УДК 004.55

НЕЧІТКЕ ПОРІВНЯННЯ РЯДКІВ

Витнова А.І., студентка, гр. АС-191
Керівник: професор каф. СПЗ Кунгурцев О.Б.
Державний університет «Одеська політехніка», УКРАЇНА

АНОТАЦІЯ. Робота присвячена розробці прикладного програмного забезпечення, що допоможе користувачу здійснити порівняння двох текстів. В основі ідеї лежать алгоритми аналізу тексту.

Вступ. Кожна людина взаємодіє з текстом, як з однією з форм подання та отримання інформації, протягом всього життя. На вулиці, в книгах, соціальних мережах, комп'ютерних іграх і т.д. Тексту дуже багато, тому люди почали створювати різні інструменти для більш легкої взаємодії з ним. «Пошук» для знаходження необхідних статей, речень, слів, …; «антиплагіат» для викладачів і т.д. Усе це має одну основу – порівняння. Саме тому потрібно розробити прикладне програмне забезпечення, що має змогу порівнювати тексти чітко і швидко.

Мета роботи. Метою роботи є покращення існуючих алгоритмів порівняння, необхідне для більш швидкого та коректного виділення термінів з тексту та їхнього порівняння.

Основна частина роботи. В ході дослідження алгоритмів нечіткого порівняння рядків[2] було обрано алгоритм «Коефіцієнт Жаккара (приватне - коеф. Танімото)», який обчислює коефіцієнт схожості за формулою:

$$K = c / (a + b – c)$$

де «a» - кількість символів в першому рядку, «b» - кількість символів у другому рядку, «c» - кількість співпадаючих символів. Але для більш коректного та швидкого порівняння символи було замінено на терміни. Якщо необхідно більш точне порівняння, доповнюєм формулу:

$$K = (c / (a + b – c)) - (d / (a + b – d))$$

де «d» - кількість співпадаючих розташувань. Це доповнення може бути важливим, якщо користувачу необхідно враховувати розташування термінів; якщо ні – цей алгоритм може дати некоректний результат, оскільки при повній перестановці термінів в однакових реченнях, схожість буде становити не більше 50%.

Для того, щоб визначити перелік сервісів, що повинна виконувати програма та її реакцію на різні вхідні дані та поведінки, була складена діаграма варіантів використання (мал. 1).

В розроблювальній системі передбачений один актор – власне користувач системи. Можна виділити 2 основних етапи роботи системи:

- обробка введеного тексту програмою для подальшого аналізу. В даному випадку при обробці кожного тексту потрібно:

1) видалити з тексту усі символи, окрім літер, цифр, пробілу та дефіса;

2) виділити усі терміни[1] та підрахувати їх кількість в кожному рядку.

- порівняння. В основі алгоритму порівняння лежить поняття термінів. Вони мають різні форми, синоніми і т.д. Тому умовно процес порівняння можна поділити на 3 етапи:

1) початкова форма слова. Система, використовуючи словник (базу даних), перероблює кожний термін у його початкову форму;

2) синоніми. Оскільки основною ідеєю програми є порівняння рядків за змістом, а не просто за символами, система також використовує словник синонімів та приводить терміни-синоніми до головного (основного) терміна;

3) алгоритм порівняння. Використання вищезазначеної формули для отримання коефіцієнту схожості рядків.

Було проведено експеримент, цілью якого було знаходження якості порівняння з синонімами та без них. З синонімами, коефіцієнт схожості буде обчислюватися на 30-35% повільніше, ніж без синонімів, але, при порівнянні «за змістом», він буде значно корисніше.

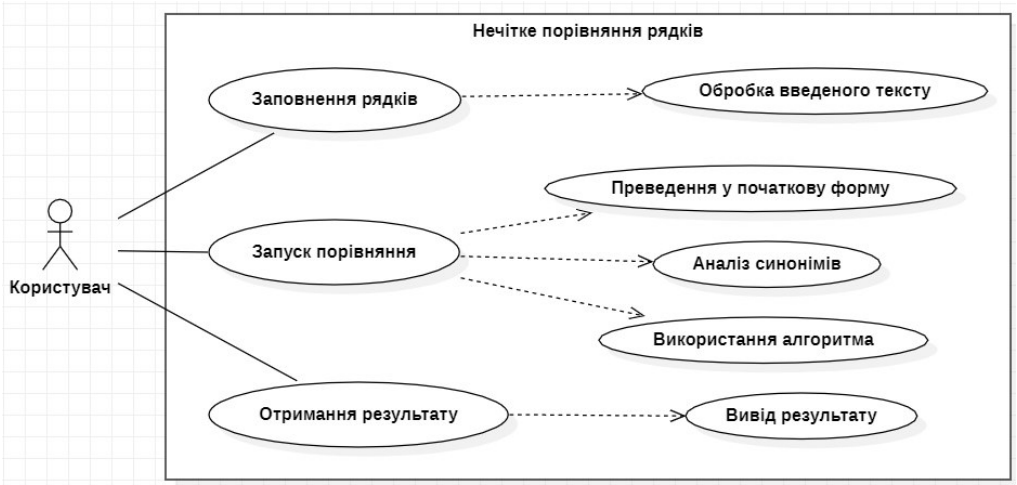

Рисунок 1 – Діаграма варіантів використання

Щоб побачити приклад роботи системи, наведена діаграма станів, що демонструє роботу системи (мал. 2). На ній видно, як взаємодіють користувач та система.

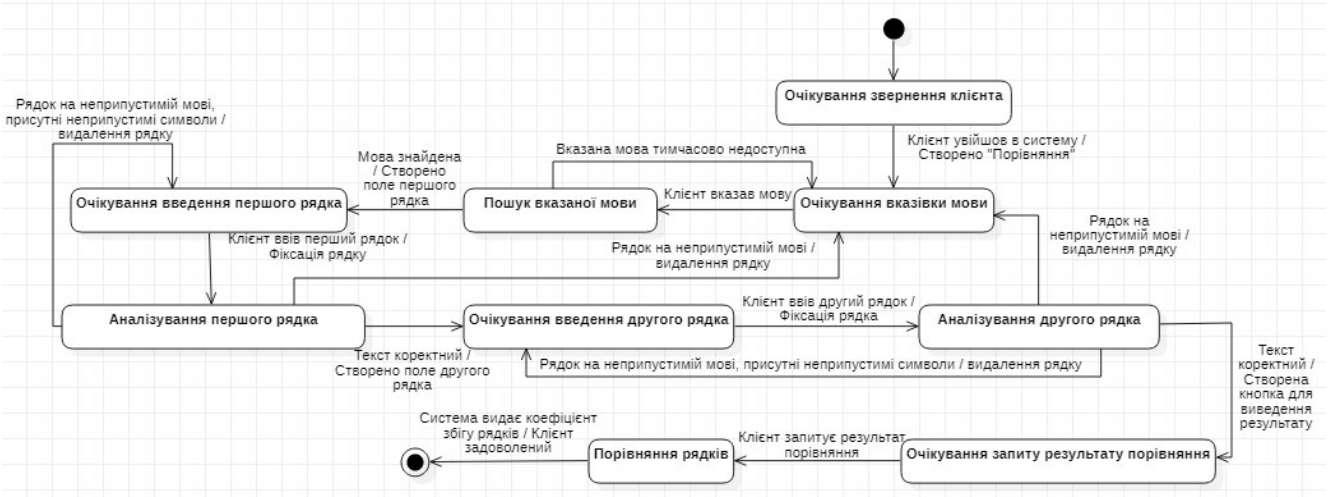

Рисунок 2 – Діаграма станів «Нечітке порівняння рядків»

Система розроблена на Java, в інструментальному середовищі «IntelliJ IDEA». Для зручного доступу та керування вмістом програми було використано інструментарій GUI – JavaFX. Усі «словники» зберігаються у базі даних MySQL, використовуючи інструмент «phpMyAdmin».

Висновки. У роботі виконані основні етапи проектування прикладного програмного забезпечення для нечіткого порівняння рядків з наведенням основних UML-діаграм, що демонструють його архітектуру. Це програмне забезпечення повинно швидко и коректно порівнювати тексти не як набір символів, а за змістом. Для цього було використано та модифіковано алгоритм «Коефіцієнт Жаккара (приватне - коеф. Танімото)».

СПИСОК ЛІТЕРАТУРИ

1. Kungurtsev O. Development of information technology of term extraction from documents in natural language / O. Kungurtsev, S. Zinovatnaya, Ia. Potochniak, M. Kutasevych // Eastern-European Journal of Enterprise Technologies. Vol 6, No 2 (96) (2018). pp. 44-51.

2. Окулов С. М. Алгоритмы обработки строк. — М.: Бином, 2013. — 255 с.