УДК 004.04

АЛГОРИТМ ШИНГЛОВ КАК МЕТОД ОПРЕДЕЛЕНИЯ УНИКАЛЬНОСТИ ТЕКСТА

Григорьев Вячеслав Владимирович
Забайкальский край, г. Чита, МБОУ «СОШ № 10» 11 класс
Научный руководитель: Брюханкова Анна Ивановна, г. Чита, МБОУ «СОШ № 10», учитель физики

Алгоритм шинглов (от англ. shingles — чешуйки) — алгоритм, разработанный в 1997 году для поиска копий и дубликатов рассматриваемого текста в веб-документе [1].

Знание алгоритма определение нечетких дублей, позволит избежать проблемы при написании текстов для поискового продвижения. Можно выделить следующие этапы, через которые проходит текст при его сравнении:

- канонизации текста;
- разбиения его на шинглы;
- вычисления, через статические функции, 84-х хэшей шинглов;
- случайной выборки значений 84 контрольных сумм;
- сравнения и определения результата [1].

Канонизация текста – приведение оригинального текста в более понятный текст, через очищение его от всех вспомогательных единиц текста (предлогов, союзов, знаков препинания, тегов и прочее), которые не должны участвовать в сравнении [2].

Разбиение текста на шинглы.

Шинглы (от англ. – чешуйки) – выделенные для сравнения из тела статьи отдельные части текста, с определенным количеством слов в его последовательности для проверки на уникальность.

Шинглы могут быть на любое количество слов – от 3 до 10. Чем шингл короче, тем точнее будет результат проверки. При назначении размера шингла в 3 слова проверка, давшая 100% уникальности, является свидетельством оригинальности текста, поскольку совпадения словосочетаний встречаются практически в любом тексте. Полученные наборы шинглов, после того как каждый из текстов разбит на под последовательности, равны количеству слов в документе минус длина шингла (-10) плюс один (+1). Схема разбиения текста на шинглы представлена на рисунке 1.

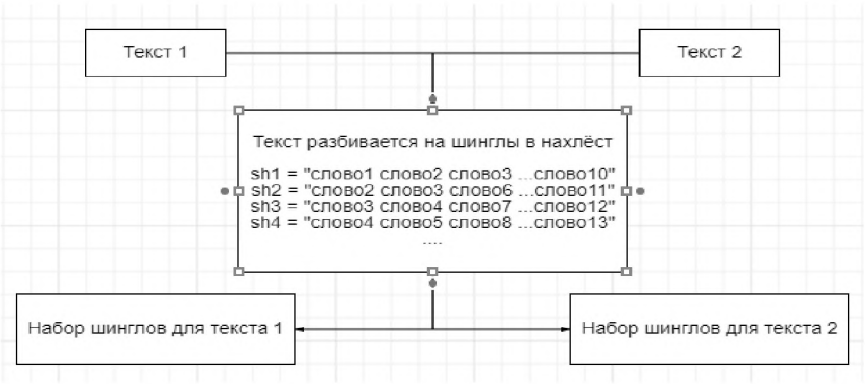

Рис. 1. Разбиение текста на шинглы

Вычисление хэшей шинглов.

Принцип алгоритма шинглов базируется на сравнении случайно выбранных контрольных сумм шинглов (под последовательностей) двух документов. Суть действия алгоритма заключается в том, чтобы найти верное количество контрольных сумм для сравнения. Завышенное число шинглов негативно отразится на результате, поскольку для сравнения будет произведено гораздо больше операций, что снизит производительность. Для облегчения текст представляется в виде таблиц с набором контрольных сумм, рассчитанных для каждого шингла по 84-м статическим хэш-функциям. Все 84 строки (для каждого из документов) охарактеризованы соответствующей контрольной суммой. Из обоих наборов случайным образом отбираются 84 значения – для каждого из документов – и сравниваются в со-

ответствии с функциями своей контрольной суммы. Иными словами, потребуется 84 операции, чтобы сравнить тексты. Схема вычисления хэшей шинглов представлена на рисунке 2.

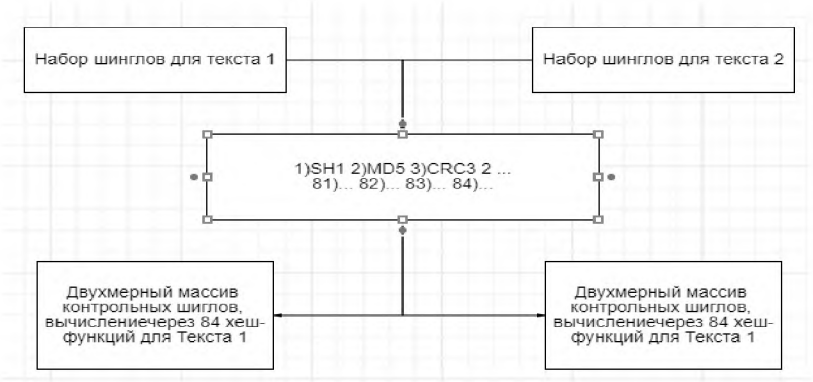

Рис. 2. Вычисление хэшей шинглов

Случайная выборка 84 значений контрольных сумм.

Для увеличения производительности при сравнении элементов каждого из 84-х выбранных массивов нужно произвести случайную выборку контрольных сумм для каждой из строк. Выбор минимального значения из каждой строки в итоге даст набор наименьших значений контрольных сумм шинглов для каждой из хэш функций.

Получение результата.

Сравнение каждого из 84 элементов обоих документов выявляет соотношение одинаковых значений, что позволяет определить уровень идентичности, или уникальности каждого из текстов рисунка 3.

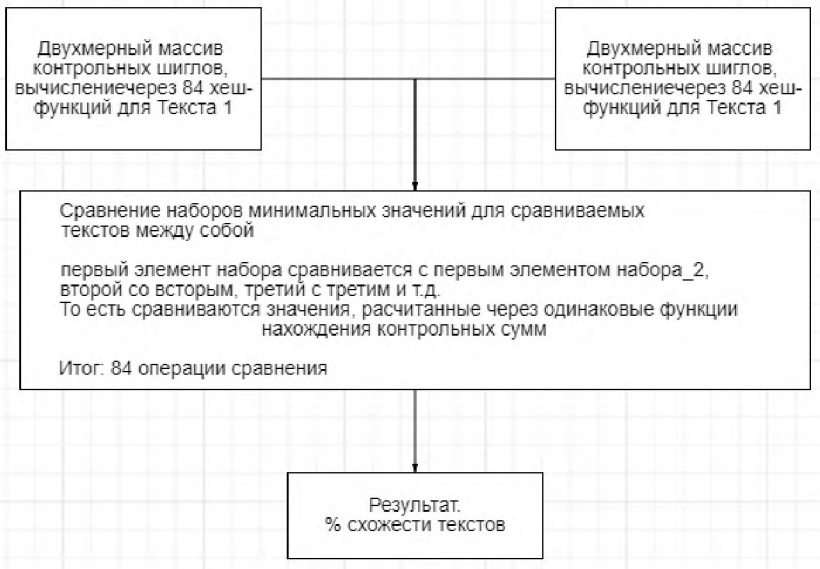

Рис. 3. Получение результата

Интерфейс представлен на рисунке 4.

Антиплагиат. Используя данный алгоритм, мною был создан один из самых быстрых и точных антиплагиатов на языке программирования Python. Скачать и узнать больше можно по ссылке рис. 5.

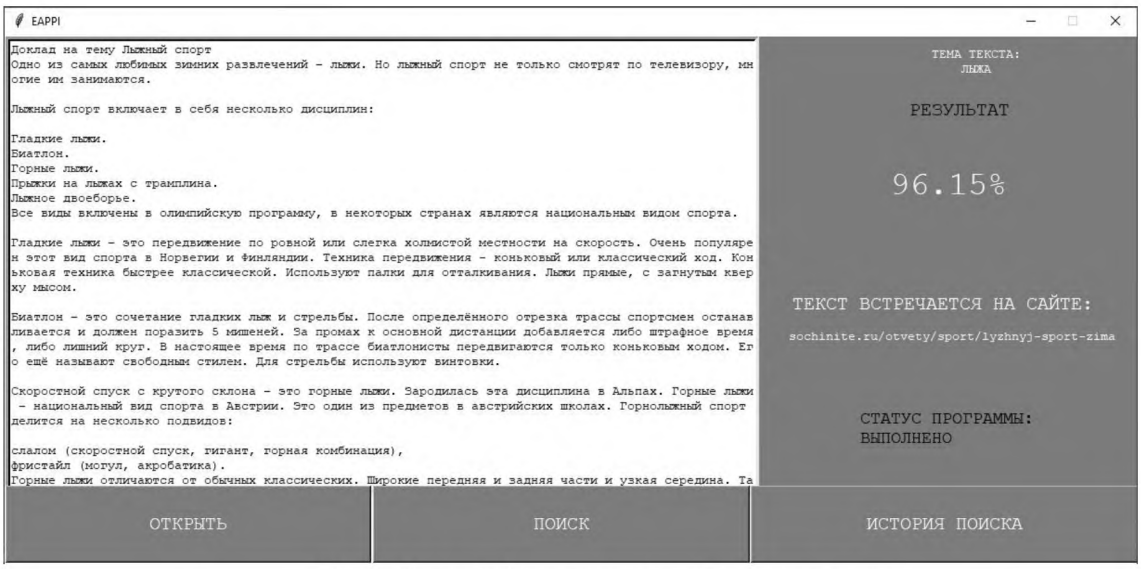

Рис. 4. Интерфейс и результат работы приложения

Рис. 5. Скачать антиплагиат (EAPIP)

Список литературы:

1. Алгоритм шинглов // Википедия. URL: https://ru.wikipedia.org/wiki/Алгоритм_шинглов
2. Алгоритм шинглов. URL: https://uniofweb.ru/wiki/algoritm-shinglov/