Министерство образования Республики Беларусь

Учреждение образования «Белорусский государственный технологический университет»

Факультет информационных технологий

Кафедра информационных систем и технологий

**Отчет к лабораторной работе**:

«Элементы теории информации. Параметры и характеристики дискретных информационных систем»

Выполнил:

студент 3 курса 7 группы

специальности ПОИБМС

Кот И.В.

Минск 2020

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.

Таким образом, простейшая информационная система состоит из трех элементов: источника сообщения, канала передачи сообщения и получателя сообщения. Отображение сообщения обеспечивается изменением какой-либо физической величины, характеризующей процесс (например, амплитуда, частота, фаза). Эта величина является информационным параметром сигнала.

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.

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

(1.1)

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 рассчитать энтропию указанного преподавателем алфавитов: белорусский и польский; в качестве входных данных принимать произвольный электронный текстовый документ на основе соответствующего алфавита; частоты появления символов алфавитов оформить в виде гистограмм.

Для расчёта энтропии алфавита необходимо было текст убрать знаки препинания и пробелы, а также привести к нижнему регистру входной текст и алфавит. 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 частоту появления каждого символа высчитаем энтропию для белорусского алфавита:

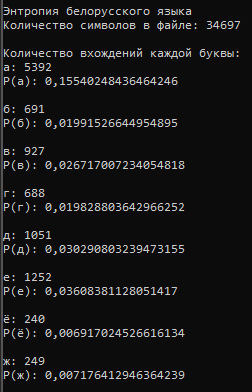


Рисунок 2.1 Расчет энтропии для белорусского языка

Теперь построим гистограмму частоты символов для этого алфавита.

Гистограмма — способ графического представления табличных данных.

Количественные соотношения показателя частоты появления символов представлены в виде прямоугольников, площади которых пропорциональны.

Частота появления символов белорусского алфавита представлена на гистограмме ниже:

Рисунок 2.2 Частота появления символов белорусского алфавита

Частота появления символов польского алфавита представлена на гистограмме ниже:

Рисунок 2.3 Частота появления символов польского алфавита

Из гистограмм видно, что в белорусском языке наиболее часто встречаются символа «а», «н», «i», в польском «a», «e», «i», «o», «z».

Теперь вычислим энтропию в бинарном алфавите.

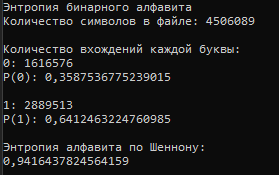


Рисунок 2.4 Энтропия бинарного алфавита по Шеннону

Используя значения энтропии алфавитов, полученных ранее, необходимо подсчитать количество информации в сообщении, состоящем из собственного ФИО.

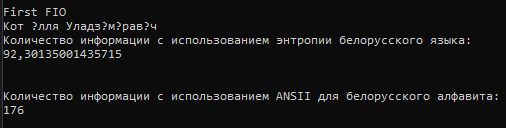


Рисунок 2.5 Количество информации в ФИО

Далее необходимо выполнить предыдущие задания, но при условии, что вероятность ошибочной передачи единичного бита сообщения составляет: 0,1; 0,5; 1,0.

Результаты изображены на следующей картинке:

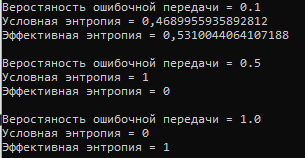


Рисунок 2.6 Энтропия при вероятности ошибочной передачи 0.1, 0.5, 1

Вывод

В данной лабораторной работе были получены практические навыки расчета информативных характеристики дискретных информационных систем. Также для решения задач было разработано консольное приложение для анализа частоты появления символов русского и английского алфавита в загруженном тексте, и была получена энтропия данных алфавитов и количество информации при передаче ФИО.