

ДОДАТОК А

Графічний матеріал кваліфікаційної роботи

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
РАДІОЕЛЕКТРОНІКИ

КВАЛІФІКАЦІЙНА РОБОТА

Методи та алгоритми семантичної
обробки веб-документів

ст. гр. СПм-20-1 Кулешов Д.О.

Керівник: доц. каф. ЕОМ Лебедєв О.Г.

Мета та завдання роботи:

2

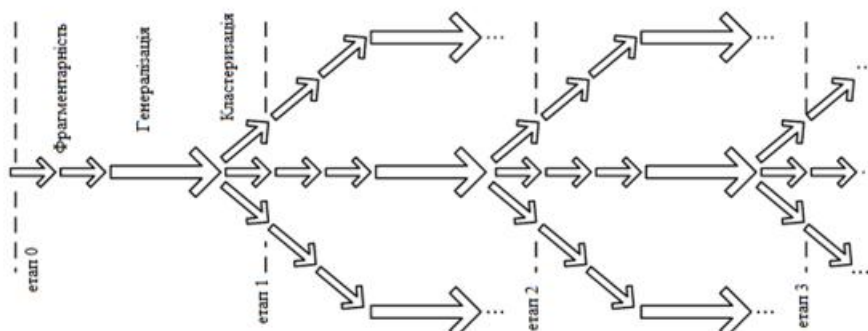
Метою кваліфікаційної роботи є аналіз методів та алгоритмів семантичної обробки веб-документів для підвищення ефективності виявлення схожих документів шляхом розробки методу оцінки подібності веб-документів на основі їх локальних параметрів з використанням попереднього розбиття на структурно-семантичні блоки.

Об'єкт дослідження: методи виявлення схожих текстів.

Завдання:

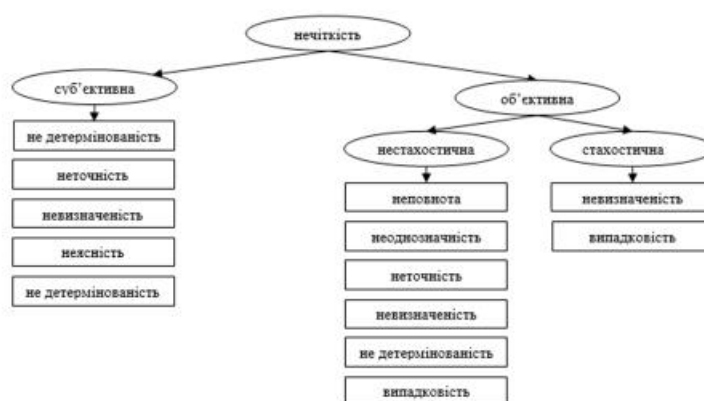
- ☐ аналіз сучасних методів виявлення схожих текстів, їх застосування до області веб-документів, виділення основних метрик схожих текстів та метрики оцінки ефективності;
- ☐ розробка математичної моделі оцінки схожості документів на рівні складових їх блоків. Створення методу оцінки схожості веб-документів на основі їх структурно-семантичних блоків;
- ☐ розробка алгоритму виділення структурно-семантичних блоків із веб-документів;
- ☐ розробка структури спеціального програмного забезпечення.

Семантичне зв'язування інформаційних фрагментів³



Схожість документів

4



Метрики схожості документів

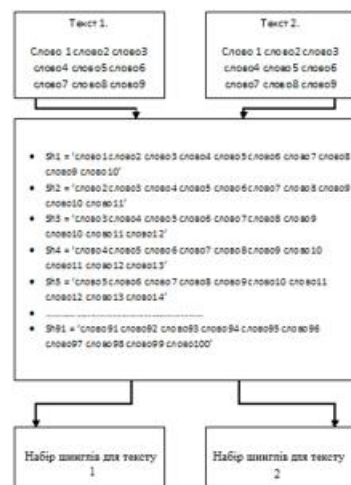
5

- символні (відстань Левінштайна) ;
- лексичні (схожість Джакарда);
- фонетичні (Phonix та Soundex);

$$\text{Resemblance}(A, B) = \frac{|S(A) \cap S(B)|}{|S(A) \cup S(B)|}$$

6

Метод визначення схожих документів. Шинглювання





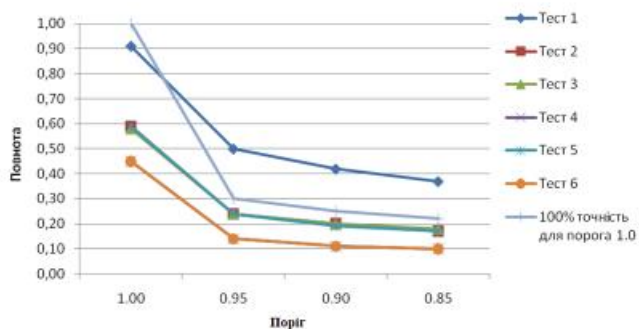
Створення єдиного відбитку на основі локальних параметрів

9

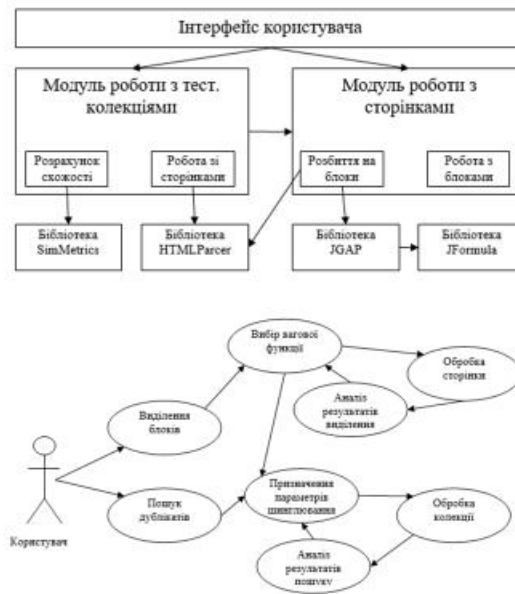
- З погляду статистичних параметрів:
 - послідовності;
 - Загальна кількість;
 - середня.
- З погляду елементів тексту:
 - абзаци; пропозиції;
 - слова (терми);
 - символи.
- З погляду розмітки:
 - із збереженням HTML-розмітки;
 - із збереженням текстової розмітки;
 - із збереженням спецсимволів; відфільтровані.

Графік повноти для тестів

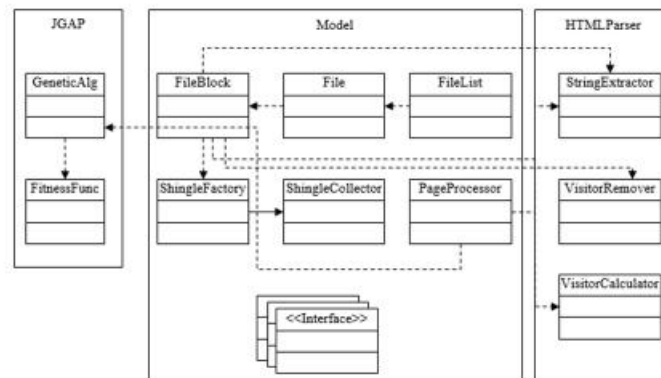
10



Структурна схема та use-case діаграма програми 11

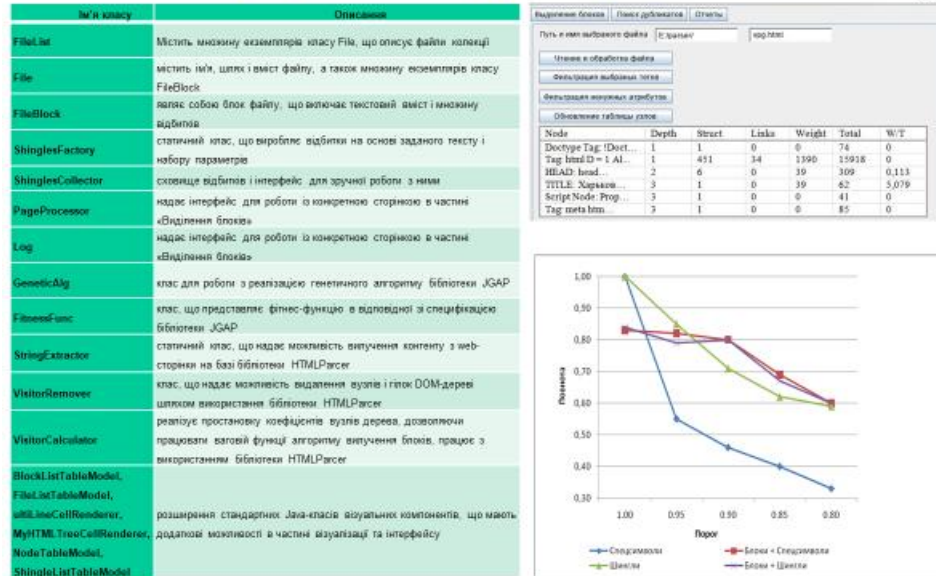


Діаграма класів додатку 12



Програмний інтерфейс та результати

13



Висновки

14

Запропоновано модель, що дозволяє оцінювати схожість документів, використовуючи інформацію про схожість їх блоків. Відповідно до моделі розроблено метод оцінки схожості документів на основі складових їх структурно-семантичних блоків, що дозволяє поліпшити якість розпізнавання дублікатів за рахунок збільшення середнього значення показника повноти. Це дозволяє отримувати більш повну вибірку схожих документів на одних і тих же наборах даних. Проведена експериментальна перевірка запропонованих алгоритмів і методів, у розробленому програмному забезпеченні.