УДК 004.021

М.В. МОГИЛЬНА, В.І. ДУБРОВІН
Національний університет «Запорізька політехніка»

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ТЕКСТУ: ЗАСТОСУВАННЯ ТА БЕЗКОШТОВНІ ПРОГРАМНІ ЗАСОБИ

Велика кількість даних, що генеруються щодня, є як можливістю, так і викликом для бізнесу. З одного боку, дані допомагають компаніям отримувати відгуки людей про продукти чи послуги. Їх можна отримати, аналізуючи електронні листи, огляди продуктів, публікації в соціальних мережах, відгуки клієнтів, звернення до служби підтримки тощо. Істотний обсяг даних зберігається у формі документів, які можуть бути різними: структурованими, частково структурованими та і неструктурованими.

З іншого боку, виникає проблема обробки цих даних. Видобування корисної інформації з величезного обсягу документів є важким завданням. Інтелектуальний аналіз тексту є важливою сферою дослідження, оскільки за допомогою нього можна видобути знання з неструктурованого тексту.

У даній статті розглянуто технологію інтелектуального аналізу тексту та їх застосування в різних сферах життя. Було проаналізовано застосування інтелектуального аналізу тексту в системі керування бібліотекою та порівняно особливості популярних інструментів технологій інтелектуального аналізу тексту.

Методами дослідження є аналіз наукових статей, у яких дослідники використовували інструменти інтелектуального аналізу тексту, порівняння програмного забезпечення з відкритим вихідним кодом.

Розглянуто методи аналізу тексту та описано основне застосування інтелектуального аналізу в системі керування бібліотекою. Було обговорено популярне безкоштовне програмне забезпечення для текстового аналізу RStudio, Python, Orange, RapidMiner з відкритим вихідним кодом, яке також використовується у машинному навчанні та науці про дані.

Стаття сприяє підвищенню рівня розуміння дослідників у сфері інтелектуального аналізу текстів. Використовуючи розглянуте програмне забезпечення, початківці зможуть прогнозувати тенденції, теми, нові концепції досліджень, знаходити дублікати текстових документів в статтях, новинах, блогах. Бібліотекарі зможуть покращити свої послуги в системі керування бібліотекою: довідкових службах, CAS, SDI.

Ключові слова: класифікація, виявлення експертних знань, видобування інформації, образи, інструменти інтелектуального аналізу тексту, застосування інтелектуального аналізу тексту, система керування бібліотекою.

M.V. MOHYLNA, V.I. DUBROVIN
Zaporizhzhia Polytechnic National University

TEXT MINING: APPLICATIONS AND FREE SOFTWARE TOOLS

The large amount of data generated every day is both an opportunity and a challenge for businesses. On the one hand, data helps companies gain insights into people's opinions about a product or service. These insights can be gained by analyzing emails, product reviews, social media posts, customer feedback, customer service calls, etc. A substantial amount of data is stored in the form of documents, which can be of different types: structured, partially structured, and unstructured.

On the other hand, there is the problem of processing this data. Extracting useful information from a huge volume of documents is a difficult task. Text mining is an important area of research because it helps extract knowledge from unstructured text.

This article discusses text mining technologies and their application in various areas of life. The article analyzes the use of text mining in a library management system and compares the features of popular text mining tools.

The methods of the study are the analysis of scientific articles in which researchers used text mining tools and the comparison of open source software.

The methods of text analysis are considered and the main application of intellectual analysis in the library management system is described. Popular free and open-source text analysis software such as RStudio, Python, Orange, RapidMiner, which is also used in machine learning and data science, was discussed.

The article helps to increase the level of understanding of researchers in the field of text mining. Using the software in this article, beginners will be able to predict trends, topics, new research concepts, and find duplicate text documents in articles, news, and blogs. Librarians will be able to improve their services in the library management system: reference services, CAS, SDI.

Keywords: classification, expert knowledge discovery, information extraction, patterns, text mining tools, text mining application, library management system.

https://doi.org/10.32782/mathematical-modelling/2022-5-2-5

Постановка проблеми

Більшість підприємств зберігають власні дані у електронному вигляді. В Інтернеті текст знаходиться в цифрових бібліотеках, блогах, соціальних мережах та електронних листах. Через великий потік інформації стає складнішим отримання вагомих знань. Для того, щоб їх отримати, потрібно все більше часу та зусиль.

Інтелектуальний аналіз тексту (ІАТ) – це процес видобування цінних образів для дослідження знань з текстових джерел. Це багатодисциплінарна сфера, яка базується на пошуку інформації, інтелектуальному аналізі даних, машинному навчанні, статистиці та комп'ютерній лінгвістиці (рис. 1). ІАТ працює з текстами природною мовою, які зберігаються в напів- або неструктурованому форматі.

Сферами застосування ІАТ є пошукові системи, системи управління взаємовідносинами з клієнтами, фільтрування електронних листів, аналіз пропозицій щодо продуктів, виявлення шахрайства та аналітика соціальних мереж. Також ІАТ використовують для сентимент-аналізу, виділяння ознак, прогнозування та аналізу тенденцій [1].

Видобування цінної інформації з набору різних документів є монотонною та стомлюючою роботою. Вибір відповідної техніки для аналізу тексту скорочує час і зусилля для пошуку відповідних образів для аналізу та прийняття рішень.

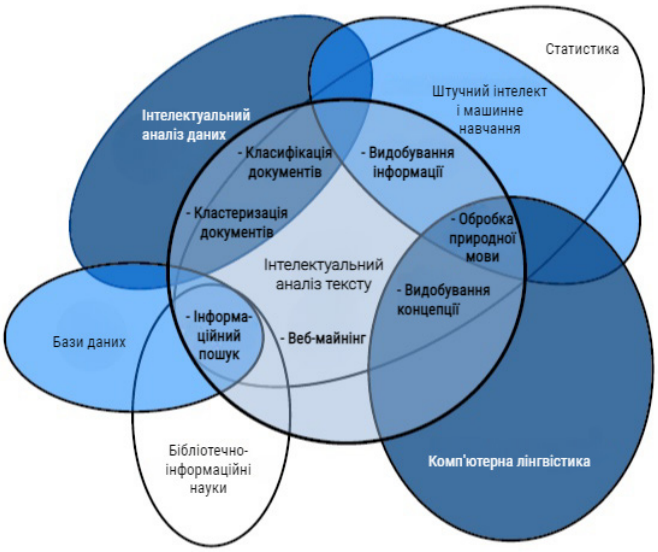

Рис. 1. Діаграма Венна – взаємодія інтелектуального аналізу тексту з іншими сферами

Аналіз останніх досліджень і публікацій

У [2] описано, що збір, видобування, попередня обробка, перетворення тексту, виділяння ознак, обрання образів та етапи оцінювання є етапами процесу інтелектуального аналізу тексту. Крім того, розглядаються різні методи ІАТ, які широко використовуються, такі як: кластеризація, категоризація, категоризація дерева рішень, а також їхнє застосування у різних сферах. В [3] висвітлено проблеми застосування і технік ІАТ. Зазначено, що складніше працювати з неструктурованим текстом, використовуючи традиційні інструменти і техніки ІАТ, ніж працювати зі структурованими або табличними даними. Описано застосування процесу інтелектуального аналізу тексту в біоінформатиці, бізнес-аналітиці та системі національної безпеки.

У [4] досліджено біомедичну базу даних MEDLINE, інтегровано структуру для розпізнавання іменованих сутностей, класифікації тексту, генерації та перевірки гіпотез, видобування зв'язків, синонімів та абревіатур. Така структура дозволила усунути зайві деталі та видобути вагому інформацію. У [5] проаналізовано текст за допомогою образів аналізу тексту та показано, що підходи, що базуються на термінах, не можуть належним чином аналізувати синоніми та полісемію. Крім того, був розроблений прототип моделі для специфікації образів з огляду

https://doi.org/10.32782/mathematical-modelling/2022-5-2-5

на призначення ваги відповідно до їх розподілу. Такий підхід допомагає підвищити ефективність процесу інтелектуального аналізу тексту. У [6] представлено систему розкриття злочинів з використанням засобів інтелектуального аналізу тексту та алгоритм виявлення зв'язків був розроблений для співвіднесення терміна з абревіатурою.

В [7] представлено підхід «зверху вниз» і «знизу вгору» для процесу видобутку тексту на основі веб-технологій. Щоб об'єднати подібні текстові документи, застосувано техніку кластеризації k-середніх для поділу «знизу вгору». Для виявлення подібності в документі використано алгоритм TF-IDF (Term Frequency- Inverse Document Frequency) для пошуку інформації щодо конкретних тем. У [8] надано огляд додатків, інструментів і проблем, які виникають під час видобування тексту. Зазначено, що документи можуть бути структурованими, напівструктурованими або неструктурованими, і що видобування корисної інформації є виснажливим завданням. Представлено загальну структуру для ІАТ на основі концепцій, яку можна візуалізувати у вигляді етапів уточнення тексту та дистиляції знань. Проміжна форма видобутку представлення сутності залежить від конкретної предметної області.

У [9] представлено інноваційні та ефективні методи виявлення закономірностей. Використано методи еволюції та виявлення образів для підвищення ефективності виявлення релевантної та відповідної інформації. Виконано фільтрацію на основі BM25 та векторної підтримки на основі корпусу маршрутизаторів та даних текстових конференцій, щоб оцінити ефективність запропонованої методики. В роботі [10] були проведені різні експерименти класифікації з використанням багатослівних ознак над текстом. Запропоновано власний метод вилучення багатослівних ознак з набору даних. Для класифікації та вилучення багатослівного тексту було здійснено поділ тексту на лінійні та нелінійні поліноміальні форми за підтримки векторного апарату, що дозволило підвищити ефективність вилучення даних.

Формулювання мети дослідження

Метою дослідження є аналіз різних методів інтелектуального аналізу тексту, які допомагають ефективно та результативно виконувати аналітику тексту з великої кількості даних; розгляд їх застосування; порівняння особливостей популярних інструментів технології ІАТ.

Викладення основного матеріалу дослідження

Існують різні методи інтелектуального аналізу текстів, які застосовуються для аналізу текстових образів та їхнього процесу видобування (рис. 2).

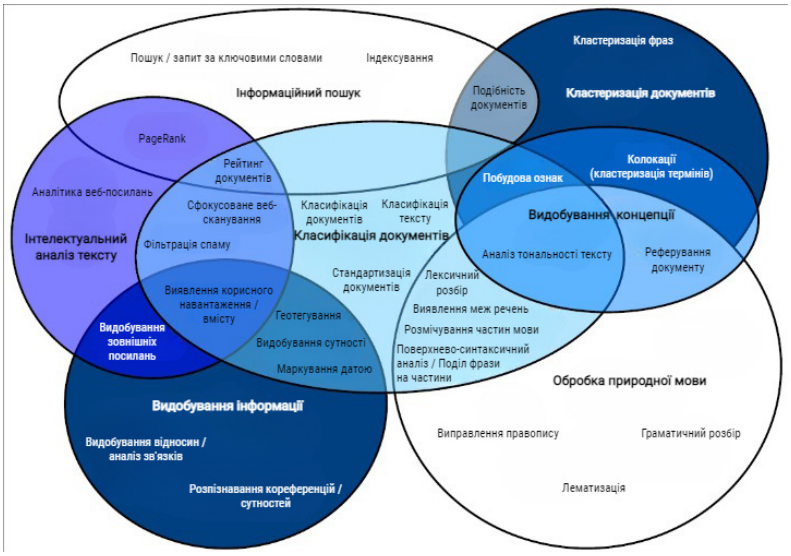

Рис. 2. Діаграма Венна – взаємозв'язок між методами інтелектуального аналізу текстів та функціональними можливостями

https://doi.org/10.32782/mathematical-modelling/2022-5-2-5

Загальний процес інтелектуального аналізу тексту складається з наступних кроків (рис. 3).

Процес очищення забезпечує охоплення справжньої суті доступного тексту та виконується для видобування коренів слів та індексування даних.

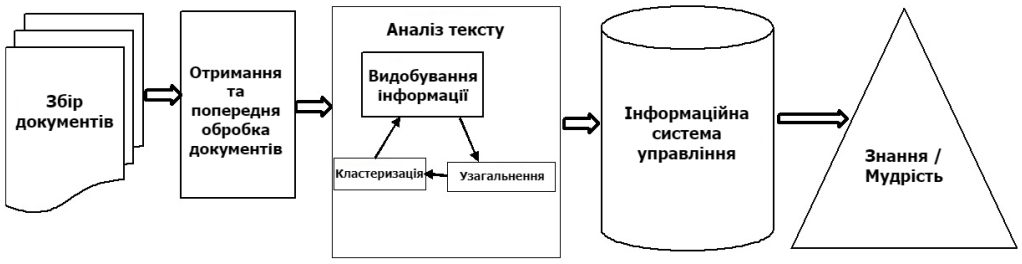

Рис. 3. Процес інтелектуального аналізу тексту

Попередня обробка тексту – це завдання перетворення вихідних даних у чітко визначені знання. Основні операції попередньої обробки текстових даних можна розділити на наступні етапи.

а) Попередня обробка тексту

• Процес токенізації: Процес розбиття потоку текстового контенту на слова, терміни, символи або інші значущі елементи, які називаються токенами. Фільтрація (стоп-слова) видаляє непотрібну інформацію, яка включає прийменники, артиклі, сполучники та ін.

• Процес лематизації: Процес, який схожий на стеммінг, але він ідентифікує словникові форми слів. Ця техніка використовується для скорочення довжини слів у тексті. У методах видобутку тексту етапи попередньої обробки відіграють важливу роль у перетворенні коренів слів у правильні корені для відповідного аналізу тексту. Процес ідентифікації коренів певних слів визначає походження слів. В основному використовують два типи походження: флективний та деривaційний. Найбільш поширеним алгоритмом визначення походження слів є алгоритм швейцара [11].

б) Процес трансформації тексту

• Перетворення тексту використовує пакет слів або модель векторного простору. Він виконує завдання вибору ознак. При цьому він зменшує розмірність, видаляючи надлишкові та нерелевантні ознаки [12]. Видаляється і послідовність слів, які часто зустрічаються, але не мають сенсу або не мають важливого змісту в колекції текстових документів.

Існують такі процеси ІАТ.

• Процес кластеризації тексту: вимірювання подібності та групування подібних текстів.

• Процес реферування (узагальнення) тексту: скорочує весь текст документів до суті.

• Процес категоризації тексту: автоматичний розподіл декількох документів на різні категорії. Є методом керованого навчання, ґрунтується на вхідних та вихідних екземплярах для класифікації нових документів. Основною метою категоризації тексту є автоматичне навчання класифікаторів на основі контрольованих і неконтрольованих категорій. Для цього можуть бути використані статистичні методи класифікації, такі як: наївний байєсівський класифікатор, класифікатор найближчого сусіда, дерево рішень та метод опорних векторів [13].

На рис. 4 зображено процес очищення тексту – видалення небажаних слів, знаків, символів, таблиць і малюнків, процес зведення модульованих слів до їх основи, кореня, які відфільтровуються до або після обробки тексту природною мовою.

На сьогодні R та Python є найбільш популярними інструментами для візуалізації ІАТ, особливо скриптового коду, а RapidMiner та Orange є інструментами графічної візуалізації без написання скриптів (табл. 1).

Дані інструменти є кросплатформними. Програми RapidMiner та Orange є зручними, оскільки користувачам потрібно лише перетягнути та з'єднати вузли разом. Вони призначені для складання бізнес-статистики, прогнозного аналізу, машинного навчання та візуалізації даних.

https://doi.org/10.32782/mathematical-modelling/2022-5-2-5

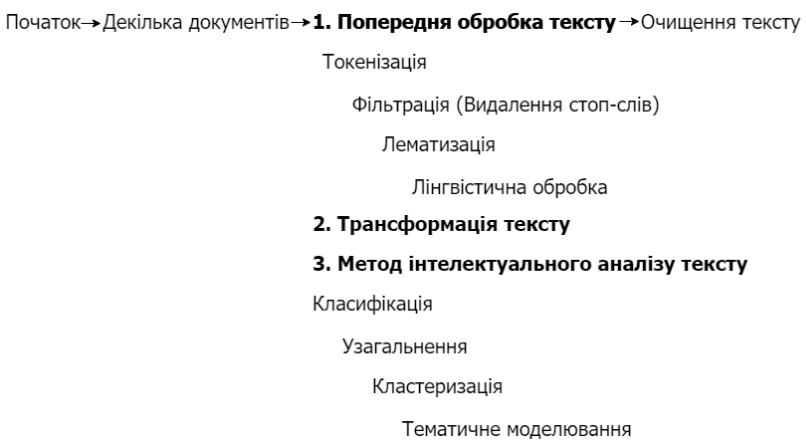

Рис. 4. Етапи попередньої обробки тексту для інтелектуального аналізу тексту

Таблиця 1

Загальна характеристика безкоштовних програмних засобів для ІАТ

Основні характеристики	RStudio	Python	RapidMiner	Orange
Початковий випуск програми	1992	1991	2006	1996
Розробник	Джозеф Аллер	Ґвідо ван Россум	Ральф Клінгенберг, Інго Мірсва та Саймон Фішер, Технічний університет Дортмунда	Люблянський університет
Мова програмування	C, Fortran, R	C, C++, Java	Java	Python, C, C++, Java
Доступна версія	3.4.3	3.9.1	9.6	3.4.5
Ліцензія	GNU з вільним відкритим вихідним кодом	безкоштовна та відкритий код	обмежена RapidMiner Studio Free	GNUv3 загальна публічна ліцензія
Головне призначення	аналітика даних, статистичний аналіз, машинне навчання	наука про дані	загальний ІАТ	ІАТ, машинне навчання, візуалізація даних, аналіз даних
Підтримка спільноти	дуже велика (~2 млн користувачів)	дані відсутні	велика (~200 тис. користувачів)	дані відсутні

Інтелектуальний аналіз тексту має багато застосувань.

1. Цифрові бібліотеки. Для визначення закономірностей і тенденцій з великої кількості сховищ на основі журналів і матеріалів використовуються численні методи та інструменти ІАТ. Ці джерела інформації допомагають у сфері досліджень і розробок.

Бібліотеки є важливим джерелом інформації для дослідників, а електронні бібліотеки прагнуть підвищити значимість своїх колекцій. Це новий метод організації інформації, що робить можливим доступ до трильйонів документів в Інтернеті.

Міжнародна цифрова бібліотека Greenstone, яка підтримує багатомовні інтерфейси, забезпечує гнучкий метод видобування документів, який працює з форматами Microsoft Word, PDF, PostScript, HTML, скриптовими мовами та електронними листами. Також підтримується видобування документів у аудіовізуальних та графічних форматах.

У процесі текстового аналізу виконуються операції вибору документів, збагачення, видобування інформації та пошук об'єктів серед документів, а також формування співвіднесення та узагальнення. Для текстового аналізу в електронних бібліотеках часто використовують інструменти GATE, Net Owl та Aylien.

В системі керування бібліотекою ІАТ також використовується для дослідження якості тексту та видобування інформації з декількох документів. Використовуючи методи інтелектуального аналізу тексту, можна:

https://doi.org/10.32782/mathematical-modelling/2022-5-2-5

а) автоматично ідентифікувати особу користувача академічної бібліотеки: ім'я, місцезнаходження, організацію, запити клієнтів, зацікавленості;

б) прогнозувати, які користувачі будуть передплачувати якийсь журнал;

в) дослідити проблеми, тенденції, а також шляхи вирішення конкретної теми дослідження;

г) прогнозувати, чи може пакет мережевих даних бути загрозою кібербезпеці;

ґ) просувати ринкові або рекламні пропозиції.

2. Академічна та дослідницька сфера. Різні інструменти та методи ІАТ використовуються у сфері освіти для аналізу освітніх тенденцій у конкретному регіоні, зацікавленості студентів у конкретній галузі та співвідношення зайнятості. Використання ІАТ у науковій сфері допомагає знайти та класифікувати в одному місці наукові роботи та відповідний матеріал з різних галузей. Використання кластеризації за методом k-середніх та інших методів допомагає виявити атрибути релевантної інформації. Можна отримати доступ до успішності студентів з різних предметів, побачити, як різні атрибути вплинули на вибір предметів.

3. Медико-біологічні науки. Медико-біологічні науки та охорона здоров'я створюють велику кількість текстових та числових даних, що стосуються обліку пацієнтів, хвороб, ліків, симптомів та методів лікування хвороб тощо.

Фільтрування відповідного та доцільного тексту з великого біологічного репозитарію для прийняття рішення є великим викликом. Медичні записи містять різну за характером інформацію, складну, довгу і технічну лексику, що робить процес пошуку знань достатньо важким.

Інструменти ІАТ у біомедичній галузі надають можливість видобувати вагомі відомості, їх асоціації та виводити взаємозв'язок між різними захворюваннями, видами та генами. Використання відповідних інструментів ІАТ медичній галузі допомагає оцінити ефективність медичних методів лікування, які показують ефективність шляхом порівняння різних захворювань, симптомів та перебігу їх лікування.

ІАТ використовується у виявленні біомаркерів, фармацевтичній промисловості, клінічному аналізі торгівлі, доклінічних дослідженнях безпечної токсичності, патентній конкурентній розвідці та ландшафтному аналізі, картографуванні генів хвороб та дослідженні цільових ідентифікацій за допомогою різних інструментів.

4. Соціальні мережі. Для аналізу додатків соціальних мереж існують програмні пакети ІАТ, які дозволяють відстежувати та аналізувати звичайний текст в Інтернеті з інтернет-новин, блогів, електронної пошти.

Інструменти ІАТ допомагають визначити та проаналізувати кількість постів, вподобань та читачів у соціальних мережах. Даний вид ІАТ показує реакцію людей на різні пости, новини і те, як вони поширюються. Він показує поведінку людей, що належать до певної вікової групи або схожих спільнот, варіативність переглядів публікації.

5. Бізнес-аналітика. ІАТ відіграє значну роль у бізнес-аналітиці, яка допомагає організаціям та підприємствам аналізувати своїх клієнтів та конкурентів для прийняття кращих рішень. Він забезпечує більш глибоке розуміння бізнесу та надає інформацію про те, як підвищити задоволеність клієнтів та отримати конкурентні переваги.

Інструменти ІАТ: IBM text analytics, Rapid miner, GATE – допомагають зробити рішення щодо організацій, які формують повідомлення про позитивні та негативні результати діяльності, зміни на ринку, що допомагають вжити заходів щодо виправлення. Вони використовуються у телекомунікаційній сфері, додатках для бізнесу та комерції, системі управління ланцюжками клієнтів.

Під час процесу ІАТ виникає багато проблем, які впливають на ефективність та результативність прийняття рішень. На етапі попередньої обробки визначаються різні правила і норми для стандартизації тексту, які роблять процес ІАТ ефективним. Перед застосуванням аналізу образів до документа необхідно перетворити неструктуровані дані в проміжну форму, але і на цьому етапі процес ІАТ має свої складнощі. Іноді реальна дані втрачають свою важливість через модифікацію в послідовності тексту.

https://doi.org/10.32782/mathematical-modelling/2022-5-2-5

Іншим важливим питанням є те, що існують небагато інструментів, які підтримують декілька мов. Для підтримки багатомовного тексту використовуються різні алгоритми та методи, що застосовуються незалежно один від одного. Багато важливих документів залишаються поза процесом видобування тексту, бо інструменти не підтримують їх.

Дані питання створюють багато проблем у процесі прийняття рішень. Реальної користі важко досягти, використовуючи лише існуючі методи та інструменти IAT, оскільки вони не завжди підтримують багатомовні документи.

Інтеграція знань предметної області є важливою сферою, оскільки вона виконує конкретні операції над конкретним корпусом і досягає бажаних результатів. У таких ситуаціях знання предметної області, з якої має бути вилучений корпус документів, повинно поєднуватися з обчислювальними можливостями, з яких має бути отримана інформація. Згідно з вимогами, експерти повинні працювати спільно з представниками різних галузей для отримання більш ефективних, точних і достовірних результатів.

Використання синонімів, багатозначних слів та антонімів у документах є проблемним для інструментів текстового аналізу, які сприймають слова в одному контексті. Важко класифікувати документи, коли колекція документів велика і сформована з різних галузей, що мають однакову тематику. Проблемними є абревіатури, що мають різне розшифрування у різному контексті. Різні концепції деталізації змінюють контекст тексту відповідно до умов і знань про предметну область.

Необхідно описати правила відповідно до сфери, які будуть використовуватися в якості стандарту в даній області і можуть бути вбудовані в інструменти IAT в якості плагіна. Розробка і розгортання плагінів для всіх сфер окремо потребує багато зусиль і часу. Для розробки плагінів потрібні глибокі знання про конкретну предметну область.

Природні мови також мають багато складнощів, що створюють проблеми в методах визначення тексту та ідентифікації зв'язків між сутностями. Слова, що мають однакове написання, але різні значення – «тепло» і «тепло». Інструменти текстового аналізу вважають обидва слова схожими, хоча одне з них є прислівником, а інше – іменником. Граматичні правила відповідно до природи та контексту все ще залишаються актуальним питанням у сфері інтелектуального аналізу текстів.

Висновки

Для отримання цінної інформації потрібний великий обсяг даних на основі тексту. Методи добування тексту використовуються для ефективного аналізу цікавої та актуальної інформації з великого обсягу неструктурованих даних. У даній статті представлено короткий огляд методів, які допомагають вдосконалити процес текстового аналізу. Для видобування корисної інформації застосовуються певні образи та послідовності, які усувають зайві для прогнозного аналізу деталі. Вибір і використання правильних методів та інструментів відповідно до предметної області допомагають зробити процес інтелектуального аналізу текстів простим і ефективним. Було порівняно безкоштовне програмне забезпечення для текстового аналізу, а саме: RStudio, Python, Orange і RapidMiner. Інтеграція знань про предметну область, деталізація різних концепцій, багатомовність тексту та неоднозначність обробки природної мови є основними проблемами та викликами, що виникають під час процесу інтелектуального аналізу текстів.

Список використаної літератури

1. Ramzan Talib, Muhammad Kashif Hanif, Shaeela Ayesha and Fakeeha Fatima. Text Mining: Techniques, Applications and Issues. *International Journal of Advanced Computer Science and Applications*, no.11(7), 2016. doi: 10.14569/IJACSA.2016.071153.
2. Liao S. H., Chu P. H., Hsiao P. Y. Data mining techniques and applications - a decade review from 2000 to 2011. *Expert Systems with Applications*, 2012, no. 39(12), pp. 11303-11311. doi: 10.1016/j.eswa.2012.02.063.

https://doi.org/10.32782/mathematical-modelling/2022-5-2-5

3. Zhong N., Li Y., Wu S.T. Effective pattern discovery for text mining. *IEEE transactions on knowledge and data engineering*, 2012, no. 1(24), pp. 30-44. doi: 10.1109/TKDE.2010.211.
4. Henriksson A., Moen H., Skeppstedt M., Daudaravicius V., Duneld M. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *Journal of biomedical semantics*, 2014, no. 1(5), p. 1, doi: 10.1186/2041-1480-5-6.
5. Laxman A., Sujatha D. Improved method for pattern discovery in text mining. *International Journal of Research in Engineering and Technology*, 2013, no. 1(2), pp. 2321-2328. doi: 10.15623/IJRET.2013.0210090.
6. Chen A.P., Zhang C.Y. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, 2014, vol. 275, pp. 314-347, 2014. doi: 10.1016/j.ins.2014.01.015.
7. Rajendra R., Saransh V. A Novel Modified Apriori Approach for Web Document Clustering. *International Journal of Computer Applications*, 2013, vol. 33, pp. 159-171. doi: 10.1007/978-81-322-2202-6_14.
8. Sumathy K. L., Chidambaram M. Text mining: Concepts, applications, tools and issues – an overview. *International Journal of Computer Applications*, 2013, no. 4 (80), pp. 29-32. doi: 10.5120/13851-1685.
9. Joby P. J., Korra J. Accessing accurate documents by mining auxiliary document information. *Advances in Computing and Communication Engineering (ICACCE)*, Second International Conference on. IEEE, 2015, pp. 634-638. doi: 10.1109/ICACCE.2015.37.
10. Wen Z., Yoshida T., Tang X. A study with multi-word feature with text classification. *Proceedings of the 51st Annual Meeting of the ISSS-2007*, Tokyo, Japan, 2007, vol. 51, p. 45.
11. Карпов І., Антоненко С. Огляд методів інтелектуального аналізу тексту. *Актуальні проблеми автоматизації та інформаційних технологій*. 2020. № 24. С. 40–46.
12. Sheela S. K., Bharathi T. Analyzing Different Approaches of Text Mining Techniques and Applications. *International Journal of Computer Science Trends and Technology*, 2018, no. 4(6).
13. Thakur K., Kumar V. An Overview of Text Mining: Application and Free Software Tools. *Library Waves*, 2020, no. 2(6), pp. 53-59.

References

1. Ramzan, T., Muhammad, K. H., Shaeela, A., & Fakeeha, F. (2016). Text Mining: Techniques, Applications and Issues. *International Journal of Advanced Computer Science and Applications*, **11** (7). doi: 10.14569/IJACSA.2016.071153.
2. Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications - a decade review from 2000 to 2011. *Expert Systems with Applications*, **39**(12), 11303-11311. doi: 10.1016/j.eswa.2012.02.063.
3. Zhong, N., Li, Y., & Wu, S.T. (2012). Effective pattern discovery for text mining. *IEEE transactions on knowledge and data engineering*, **1** (24), 30-44. doi: 10.1109/TKDE.2010.211.
4. Henriksson, A., Moen, H., Skeppstedt, M., Daudaravicius, V., & Duneld, M. (2014). Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *Journal of biomedical semantics*, **1** (5), 1, doi: 10.1186/2041-1480-5-6.
5. Laxman, A., & Sujatha, D. (2013). Improved method for pattern discovery in text mining. *International Journal of Research in Engineering and Technology*, **1** (2), 2321-2328. doi: 10.15623/IJRET.2013.0210090.
6. Chen, A.P., & Zhang, C.Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, **275**, 314-347. doi: 10.1016/j.ins.2014.01.015.

https://doi.org/10.32782/mathematical-modelling/2022-5-2-5

7. Rajendra, R., & Saransh, V. (2013). A Novel Modified Apriori Approach for Web Document Clustering. *International Journal of Computer Applications*, **33**, 159-171. doi: 10.1007/978-81-322-2202-6_14.
8. Sumathy, K. L., & Chidambaram, M. (2013). Text mining: Concepts, applications, tools and issues – an overview. *International Journal of Computer Applications*, **4** (80), 29-32. doi: 10.5120/13851-1685.
9. Joby, P.J., & Korra, J. (2015). Accessing accurate documents by mining auxiliary document information. *Advances in Computing and Communication Engineering (ICACCE)*, Second International Conference on. IEEE, 634-638. doi: 10.1109/ICACCE.2015.37.
10. Wen, Z., Yoshida, T., & Tang, X. (2007). A study with multi-word feature with text classification. *Proceedings of the 51st Annual Meeting of the ISSS-2007*, Tokyo, Japan, **51**, 45.
11. Karpov, I., & Antonenko, S. (2020). Ohliad metodiv intelektualnoho analizu tekstu. *Aktualni problemy avtomatyzatsii ta informatsiinykh tekhnolohii*, **24**, 40–46.
12. Sheela, S. K., & Bharathi, T. (2018). Analyzing Different Approaches of Text Mining Techniques and Applications. *International Journal of Computer Science Trends and Technology*, **4** (6).
13. Thakur, K., & Kumar, V. (2020). An Overview of Text Mining: Application and Free Software Tools. *Library Waves*, **2** (6), 53-59.

Могильна Марія Володимирівна – студентка кафедри програмних засобів Національного університету «Запорізька політехніка», e-mail: mariiamohylna@gmail.com, ORCID: 0000-0001-7190-8698.

Дубровін Валерій Іванович – к.т.н., професор кафедри програмних засобів Національного університету «Запорізька політехніка», e-mail: vdubrovin@gmail.com, ORCID: 0000-0002-0848-8202.

https://doi.org/10.32782/mathematical-modelling/2022-5-2-5