Учреждение образования

«Белорусский государственный технологический университет»

**Кафедра информационных систем и технологий**

**«Отчёт по лабораторной работе №2»**

“Элементы теории информации. Параметры и характеристики дискретных информационных систем”

**Выполнил:** студент 3 курса

4 группы специальности ПОИТ

Мергель Каролина Андреевна

**Проверил:** преподаватель

Копыток

Минск 2022

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.

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.

1. **Практическая часть**

В данной лабораторной работе необходимо рассчитать энтропию указанного преподавателем алфавитов: английский или русский; в качестве входных данных принимать произвольный электронный текстовый документ на основе соответствующего алфавита; частоты появления символов алфавитов оформить в виде гистограмм.

Для более точного анализа данных приложение автоматически убирает пробелы, точки, запятые и прочие символы-сепараторы, а также переводит текст в нижний регистр. Далее подсчитывается количество повторений для каждого символа алфавита в тексте, затем каждый из полученных значений делим на общее количество символов в тексте.

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

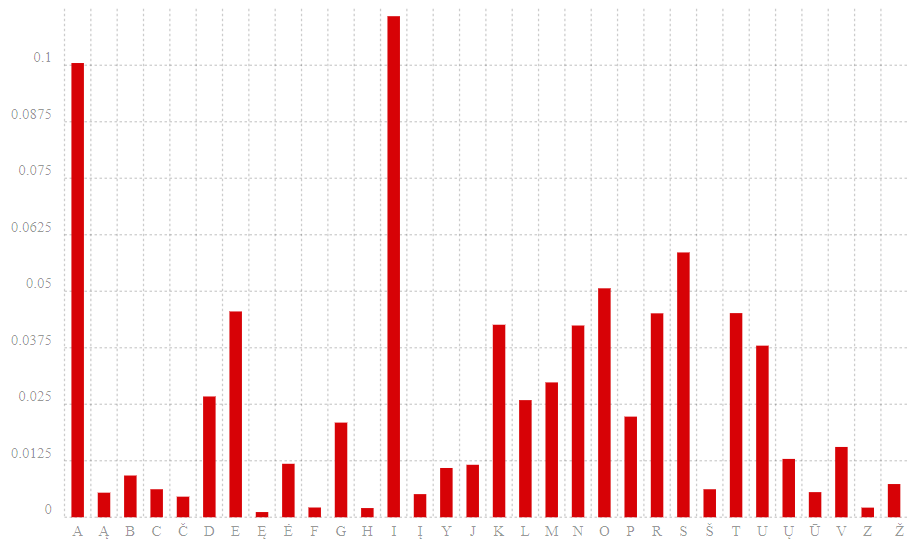


Рисунок 2.1 – расчет энтропии для литовского языка

|  |  |
| --- | --- |
| Всего символов текста: | 40829 |
| Энтропия алфавита: | 3.807540 |

На рисунке 2.1 представлена энтропия для литовского языка и частота появления каждого символа из алфавита.

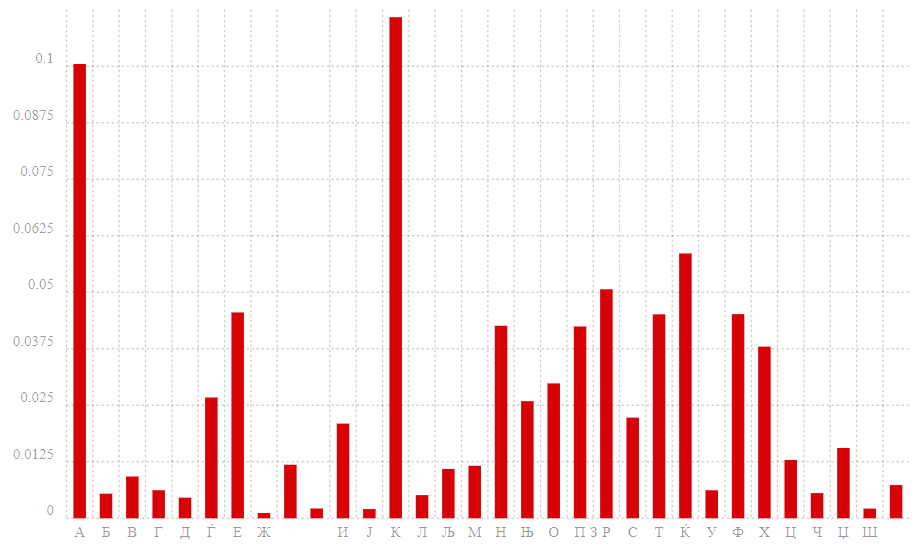


Рисунок 2.2 – энтропия македонского языка

|  |  |
| --- | --- |
| Всего символов текста: | 38449 |
| Энтропия алфавита: | 3.536094 |

На рисунке 2.2 представлена энтропия для македонского языка и частота появления каждого символа из алфавита.

Следующим заданием было расчёт энтропии для текста, представленного в бинарном виде.

Двоичный канал передачи информации является дискретным – он основан на алфавите, состоящем из двух символов: 0 и 1 – A {0,1}. Используя (1.1), вычислим энтропию этого алфавита:

(1.4)

Так как в последующем двоичный формат записи символа мы будем получать из кодировки ASCII, на вход программы нужно подать последовательность любых символов, которые имеются в таблице ASCII. Программа переведет посимвольно в двоичную систему счисления и посчитает энтропию.

Для задания были использованы два файла из предыдущего задания, для этого символы были переведены в символы ASCII и для литовского windows-1251 и посимвольно преобразованы в двоичный вид.

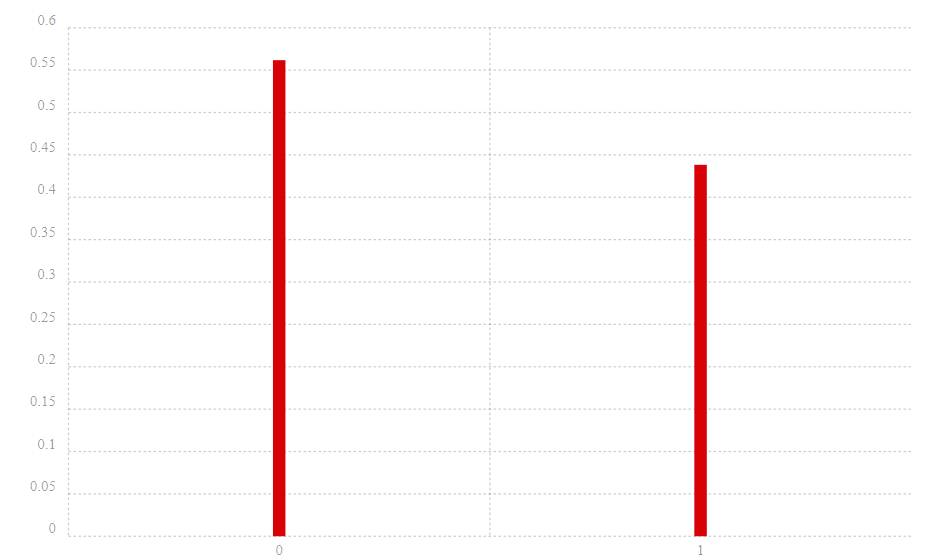
Посчитаем энтропию для текста на литовском языке переведя в двоичный формат. 

Рис 2.3 – энтропия двоичного алфавита (символы литовского)

|  |  |
| --- | --- |
| Всего символов текста: | 283394 |
| Энтропия алфавита: | 0.988869 |

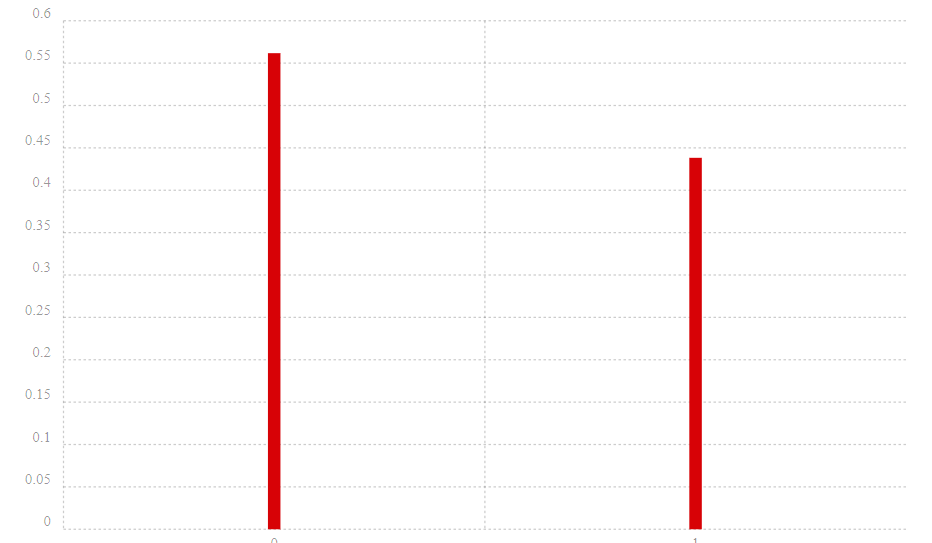
Посчитаем тоже самое, но для македонского. 

Рис 2.4 - энтропия двоичного алфавита (символы македонского)

|  |  |
| --- | --- |
| Всего символов текста: | 290003 |
| Энтропия алфавита: | 0.988975 |

В следующем задании подсчитать количество информации в сообщении из собственных фамилии, имени и отчества (на основе исходного алфавита – (а) и в кодах ASCII – (б));

Строки используемые для расчетов (пробелы исключаются)

" Mergel Karolina Andreevna "

" Мергел Каролина Андреевна "

**Результат расчетов:**

Размер информации на литовском: 115

Размер информации на литовском ASCII: 184

Размер информации на македонском: 113.946508

Размер информации на македонском ASCII: 184

Далее необходимо подсчитать количество информации при условии, что вероятность ошибочной передачи единичного бита сообщения составляет: 0.1; 0.5; 1.0.

Условной энтропией Источника дискретного сообщения X в ДСК называем величину

Примем p – количество ошибок, q – количество правильной информации, по формуле:

H(x|y) = - q log2 q - p log2 p;

Посчитаем количество потерянной информации для каждого случая:

H(x|y) = - 0.1 log2 0.1 – 0.9 log2 0.9 = 0.46899

He =1 - 0.46899 = 0.53111

23\*0.53111=12.21553

H(x|y) = - 0.5 log2 0.5 – 0.5 log2 0.5 = 1

He =1 - 1 = 0

0

H(x|y) = - 1 log2 1 – 0 log2 0 = NaN

**Вывод:** В данной лабораторной работе рассчитали энтропию для двух языков, для данного варианта македонского и литовского. Также энтропию для двоичного алфавита при больших значениях энтропия в этом алфавите приближается всегда к единице. При расчетах в третьем задании мы посчитали количество информации для конкретных алфавитов и информацию, которая нужна для кодировки этих символов в ASCII, исходя из результатов мы видим, что у нас присутствует избыточная информация, особенно для латинских символов, так как последний бит никогда не используется. В последнем задании можно сделать вывод о том, что самый плохой исход при передаче информации получить половину некорректных данных, так как нам придется исправлять в любом случае ровно половину неправильных значений.