Учреждение образования

«Белорусский государственный технологический университет»

**Лабораторная работа № 2**

Элементы теории информации. Параметры и характеристики дискретных информационных систем

Выполнила:

Студент 3 курса 6 группы ФИТ

Батурель Евгений Дмитриевич

**2022 г.**

1. **Теоретические сведения**

Передача информации (данных) осуществляется между двумя абонентами, называемыми источником сообщения (ИcС) и получателем сообщения (ПС). Третьим элементом информационной системы является канал (среда) передачи, связывающий ИсС и ПС.

Отображение сообщения обеспечивается изменением какой-либо физической величины, характеризующей процесс (например, амплитуда, частота, фаза). Эта величина является *информационным* *параметром сигнала* (в общем случае–информационной системы).

ИсС и ПС обмениваются информацией в технических системах в виде сигналов, сформированных на основе определенного алфавита. Характеристикой алфавита является его мощность, N – количество символов, на основе которых формируется сообщение. Например, мощность английского алфавита – 26 символов, русского – 33 символа, мощность алфавита, на основе которого функционируют и взаимодействуют между собой компьютеры, составляет 2 символа (0 и 1).

Сигналы, как и сообщения, могут быть непрерывными и дискретными. Информационный параметр непрерывного сигнала с течением времени может принимать любые мгновенные значения в определенных пределах. Непрерывный сигнал часто называют аналоговым, а каналы и устройства, функционирующие на основе такого типа сигналов – аналоговыми. Дискретный сигнал (устройство или канал передачи) характеризуется конечным числом значений информационного параметра.

Дискретные сообщения состоят из последовательности дискретных знаков. Часто этот параметр принимает всего два значения (0 или 1).

Сообщение или канал его передачи на основе этих двух значений сигнала называют двоичным или бинарным. Построение сигнала по определенным правилам, обеспечивающим соответствие между сообщением и сигналом, называют кодированием.

Кодирование в широком смысле– преобразование сообщения в сигнал.

Кодирование в узком смысле – представление исходных знаков, называемых символами, в другом алфавите с меньшим числом знаков. Оно осуществляется с целью повышения надежности и преобразования сигналов к виду, удобному для передачи по каналам связи. Последний тип кодирования относится к так называемой прикладной теории кодирования информации, занимающейся поиском и реализацией методов и средств обнаружения несоответствий (ошибок) между переданным и принятым сообщениями.

В произвольном сообщении символы алфавита могут появляться с различной вероятностью. Если длина сообщения достаточно велика, то статистический анализ этого сообщения позволит получить вероятностные характеристики данного алфавита. Очевидно, что отличные символы в произвольном сообщении (особенно при N> 2) появляются с различной вероятностью.

Информационной характеристикой алфавита (источника сообщений на основе этого алфавита) является энтропия.

Этот термин применительно к техническим системам был введен К. Шенноном и Р. Хартли.

(1.1)

Энтропию алфавита А{} по Шеннону рассчитывают по следующей формуле:

где P() – вероятность P(ξ=); – элемент алфавита,.

С физической точки зрения энтропия показывает, какое количество информации приходится в среднем на один символ алфавита. Частным случаем энтропии Шеннона является энтропия Хартли. Дополнительным условием при этом является то, что все вероятности одинаковы и постоянны для всех символов алфавита.

С учетом этого формулу (1.1) можно преобразовать к виду:

(1.2)

Например, энтропия Хартли для латинского (английского) алфавита составляет 4,7 бит. Если подсчитать энтропию Шеннона и энтропию Хартли для одного и того же алфавита, то они окажутся неравными. Это несовпадение указывает на избыточность любого алфавита (при N> 2).

Сообщение M, которое состоит из n символов, должно характеризоваться определенным количеством информации I(M):

(1.3)

Здесь Н(А)– энтропия алфавита с соответствующим распределением вероятностей р().

Нетрудно предположить и просто убедиться, что количество информации в сообщении, подсчитанное по Шеннону, не равно количеству информации, подсчитанному по Хартли. На основе этого парадокса строятся и функционируют все современные системы сжатия (компрессии) информации.

Любые сообщения характеризуются информационной избыточностью, что позволяет сжимать их без потери информации.

Двоичный канал передачи информации является дискретным – он основан на алфавите, состоящем из двух символов: 0 и 1 – A {0,1}.

Полагая, что сообщение М состоит только из единиц (М = 11…1) и имеет длину n: 111…11, вероятность того, что произвольный символ равен единице, составляет единицу (Р(1) = 1); другая вероятность – Р(0) = 0 для . Здесь имеет место использование моноалфавита: алфавита, состоящего из одного символа.

Физический смысл понятия информации в теории Шеннона: информацией является лишь такое сообщение, которое снимает некоторую неопределенность, т. е. содержит новые для получателя данные.

1. **Практическая часть**

В данной лабораторной работе необходимо было создать приложение для расчета и анализа параметров и информативных характеристик дискретных ИС, с помощью которого:

*а) рассчитать энтропию указанных преподавателем алфавитов: один – на латинице, другой – на кириллице (по формуле (2.1) перейти от частоты появления каждого символа алфавита к соответствующей вероятности); в качестве входного может быть принят произвольный электронный текстовый документ на основе соответствующего алфавита; частоты появления символов алфавитов оформить в виде гистограмм (можно воспользоваться приложением MS Excel);*

При выполнении лабораторной работы было выбрано два текста: один на украинском языке, другой на голандском. При расчете частоты появления символов из текста убирались ненужные символы, такие как точки, запятые, тире и т. д.

Для нахождения энтропии алфавитов необходимо сначала найти частоты появления символов соответствующих алфавитов. Так, на рисунке 2.1 представлены частоты появления латинского и русского алфавитов.

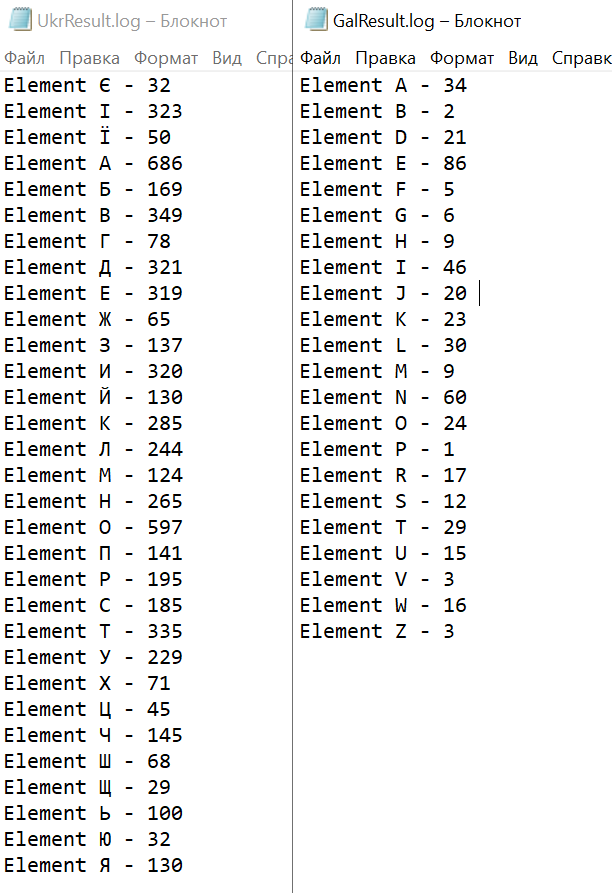


Рисунок 2.1 – Частоты появления символов алфавитов

Программный код для расчета частоты появления символов представлен на рисунке 2.2.

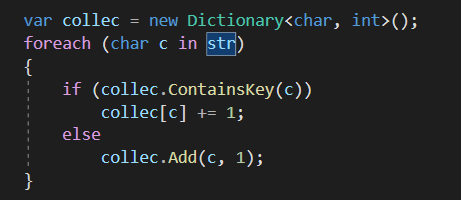


Рисунок 2.2 – Функция расчета частоты появления символов алфавитов

Гистограмма частот появления букв украинского алфавита представлена на рисунке 2.3.



Рисунок 2.3 – Гистограмма частот появления символов украинского алфавита

Используя формулу Шеннона (формула 1.1) и полученные числа подсчитываем энтропию (рисунок 2.4):



Рисунок 2.4 – Частоты появления символов алфавитов

Программная реализация функции расчета энтропии представлена на рисунке 2.5. Параметр collec – это коллекция частот появления символов алфавита.

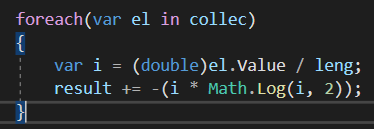


Рисунок 2.5 – Реализация нахождения энтропии

*б) для входных документов, представленных в бинарных кодах, определить энтропию бинарного алфавита;*

Двоичный канал передачи информации является дискретным – он основан на алфавите, состоящем из двух символов: 0 и 1 – A{0,1}.

Для представления исходного текста в бинарном виде была написана следующая функция, представленная на рисунке 2.6.

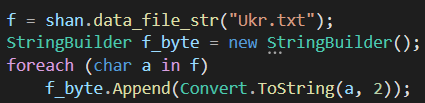


Рисунок 2.6 – Функция представления текста в бинарном виде

В данном случае алфавит состоит только из двух символом, найдем частоту их появления (рисунок 2.7).

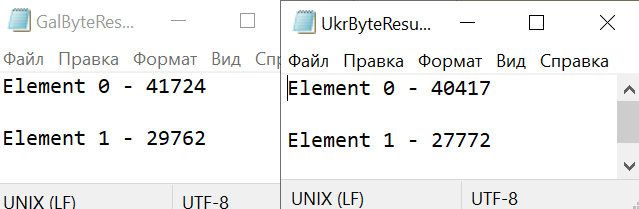


Рисунок 2.7 – Частоты появления символов бинарных символов

Воспользовавшись вышеуказанными функциями, найдем энтропию алфавитов (рисунок 2.8).



Рисунок 2.8 – Энтропия бинарных алфавитов

*в) используя значения энтропии алфавитов, полученных в пунктах (а) и (б), подсчитать количество информации в сообщении, состоящем из собственных фамилии, имени и отчества (на основе исходного алфавита – (а) и в кодах ASCII – (б)); объяснить полученный результат;*

Для вычисления количества информации в сообщении была написана функция, представленная на рисунке 2.9.

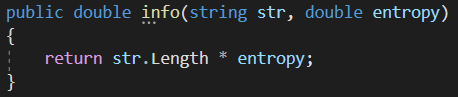


Рисунок 2.9 – Функция нахождения информативности

Функция принимает в себя 2 параметра: сообщение, информативность которого необходимо найти, и энтропию алфавита. Полученная информативность для собственных фамилии, имени и отчества представлена на рисунке 2.10.

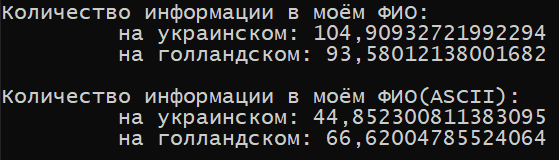


Рисунок 2.10 – Информативность сообщений

*г) выполнить задание пункта (в) при условии, что вероятность ошибочной передачи единичного бита сообщения составляет: 0,1; 0,5; 1,0.*

Реализация функции подсчёта ошибочной передачи единичного бита представлена на рисунке 2.11:

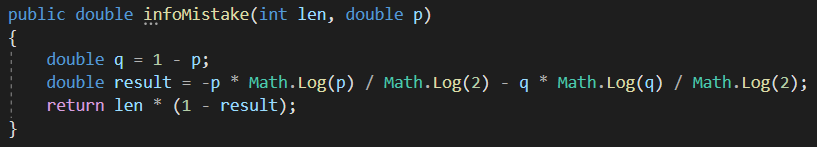
**

Рисунок 2.11 – Функция подсчёта ошибочной передачи единичного бита

Функция принимает два параметра: первый – длина сообщения, второй - вероятность ошибочной передачи единичного бита.

Полученные результаты отображены на рисунке 2.12:

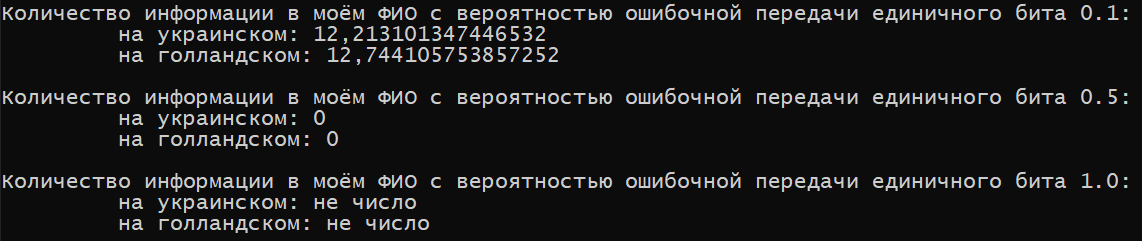


Рисунок 2.12 – Информативность с вероятностями ошибок

Такие результаты показывают, что, в случае вероятности ошибки равной 0,5, сообщение не несёт никакой информации, а с вероятностью 1.0 – вовсе выходит за предел определённости.

**Вывод:** в данной лабораторной работе я приобрел практические навыки расчета и анализа параметров и информативных характеристик дискретных информационных систем, выводы результатов моих действий были описаны в данном отчёте. Также я закрепил теоретические знания по основам теории информации и разработал приложение для расчета и анализа параметров и информативных характеристик дискретных информационных систем.