

ГУАП

КАФЕДРА № 42

ОТЧЕТ  
ЗАЩИЩЕН С ОЦЕНКОЙ  
ПРЕПОДАВАТЕЛЬ

доцент, канд. техн. наук  
\_\_\_\_\_  
должность, уч. степень, звание

\_\_\_\_\_  
подпись, дата

В. А. Миклуш  
\_\_\_\_\_  
инициалы, фамилия

ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ № 1

ВЫЧИСЛЕНИЕ СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК ТЕКСТОВОЙ  
ИНФОРМАЦИИ

по курсу:

ТЕОРИЯ ИНФОРМАЦИИ, ДАННЫЕ, ЗНАНИЯ

РАБОТУ ВЫПОЛНИЛ

СТУДЕНТ гр. № 4326

\_\_\_\_\_  
подпись, дата

Г. С. Томчук  
\_\_\_\_\_  
инициалы, фамилия

Санкт-Петербург 2025

## 1 Цель работы

Цель работы: анализ текстовой информации. Применение статистики для анализа текстов.

## 2 Краткое описание задания

1. Определить количество информации (по Хартли), содержащееся в заданном сообщении;
2. Построить таблицу распределения частот символов, характерных для заданного сообщения. Производится так называемая частотная селекция, текст сообщения анализируется как поток символов и высчитывается частота встречаемости каждого символа. Сравнить с имеющимися данными, в зависимости от языка сообщения;
3. На основании полученных данных определить среднее и полное количество информации, содержащееся в заданном сообщении;
4. Оценить избыточность сообщения.

## 3 Вариант задания

В таблице 1 представлен вариант задания.

Таблица 1 – Вариант задания

№ п/п	Алфавит	Текст
19.	Русский	В чужом глазу соринку видим, а в своём бревна не видим. Что хорошо для вторника, не всегда подходит для среды Повадился кувшин по воду ходить, не там ему голову сложить, а там ему полным быть

## 4 Ход работы

### 4.1 Количество информации по Хартли

Для определения количества информации по Хартли используем формулу:

$$I_H = L \cdot \log_2 m, \quad (1)$$

где  $L$  — длина сообщения (в символах, включая пробелы),  $m$  — мощность алфавита. Мощность алфавита (с учётом нормализации: нижний регистр, без знаков препинания, слияния «ё» и «е», «ъ» и «ь» и пробела как отдельного символа)  $m = 32$ . Длина сообщения  $L = 186$  символов. По формуле (1):

$$I_H = L \cdot \log_2 m = 186 \cdot \log_2 32 = 186 \cdot 5 = 930 \text{ бит.}$$

## 4.2 Таблица распределения частот символов

В таблице 2 приведены частоты встречаемости знаков алфавита для русского языка.

Таблица 2 – Частота знаков алфавита для русского языка

Буквы	Частота букв	Буквы	Частота букв	Буквы	Частота букв
пробел	0,145	к	0,029	ч	0,013
о	0,095	м	0,026	й	0,001
е	0,074	д	0,026	х	0,009
а	0,064	п	0,024	ж	0,008
и	0,064	у	0,021	ю	0,007
т	0,056	я	0,019	ш	0,006
н	0,056	ы	0,016	ц	0,004
с	0,047	з	0,015	щ	0,003
р	0,041	ь,Ъ	0,015	э	0,003
в	0,039	б	0,015	ф	0,002
л	0,036	г	0,014		

В таблице 3 представлена информация о распределении частот символов в сообщении по заданию после нормализации. Для расчетов и получения всех значений была написана программа на языке Python (см. Приложение А).

Таблица 3 – Частотная таблица символов

Символ	Кол-во	Доля (p)	%
пробел	34	0.1828	18.28%
о	18	0.0968	9.68%
в	12	0.0645	6.45%
д	11	0.0591	5.91%
и	11	0.0591	5.91%
а	9	0.0484	4.84%
е	9	0.0484	4.84%
м	9	0.0484	4.84%
н	8	0.043	4.30%
т	8	0.043	4.30%
у	8	0.043	4.30%
л	7	0.0376	3.76%
с	6	0.0323	3.23%
р	5	0.0269	2.69%

Символ	Кол-во	Доля (p)	%
п	4	0.0215	2.15%
г	3	0.0161	1.61%
к	3	0.0161	1.61%
х	3	0.0161	1.61%
ы	3	0.0161	1.61%
ь	3	0.0161	1.61%
я	3	0.0161	1.61%
б	2	0.0108	1.08%
ж	2	0.0108	1.08%
ч	2	0.0108	1.08%
ш	2	0.0108	1.08%
з	1	0.0054	0.54%

На рисунке 1 изображена гистограмма вероятностей появления символов по эмпирическим (таблица 3) и теоретическим (таблица 2) данным.

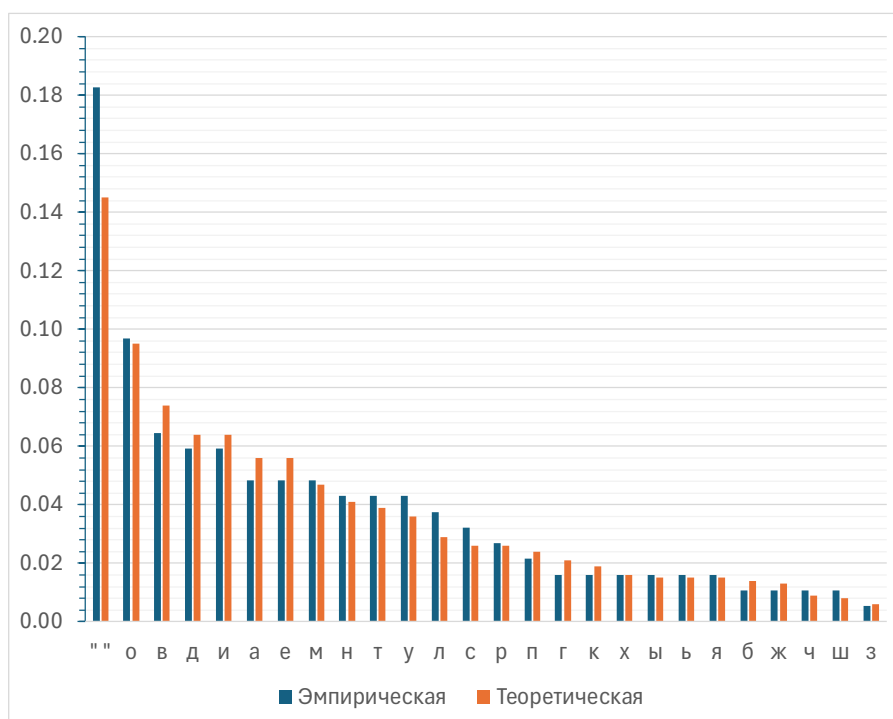


Рисунок 1 — Гистограмма вероятностей появления символов

Сравним экспериментальные данные с теоретическими:

- «о»: 9.68% (экспер.) и 9.5% (теоретич.) — близко;
- «в»: 6.45% и 3.9% — выше среднего;
- «е»: 4.84% и 7.4% — ниже среднего;
- «т»: 4.3% и 5.6% — близко;

- пробел: 18.28% и 14.5% — заметно выше, так как текст состоит из коротких пословиц.

Как видно из сравнения, эмпирические частоты в целом близки к теоретическим, но есть отклонения, обусловленные ограниченным объёмом текста и особенностями стиля.

### 4.3 Среднее и полное количество информации

Эмпирическая энтропия на символ:

$$H = -\sum p_i \log_2 p_i \approx 4.2271 \text{ бит/символ.}$$

Полное количество информации:

$$I = H \cdot L = 4.2271 \cdot 186 \approx 786.24 \text{ бит.}$$

Итого, среднее количество информации на символ — 4.23 бит, полное количество информации в сообщении — 786.24 бит.

### 4.4 Избыточность сообщения

Избыточность определяется как:

$$D = 1 - \frac{H}{H_0}, \quad (2)$$

где энтропия оптимальных сообщений  $H_0 = \log_2 m = 5$  бит/символ.

Подставим значения в (2):

$$D = 1 - \frac{4.2271}{5} \approx 0.1546 \approx 15.5\%.$$

Итого, избыточность сообщения составляет 15.5%.

## 5 Выводы

В ходе выполнения лабораторной работы было исследовано сообщение по варианту № 19.

1. По методу Хартли количество информации составило 930 бит, что соответствует максимально возможному значению при равновероятном распределении символов в алфавите.
2. Построена таблица частотного распределения символов сообщения. Сравнение с табличными данными показало близость эмпирических частот к усреднённым для русского языка, но выявлены отличия: в частности, в тексте повышена доля пробелов и буквы «в», а частота

буквы «е» оказалась ниже. Эти отклонения объясняются ограниченным объёмом выборки и спецификой текста (пословицы).

3. Среднее количество информации на символ составило 4.23 бит, а полное количество информации в сообщении — 786.24 бит, что меньше значения по Хартли из-за неравномерного распределения символов.
4. Избыточность сообщения оценена в 15.5%, что подтверждает наличие статистической структуры и характерных закономерностей естественного языка.

Сообщение содержит меньше информации, чем максимально возможное значение по Хартли, что связано с неравномерным распределением символов и особенностями языка. Это подтверждает теоретические положения о наличии избыточности в текстах на естественных языках.

## **СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Миклуш, В.А. Основы теории информации: Учебно-методическое пособие / В.А. Миклуш, В.А. Ушаков. — СПб: ГУАП, 2024. — 41 с.
2. Шеннон, К. Э. Работы по теории информации и кибернетике / К. Э. Шеннон. — М.: Иностранная литература, 1963. — 830 с.
3. Колодуб, В. Д. Теория информации: учебник для вузов / В. Д. Колодуб. — М.: ФОРУМ, 2019. — 352 с.
4. Ковалёв, В. А. Теория информации и кодирование: учебное пособие / В. А. Ковалёв. — СПб.: Питер, 2020. — 368 с.

## ПРИЛОЖЕНИЕ А

```
import collections
from math import log2

text = "В чужом глазу соринку видим а в своем бревна не видим Что хорошо для вторника
не всегда подходит для среды Повалился кувшин по воду ходить не там ему голову
сложить а там ему полным быть"

char_counts = collections.Counter(text.lower())
total_char = len(text)
entropy = -sum(
    (char_counts[ch] / total_char) * log2(char_counts[ch] / total_char)
    for ch in char_counts
)
m = 32
H0 = log2(m)
total_info_hartley = total_char * H0
total_info = entropy * total_char
redundancy = 1 - (entropy / H0)

char_counts = sorted(char_counts.items(), key=lambda x: (-x[1], x[0]))
char_stats = list(
    map(lambda x: (x[0], x[1], x[1] / total_char, x[1] / total_char * 100),
    char_counts)
)

for char, n, p, percent in char_stats:
    print(f"{char} {n} {p:.4f} {percent:.2f}%")

for x in [total_char, entropy, H0, total_info_hartley, total_info, redundancy]:
    print(x)
```