ГУАП

КАФЕДРА № 42

ОТЧЕТ   
ЗАЩИЩЕН С ОЦЕНКОЙ

ПРЕПОДАВАТЕЛЬ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| канд. техн. наук, доцент |  |  | |  | 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. № | 4326 |  |  |  | Г. С. Томчук |
|  |  |  | подпись, дата |  | инициалы, фамилия |

Санкт-Петербург 2025

СОДЕРЖАНИЕ

[1 Цель работы 3](#_Toc208919365)

[2 Краткое описание задания 3](#_Toc208919366)

[3 Вариант задания 3](#_Toc208919367)

[4 Ход работы 3](#_Toc208919368)

[4.1 Количество информации по Хартли 3](#_Toc208919369)

[4.2 Таблица распределения частот символов 4](#_Toc208919370)

[4.3 Среднее и полное количество информации 5](#_Toc208919371)

[4.4 Избыточность сообщения 5](#_Toc208919372)

[5 Выводы 5](#_Toc208919373)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 7](#_Toc208919374)

1. Цель работы

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

1. Краткое описание задания
2. Определить количество информации (по Хартли), содержащееся в заданном сообщении;
3. 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. Сравнить с имеющимися данными, в зависимости от языка сообщения;
4. На основании полученных данных определить среднее и полное количество информации, содержащееся в заданном сообщении;
5. Оценить избыточность сообщения.
6. Вариант задания

В таблице 1 представлен вариант выполненного задания.

Таблица – 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 |

1. Ход работы
   1. Количество информации по Хартли

Для определения количества информации по Хартли используем формулу:

,

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) Длина сообщения символов. По формуле Хартли:

* 1. Таблица распределения частот символов

В таблице 2 представлена информация о распределении частот символов в сообщении после нормализации.

Таблица – 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 | 0.0430 | 4.30% |
| т | 8 | 0.0430 | 4.30% |
| с | 6 | 0.0323 | 3.23% |
| л | 6 | 0.0323 | 3.23% |
| р | 4 | 0.0215 | 2.15% |
| х | 4 | 0.0215 | 2.15% |
| б | 3 | 0.0161 | 1.61% |
| п | 3 | 0.0161 | 1.61% |
| у | 3 | 0.0161 | 1.61% |
| г | 2 | 0.0108 | 1.08% |
| к | 2 | 0.0108 | 1.08% |
| ч | 1 | 0.0054 | 0.54% |
| ж | 1 | 0.0054 | 0.54% |
| з | 1 | 0.0054 | 0.54% |
| й | 1 | 0.0054 | 0.54% |
| ш | 1 | 0.0054 | 0.54% |
| ы | 1 | 0.0054 | 0.54% |
| я | 1 | 0.0054 | 0.54% |

Сравним полученные данные с табличными:

* «о»: 9.68% (полученное) и 9.5% (табличное) — близко;
* «в»: 6.45% и 3.9% — выше среднего;
* «е»: 4.84% и 7.4% — ниже среднего;
* 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 табличным, но есть отклонения, обусловленные ограниченным объёмом текста и особенностями стиля.

* 1. Среднее и полное количество информации

Эмпирическая энтропия на символ:

.

Полное количество информации:

.

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

* 1. Избыточность сообщения

Избыточность определяется как:

,

где бит/символ.

.

Итого, избыточность сообщения составляет 15,5%.

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