Учреждение образования

«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ»

ЭЛЕМЕНТЫ ТЕОРИИ ИНФОРМАЦИИ. ПАРАМЕТРЫ И ХАРАКТЕРИСТИКИ ДИСКРЕТНЫХ ИНФОРМАЦИОННЫХ СИСТЕМ

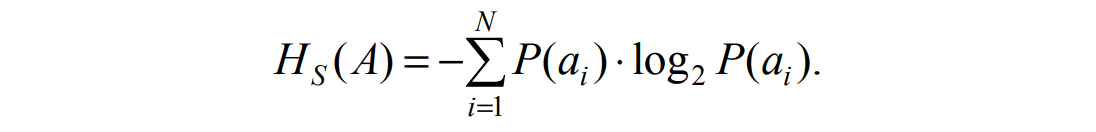
Студент: Яшный Н. С.

ФИТ 3 курс 4 группа

Минск 2024

# Энтропия алфавита

Информационной характеристикой алфавита (источника сообщений на основе этого алфавита) является энтропия. Энтропию алфавита *А =* {*ai*} по К. Шеннону рассчитывают по следующей формуле:

(1.1)

С физической точки зрения энтропия алфавита показывает, какое количество информации приходится в среднем на один символ алфавита.

Для вычисления энтропии в начале необходимо подсчитать количество появлений каждого символа в исходной строке. Для этого используется функция, представленная на рисунке 1.1.

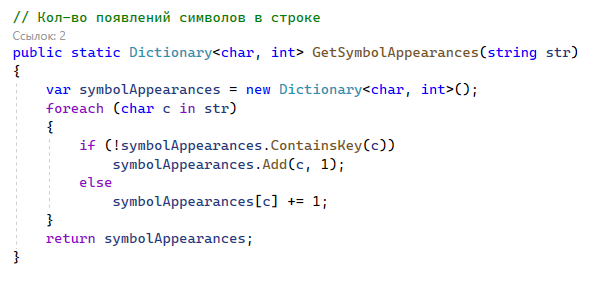


Рисунок 1.1 – Функция подсчёта кол-ва появлений символов

Далее, для вычисления энтропии, необходимо перейти от количества появлений символов к вероятности их появления. Для этого необходимо кол-во появлений каждого символа разделить на кол-во символов в строке. Далее эту вероятность *P*(*ai*)можно подставлять в формулу (1.1). Функция, вычисляющая энтропию Шеннона, представлена на рисунке 1.2.

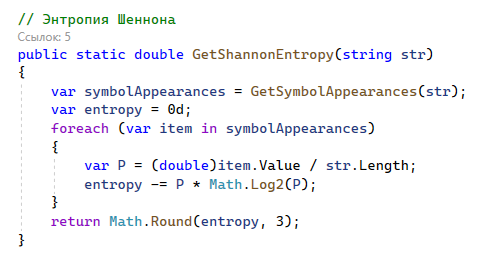


Рисунок 1.2 – Функция вычисления энтропии Шеннона

# Кириллица

В качестве кириллического алфавита был выбран белорусский. Электронный текстовый документ на основе алфавита содержится в файле belarusian.txt, представленном на рисунке 1.3.

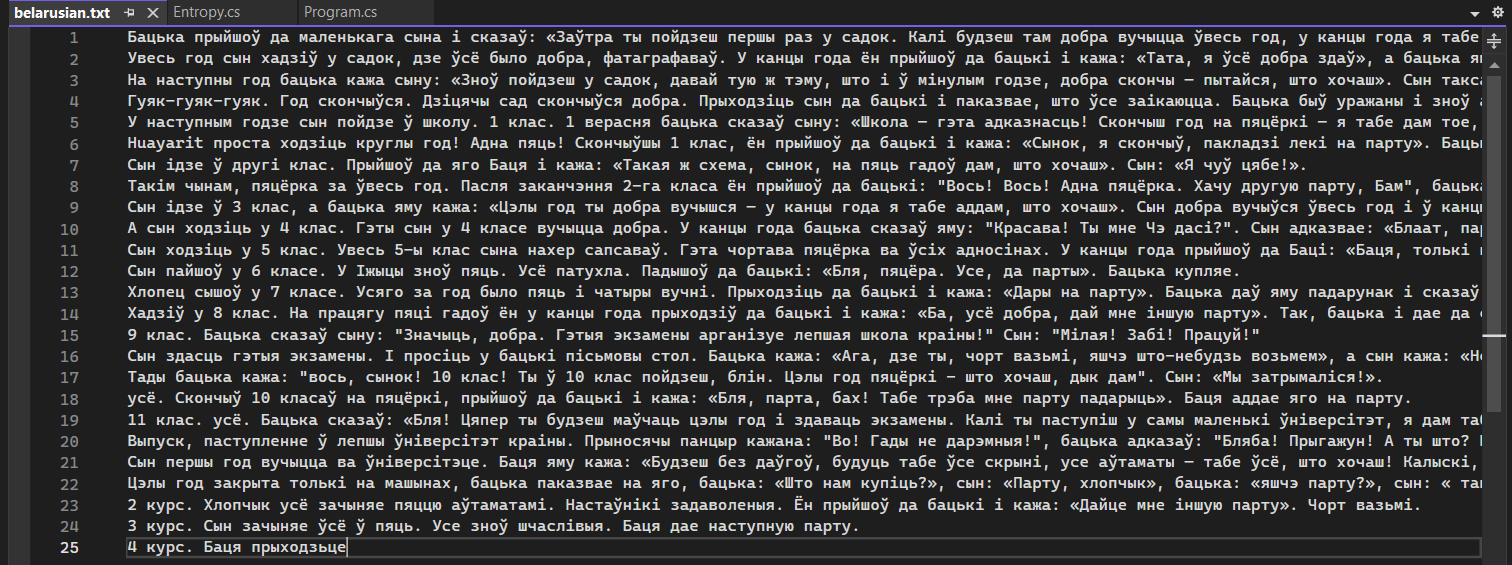


Рисунок 1.3 – Текст на белорусском алфавите

# Латиница

В качестве алфавита на латинице был выбран польский алфавит. Электронный документ, в котором содержится текст на исходном польском алфавите для подсчета энтропии алфавита, содержится в файле polish.txt, представленном на рисунке 1.4.

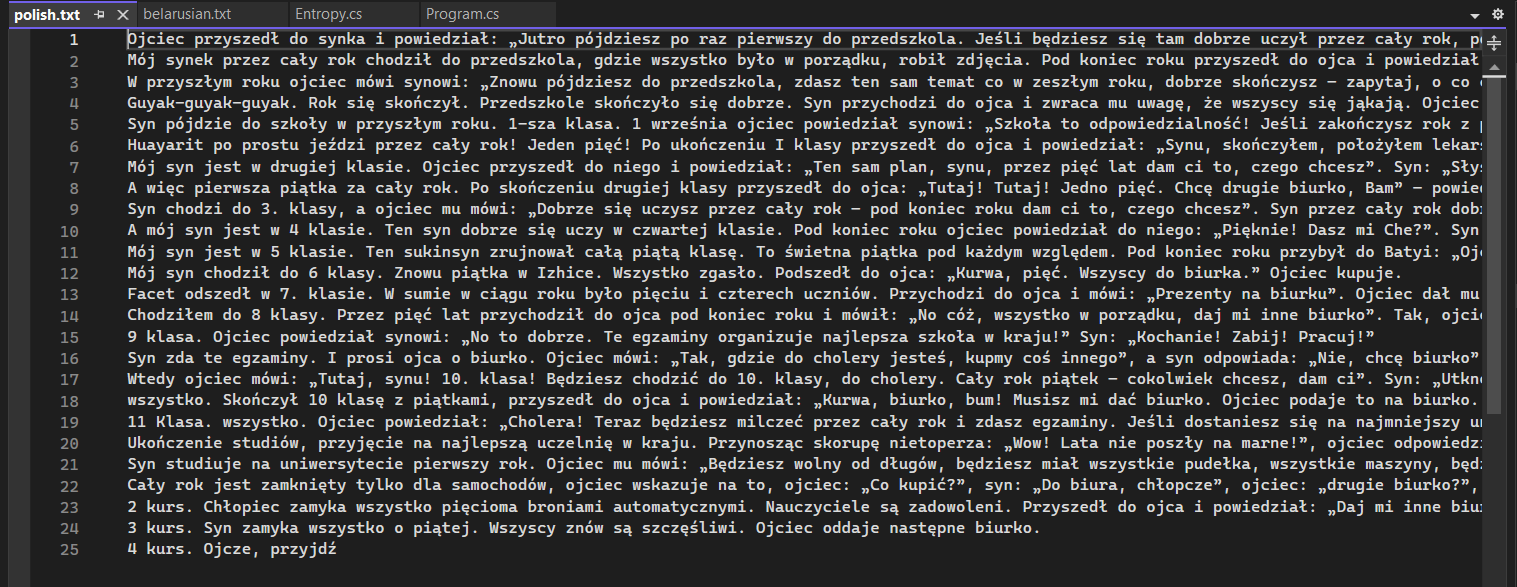


Рисунок 1.4 – Текст на польском алфавите

# Вычисление энтропии алфавита

Графики, отражающие частоты появления символов в тексте на белорусском и польском алфавитах, представлены на рисунке 1.5.

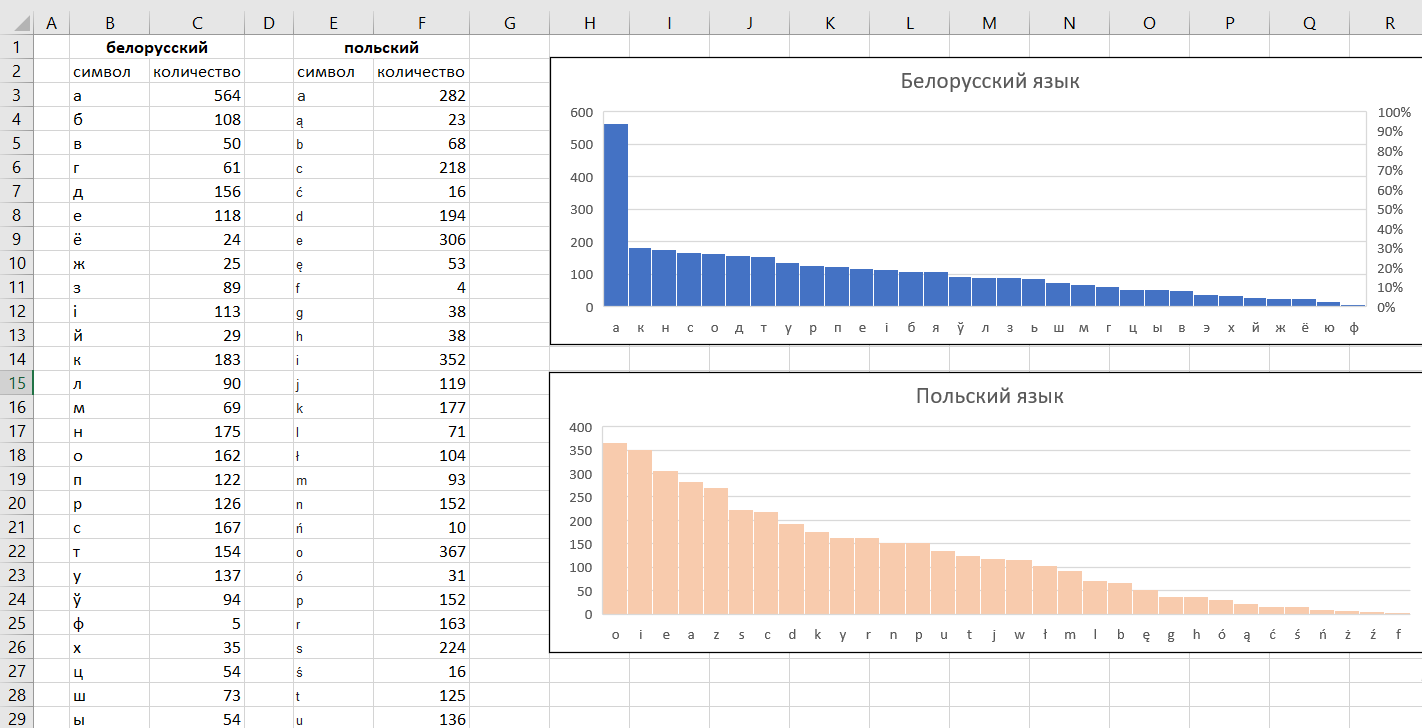


Рисунок 1.5 – Частоты появления символов в алфавитах

Из этих данных можно получить вероятности появления каждого символа *P*(*ai*)*.* При подстановке их в формулу (1.1) получаем следующие значения энтропии для белорусского, польского и бинарного алфавитов, продемонстрированные на рисунке 1.6:

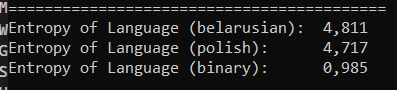


Рисунок 1.6 – Вывод функции подсчёта энтропии

Энтропия белорусского и польского алфавитов равны соответственно 4,811 бит и 4,717 бит. Энтропия бинарного алфавита, которая всегда примерно равна 1 биту, в данном случае равняется 0,985 бит.

# Количество информации

При известной энтропии алфавита количество информации вычисляется по следующей формуле:

(1.2)

То есть, необходимо умножить полученную энтропию на количество символов в сообщении. Программный код вычисления количества информации сообщения представлен на рисунке 2.1.

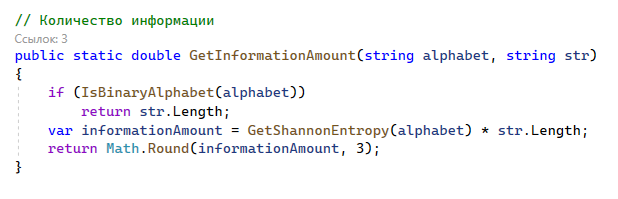


Рисунок 2.1 – Функция количества информации

Если алфавит является бинарным, то за количество информации принимается длина строки, так как энтропия равняется 1. В противном случае, энтропия выбранного алфавита умножается на количество символов в сообщении, количество информации в котором мы желаем найти.

Вывод функции представлен на рисунке 2.2.

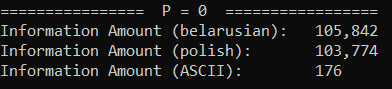


Рисунок 2.2 – Вывод количества информации

Как видно из данного рисунка, количество информации в сообщении, состоящем из ФИО в кодировке ASCII, равняется 176 битам, то есть равняется количеству символов в исходном сообщении, записанном в бинарном виде.

# Эффективная энтропия

Если сообщение может передаться ошибочно, то для вычисления количества информации используется формула эффективной энтропии:

(1.3)

где *H*(*Y | X*) – условная энтропия:

(1.4)

Программный код функции для вычисления эффективной энтропии представлен на рисунке 3.1.

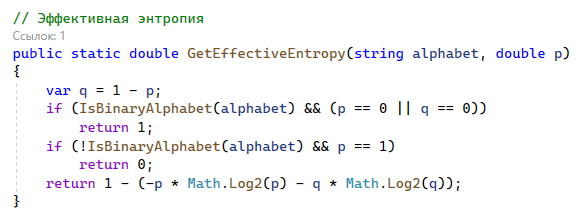


Рисунок 3.1 – Функция эффективной энтропии

В бинарном алфавите эффективная энтропия и, соответственно, количество информации, будет равняться длине строки в случае, если вероятность ошибки равняется 0 или 1. В небинарных же алфавитах, при вероятности ошибки, равной 1, эффективная энтропия и количество информации будет равняться 0.

В таком случае, функция вычисления количества информации станет выглядеть следующим образом:

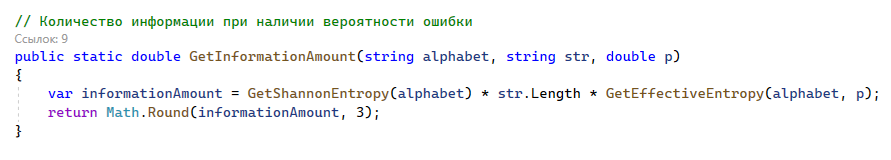


Рисунок 3.2 – Функция кол-ва информации при наличии ошибок

Вывод функции вычисления количества информации при вероятности ошибки, отличной от нулевой, представлен на рисунке 3.3.

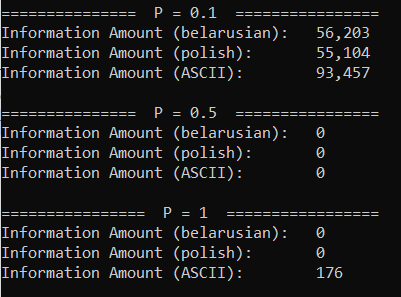


Рисунок 3.3 – Вывод кол-ва информации при *p>0*

**Вывод:** При вероятности ошибки *p=*1 количество информации в бинарном алфавите равняется количеству информации при *p=*0. Это связано с тем, что все биты исходного сообщения инвертируются, то есть заменяются на обратные, поэтому из такого сообщения можно получить исходную информацию. В небинарном же алфавите при *p=*1 определить исходное сообщение не представляется возможным ввиду того, что мощность алфавита больше 2, и каждый символ заменяется на некоторый неизвестный символ из алфавита, поэтому количество информации в таком случае равняется 0.

Ответы на вопросы

**1. Что такое алфавит источника сообщения?**

Алфавит источника сообщения – это конечная или бесконечная совокупность символов и знаков, формирующих сообщение.

**2. Что такое мощность алфавита источника сообщения?**

Мощность алфавита – это количество символов, составляющих алфавит.

**3. Какова мощность алфавита белорусского языка?**

32

**4. Какова мощность алфавита русского языка?**

33

**5. Какова мощность алфавита «компьютерного» языка?**

2

**6. Что такое энтропия алфавита?**

Энтропия алфавита – это информационная характеристика алфавита. Энтропия характеризует количество информации, приходящееся в среднем на один символ алфавита.

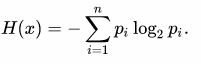
**7. Что такое энтропия сообщения?**

количество информации, приходящееся в среднем на один символ сообщения.

**8. От чего зависит энтропия алфавита?**

Зависит только от распределения вероятностей, а не от алфавита

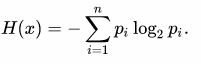
**9. Записать формулу для вычисления энтропии.**



**10. Что нужно знать для вычисления энтропии алфавита?**

Сумма всех элементов и вероятность появления каждого элемента

**11. Как рассчитываются энтропия Шеннона и энтропия Хартли? В чем принципиальное различие между этими характеристиками? Дайте толкование физического смысла энтропии.**

 (шенона)

где pi – вероятность i-ого события.

*, (хартли)*

где N – мощность алфавита.

Энтропия по Хартли рассчитывается с учетом того, что алфавит состоит из символов с равными вероятностями появления. Для расчета энтропии по Шеннону необходимо вычислить отдельно вероятность появления каждого символа алфавита.

**12. Пояснить назначение знака «минус» в формулах (2.1) и (2.4).**

**13. Что такое избыточность алфавита и избыточность сообщений, сформированных в компьютерных системах? Принцип действия каких систем основан на существовании данной избыточности?**

Избыточность – термин, означающий превышение количества информации, используемой для передачи или хранения сообщения, над его информационной энтропией. Возникает, когда энтропия Шеннона и энтропия Хартли не равны. Неполно используется алфавит.

**14. Расположить в порядке возрастания энтропии известные Вам алфавиты.**

**15. Вычислить энтропию алфавита белорусского (русского) языка.**

4,811

**16. Вычислить энтропию Шеннона бинарного алфавита, если вероятность появления в произвольном документе на основе этого алфавита одного из символов составляет 0,25, другого – 0,75; либо 0 и 1,0; либо 0,5 и 0,5.**

H = - (0,25 \* log2(0,25) + 0,75 \* log2(0,75))

H ≈ 0,811 бит.

H = - (0 \* log2(0) + 1,0 \* log2(1,0))

H = 0.

H = - (0,5 \* log2(0,5) + 0,5 \* log2(0,5))

H ≈ 1 бит.

**17. Чему равна энтропия алфавита по Хартли, если мощность этого алфавита равна: а) 1 символ; б) 2 символа; в) 8 символов?**

H = log2(1) = 0.

H = log2(2) = 1.

H = log2(8) = 3.