Issue 14, 2023

https://doi.org/10.23939/sisn2023.14.235

УДК 004.9

ІНТЕЛЕКТУАЛЬНА СИСТЕМА ВИЯВЛЕННЯ ПЛАГІАТУ В ТЕХНІЧНИХ ТЕКСТАХ

Юрій Геряк[1], Андрій Берко[2]

Національний університет "Львівська політехніка", кафедри інформаційних систем та мереж
вул. С. Бандери, 12, Львів, Україна
[1] yurii.heriak.mnitm.2021@lpnu.ua, ORCID: 0009-0008-3251-2007
[2] andrii.y.berko@lpnu.ua, ORCID: 0000-0003-2892-9519

Автори розробили наукове обґрунтування, виконали проєктування та розроблення інтелектуальної системи виявлення плагіату в технічних текстах. В роботі визначено проблему плагіату в сучасному світі та її актуальність, проаналізовано останні дослідження та публікації, які стосуються новітніх методів застосування інтелектуальних інформаційних технологій для виявлення плагіату. Обґрунтовано потребу і доцільність розроблення та вдосконалення інтелектуальних інформаційних технологій виявлення плагіату, а також застосування різних методів ідентифікації збігів у текстах для подальшого розвитку таких технологій. Розроблено загальний алгоритм виявлення плагіату в технічних текстах на основі методу векторного порівняння. Практичним результатом дослідження є розроблення інтелектуальної системи виявлення плагіату в технічних текстах та підтвердження її працездатності із застосуванням на конкретних прикладах технічних текстів.

Ключові слова: плагіат; машинне навчання; інтелектуальна система; текстові документи; векторне порівняння.

Вступ та постановка проблеми

В сучасному світі інформаційних технологій, де доступ до інформації є швидким і необмеженим, питання інтелектуальної власності та плагіату стають все актуальнішими. Сучасні інформаційні технології дають змогу легко спілкуватися, обмінюватися знаннями та використовувати інформацію з різних джерел, але разом з цим зростає й ризик порушення авторських прав та поширення некоректної інформації.

Однією із серйозних проблем, пов'язаних із використанням інформаційних технологій, є плагіат – копіювання або незаконне використання чужого матеріалу без дозволу або належної атрибуції. Плагіат вважають істотною проблемою наукової та технічної спільноти, оскільки він не тільки порушує права власності, але й вводить читача в оману, спотворюючи оригінальність автора. Завдяки швидкому доступу до великої кількості інформації та легкості її поширення плагіат стає все поширенішим явищем. Розглянемо основні проблеми, пов'язані з виявленням плагіату.

Складність виявлення. Зі зростанням обсягу доступної інформації стає складніше виявити плагіат. Автори можуть використовувати різні способи приховування копіювання, змінюючи послідовність слів, перефразовуючи фрази або вносячи незначні зміни у контент. Це ускладнює автоматичну перевірку та потребує використання складних алгоритмів та технологій для виявлення плагіату.

Різноманітність джерел. Можливий плагіат з різних джерел, ураховуючи інтернет, книги, журнали, статті та навіть інші наукові роботи. Для виявлення плагіату потрібен доступ до широкого спектра джерел і засобів пошуку.

Поліморфізм текстів. Автори можуть використовувати різні стилі письма, граматичні конструкції та лексику, що ускладнює автоматичне порівняння текстів для виявлення плагіату. Додатково використання різних мов або переклади також можуть приховувати фактичний плагіат.

Розвиток нових технологій. Разом зі швидким розвитком інформаційних технологій, автори плагіату також набувають нових можливостей, щоб приховати копіювання – використання автоматичного перекладу, генерації текстів тощо.

З огляду на перелічені чинники та складності виявлення плагіату, можна зазначити, що ця проблема загострилась у сучасному світі, а методи її вирішення потребують постійних та проактивних удосконалень, щоб виявляти відомі та запобігати новим способам застосування плагіату в науковому середовищі та інших сферах діяльності, де інтелектуальна власність має значення.

Аналіз останніх досліджень та публікацій

Аналізуючи дослідження проблеми плагіату в Україні, можна стверджувати, що він є найпоширенішим видом академічної недоброчесності, передбачає продаж та використання робіт і дисертацій на комерційному рівні. В українській академічній спільноті поширення набув плагіат як “акт шахрайства, який включає купівлю, викрадення та позичення чужих ідей та їхнє навмисне чи випадкове видавання за свої” [1]. Ці факти підтверджують дані опитування, згідно з яким 93 % опитаних студентів практикують плагіат у певній формі, і більша частина дисертацій порушують стандарти академічної доброчесності, містять плагіат [2]. На рис. 1 вказано найпоширеніші варіанти плагіату, які назвали студенти.

Згідно з іншими опитуваннями, які виконали у 2015 р. Східноукраїнський фонд соціальних досліджень та Інститут соціально-гуманітарних досліджень Харківського національного університету ім. В. Н. Каразіна для проєкту “Академічна культура українського студентства: основні чинники формування та розвитку”, найпоширенішою формою академічної недоброчесності є використання наукових робіт з мережі Інтернет без бібліографічних посилань на них. Це підтверджують 75 % опитаних викладачів та 60 % студентів. 41 % викладачів та 39 % студентів зазначили, що поширеною формою плагіату є привласнення робіт однокурсників чи випускників. Крім того, достатньо поширеною формою плагіату в Україні є купівля робіт і дисертацій у приватних компаніях, науково-викладацького складу або інших студентів [3].

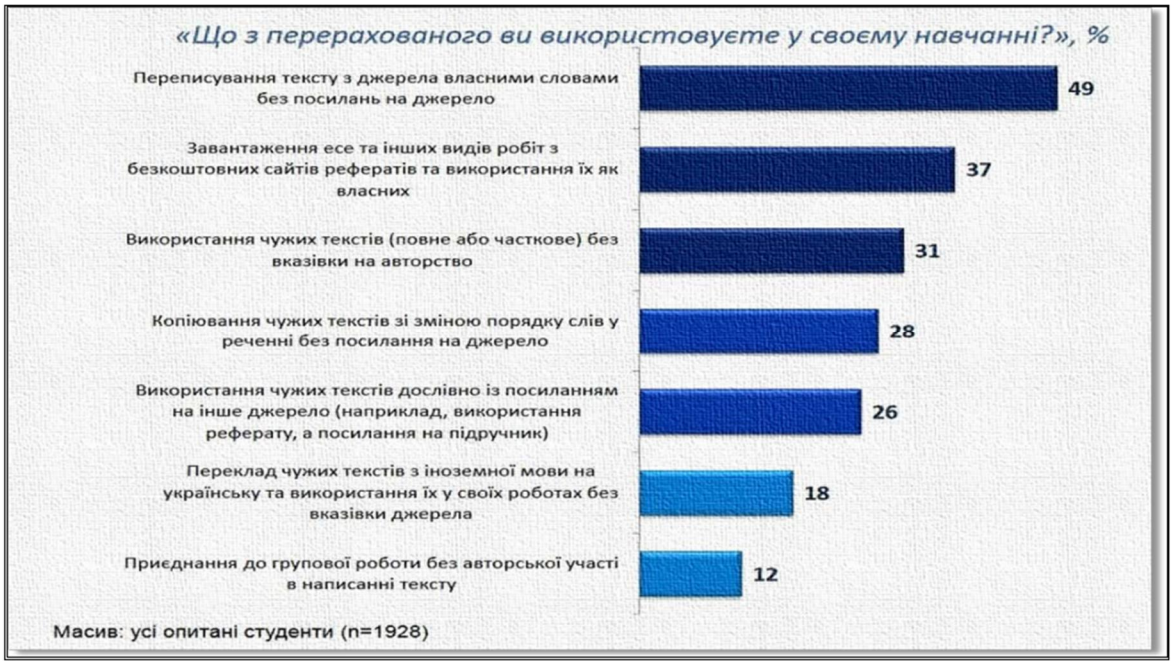


Рис. 1. Практика використання плагіату в студентському навчанні

Феномен плагіату постійно цікавив багатьох дослідників, які розробили ефективні системи для його виявлення. Але основною проблемою плагіату є різноманітність ступенів його складності, як на лексичному чи синтаксичному, так і на семантичному рівні. Багато досліджень ґрунтуються на статистичних методах виявлення плагіату. Санчес Вега у своїй роботі [4] ввів поняття автоматичного підходу до ідентифікації збігів, який ґрунтується на характеристиках рівнів символів. Ключовим відкриттям було дослідження тексту за змістом та стилістикою, що дало змогу виявляти подібні фрагменти. Основним недоліком цього методу є нехтування семантичним аспектом.

Sanchez-Perez та його колеги у роботі "A winning approach to text alignment for text reuse detection" [5] запропонували систему, яка виявляла текстові збіги між підозрілим документом та джерелом. Вони зробили внесок у використання алгоритму TF-ISF (Term Frequency – Inverse Sentence Frequency) на реченнях для вилучення можливих випадків збігу. Також об'єктом дослідження був рекурсивний алгоритм для пошуку суміжних випадків найбільш плагіатних фрагментів та вирішення проблеми їх накладання. Додатковою опцією системи була конфігурація системи залежно від типу плагіату. Недолік системи – те, що неможливо аналізувати лінгвістичний аспект.

M. Roostaee, S. M. Fakhrahmad, M. H. Sadreddini запропонували дворівневу стратегію порівняння [6], щоб досягти правильного вирівнювання фрагментів подібності текстового джерела та досліджуваного документа. Цей підхід враховує як синтаксичні, так і семантичні аспекти. На першому рівні застосовують модель векторного простору, яка залежить від способу визначення ваги слова та його мультимовності, на підставі словника найменших схожих пар фрагментів. Далі, на другому рівні, використовують граф представлень слів, за допомогою якого здійснюють попарні порівняння. Обмеженням цього підходу є розпізнавання коротких випадків.

У статті "A new hybrid technique for detection of plagiarism from text documents" авторів L. Ahuja, V. Gupta, R. Kumar [7] описано підхід, оснований на вилученні синтаксичних та семантичних знань із текстових документів. Ці знання вираховують як сукупність лінгвістичних аспектів: схожість шляху та оцінка глибини з різними вагами. Крім того, для оцінки синтаксичних та семантичних складових тексту використано матрицю подібності. Цей підхід не ґрунтується на методах машинного навчання і не завжди може виявити складні випадки плагіату, як, наприклад, переклад. Gharavi та ін. [8] розробили масштабовану та незалежну від мови систему для виявлення плагіату, основану на вкладанні тексту із синтаксичною та семантичною інформацією для порівняння. Порівняння виконували на рівні речень з метою виділення пар підозрілих та оригінальних речень з найвищим рівнем схожості. У цій системі запропоновано три методи налаштування параметрів. Перший метод, відомий як налаштування порога в автономному режимі, передбачав кілька тестів на навчальному наборі даних з різними значеннями параметрів. Другий метод називався налаштуванням порога в автономному режимі з обфускацією. У цьому аналізі тип приховування розглядався як особливий. Третій метод – налаштування порога в режимі онлайн; параметри змінювалися з одного типу на інший залежно від стандартного відхилення або медіани абсолютного відхилення для всіх значень схожості Джаккарда між кожною виявленою парою. Результат показав, що тип обфускації все ще потребує додаткового вивчення.

Altheneyan та ін. [9] створили автоматичну систему виявлення плагіату, основану на класифікаторі методу опорних векторів (SVM) з лексичними, синтаксичними та семантичними ознаками. Залежно від використаного ядра створено два прототипи: лінійне ядро (PlagLinSVM) та радіусне ядро (PlagRbfSVM). Реалізація системи відбувалась переважно у два етапи: абзацний та реченнєвий. На абзацному етапі виявлялися пари схожих абзаців між підозрілими та джерелом за допомогою порівняння загальних уніграм і біграм. Після цього на реченнєвому етапі виявлялися пари схожих речень на основі загальних уніграмів та метеор-оцінок заздалегідь заданою умовою. Однак, якщо умова не виконується, класифікатор SVM визначає, чи речення схожі, чи ні. Наприкінці су-

міжні випадки розширюються, утворюючи великі плагіатні фрагменти між підозрілими та джерелом. Недоліком цієї системи є те, що вона не досліджувала кілька евристик злиття та техніки глибокого навчання.

Узагальнюючи аналіз останніх досліджень та публікацій, які стосуються теми виявлення плагіату за допомогою інтелектуальних систем, можна дійти висновку, що дослідження виявлення плагіату свідчать про постійну потребу в розробленні та вдосконаленні інтелектуальних інформаційних систем, використанні різноманітних ознак та методів, а також подальшому розвитку технологій для боротьби із плагіатом у сучасному світі.

Формулювання цілі статті

Мета дослідження – обґрунтувати проєктування та реалізацію інтелектуальної системи виявлення плагіату в технічних текстах, зокрема, з інформаційних технологій. Для досягнення мети роботи поставлено такі завдання: формулювання проблеми виявлення плагіату в текстах з використанням інформаційних технологій та аналіз відомих методів, підходів, публікацій, досліджень стосовно стану її вирішення; вибір методів та засобів розроблення рішення проблеми на основі алгоритмів та моделей машинного навчання; розроблення та імплементація програмного забезпечення для побудови інтелектуальної системи виявлення плагіату в технічних текстах; експериментальне дослідження розробленої системи та аналіз її функціонування; аналіз результатів дослідження та формулювання висновків щодо можливості використання розробленої системи в практичній діяльності для контролю за оригінальністю текстів у сфері інформаційних технологій.

Результатом роботи стала інтелектуальна система, яка дає змогу поліпшити якість контролю за оригінальністю технічних текстів, насамперед, у галузі інформаційних технологій, зменшити кількість випадків плагіату та, відповідно, зберегти репутацію авторів і видавництв.

Крім того, розроблення такої інформаційної системи буде корисним для освітніх та наукових закладів, дасть змогу підвищити ефективність перевірки на плагіат у студентських та наукових роботах. Основні результати роботи можна використати в інших сферах, таких як журналістика, література, маркетинг тощо.

Основні результати дослідження

Більшість наукових праць є продовженням результату попередніх досліджень, а отже, цитування або посилання на джерела є історично прийнятним явищем. Але часто науковці заходять занадто далеко і прямо копіюють без зазначення авторства.

Для розв’язання задачі інтелектуального виявлення плагіату в текстах з інформаційних технологій проаналізовано та оцінено три популярні методи порівняння текстів

Перший метод – **порівняння текстів за допомогою хеш-функцій**. Цей метод полягає в обчисленні хеш-функцій для різних фрагментів тексту та порівнянні отриманих значень для визначення ступеня подібності між текстами. Якщо певний відсоток значень хеш-функцій для двох текстів збігається, то можна стверджувати про істотну спільність цих текстів.

Хеш-функція – функція, яка перетворює вхідні дані будь-якого розміру на вихідний рядок фіксованого розміру. Хеш-функції зазвичай застосовують для шифрування та перевірки цілісності даних. Для порівняння текстів за допомогою хеш-функцій зазвичай використовують хеш-функції з розміром вихідного рядка близько до 128 біт [19].

Щоб порівняти два тексти за допомогою хеш-функцій, необхідно спочатку обчислити хеш-суму кожного тексту за допомогою вибраної хеш-функції. Потім порівнюють хеш-суми. Якщо вони збігаються, це може вказувати на наявність плагіату. На цьому методі ґрунтується алгоритм шинглів.

Хоча іноді цей метод доволі ефективний, він не завжди дає точний результат. У деяких випадках у двох текстів подібні структурі та зміст, але різні хеш-суми через незначні відмінності в символах або форматуванні. Хеш-функції можна використовувати також як частину складніших методів порівняння текстів, таких як метод Левенштейна [5].

Другий метод – **використання методів машинного навчання**. цей метод полягає в створенні моделі машинного навчання, яка буде навчена виявляти плагіат на основі вхідних даних. Для цього можна використовувати різноманітні алгоритми машинного навчання, такі як класифікація, кластеризація, нейронні мережі тощо.

Переваги використання методів машинного навчання для виявлення плагіату такі [20]:

· Висока точність. Методи машинного навчання можуть дуже точно виявляти плагіат, оскільки використовують складні математичні алгоритми для аналізу текстів.

· Ефективність. За допомогою методів машинного навчання можна швидко обробляти значні обсяги текстової інформації, що важливо у разі оброблення великої кількості документів.

· Гнучкість. Методи машинного навчання можна налаштовувати для різних типів документів і завдань, що дає змогу використовувати їх у різних контекстах.

До недоліків використання методів машинного навчання для виявлення плагіату зараховують:

· Вимоги до даних. Ці методи потребують значної кількості даних для навчання, тому для створення моделі може бути необхідна велика кількість документів.

· Складність. Методи машинного навчання потребують високої кваліфікації в галузі математики та програмування, що складно для людей без необхідної підготовки.

· Чутливість до якості даних. Методи машинного навчання чутливі до якості даних, тому якість результатів може залежати від якості даних, на яких вони навчалися.

Метод векторного порівняння є одним з методів розв'язання задачі виявлення плагіату в текстах. Оснований на перетворенні текстових документів на вектори та порівнянні їх між собою [22].

Для семантичного пошуку текстів ми використали процесори природних мов, такі як модель векторного подання речень (*Sentence Embedding Model*) та **модель середніх векторів слів** (*Average Word Embeddings Model*).

Модель середніх векторів слів – простий та ефективний засіб для отримання векторного зображення тексту. Для цього потрібно взяти векторне подання кожного слова у тексті та обчислити середнє значення векторів усіх слів у тексті. У результаті отримують один вектор, який зображає текст і яким можна скористатись для класифікації тексту, внесення тексту до певної категорії та інших завдань оброблення природної мови.

Векторне подання слів у моделі середніх векторів слів можна створити за допомогою різних алгоритмів, таких як Word2Vec або GloVe. Ці алгоритми навчають модель на великому корпусі текстів і створюють вектори, які відображають семантичну близькість між словами.

Одна із основних переваг моделі середніх векторів слів – її простота та швидкість. Вона придатна для оброблення значних обсягів тексту, її виконання не потребує великої кількості ресурсів.

Однак модель середніх векторів слів також має певні обмеження. Наприклад, вона не враховує послідовності слів у тексті та контекстуальної семантики слів. Також вона може бути менш ефективною для текстів, які містять багато фраз та ідіом, де значення тексту складніше одержати за допомогою простого усереднення векторів слів [21].

Результати порівняльного оцінювання перелічених методів подано у таблиці.

Аналіз та оцінювання методів порівняння текстів

Метод	Переваги	Недоліки
Порівняння текстів за допомогою хеш-функцій	1. Простота формалізації та алгоритмізації 2. Прості критерії порівняння текстів 3. Для виявлення збігів і розбіжностей опрацьовують не сам текст, а хеш-значення 4. Значне зменшення обсягів даних для порівняння	1. Значна частина рядків тексту можуть мати невеликі хеш-значення 2. Хеш-значення різних рядків можуть бути однаковими 3. Для встановлення збігів і розбіжностей необхідне додаткове порівняння підрядків 4. Великі часові затрати 5. Невисока точність результатів порівняння
Порівняння текстів за допомогою методів машинного навчання	1. Універсальність 2. Висока точність 3. Ефективність 4. Гнучкість	1. Складні вимоги до вхідних текстів 2. Складність реалізації 3. Чутливість до якості даних
Метод векторного порівняння	1. Можливість застосування різних алгоритмів 2. Простота реалізації 3. Можливості машинного навчання 4. Універсальність 5. Відсутність спеціальних вимог до текстів	1. Менша точність порівняння 2. Можливість виникнення помилок 3. Більші обсяги даних для порівняння текстів

На основі виконаного аналізу та оцінювання для реалізації проєкту інтелектуальної системи виявлення плагіату в технічних текстах ми вибрали комбінацію методів порівняння текстів за допомогою хеш-функції та методу векторного порівняння із використанням середніх векторів слів.

Загальний алгоритм функціонування системи виявлення плагіату, побудованої на застосуванні цього методу, передбачає виконання такої послідовності кроків.

Крок 1: побудова векторів для кожного текстового документа, які будемо порівнювати. Для цього використовують такі методи як, "мішок слів" (*bag of words*), "мішок слів з TF-IDF" (term *frequency-inverse document frequency*) або побудову "N-грами" (*N-grams*).

Крок 2: порівняння векторів між собою, яке виконують за допомогою метрик подібності, таких як косинусна подібність *(cosine similarity*), евклідова відстань (*Euclidean distance*) або відстань Левенштейна.

Крок 3: обчислення результатів порівняння – числових мір подібності між текстовими документами, які подано у вигляді відсотків або чисел із певним інтервалом.

Крок 4: формулювання рішення про наявність плагіату на основі результатів порівняння текстів.

Проєктування інтелектуальної системи виявлення плагіату в технічних текстах

Основний набір функціональних вимог системи виявлення плагіату в технічних текстах подано на діаграмі варіантів використання (рис. 2). Система підтримуватиме два види користувачів – звичайний Користувач та Адміністратор, який має ті самі варіанти взаємодії з системою, що і Користувач, окрім можливості маніпулювання текстами та користувачами (рис. 2).

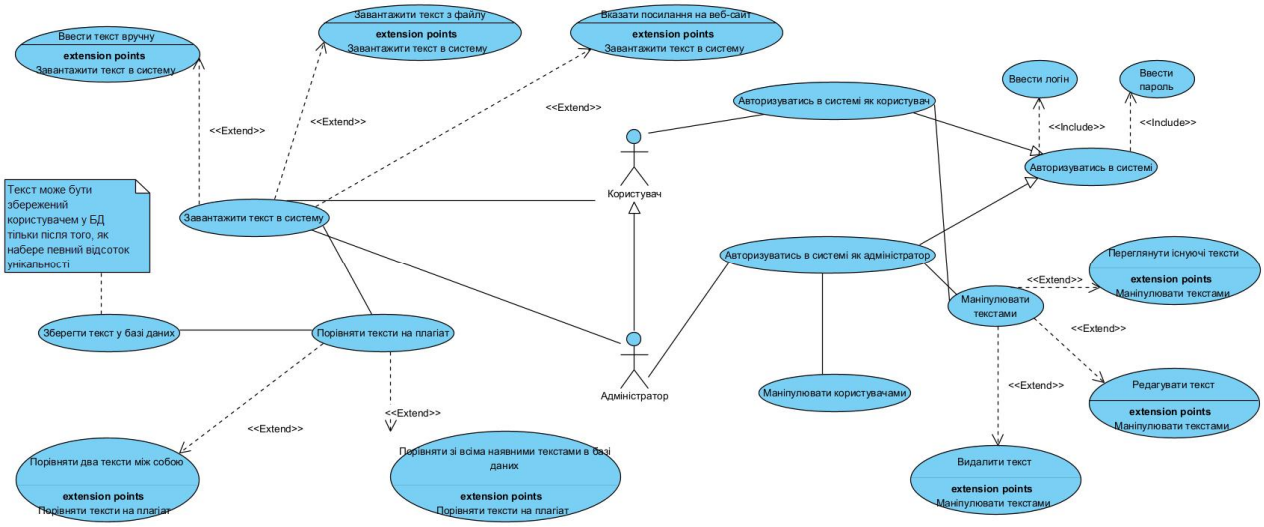


Рис. 2. Діаграма варіантів використання інтелектуальної системи виявлення плагіату в технічних текстах

Згідно із побудованою моделлю роль *Користувач* передбачає виконання таких функцій:

1) завантаження тексту (ручним способом, з файлу, із вебсайту);
2) збереження тексту в базі даних (репозиторії текстів) системи;
3) порівняння завантаженого тексту (із заданим текстом або з текстами в базі даних);
4) отримання та інтерпретація результатів порівняння текстів.

З роллю *Адміністратор* пов'язано виконання перелічених вище функцій, а також додатково функції адміністрування системи, які дають змогу:

1) реєструвати нових користувачів системи;
2) надавати користувачам та вилучати права і повноваження;
3) отримувати й аналізувати статистику використання системи та активності користувачів.

Загальну структуру програмного забезпечення розроблюваної інтелектуальної інформаційної системи виявлення плагіату в технічних текстах описує діаграма класів (рис. 3). Ключові класи цієї системи випливають з чотирьох основних завдань, які виконує система: авторизація і контроль доступу, адміністрування профілів користувачів, адміністрування і маніпулювання базою даних збережених текстів, порівняння текстів (рис. 3). Такими класами в структурі програмного забезпечення системи виявлення плагіату в технічних текстах, зокрема, є:

1) користувач;
2) текст;
3) репозиторій користувачів;
4) репозиторій текстів;
5) об'єднаний репозиторій (база даних текстів і користувачів);
6) порівнювач текстів.

Окрім цих елементів, структура системи передбачає використання чотирьох контролерів, а саме:

1) контролер тексту;
2) контролер порівняння;
3) контролер користувачів;
4) контролер авторизації.

Описання взаємодії об'єктів на високому рівні абстракції інтелектуальної системи виявлення плагіату в текстах з інформаційних технологій подано за допомогою діаграми кооперацій (рис. 4). Діаграма дає змогу зрозуміти послідовний процес взаємодії об'єктів, які викликають певні функції та виконують операції у межах системи.

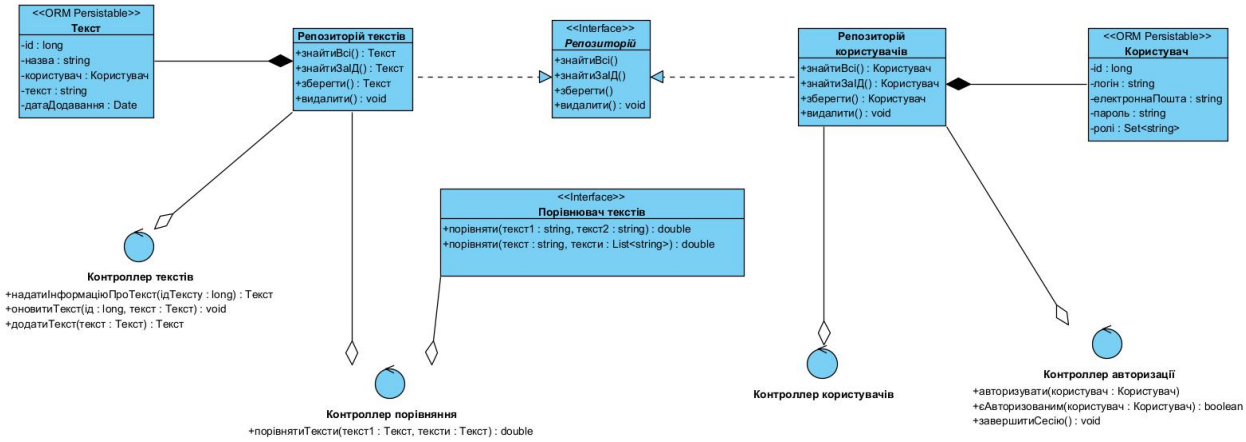


Рис. 3. Діаграма класів інтелектуальної системи виявлення плагіату в технічних текстах

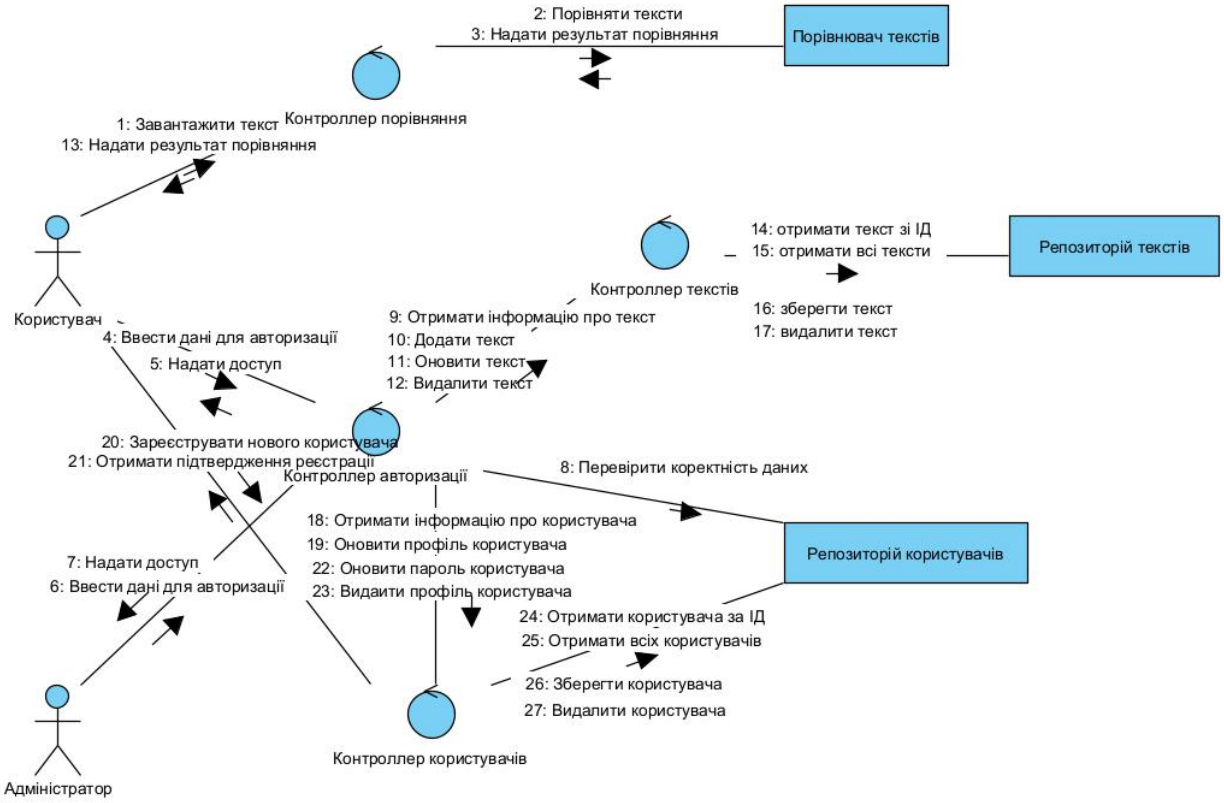


Рис. 4. Діаграма кооперацій інтелектуальної системи виявлення плагіату в технічних текстах

Як показано на рис. 4, *Користувач* системи реалізує свої функції, взаємодіючи з такими елементами системи, як *Репозиторій текстів* та *Порівнювач текстів*, через контролери авторизації, порівняння та контролер текстів. *Адміністратор* додатково взаємодіє із *Репозиторієм користувачів* через контролер авторизації та контролер користувачів.

Результати моделювання бізнес-процесів та послідовності дій, які необхідно виконати користувачу інтелектуальної системи виявлення плагіату в технічних текстах, подано у формі діаграм діяльності [15]. На рис. 5 зображено діяльність системи, коли користувач починає взаємодіяти з нею. Користувачу надано два способи взаємодії: завантажити текст для перевірки на плагіат або ввести авторизаційні дані – логін та пароль.

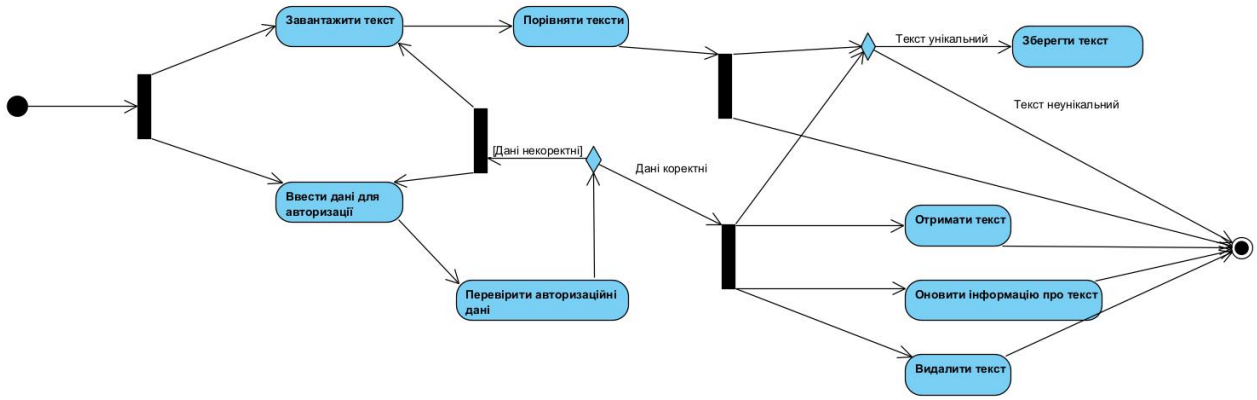


Рис. 5. Діаграма діяльності інтелектуальної системи виявлення плагіату в технічних текстах

Подальшими діями користувача системи, відповідно до побудованої моделі, є дії з текстами для визначення та вимірювання плагіату:

1) завантаження тексту;
2) маніпулювання текстом – отримання, оновлення, видалення;
3) порівняння тексту із відповідником;
4) збереження тексту та результатів порівняння на плагіат.

Остаточну структуру розробленої інтелектуальної системи виявлення плагіату в технічних текстах на рівні компонентів та зв'язків між ними описано у формі діаграми компонентів [13]. У контексті цієї системи її структуру побудовано довкола трьох основних компонент – програми *PlagDetector*, сервісу векторного порівняння *plagiarism-detector* та *Бази Даних*. Компонент *PlagDetector* залежить від двох інших компонентів: *Web-сервер* та *Web-інтерфейс*. Крім цього, на діаграмі компонентів показано інтерфейси взаємодії компонентів та їх взаємодію із зовнішніми системами. Як бачимо (рис. 6), основна програма взаємодіє із *plagiarism-detector* сервісом через REST API, а *plagiarism-detector* узалежнений від зовнішньої векторної бази даних, яка розміщена на сервісі *Pinecode*.

На основі цієї діаграми компонентів виконано завдання розроблення, аналізування і тестування системи виявлення плагіату в технічних текстах.

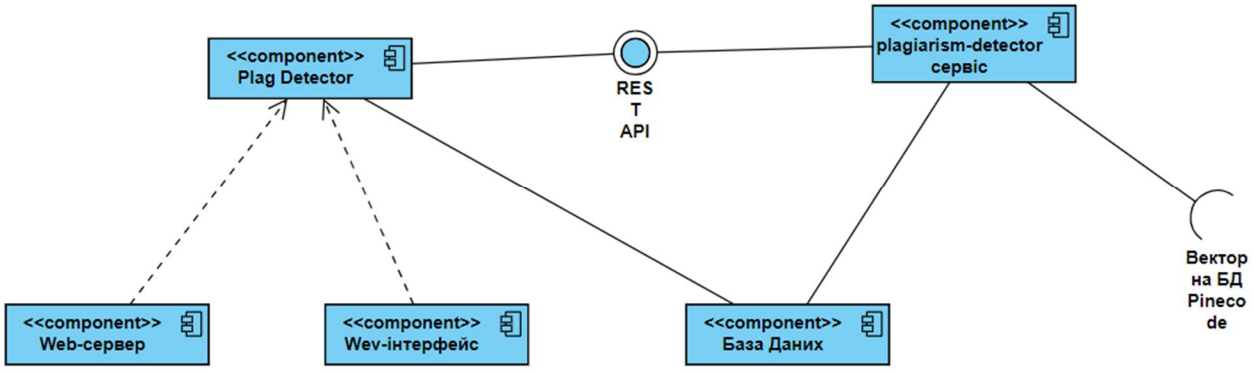


Рис. 6. Діаграма діяльності інтелектуальної системи виявлення плагіату в технічних текстах

Результатом виконаної роботи є інтелектуальна інформаційна система, яка має змогу виявляти плагіат у текстах, використовуючи два методи – метод порівняння текстів за допомогою хеш-функцій, на якому ґрунтується алгоритм шинглів, що є основою функціональності порівняння двох

текстів між собою. Другий метод, який використовує система, – метод векторного порівняння, для реалізації якого застосовують методи аналізу даних та оброблення природних мов. Система розроблена як вебсайт, який надає користувачеві змогу завантажити текст для порівняння одразу після завантаження головної сторінки (рис. 7). Для зберігання текстів у базі даних системи користувач повинен авторизуватись.

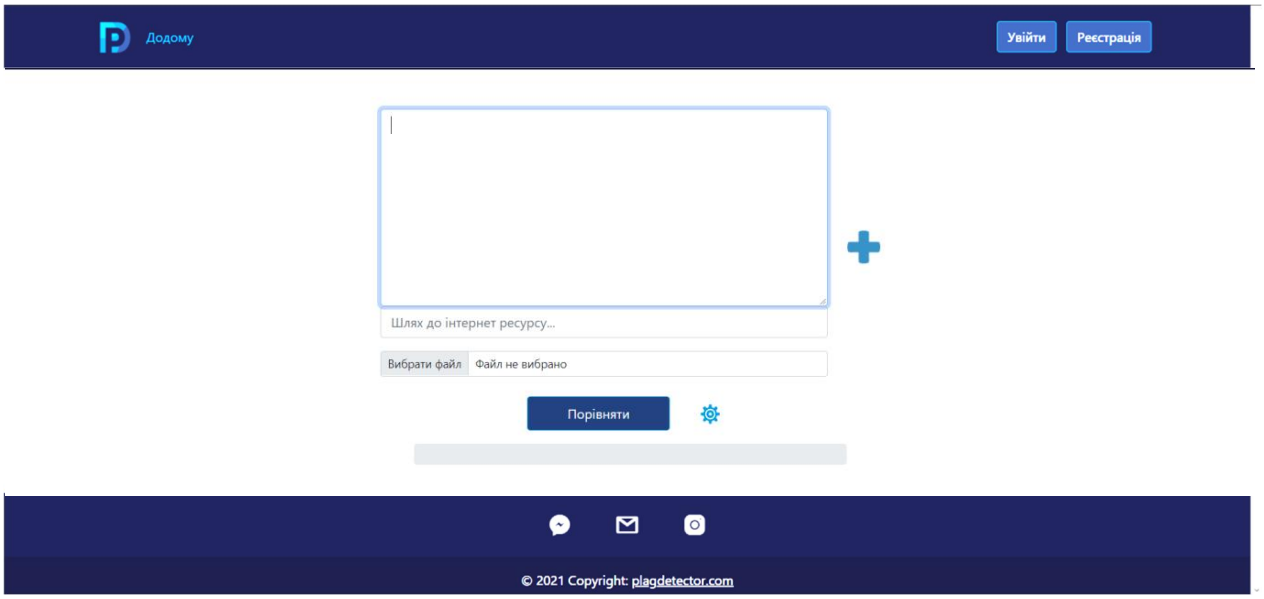
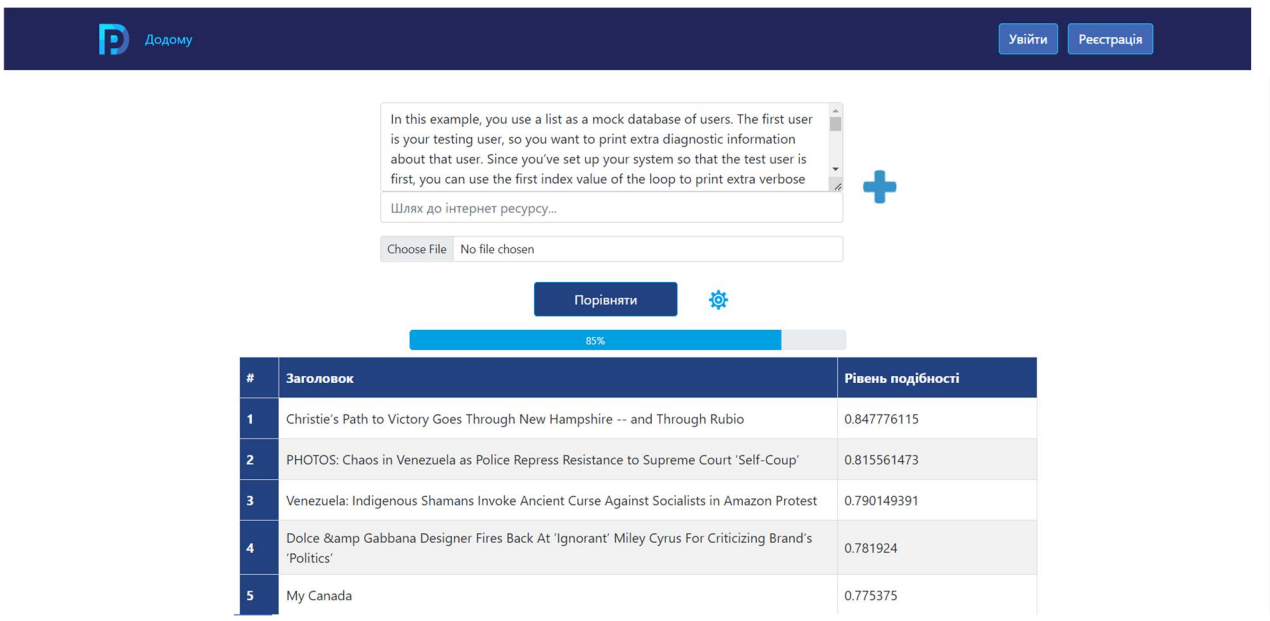


Рис. 7. Головна сторінка системи PlagDetector

Після того, як виконано порівняння одного тексту з усіма наявними в базі даних, користувачу надають звіт про десять найподібніших текстів у табличній формі та шкалу, яка відображає рівень найбільших збігів (рис. 8).



#	Заголовок	Рівень подібності
1	Christie's Path to Victory Goes Through New Hampshire -- and Through Rubio	0.847776115
2	PHOTOS: Chaos in Venezuela as Police Repress Resistance to Supreme Court 'Self-Coup'	0.815561473
3	Venezuela: Indigenous Shamans Invoke Ancient Curse Against Socialists in Amazon Protest	0.790149391
4	Dolce &amp Gabbana Designer Fires Back At 'Ignorant' Miley Cyrus For Criticizing Brand's 'Politics'	0.781924
5	My Canada	0.775375

Рис. 8. Результат порівняння тексту з усіма наявними в базі даних

Користувач також може додати ще одне поле для тексту, щоб порівняти два тексти між собою.

Висновки

Результатом роботи є обґрунтування, проєктування та розроблення інтелектуальної системи виявлення плагіату в технічних текстах. У дослідженні здійснено огляд поточного стану вирішуваної проблеми та проаналізовано вже відомі дослідження на тему виявлення плагіату за допомогою інтелектуальних інформаційних технологій. Беручи до уваги велику кількість відомих засобів вирішення цієї проблеми, постійне зростання кількості відкритих текстів у мережі Інтернет та способів приховати плагіат, можна стверджувати про певну доцільність реалізації такої системи. Загальний алгоритм виявлення плагіату в технічних текстах побудовано із використанням методу порівняння текстів за допомогою хеш-функцій та методу векторного порівняння. Поєднання переваг кожного з методів дало змогу спростити виявлення плагіату в технічних текстах, формалізувати й алгоритмізувати основні кроки перетворення текстів, істотно зменшити обсяги даних, які використовують для порівняння текстів.

Загалом, розроблена система може стати комплексом засобів та заходів для вирішення проблеми плагіату. Подальший розвиток досліджень у цьому напрямі передбачає, насамперед:

- розширення бази текстів для порівняння за рахунок україномовних публікацій у галузі інформаційних технологій;
- підвищення точності результатів виявлення плагіату;
- розроблення інтерфейсів для доступу до текстів, розміщених у зовнішніх ресурсах.

Наукові на практичні результати застосовні як для аналізу ідентичності технічних текстів, зокрема, в галузі інформаційних технологій, так і для поширення на інші галузі – навчання, наукову та академічну діяльність, журналістику, засоби масової інформації, вебресурси тощо.

Список літератури

1. Todorova N. Yu. Fighting Plagiarism: Cultural Patterns and Pedagogical Implications in the EAP/ESP Context. Academia.edu. Platform for academics to share research papers. URL: https://www.academia.edu/1459171/Fighting_Plagiarism_Cultural_ Patterns_and_Pedagogical_Implications_in_the_EAP_ESP_Context

2. Academic dishonesty - cheating and plagiarism in Ukrainian higher education освітянам. OECD Reviews of Integrity in Education: Ukraine 2017. URL: http://www.keepeek.com/DigitalAsset-Management/oecd/education/oecd-reviews-of-integrity-in-education-ukraine-2017/academic-dishonestycheating-and-plagiarism-in-ukrainian-higher-education_9789264270664-13-en#page1

3. Академічна культура українського студентства: основні чинники формування та розвитку. Східноукраїнський Фонд соціальних досліджень. URL: http://fond.sociology.kharkov.ua/images/docs/academ_cult/material.pdf

4. Sánchez-Vega F., Villatoro-Tello E., Montes-y-Gómez M., Rosso P., Stamatatos E., Villaseñor-Pineda L. (2019). Paraphrase plagiarism identification with character-level features. Pattern Anal Appl 22(2):669–681. DOI: 10.1007/s10044-017-0674-z

5. Sanchez-Perez M., Sidorov G., and Gelbukh A. (2014). A winning approach to text alignment for text reuse detection at PAN 2014– notebook for PAN at CLE", In: Cappellato L., Ferro N., Halvey M., Kraaij W. (eds.) CLEF 2014 evaluation labs and workshop-working notes papers, 15–18 September, CEUR-WS.org, Shefeld, UK, 1004–1011

6. Roostaee M., Fakhrahmad S. M., Sadreddini M. H. (2020). Cross-language text alignment: A proposed two-level matching scheme for plagiarism detection. Expert Syst Appl 160:113718. DOI: 10.1016/j.eswa.2020.113718

7. Ahuja L., Gupta V., Kumar R. (2020). A new hybrid technique for detection of plagiarism from text documents. Arab J Sci Eng 45(12):9939–9952. DOI: 10.1007/s13369-020-04565-9

8. Gharavi E., Veisi H., Rosso P. (2020). Scalable and language-independent embedding-based approach for plagiarism detection considering obfuscation type: no training phase. Neural Comput Appl 32(14):10593–10607. DOI: 10.1007/s00521-019-04594-y

9. Altheneyan A. S., Menai M. E. B. (2020). Automatic plagiarism detection in obfuscated text. Pattern Anal Appl 23(4):1627–1650. DOI: 10.1007/s10044-020-00882-9

10. Van Son N., Huong L. T., Thanh N. C. (2021). A two-phase plagiarism detection system based on multi-layer lstm networks. IAES Int J Artif Intel 10(3):636–648. DOI: 10.11591/ijai.v10.i3.pp636-648

11. Martin Fowler. “UML Distilled: A Brief Guide to the Standard Object Modeling Language”. Addison-Wesley Professional, 2003.

12. Глибовець М. М., Глибовець А. М., Поляков М. В. Інтелектуальні мережі. Дніпропетровськ: Нова ідеологія, 2014. 462 с.

13. Rubin D. and Bernard P. UML 2.0 in a Nutshell, O'Reilly Media, 2005.

14. Квєтний Р. Н., Богач І. В., Бойко О. Р., Софина О. Ю., Шушура О. М. (2012). Комп’ютерне моделювання систем та процесів. Методи обчислень. Частина 1: навч. посіб. / за заг. ред. Р. Н. Квєтного. Вінниця: ВНТУ. 193 с.

15. Литвин В. В., Шаховська Н. Б. Проектування інформаційних систем. Львів: Магнолія-2006. 380 с.

16. Катренко А. В. (2001). Застосування технологій та інструментальних засобів проектування інформаційних систем. Вісник Національного університету “Львівська політехніка”. № 438 : Інформаційні системи та мережі. С. 48–63. Бібліографія: 8 назв.

17. Berlinck R. G. S. (2011). The academic plagiarism and its punishments – a review. Brazilian Journal of Pharmacognosy, No. 21(3), 365–372. DOI: 10.1590/S0102-695X2011005000099

18. Петренко В. С. (2013). Поняття та види плагіату. Часопис цивілістики. Вип. 14, 128–131.

19. Mutiara A. B. and Agustina S. “Anti Plagiarism Application with Algorithm Karp-Rabin at Thesis in Gunadarma University”, arXiv Prepr, 2008. DOI: 10.13140/RG.2.1.3138.2802

20. Hunt, E., Janamsetty, R., Kinares, C., Koh, C., Sanchez, A., Zhan, F., ... & Oh, P. (2019, November). Machine learning models for paraphrase identification and its applications on plagiarism detection. In 2019 IEEE International Conference on Big Knowledge (ICBK), 97–104. DOI: 10.1109/ICBK.2019.00021

21. Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015, June). From word embeddings to document distances. In International conference on machine learning, 957–966. PMLR.

22. Kaytez, F., Taplamacioglu, M. C., Cam, E., & Hardalac, F. (2015). Forecasting electricity consumption: A comparison of regression analysis, neural networks and least squares support vector machines. International Journal of Electrical Power & Energy Systems, 67, 431–438. DOI: 10.1016/j.ijepes.2014.12.036

References

1. Todorova N. Yu. Fighting Plagiarism: Cultural Patterns and Pedagogical Implications in the EAP/ESP Context. Academia.edu. Platform for academics to share research papers. URL: https://www.academia.edu/1459171/Fighting_Plagiarism_Cultural_Patterns_and_Pedagogical_Implications_in_the_EAP_ESP_Context

2. Academic dishonesty – cheating and plagiarism in Ukrainian higher education. OECD Reviews of Integrity in Education: Ukraine 2017. URL: http://www.keepeek.com/DigitalAsset-Management/oecd/education/oecd-reviews-of-integrity-in-education-ukraine-2017/academic-dishonestycheating-and-plagiarism-in-ukrainian-higher-education_9789264270664-13-en#page1

3. Academic Culture of Ukrainian Students: Key Factors of Formation and Development. East Ukrainian Foundation of Social Research. URL: http://fond.sociology.kharkov.ua/images/docs/academ_cult/material.pdf

4. Sánchez-Vega F., Villatoro-Tello E., Montes-y-Gómez M., Rosso P., Stamatatos E., Villaseñor-Pineda L. (2019) Paraphrase plagiarism identification with character-level features. Pattern Anal Appl 22(2): 669–681. DOI: 10.1007/s10044-017-0674-z

5. Sanchez-Perez M., Sidorov G. and Gelbukh A. (2014). A winning approach to text alignment for text reuse detection at PAN 2014– notebook for PAN at CLE”, In: Cappellato L., Ferro N., Halvey M., Kraaij W. (eds) CLEF 2014 evaluation labs and workshop-working notes papers, 15–18 September, CEUR-WS.org, Shefeld, UK, 1004–1011.

6. Roostaee M., Fakhrahmad S. M., Sadreddini M. H. (2020). Cross-language text alignment: A proposed two-level matching scheme for plagiarism detection. Expert Syst Appl 160:113718. DOI: 10.1016/j.eswa.2020.113718

7. Ahuja L., Gupta V., Kumar R. (2020). A new hybrid technique for detection of plagiarism from text documents. Arab. J. Sci. Eng. 45(12):9939–9952. DOI: 10.1007/s13369-020-04565-9

8. Gharavi E., Veisi H., Rosso P. (2020). Scalable and language-independent embedding-based approach for plagiarism detection considering obfuscation type: no training phase. Neural Comput Appl 32(14):10593–10607. DOI: 10.1007/s00521-019-04594-y

9. Altheneyan A. S., Menai M. E. B. (2020). Automatic plagiarism detection in obfuscated text. Pattern Anal. Appl. 23(4):1627–1650. DOI: 10.1007/s10044-020-00882-9

10. Van Son N., Huong L. T., Thanh N. C. (2021). A two-phase plagiarism detection system based on multi-layer lstm networks. IAES Int. J. Artif. Intel. 10(3):636–648. DOI: 10.11591/ijai.v10.i3.pp636-648

11. Fowler Martin (2003). UML Distilled: A Brief Guide to the Standard Object Modeling Language. Addison-Wesley Professional.

12. Hlybovets M. M., Hlybovets A. M., Polyakov M. V. (2014). Intelligent Networks. Dnipropetrovsk: Nova Ideolohiya. 462 p.

13. Rubin D. and Bernard P. (2005). UML 2.0 in a Nutshell, O'Reilly Media.

14. Kvetny R. N., Bohach I. V., Boyko O. R., Sofina O. Yu., Shushura O. M. (2012). Computer Modeling of Systems and Processes. Computation Methods. Part 1: educational guide/ edited by R. N. Kvetny. innytsia: VNTU. 193 p.

15. Lytvyn V. V., Shakhovska N. B. Information Systems Design. Lviv: Magnolia-2006. 380 p.

16. Katrenko A. V. (2001). Application of Design Technologies and Tools for Information Systems. Bulletin of Lviv Polytechnic National University, No. 438 : Information Systems and Networks, 48–63. Bibliography: 8 references.

17. Berlinck R. G. S. (2011). The academic plagiarism and its punishments – a review. Brazilian Journal of Pharmacognosy, Vol. 21(3), 365–372. DOI: 10.1590/S0102-695X2011005000099

18. Petrenko V. S. (2013). Concepts and Types of Plagiarism. Journal of Civil Law, Vol. 14. 128–131.

19. Mutiara A. B. and Agustina S. (2008). Anti Plagiarism Application with Algorithm Karp-Rabin at Thesis in Gunadarma University , arXiv Prepr. DOI: 10.13140/RG.2.1.3138.2802

20. Hunt, E., Janamsetty, R., Kinares, C., Koh, C., Sanchez, A., Zhan, F., ... & Oh, P. (2019, November). Machine learning models for paraphrase identification and its applications on plagiarism detection. In 2019 IEEE International Conference on Big Knowledge (ICBK), 97–104. IEEE. DOI: 10.1109/ICBK.2019.00021

21. Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015, June). From word embeddings to document distances. In International conference on machine learning, 957–966. PMLR.

22. Kaytez, F., Taplamacioglu, M. C., Cam, E., & Hardalac, F. (2015). Forecasting electricity consumption: A comparison of regression analysis, neural networks and least squares support vector machines. International Journal of Electrical Power & Energy Systems, 67, 431–438. DOI: 10.1016/j.ijepes.2014.12.036

INTELLIGENT SYSTEM FOR DETECTING PLAGIARISM IN TECHNICAL TEXTS

Yurii Heriak[1], Andrii Berko[2]

Lviv Polytechnic National University, Information Systems and Networks Department,
12, S. Bandery str., Lviv, Ukraine
[1] yurii.heriak.mnitm.2021@lpnu.ua, ORCID 0009-0008-3251-2007
[2] andrii.y.berko@lpnu.ua, ORCID 0000-0003-2892-9519

The authors of the article developed a scientific reasoning, designed, and developed an intelligent system for detecting plagiarism in technical texts. The work defines the problem of plagiarism in the modern world and its relevance and analyzes the latest research and publications devoted to the latest methods of using intelligent information technologies to detect plagiarism. The need and expediency of developing and improving intellectual information technologies for detecting plagiarism, as well as the use of various methods of identifying matches in texts for the further development of such technologies, are substantiated. The authors developed a general algorithm for detecting plagiarism in technical texts based on the vector comparison method. The practical result of the study is the development of an intelligent system for detecting plagiarism in technical texts and confirmation of its efficiency by applying it to specific examples of technical texts.

Key words: plagiarism; machine learning; intelligent system; text documents; vector comparison.