, primarily the high probability of synthesizing unreliable information.

The proposed approach suggests augmenting data on water bodies of a river network, considering the propagation of reliable information about one water body to others located upstream or downstream or otherwise connected to them. To formalize and automate this process, a new formalization of the river network in the form of a multi-related geoinformation system of named entities (MGISNE) is proposed, which involves identifying named entities among all objects and then establishing spatial relationships between them. Examples of MGISNE are described, including hydrographic or ecological networks, networks of administrative entities, and others. The previously proposed recursive algorithm for referencing water body data with named entities in MGISNE is improved, and its formalized description is developed. After referencing texts with water bodies, the augmentation of the texts is proposed with subsequent verification of the results in a semi-automated manner, which can later be made more automated.

The results of the proposed method, algorithm, and approaches in the WISEST-SBB system are characterized, demonstrating their effectiveness. The findings of this work can be extended to other types of MGISNE, both for basins of other rivers and systems of a different character.

Keywords: *text augmentation, natural language processing, NLP, named entities, spatial data, multi-related geoinformation systems, analytical web systems, intelligent technology.*

Mokin Vitalii B. — Dr. Sc. (Eng.), Professor, Head of the Chair of System Analysis and Information Technology, e-mail: vbmokin@vntu.edu.ua ;

Bondalietov Kostiantyn O. — Post-Graduate Student of the Chair of System Analysis and Information Technology, e-mail: bondaletov.k@gmail.com ;

Kryzhanovsky Yevghenii M. — Cand. Sc. (Eng.), Associate Professorof the Chair of System Analysis and Information Technology, e-mail: kruzhan@gmail.com ;

Karavaiev Vadim O. — Student of the Department of Intelligent Information Technologies and Automation, e-mail: karavaevvadim1999@gmail.com