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 6 группы

специальности ПОИТ

Капорцев О. А.

Минск 2021

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 разработать программу, которая бы позволяла выполнять следующие задачи:

* Вычисляла энтропию указанного преподавателем алфавитов: немецкого и сербского. А также оформлять частоты появления символов заданного алфавита в виде гистограммы. В качестве входных данных принимать произвольный электронный текстовый документ на основе соответствующего алфавита.
* Вычисляла энтропию бинарного алфавита.
* Производила подсчет количества информации для заданного текста, используя энтропию бинарного алфавита и по заданию.
* Производила подсчет количества информации, если вероятность доставления информации с ошибкой составляет: 0.1, 0.5 и 1.0.

В качестве программы было выбрано веб-приложение. Для более точного анализа данных сервер автоматически убирает пробелы, точки, запятые и прочие символы-сепараторы, а также переводит текст в нижний регистр. Далее создаётся ассоциативный массив, состоящий из символов заданного алфавита, в котором ключи – символы алфавита, а значения – количество повторений исходного текста. После подсчёта рассчитывается процент каждого символа в исходном тексте.

Наконец, используя формулу Шеннона (формула 1.1) и полученные числа, подсчитываем энтропию:

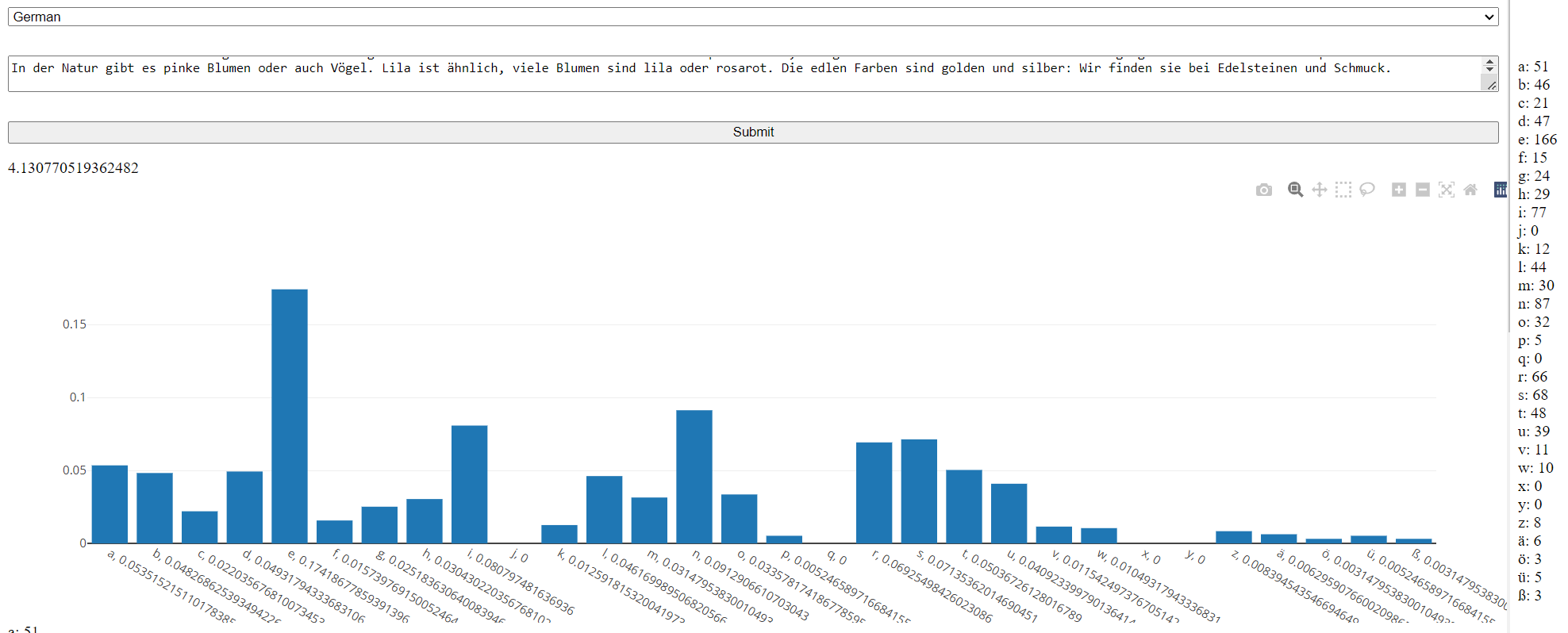


Рисунок 2.1 - Расчет энтропии немецкого языка

На Рисунке 2.1 представлен расчет энтропии немецкого языка для текста на клиенте веб-приложения. Для этого имеются форма, заполнение которой приводит к отправке запроса на сервер, который по алгоритму, описанному выше, рассчитывает энтропию и высылает данные для построения гистограммы – способа графического представления табличных данных.

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

Гистограммы были оформлены на клиенте веб-приложения.

Частота появления символов сербского алфавита представлена на Рисунке 2.2:

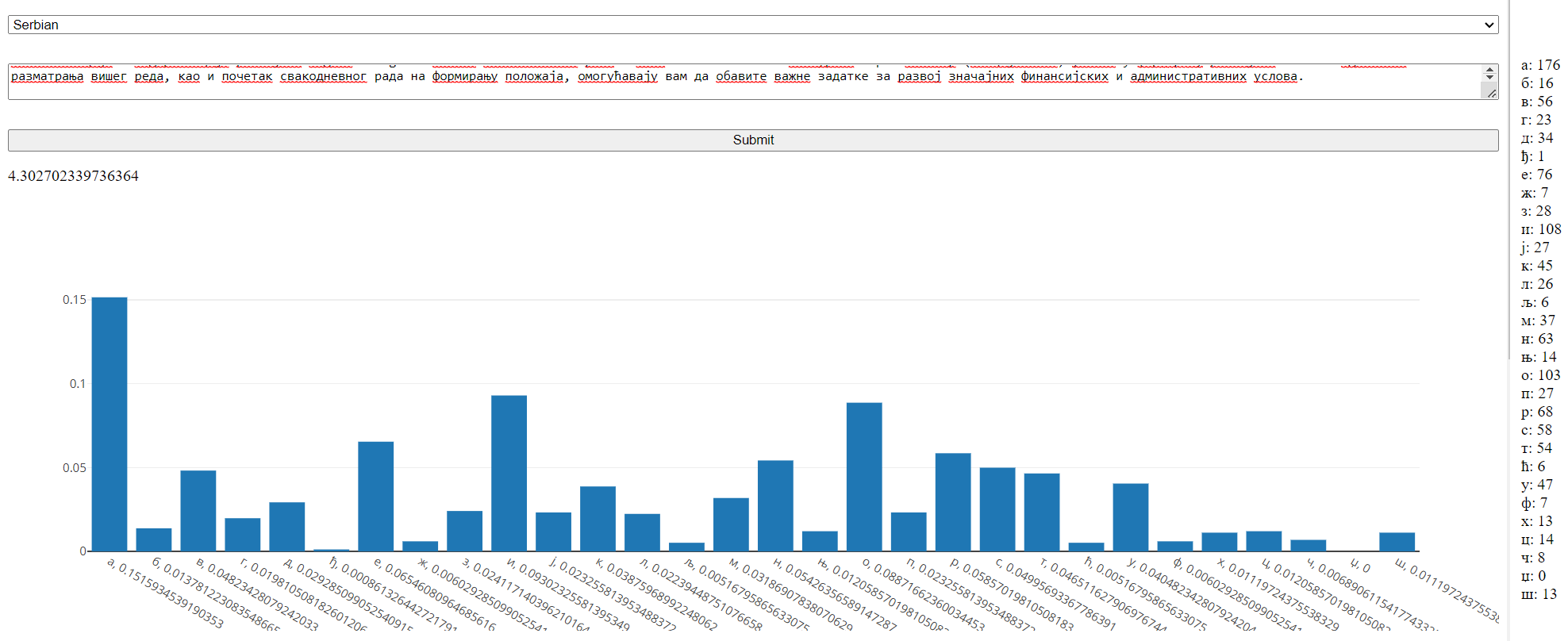


Рисунок 2.2 – Расчет энтропии сербского языка

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.

Работа приложения над выполнением этого задания:

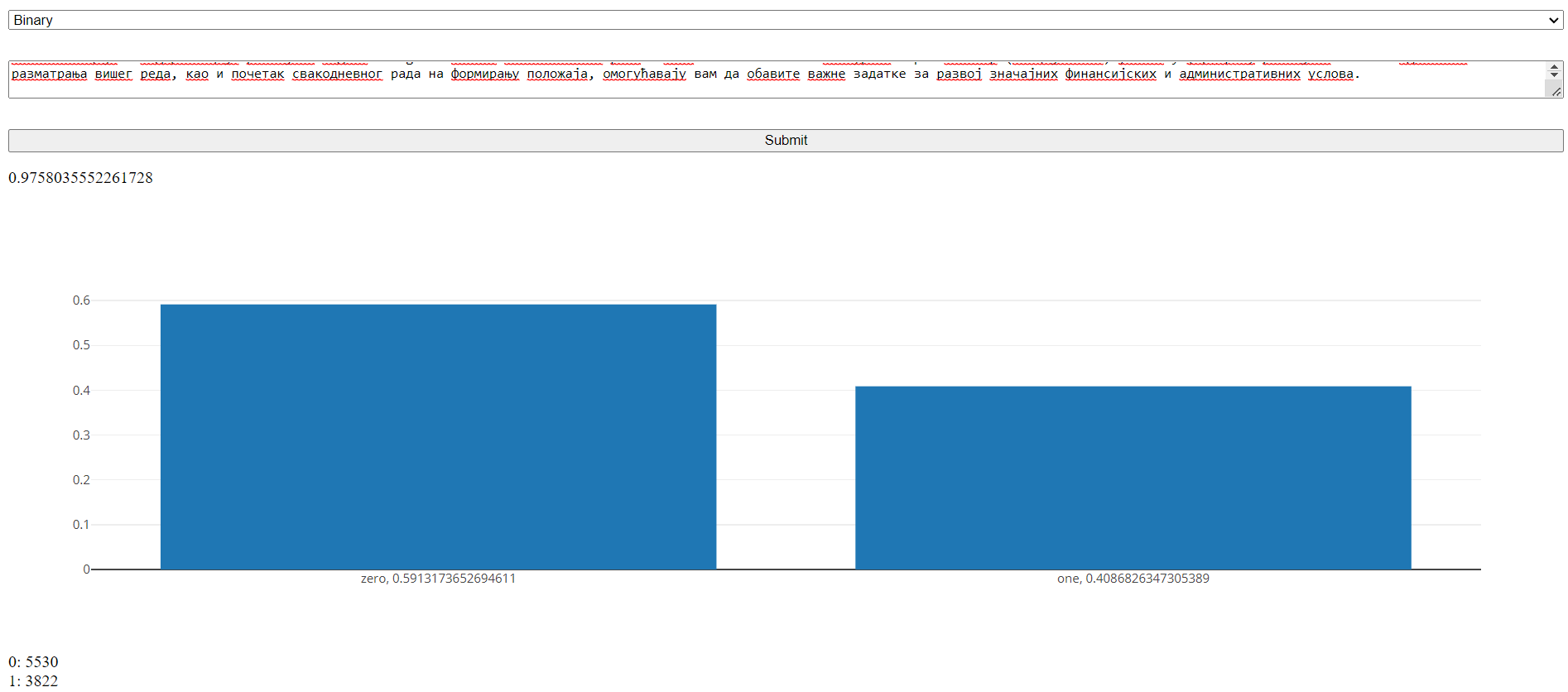


Рисунок 2.3 - Энтропия бинарного алфавита

Используя значения энтропии алфавитов, полученных в заданиях ранее, необходимо было подсчитать количество информации в сообщении, состоящем из собственных фамилии, имени и отчества. А также посчитать количество информации при вероятности ошибки 0.1, 0.5 и 1.0. Используя вероятность ошибки, мы можем высчитать условную энтропию согласно данной формуле:

Где P(p) – вероятность ошибки, P(q) = 1– P(p). Далее, отнимая энтропию от условной, мы получаем эффективную энтропию. Умножая её на количество символов текста, мы получаем фактическое количество информации, передаваемое по каналу.

На Рисунке 2.4 приведены итоги расчётов для заданного сообщения:

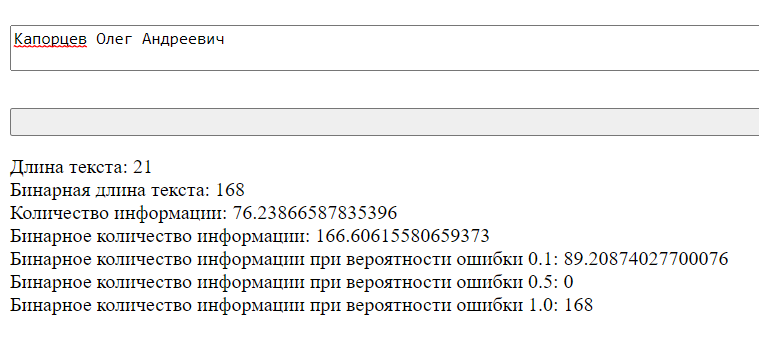


Рисунок 2.4 – Количество информации ФИО

Так как мы высчитывали количество информации при наличии ошибок для бинарного алфавита, то при вероятности 1.0 всё количество должно равняться нулю. Однако, так как у нас бинарный алфавит, то мы можем сделать операцию XOR, которая инвертирует каждый бит, и мы сможем получить исходное сообщение. Поэтому количества информаций при вероятности ошибок 1.0 и 0.0 равны. При вероятности же 0.5 мы не можем выяснить какой из битов правильный, а какой нет – следовательно, количество информации будет 0.

**Вывод**

В данной лабораторной работе я приобрел практические навыки расчета и анализа параметров и информативных характеристик дискретных информационных систем. Также закрепил теоретические знания по основам теории информации и разработал приложение для расчета и анализа параметров и информативных характеристик дискретных информационных систем.