УДК 004.4

М. Я. Мельник
Л. М. Куперштейн

ДОСЛІДЖЕННЯ МЕТОДІВ ВИЯВЛЕННЯ ТЕКСТОВОГО ПЛАГІАТУ

Вінницький національний технічний університет

Анотація
Розроблено програмний засіб для перевірки текстів на плагіат шляхом застосування комплексного підходу. Аналіз поданих текстів виконується декількома алгоритмами. Оцінено ефективність роботи алгоритмів, оцінено час їх виконання в залежності від кількості файлів у локальній бібліотеці.
Ключові слова: кібербезпека, аналіз текстів, перевірка текстів на плагіат, веб-застосунок.

Abstract
A plagiarism detection software has been developed. Text analysis tool uses a number of alghorythms to detect plagiarism. The algorithms are evaluated and their execution time is estimated depending on the number of files in the local library..
Keywords: cybersecurity, plagiarism detection tool, anti-plagiarism, web app.

Вступ

Розвиток технологій та засобів обміну інформацією, а також можливість будь-ким публікувати від свого імені наукові роботи породжує важливу проблему у спільноті – текстовий плагіат.

Наука сприяє створенню нових галузей виробництва, вона є рушійною силою суспільства [1], але її розвиток не може бути повноцінним та бесперешкодним, якщо видання, журнали та сайти будуть наповнюватись однаковою інформацією, що не несе в собі ніякої новизни – плагіатом.

На сьогоднішній день для захисту від текстового плагіату використовуються методи порівняння документів, нечтіких дублікатів: метод шинглів, Moodle Crot, варіації методу частоти слова та відстані Левенштейна. Перед поданням до аналзіу тексти проходять попередню обробку: стемінг та лематизацію [2].

Метою роботи є дослідження методів виявлення текстового плагіату та перевірка їх ефективності для порівняння документів, а також визначення швидкодії алгоритмів залежно від кількості файлів у локальній базі даних.

Результати дослідження

Під час дослідження було розроблено програмний засіб для перевірки текстів на наявність плагіату, що використовує для порівняння поданих текстів локальну базу документів. Інтерфейс програмного засобу реалізований у вигляді веб-застосунку. Було визначено переваги та недоліки кожного з методів та оцінено ефективність підходів за критерієм часу виконання аналізу [2].

Оцінка ефективности алгоритму виконувалась за такими параметрами як кількість операций, сумарний час виконання операцій, середня довжина ланцюжків тексту, поданого до алгоритму, кількість слів у базі даних, а також визначений процент збігу документів.

Операція – кількість необхідних дій для повного проходу алгоритму. Враховуються перестановки, сортування, ітерації циклів.

Довжина тексту – загальна кількість шинглів [3], рядків, ланцюжків тексту, в залежності від того: як оперує даними алгоритм.

Визначений процент збігу документів – процент відношення входження змісту поданого на аналіз документу до існуючого в базу даних.

Під час тестування було завантажено до бази даних 2 документи: один майже точно повторював поданий текст, другий документ не мав ніякого відношення до нього.

Найбільша повнота виявлення плагіату у методі шинглів – 85,94%. У методі відстані Левенштейна відсоток виявлення низький – 26.61%. Метод Moodle Crot – 44,86%, метод частоти слова і косинусів – 73,49%.

По кількості операцій метод частоти слова і косинусів виявився лідером: 201285446 операцій на загальній кількості слів у базі 10120. Метод шинглів – 43638 операцій, метод Moodle Crot – 88848 операцій на такій ж самій кількості слів у базі.

Відповідно, метод шинглів виявився найшвидшим: 2.0381 секунд зайняв процесс порівняння поданого тексту із базою, у метода Moodle Crot – 4.2989с, у метода частоти слова і косинусів – 16.5569с. Метод відстані Левенштейна виконувався – 16.0502 с.

Далі досліджено зміни часу виконання алгоритмів при збільшенні кількості файлів у базі даних. Нас цікавить лише залежність часу роботи алгоритмів від обсягу даних, тому процент виявленого плагіату оцінюватись не буде.

Для оцінки залежності часу роботи алгоритмів від обсягу бази даних були проведені тестові порівняння документів, на кожному кроці кількість файлів, а отже і обсяг, у локальній бібліотеці збільшувався. На вхід для прозорості експерименту подавався один і той ж самий файл, а до таблиці додавались середні значення роботи алгоритмів. Результати відображені на рисунку 1.

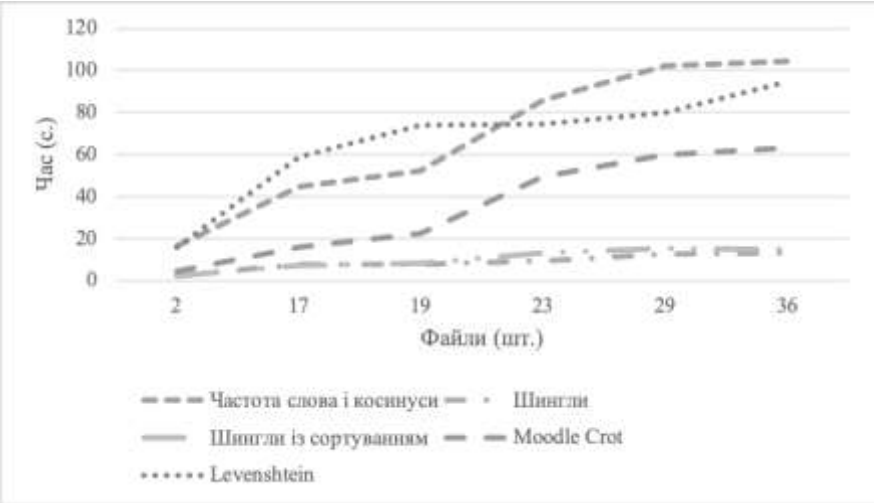

Рисунок 1 – Графік залежності часу роботи алгоритмів від кількості файлів у базі даних

Отримавши достатню кількість даних про поведінку алгоритмів, остаточні висновки про їх роботу будуть такими:

1) Базовий метод шинглів виявився, лідером серед алгоритмів, що виявляють запозичення, навіть коли подані тексти були змінені. Даний алгоритм найкраще демонструє пошук нечітких дублікатів.

2) Метод шинглів із сортуванням знаходить низький відсоток входжень у випадках, коли документ був змінений хоча б частково. Він є повільнішим, ніж його проста варіація через виконання операцій сортування перед формуванням шинглів.

3) Метод Відстані Левенштейна за результатами роботи демонструє ті ж самі риси, що і метод шинглів із сортуванням. Виконується великий обсяг операцій, що впливає на час виконання.

4) Метод частоти слова і косинусів демонструє задовільну для нього повноту виявлення як при зміненому документі, так і при поданні дублікату. Даний метод виявився найповільнішим у розробленій системі.

5) Метод Moodle Crot має властивості алгоритму шинглів, але він, розбиваючи текст не на ланцюжки по n слів, а літер, виконує набагато більше операцій, що робить його очікувано повільним. Демонструє невелику повноту виявлення у випадках, коли документ був змінений частково..

Виходячи з результатів проведеного тестування, можна сказати, що обрані методи ефективні для порівняння документів та виявлення плагіату. Деякі алгоритми, наприклад, метод Левенштейна, най-

краще використовувати для виявлення чітких дублікатів документів, а інші, такі як шингли – для виявлення нечітких.

Опираючись на час виконання порівнянь документів зазначеними алгоритмами, можна зробити висновок, що відношення часу виконання до повноти виявлення у деяких окремих методах виявилось не найкращим. Для вирішення даної проблеми потрібно реалізувати розпаралеллювання операцій та, можливо, оптимізувати самі алгоритми для підвищення швидкості аналізу.

Висновки

У процесі дослідження проведено тестування програмного засобу та визначено ефективність роботи алгоритмів, виконано їх порівняння та визначена швидкодія в залежності від кількості документів у базі даних. Визначено та порівняно точність алгоритмів під час порівняння документів, а також запропоновано рекомендації щодо їх подальшого вдосконалення.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Методичні вказівки до проведення практичних занять та до виконання самостійної й індивідуальної роботи з дисципліни „Основи науково-дослідної роботи / Укладачі: А. О. Азарова, В. В. Карпінець. – Вінниця: ВНТУ, 2013. – 44 с.
2. Куперштейн Л.М. Сучасні тенденції перевірки тексту на плагіат / Куперштейн Л.М., Мельник М.Я. // Тези доповідей Міжнародного форуму з інформаційних систем і технологій «INFOS-2019» м. Харків, 24-27 квітня 2019 року. – Харків, – С. 44-46.
3. Подходы к сравнению версий файлов - Режим доступу: https://habr.com/post/65944/ - Дата доступу: 01.03.20.
4. Чиркин Е.С. Системы автоматизированной проверки на неправомерные заимствования // Вестник ТГУ. – 2013. − №12. − С. 164-171.

Мельник Максим Ярославович — студент групи 1БС-19м, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: aktotut@pm.me

Куперштейн Леонід Михайлович — кандидат технічних наук, доцент кафедри захисту інформації, Вінницький національний технічний університет, м. Вінниця

Melnyk Maksym Y. — Student of 1BS-19m, Faculty of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, email: aktotut@pm.me

Kupershtein Leonid M. — Candidate of Technical Sciences, Docent of the Information department, Vinnytsia National Technical University, Vinnytsia, email: kupershtein.lm@gmail.com