**Лабораторная работа №5**

**Реализация алгоритма обнаружения нечетких дубликатов.**

***Цель работы:*** научиться реализовывать на выбранном языке программирования алгоритмы обнаружения нечетких дубликатов на основе алгоритма шинглов.

**Краткие теоретические сведения**

Проблема обнаружения нечетких дубликатов является одной из наиболее важных и трудных задач анализа данных и поиска информации. Актуальность этой проблемы определяется разнообразием приложений, в которых необходимо учитывать «похожесть», например, текстовых документов — это и улучшение качества индекса и архивов поисковых систем за счет удаления избыточной информации, и объединение новостных сообщений в сюжеты на основе сходства этих сообщений по содержанию, и фильтрация спама (как почтового, так и поискового), и установление нарушений авторских прав при незаконном копировании информации (проблема плагиата или копирайта), и ряд других.

Алгоритм шинглов — алгоритм, разработанный для поиска копий и дубликатов рассматриваемого текста в веб-документе, мощный инструмент, призванный бороться с проявлениями плагиата в интернете.

Реализация алгоритма подразумевает несколько этапов:

* канонизация текстов;
* разбиение текста на шинглы;
* нахождение контрольных сумм;
* поиск одинаковых подпоследовательностей.

Контрольная сумма

В самом общем своем виде контрольная сумма представляет собой некоторое значение, построенное по определенной схеме на основе кодируемого сообщения.

Алгоритм CRC базируется на свойствах деления с остатком двоичных многочленов, то есть [многочленов над конечным полем](http://ru.wikipedia.org/wiki/%D0%9C%D0%BD%D0%BE%D0%B3%D0%BE%D1%87%D0%BB%D0%B5%D0%BD_%D0%BD%D0%B0%D0%B4_%D0%BA%D0%BE%D0%BD%D0%B5%D1%87%D0%BD%D1%8B%D0%BC_%D0%BF%D0%BE%D0%BB%D0%B5%D0%BC) GF(2N). Значение CRC является по сути [остатком от деления](http://ru.wikipedia.org/wiki/%D0%94%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D1%81_%D0%BE%D1%81%D1%82%D0%B0%D1%82%D0%BA%D0%BE%D0%BC) многочлена, соответствующего входным данным, на некий фиксированный [порождающий многочлен](http://en.wikipedia.org/wiki/Generator_(mathematics)).

Каждой конечной последовательности [битов](http://ru.wikipedia.org/wiki/%D0%91%D0%B8%D1%82) a0,a1,…,aN-1[взаимно однозначно](http://ru.wikipedia.org/wiki/%D0%91%D0%B8%D0%B5%D0%BA%D1%86%D0%B8%D1%8F) сопоставляется двоичный полином, последовательность коэффициентов которого представляет собой исходную последовательность. Например, последовательность битов 1011010 соответствует многочлену:

Значение контрольной суммы в алгоритме с порождающим многочленом G(x) степени N определяется как битовая последовательность длины N, представляющая многочлен R(x), получившийся в остатке при делении многочлена P(x), представляющего входной поток бит, на многочлен G(x):

где

 R(x) — многочлен, представляющий значение CRC.

 P(x) — многочлен, коэффициенты которого представляют входные данные.

G(x) — порождающий многочлен.

N — степень порождающего многочлена.

Рассмотрим общую схему алгоритма расчета CRC:

1. Выбрать полином P, в результате автоматически становится известна его степень N.
2. Добавить к исходной двоичной последовательности сообщения N нулевых битов. Это добавление делается для гарантированной обработки всех битов исходной последовательности.
3. Выполнив деление дополненной N нулями исходной строки S на полином P по правилам CRC арифметики. Запомнить остаток от деления, который и будет являться CRC.

**Задание к лабораторной работе**

Написать программу на выбранном языке программирования, реализующую поиск нечетких дубликатов заданных текстов описанным выше алгоритмом. Программа должна запрашивать имена входных файлов и выводить схожие документы и степень их схожести (в процентах).

**Контрольные вопросы**

1. Что такое нечеткий дубликат?
2. Что такое хеширование?
3. Что такое контрольная сумма?
4. Принципы работы алгоритма шинглов.
5. Принципы работы алгоритма CRC.
6. Что такое CRC-арифметика?
7. Что такое канонизация текстов?
8. Как формировать шинглы?
9. Как выбирать порождающий полином?