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 заданного сообщения. Производится так называемая частотная селекция, текст сообщения анализируется как поток символов и высчитывается частота встречаемости каждого символа. Сравнить с частоты символов с эталонной, в зависимости от языка сообщения (таблица 1);

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»

Эталонная частота для европейских языков:

Таблица 1 – Таблица частот букв европейских языков

| **Буква алфавита** | **Французский язык** | **Немецкий**  **язык** | **Английский язык** | **Испанский**  **язык** | **Итальянский язык** |
| --- | --- | --- | --- | --- | --- |
| **A** | 7,68 | 5,52 | 7,96 | 12,90 | 11,12 |
| **B** | 0,80 | 1,56 | 1,60 | 1,03 | 1,07 |
| **C** | 3,32 | 2,94 | 2,84 | 4,42 | 4,11 |
| **D** | 3,60 | 4,91 | 4,01 | 4,67 | 3,54 |
| **E** | 17,76 | 19,18 | 12,86 | 14,15 | 11,63 |
| **F** | 1,06 | 1,96 | 2,62 | 0,70 | 1,15 |
| **G** | 1,10 | 3,60 | 1,99 | 1,00 | 1,73 |
| **H** | 0,64 | 5,02 | 5,39 | 0,91 | 0,83 |
| **I** | 7,23 | 8,21 | 7,77 | 7,01 | 12,04 |
| **J** | 0,19 | 0,16 | 0,16 | 0,24 | – |
| **K** | – | 1,33 | 0,41 | – | – |
| **L** | 5,89 | 3,48 | 3,51 | 5,52 | 5,95 |
| **M** | 2,72 | 1,69 | 2,43 | 2,55 | 2,65 |
| **N** | 7,61 | 10,20 | 7,51 | 6,20 | 7,68 |
| **O** | 5,34 | 2,14 | 6,62 | 8,84 | 8,92 |
| **P** | 3,24 | 0,54 | 1,81 | 3,26 | 2,66 |
| **Q** | 1,34 | 0,01 | 0,17 | 1,55 | 0,48 |
| **R** | 6,81 | 7,01 | 6,83 | 6,95 | 6,56 |
| **S** | 8,23 | 7,07 | 6,62 | 7,64 | 4,81 |
| **T** | 7,30 | 5,86 | 9,72 | 4,36 | 7,07 |
| **U** | 6,05 | 4,22 | 2,48 | 4,00 | 3,09 |
| **V** | 1,27 | 0,84 | 1,15 | 0,67 | 1,67 |
| **W** | – | 1,38 | 1,80 | – | – |
| **X** | 0,54 | – | 0,17 | 0,07 | – |
| **Y** | 0,21 | – | 1,52 | 1,05 | – |
| **Z** | 0,07 | 1,17 | 0,05 | 0,31 | 1,24 |

# **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 не равновероятных 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*

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

H — энтропия, рассчитанная по формуле 3.

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.

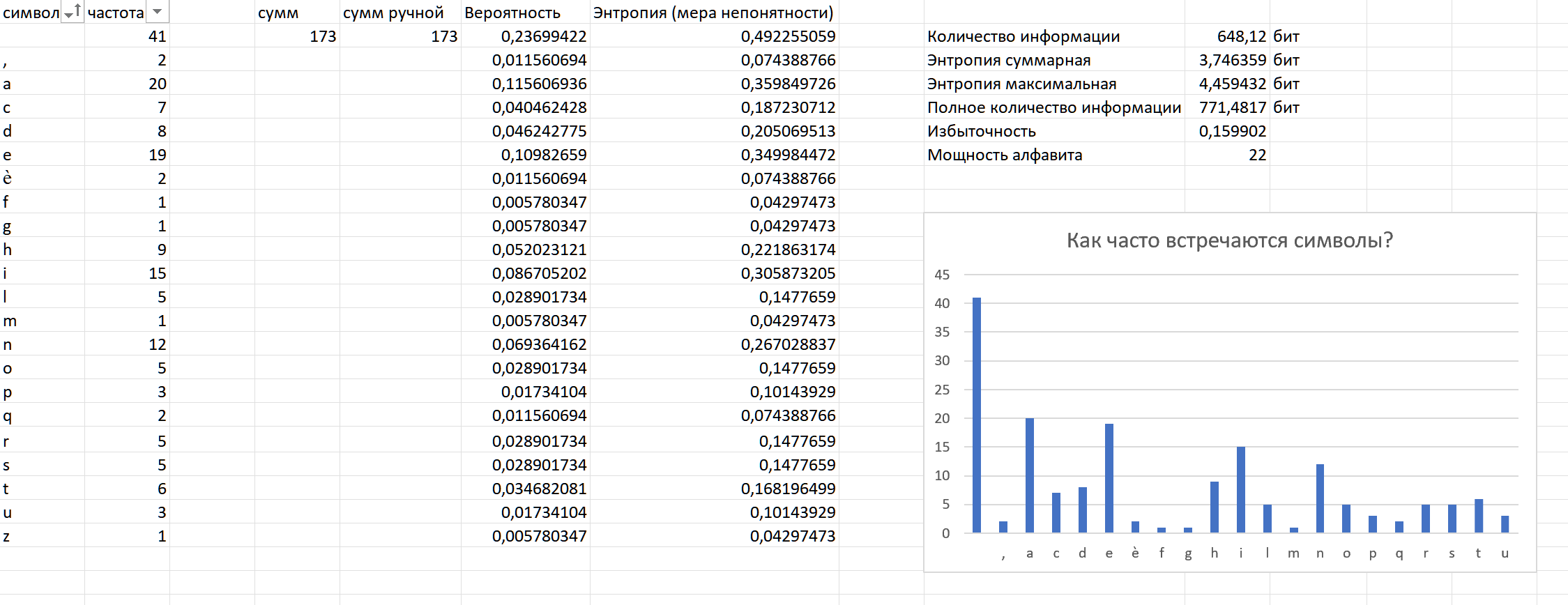


Рисунок 1 – Выполнение работы в Excel

Получив график частот появления символов, сравним их с данными из таблицы 1, построив такой же график.

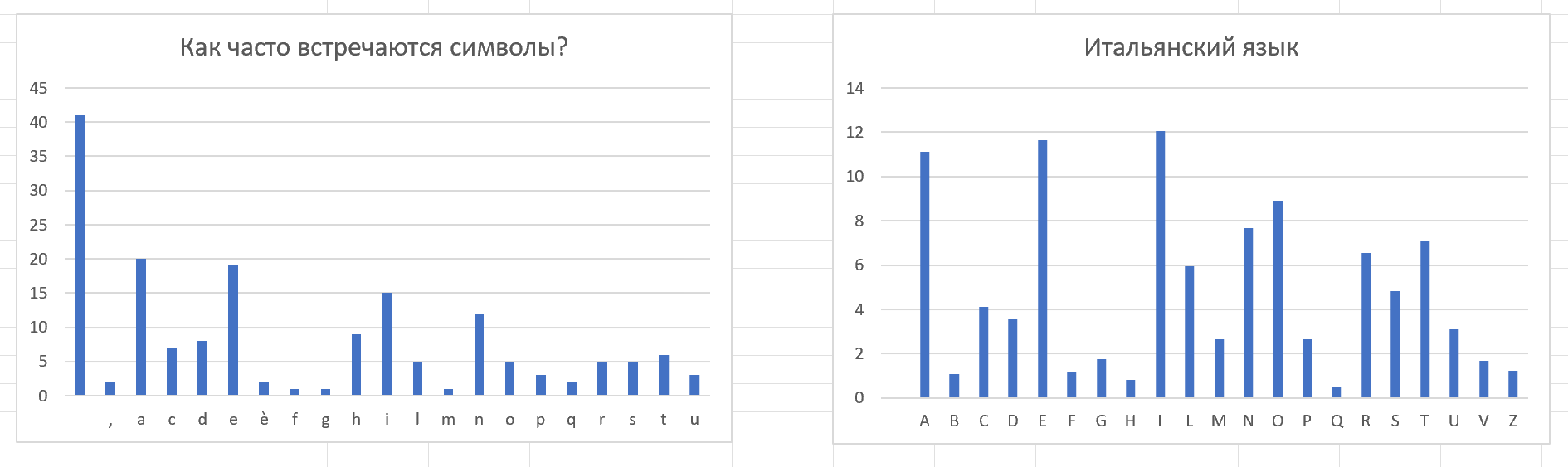


Рисунок 2 – Сравнение полученной статистики и эталонной

Как мы можем видеть, если не учитывать символы пробел и запятую, некоторые буквы у нас в тексте отсутствуют (b, v, z), некоторые буквы отсутствуют в эталонной статистике в силу специфики итальянского языка (è), у некоторых частоты отличаются (c и d, e и a, l и h, n и o), а остальные примерно совпадают. Расхождение можно объяснить малым объёмом анализируемого текста.

# **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:

1. Определено количество информации по Хартли для алфавита, используемого в сообщении – 4,459431619 бит. Это значение представляет собой максимально возможную энтропию для системы с данным количеством равновероятных символов.

2. 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. Проведя сравнение с таблицей частоты букв в итальянском языке, можем понять, что в основном данные совпадают, но есть расхождение в некоторых буквах, некоторые же в полученной нами статистике вообще отсутствуют, что можно объяснить малым объёмом анализируемого текста.

3. Рассчитано среднее и полное количество информации в сообщении – 648,1200362 бит и 771,48167 бит. Реальная энтропия (3,746358591 бит) оказалась ниже максимальной (4,459431619 бит) (по Хартли), что свидетельствует о наличии статистических связей между символами и их не равновероятном появлении.

4. Оценена избыточность сообщения. Расчет показал значительную избыточность – 0,159902223 (доля) – 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 надёжности 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.