Учреждение образования

«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ»

Элементы теории информации. Информативность данных в различных кодировках

Студент: Голодок А. Ю.

ФИТ 3 курс 4 группа

Преподаватель: Сазонова Д. В.

Минск 2023

1. **Энтропия**

Информационной характеристикой алфавита (источника сообщений на основе этого алфавита) является энтропия. С физической точки зрения энтропия алфавита показывает, какое количество информации приходится в среднем на один символ алфавита. Энтропию алфавита ***А*** = {*ai*} по К. Шеннону рассчитывают по следующей формуле:

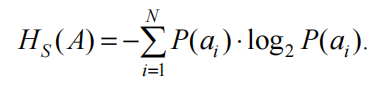


Рисунок 1.1 – формула энтропии

Чтобы вычислить энтропию, необходимо подготовить структуру данных, для хранения информации о каждой букве и количестве её вхождений в исходную строку. Программное решение представлено на рисунке 1.2.

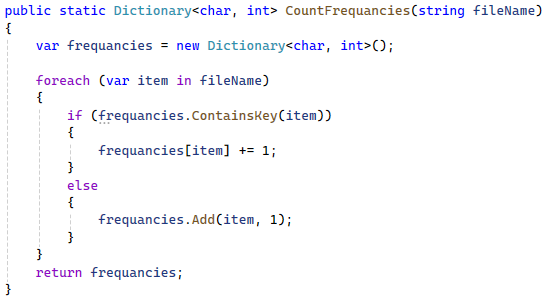


Рисунок 1.2 – функция подсчета частоты появления буквы

Затем вычисляется значение энтропии при помощи функции, представленной на рисунке 1.3.

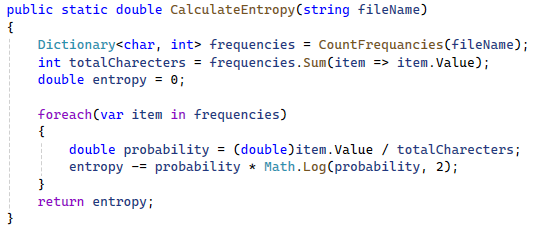


Рисунок 1.3 – функция вычисления энтропии Шеннона

* 1. **Кириллица**

Для исследования кириллического алфавита был выбран текст на русском языке. Исходный текст представлен на рисунке 1.4.

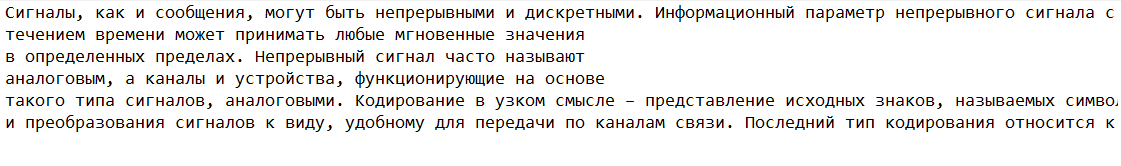


Рисунок 1.4 – исходный текст русского алфавита

* 1. **Латиница**

В качестве исходного текста для исследования энтропии латиницы был выбран текст на английском языке, представленный на рисунке 1.5.

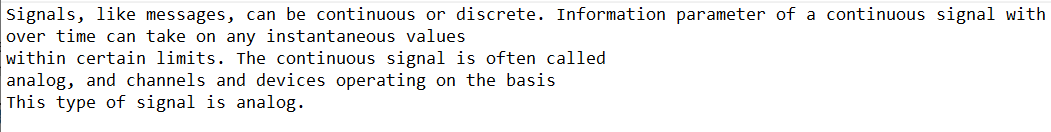


Рисунок 1.5 – исходный текст английского алфавита

* 1. **Частота появления**

Графики, отражающие частоты появления символов в тексте русского и английского алфавита, представлены на рисунке 1.6.

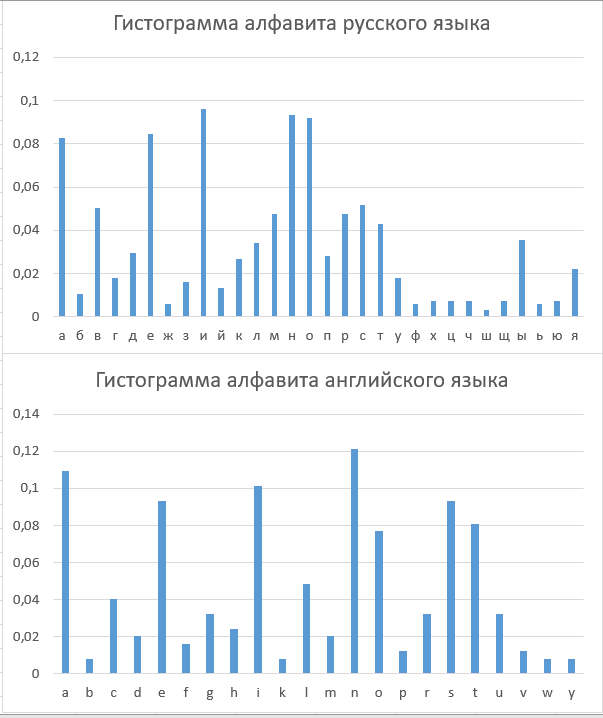


Рисунок 1.6 – гистограммы частоты появления

На основе этих данных, получены значения энтропии для русского и английского алфавитов. Результат представлен на рисунке 1.7.



Рисунок 1.7 – энтропия алфавитов

Вычислена энтропия русского и английского алфавита в бинарном формате. Полученные значения представлены на рисунке 1.8.



Рисунок 1.8 – энтропия алфавитов в бинарном представлении

1. **Количество информации**

Количество информации рассчитывается по формуле, представленной на рисунке 2.1.



Рисунок 2.1 – формула количества информации

В ней значение *H(A)* это значение энтропии алфавита, а *k* – количество символов исходного текста.

Расчет количества информации при помощи программы показано на рисунке 2.2.

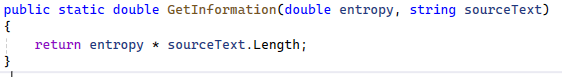


Рисунок 2.2 – функция расчета количества информации

Вычислено количество информации для текста, состоящего из ФИО на русском языке и в бинарном формате. Результат представлен на рисунке 2.3.



Рисунок 2.3 – количество информации в ФИО на русском и в бинарном формате

1. **Эффективная энтропия**

Если переданное сообщение может содержать ошибки, количество информации определятся по формуле эффективной энтропии, представленной на рисунке 3.1, 3.2.



Рисунок 3.1 – формула эффективной энтропии бинарного алфавита

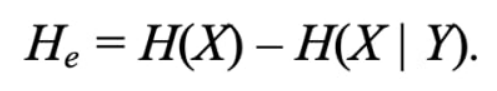


Рисунок 3.2 – формула эффективной энтропии не бинарного алфавита

В данной формуле *H(Y* | *X)* – условная энтропия, вычисляемая по формуле на рисунке 3.3.



Рисунок 3.3 – формула условной энтропии

Здесь *p –* вероятность передачи ошибочного символа.

Для программного расчета эффективной энтропии разработана функция, представленная на рисунке 3.4.

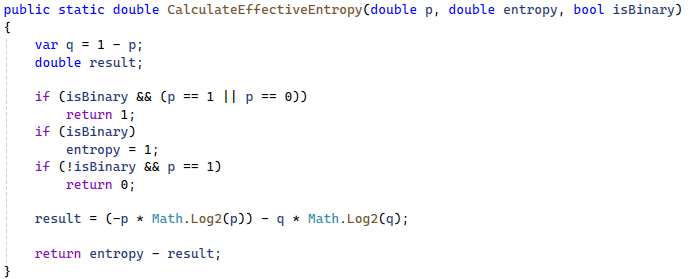


Рисунок 3.4 – функция расчета эффективной энтропии

В качестве практического примера рассчитана эффективная энтропия для текста русского алфавита и его двоичного представления. Полученные вычисления представлены на рисунке 3.5.

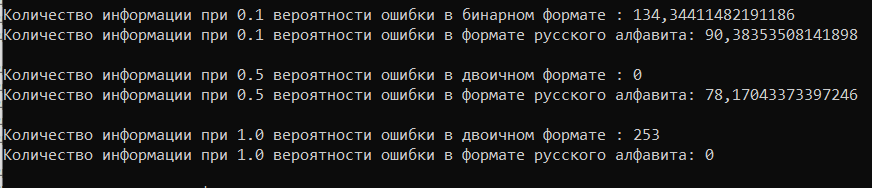


Рисунок 3.5 – количество информации при вероятностях ошибочной передачи