УДК 004.021

АНАЛІЗ СЕРВІСІВ ДЛЯ ПОРІВНЯННЯ ТЕКСТУ ТА КОДІВ

Глущенко В.В., МгІТ-1-21*, магістрант, yovmyn@gmail.com*
Астістова Т.І. – к.т.н., доц., *astistova.ti@knutd.edu.ua*
Київський національний університет технологій та дизайну

Метою роботи є дослідити та проаналізувати існуючі сервіси порівняння тексту та кодів, розробити варіант автоматизованої системи зберігання та аналізу кодів та текстів, який би перейняв переваги та врахував недоліки існуючих на ринку систем.

Для виконання поставлених задач доведеться реалізувати 3 методи перевірки коду на їх ідентичність:

1. Text-based метод. Цей метод являє собою просту перевірку коду слово в слово та видає відсоток слів вхідного тексту які співпадають з словами іншого вхідного тексту в тих самих місцях. Вихідний код програми не піддається ніяким змінам і представлен у вигляді символьного рядка
2. Metrics-based метод. Метод представляє з себе пошук ключових слів та розбиття їх на категорії з виводом кількості слів в конкретних категоріях, таких як: ключові слова, цикли, змінні, оператори. Але на відміну від Text-based методу просте перейменування функцій і змінних, а так само незначні маніпуляції з вихідним кодом не вплинуть на роботу цього методу. У якості відправної точки для виявлення співпадінь вибираються деякі кількісні характеристики програми. Вихідний код програми теж не зазнає жодних змін. Алгоритм реалізації Metrics-based показаний на рисунку1.
3. Token-based метод - у цьому підході весь текст перетворюється в послідовність лексем, яка потім сканується для пошуку дублікатів. При такому підході зберігаються всі суттєві і пропускаються всі поверхневі деталі коду. Метод представляє з себе перезапис усіх важливих частин у вигляді зазначених символів, після чого перевіряє між собою токанізовані тести методом Text-based.

Аналізуючи існуючи інтернет-сервіси, ми прийшли до висновку, що їх недоліки пов'язані з частими хибними спрацювання, а саме , щойно написаний власноруч текст буде унікальним всього на 50%, існують обмеження по кількості символів та присутні технічні збої.

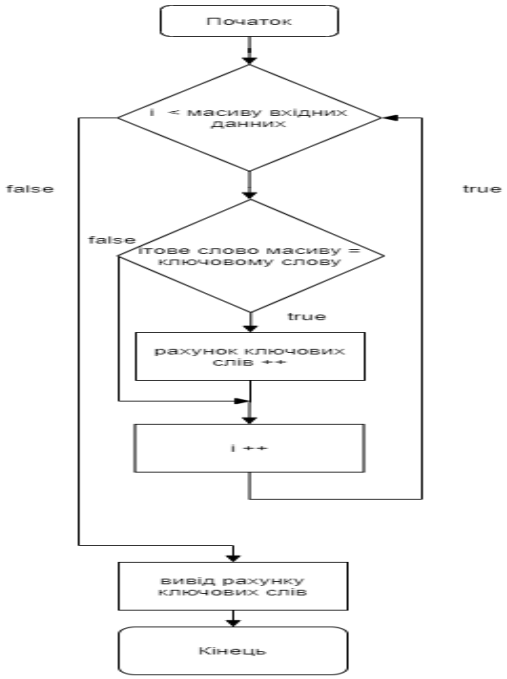

Рисунок 1 – Алгоритм реалізації методу Metrics-based

Висновок. Результатом аналізу алгоритмів та методів перевірки коду та існуючих сайтів є розробка програмного сервісу, що відрізняється більш точним алгоритмом пошуку запозичених фрагментів та підготовки тексту до порівняння, що гарантує більш високу якість порівняння. Оптимізовані шляхи зрівняння дають більш високу швидкість роботи та отримати результат**.** Зручний інтерфейс дасть змогу поліпшити процес порівняння текстів та кодів.

Література

1. Jeffrey E.F. Friedl Mastering Regular Expressions, 3rd Edition - O'Reilly Media, 2016. - 544

2. Перебийніс В.С.. Математична лінгвістика. Українська мова: енциклопедія / В. С. Перебийніс – К. : Українська енциклопедія, 2000. – ISBN 966-7492-07-9.

3. Астістова Т.І Розробка автоматизованої системи аналізу текстів «Антиплагіат» / Т.І. Астістова, В.О.Керіб. // Інформаційні технології в науці, виробництві та підприємництві: зб. наук. праць молодих вчених, аспірантів, магістрів кафедри інформаційних технологій проектуванн. – К. : КНУТД, 2017. С. 118–121 – ISBN 978-966.

4. Пасічника В. В. Математична лінгвістика . Квантитативна лінгвістика / В. В. Пасічник, Ю. М. Щербина, В. А. Висоцька, Т. В. Шестакевич / навч. посіб. Кн.1: Новий Світ – 2000, 2016. – 359 с.

5. Ланде Д. В. Елементи комп'ютерної лінгвістики в правовій інформатиці // Д.В.Ланде – К.: НДІІП НАПрН України, 2014. – 351 с. – ISBN 978-966-2344-33-2.