ГУАП

КАФЕДРА № 42

ОТЧЕТ   
ЗАЩИЩЕН С ОЦЕНКОЙ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ПРЕПОДАВАТЕЛЬ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Доцент, канд. техн. наук |  |  |  | В.А. Миклуш |
| должность, уч. степень, звание |  | подпись, дата |  | инициалы, фамилия |

|  |
| --- |
| ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ №1 |
| Вычисление статистических характеристик текстовой информации |
| по курсу: Теория информации, данные, знания |
|  |

РАБОТУ ВЫПОЛНИЛ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| СТУДЕНТ ГР. № | 4329 |  |  |  | Д.С. Шаповалова |
|  |  |  | подпись, дата |  | инициалы, фамилия |

Санкт-Петербург 2025

# 1. Цель работы:

Анализ текстовой информации. Применение статистики для анализа текстов.

# 2. Задание:

Постановка задачи.

1. Определить количество информации (по Хартли), содержащееся в заданном сообщении;

2. Построить таблицу распределения частот символов, характерных для заданого сообщения. Производится так называемая частотная селекция, текст сообщения анализируется как поток символов и высчитывается частота встречаемости каждого символа.

3. На основании полученных данных определить среднее и полное количество информации, содержащееся в заданном сообщении;

4. Оценить избыточность сообщения.

# 3. Исходные данные

Исходный текст выбран под вариантом 17 и представлен на итальянском языке: «Chi ha i denti non ha il pane e chi ha il pane non ha i denti Chi tante male azioni fa, una grossa ne aspetta Dare a Cesare quel che è di Cesare, dare a Dio quel che è di Di»

# 4. Теоретические сведения:

В основе работы лежат понятия из теории информации, разработанной Клодом Шенноном.

**1. Количество информации по Хартли (для равновероятных событий)**

I — количество информации в сообщении (в битах).

n — количество символов в сообщении.

m — мощность алфавита (общее количество различных символов, которые могут появиться в сообщении).

Результат показывает, сколько бит информации несет один символ из данного алфавита, если все символы равновероятны.

Смысл: Эта формула работает в "идеальном" случае, когда никакие символы не имеют преимущества перед другими. Она отвечает на вопрос: "Сколько информации мы получили, узнав, что произошло одно из m равновероятных событий n раз подряд?".

**2. Энтропия Шеннона (для неравновероятных событий)**

H — энтропия (среднее количество информации, приходящееся на один символ алфавита, в битах). Это более реальный показатель, чем формула Хартли, так как он учитывает разную частоту символов.

 — вероятность появления i-го символа в сообщении. На практике она вычисляется как частота:  = (количество раз, когда встретился символ i) / n.

Σ — Нужно просуммировать выражение  для всех уникальных символов алфавита, встречающихся в сообщении.

Минус перед суммой нужен, чтобы результат был положительным, так как log₂() для вероятностей (меньших 1) всегда отрицателен.

Смысл: Энтропия измеряет "степень неопределенности" или "информационную насыщенность" источника данных. Чем выше энтропия, тем больше информации несет каждый символ. Максимальна она тогда, когда все символы равновероятны (и тогда H совпадает с log₂(m)).

**3. Максимальная энтропия**

 — максимально возможная энтропия для алфавита с мощностью m. Это частный случай формулы H, когда все  равны (т.е. ).

**4. Полное количество информации в сообщении (по Шеннону)**

I\_общ — общее количество информации во всем сообщении с учетом реального распределения частот символов.

H — энтропия, рассчитанная по формуле 2.

n — длина сообщения.

**5. Избыточность алфавита / сообщения**

D — избыточность (безразмерная величина, обычно выражается в процентах).

Смысл: Избыточность показывает, какая часть "информационной емкости" алфавита не используется из-за неравномерного распределения символов. Высокая избыточность (например, в естественных языках, где 40-50%) связана с наличием правил (грамматика, синтаксис) и частотных закономерностей. Именно из-за избыточности возможны сжатие данных и исправление ошибок.

4. Ход работы:

Для выполнения работы был выбран Excell. По формулам (1) – (4) были посчитаны: количество информации, энтропия для каждого символа, суммарная энтропия, максимальная энтропия (обычная), полное количество информации, а также определена мощность алфавита. Результаты расчетов представлены на рисунке 1.

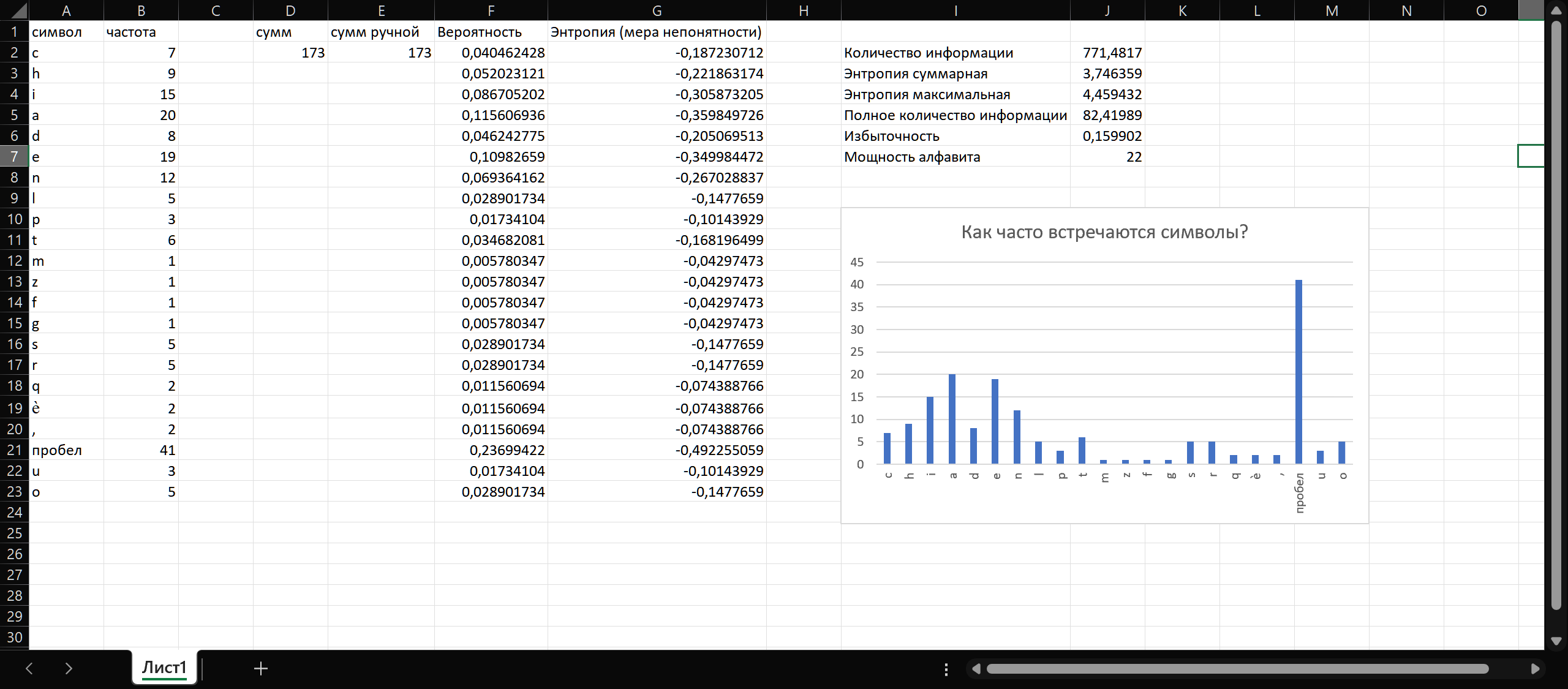


Рисунок 1 – Выполнение работы в Excel

# 5. Вывод:

В ходе выполнения лабораторной работы были успешно решены поставленные задачи по статистическому анализу текстовой информации. На примере заданного текста на итальянском языке были проведены следующие действия:

Определено количество информации по Хартли для алфавита, используемого в сообщении. Это значение представляет собой максимально возможную энтропию для системы с данным количеством равновероятных символов.

Построена таблица распределения частот символов для конкретного текста. Проведенный частотный анализ показал, что распределение вероятностей появления букв в анализируемом сообщении является неравномерным, что характерно для естественных языков. Можем заметить, что наиболее лидирует символ «пробел», что говорит о важности пауз в речи.

Рассчитано среднее и полное количество информации (энтропия) в сообщении. Реальная энтропия оказалась существенно ниже максимальной (по Хартли), что свидетельствует о наличии статистических связей между символами и их неравновероятном появлении.

Оценена избыточность сообщения. Расчет показал значительную избыточность текста, которая является типичной для естественного языка. Высокая избыточность объясняется наличием устойчивых статистических закономерностей (частот букв, биграмм и т.д.) и необходима для обеспечения помехоустойчивости и надежности восприятия информации.

Таким образом, работа подтвердила ключевые теоретические положения теории информации: реальные тексты обладают значительной избыточностью из-за неравномерного распределения вероятностей символов и статистических связей между ними. Это свойство естественных языков, обеспечивающее их устойчивость к ошибкам и помехам при передаче и хранении.