**Лабораторная №1: Количественная мера информации**

**Цель работы: экспериментальное изучение количественных аспектов информации**

**Лабораторное задание:**

**1. Определить количество информации, содержащееся в заданном сообщении, при условии, что значениями являются буквы кириллицы. Количество информации необходимо определять по формуле Хартли и Шеннона.**

Сообщение: Горовой Юрий Михайлович участвовал в ежегодной научной конференции студентов, стремящихся стать высококлассными профессионалами в области информационных технологий.

**Комментарий:**

Определим на данном этапе длину сообщения (с учётом пробелов, но без учёта знаков пунктуации). Букву «Ё» считать за «Е».

Длину быстро сосчитать с помощью текстового редактора (например, в Word) – выделить текст сообщения и внизу экрана посмотреть, сколько символов есть в тексте.

*В нашем случае длина сообщения n = 162.*

**2. Построить таблицу распределения частот символов, характерные 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 таблице 1.**

Таблица 1 – Частоты символов (заданные)

|  |  |  |  |
| --- | --- | --- | --- |
| - (пробел)              0.175 | О             0.090 | Е, Ё         0.072 | А            0.062 |
| И             0.062 | Т             0.053 | Н             0.053 | С             0.045 |
| Р             0.040 | В             0.038 | Л             0.035 | К             0.028 |
| М            0.026 | Д             0.025 | П             0.023 | У             0.021 |
| Я             0.018 | Ы            0.016 | З             0.016 | Ь, Ъ        0.014 |
| Б             0.014 | Г             0.013 | Ч             0.012 | Й             0.010 |
| Х             0.009 | Ж            0.007 | Ю           0.006 | Ш            0.006 |
| Ц             0.004 | Щ            0.003 | Э             0.003 | Ф            0.002 |

Составим таблицу 2 с получившимися частотами символов.

Чтобы найти частоту встречаемости символа, необходимо сосчитать, сколько раз встречается данный символ. Затем полученное значение разделить на общее количество символов в сообщении (n), которое мы определили на предыдущем шаге.

|  |  |  |
| --- | --- | --- |
| Символ | Количество | **Частота** |
| А | 10 | 0,062 |
| Б | 1 | 0,006 |
| В | 8 | 0,049 |
| Г | 3 | 0,018 |
| Д | 2 | 0,012 |
| Е, Ё | 8 | 0,049 |
| Ж | 1 | 0,006 |
| З | 0 | 0,000 |
| И | 13 | 0,080 |
| Й | 6 | 0,037 |
| К | 3 | 0,018 |
| Л | 6 | 0,037 |
| М | 5 | 0,031 |
| Н | 12 | 0,074 |
| О | 19 | 0,117 |
| П | 1 | 0,006 |
| Р | 6 | 0,037 |
| С | 11 | 0,068 |
| Т | 8 | 0,049 |
| У | 3 | 0,018 |
| Ф | 3 | 0,018 |
| Х | 4 | 0,025 |
| Ц | 2 | 0,012 |
| Ч | 3 | 0,018 |
| Ш | 0 | 0 |
| Щ | 1 | 0,006 |
| Ъ, Ь | 1 | 0,006 |
| Ы | 3 | 0,0018 |
| Э | 0 | 0 |
| Ю | 1 | 0,006 |
| Я | 2 | 0,012 |
| Пробел | 14 | 0,086 |

Далее сравниваем получившиеся частоты с заданными. Делаем вывод на основании сравнения. В качестве вывода можно указать, что рассматривался слишком маленький текст. К тому же в разных текстах (например, обзор какого-то продукта, техническая информация о какой-нибудь установке на предприятии) может сильно отличаться частотой некоторых символов (букв).

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

Для начала найдём количество информации по Хартли:

I=log2N = n\*log2 (m)

m – количество букв в алфавите.

n – количество букв в сообщении.

N – возможное количество различных сообщений.

n = 162

m = 32

I = 162\*log2(32) = 162\*5 = 810

Найдём среднее количество информации в сообщении.

Среднее количество информации в сообщении в нашем случае – энтропия.

Она вычисляется по следующей формуле

В нашем случае H = 4,238.

Далее определим полное количество информации по формуле Шеннона.

kв данном случае – количество символов в сообщении.

Часть с суммой мы высчитывали во время расчёта энтропии. В итоге получаем:

I = 162 \* 4,238 = 686, 556

Знак минус мы убрали из формулы, так как мы использовали его при расчёте энтропии. Минус нужен для того, чтобы получился положительный результат (отрицательное число может дать логарифм).

**4. Оценить избыточность сообщения.**

Вычислим избыточность информациипо формуле:

D = 1 – H0/Hmax

H0 – энтропия, которую мы нашли ранее.

Hmax – максимальная энтропия.

Она вычисляется по следующей формуле:

Hmax = log2(m), где m – количество символов в алфавите.

Hmax = log2(32) = 5

Находим избыточность.

D = 1 – 4,238/5 = 0,1524

Далее необходимо оценить избыточность. Например, у русского и европейского языков избыточность в общении составляет примерно 0,5.

Избыточность при передаче сообщений имеет свои положительные и отрицательные стороны. Увеличение избыточности приводит к увеличению времени передачи сообщений, излишней загрузке каналов связи. За определенный промежуток времени по каналу передается меньшее количество информации, чем это возможно; поэтому одной из задач теории информации и техники кодирования является задача сокращения избыточности.

**ВАШ ТЕКСТ:** \*Ваше ФИО\* участвовал(а) в семьдесят второй Всероссийской научно-технической конференции студентов, магистрантов и аспирантов с международным участием, посвященной семидесяти пятилетию ЯГТУ.

**Возможные вопросы на защиту лабораторной:**

1. Что такое количество информации? В чём измеряется количество информации?

2. В чём отличие между подсчётом количества информации по методу Хартли и по методу Шеннона? Когда применяются эти методы?

3. Что такое среднее и полное количество информации?

4. Что такое избыточность источника информации?

**Возможные практические задания на защиту лабораторной:**

1. Бросаются одновременно две игральные кости. Определить количество информации, содержащееся в сообщении о том, что произведение чисел выпавших очков четно

Могут быть другие несложные задачи на поиск количества информации. Например, определить количество информации в ответе человека, которого спросили про число в дате его рождения, задачи с кубиками и т.п.