

Занятие №1.

Информация. Измерение количества информации.

Теоретическая часть

Для определения количества информации, содержащейся в сообщении длины n , используют **формулу Хартли**.

Имеется алфавит A , из букв которого составляется сообщение:

$$|A| = m$$

Количество возможных вариантов разных сообщений:

$$N = m^n$$

Где: N - возможное количество различных сообщений, шт; m - количество букв в алфавите, шт; n - количество букв в сообщении, шт.

Пример: Алфавит состоит из 2-х букв B и X , длина сообщения 3 буквы - таким образом $m=2$, $n=3$. При выбранных нами алфавите и длине сообщения можно составить $N=m^n=2^3=8$ разных сообщений "BBB", "BBX", "BXB", "BXX", "XBB", "XBX", "XXB", "XXX" - других вариантов нет.

Формула Хартли определяется:

$$I = \log_2 N = n \log_2 m.$$

Где: I - количество информации, бит (\log – логарифм, формула читается: логарифм эн по основанию два).

При равновероятности символов $p=1/m$, $m=1/p$ формула Хартли выглядит так:

$$I = \log_2 N = \log_2 (1 / p) = - \log_2 p.$$

Формула Хартли была предложена Ральфом Хартли в 1928 году как один из научных подходов к оценке сообщений. Допустим, нам требуется что-либо найти или определить в той или иной системе. Есть такой способ поиска как «деление пополам». Например, кто-то загадывает число от 1 до 100, а другой должен отгадать его, получая лишь ответы «да» или «нет». Задается вопрос: число меньше? Ответ и «да» и «нет» сократит область поиска вдвое. Далее по той же схеме диапазон снова делится пополам. В конечном итоге, загаданное число будет найдено.

Посчитаем сколько вопросов надо задать, чтобы найти задуманное число.

Допустим загаданное число 27. Начали:

Больше 50? Нет

Больше 25? Да

Больше 38? Нет

Меньше 32? Да

Меньше 29? Да

Больше 27? Нет

Это число 26? Нет

Ура! если число не 26 и не больше 27, то это явно 27.

Чтобы угадать методом «деления пополам» число от 1 до 100 нам потребовалось 7 вопросов. Кто-то может задаться вопросом: а почему именно так надо задавать вопросы? Ведь, например, можно просто спрашивать: это число 1? Это число 2? И т.д. Но тогда вам потребуется намного больше вопросов (возможность того, что вы телепат, и угадаете с первого раза не рассматривается). «Деление пополам» самый короткий рациональный способ найти число. Объем информации заложенный в ответ «да» или «нет» равен одному биту. Действительно,

ведь бит может быть в состоянии 1 или 0. Итак, для угадывания числа от 1 до 100 нам потребовалось семь бит (семь ответов «да» - «нет»).

$$N = 2^i$$

Такой формулой можно представить, сколько вопросов (бит информации) потребуется, чтобы определить одно из возможных значений. N – это количество значений, а i – количество бит. Например, в нашем примере 100 меньше чем 27, однако больше, чем 26. Да, нам могло потребоваться и всего 6 вопросов, если бы загаданное число было бы 28.

Задача: При бросании игральной кости, выпадает одно из шести чисел (1-6). Определить сколько бит информации содержит сообщение о том, что выпало число 5.

Решение:

Такое сообщение содержит $I = \log_2 6 = 2,5849625$ бита информации.

Но не все ситуации имеют одинаковые вероятности реализации. Существует много таких ситуаций, у которых вероятности реализации различаются. Например, если бросают несимметричную монету, пирамидку или "правило бутерброда".

В 1948 г. американский инженер и математик Клод Шеннон предложил формулу для вычисления количества информации для событий с различными вероятностями.

Если I - количество информации,

N - количество возможных событий,

p_i - вероятности отдельных событий,

то количество информации для событий с различными вероятностями можно определить по формуле:

$$I = -(p_1 \log_2 p_1 + p_2 \log_2 p_2 + \dots + p_k \log_2 p_k) = - \sum p_i \log_2 p_i,$$

где i принимает значения от 1 до N (Σ – значок суммы).

Формулу Хартли теперь можно рассматривать как частный случай формулы Шеннона:

$$I = - \sum 1 / N \log_2 (1 / N) = \log_2 N.$$

При равновероятных событиях получаемое количество информации максимально.

Задача: Определить количество информации, получаемое при реализации одного из событий, если бросают

а) несимметричную четырехгранную пирамидку;

б) симметричную и однородную четырехгранную пирамидку.

Решение:

а) Будем бросать несимметричную четырехгранную пирамидку.

Вероятность отдельных событий будет такова:

$$p_1 = 1 / 2,$$

$$p_2 = 1 / 4,$$

$$p_3 = 1 / 8,$$

$$p_4 = 1 / 8,$$

тогда количество информации, получаемой после реализации одного из этих событий, рассчитывается по формуле:

$$I = -(1/2 \log_2 1/2 + 1/4 \log_2 1/4 + 1/8 \log_2 1/8 + 1/8 \log_2 1/8) = 1/2 + 2/4 + 3/8 + 3/8 = 14/8 = 1,75 \text{ (бит)}.$$

б) Теперь рассчитаем количество информации, которое получится при бросании симметричной и однородной четырехгранной пирамидки: $I = \log_2 4 = 2 \text{ (бит)}$.

При передаче и хранении информации с помощью различных технических устройств информацию следует рассматривать как последовательность знаков (цифр, букв, кодов цветов точек изображения), не рассматривая ее содержание.

Считая, что алфавит (набор символов знаковой системы) - это событие, то появление одного из символов в сообщении можно рассматривать как одно из состояний события. Если появление символов равновероятно, то можно рассчитать, сколько бит информации несет каждый символ. Информационная емкость знаков определяется их количеством в алфавите. Чем из большего количества символов состоит алфавит, тем большее количество информации несет один знак. Полное число символов алфавита принято называть мощностью алфавита.

Молекулы ДНК (дезоксирибонуклеиновой кислоты) состоят из четырех различных составляющих (нуклеотидов), которые образуют генетический алфавит. Информационная емкость знака этого алфавита составляет:

$$4 = 2^i, \text{ т.е. } i = 2 \text{ бит}.$$

Каждая буква русского алфавита (если считать, что ё=е) несет информацию 5 бит ($32 = 2^5$).

При таком подходе в результате сообщения о результате бросания кубика, получим различное количество информации, Чтобы его подсчитать, нужно умножить количество символов на количество информации, которое несет один символ.

Количество информации, которое содержит сообщение, закодированное с помощью знаковой системы, равно количеству информации, которое несет один знак, умноженному на число знаков в сообщении.

По умолчанию (если в задаче не указано специально) при решении задачи указывается 256-символьный алфавит – таблица ASCII (мощность алфавита = 256 символов). Значит, на один символ (букву, цифру, знак, знак препинания, пробел) приходится 8 бит информации или 1 байт.

Задача: Определить количество информации, которое содержится на печатном листе бумаги (двусторонняя печать), если на одной стороне уместается 40 строк по 67 символов в строке.

Решение:

Определим количество символов на одной стороне листа:

$$40 \text{ строк} * 67 \text{ символов} = 2680 \text{ символов}$$

Определим количество символов на 2-х сторонах листа:

$$2680 \text{ символов} * 2 = 5360 \text{ символов}$$

$$\text{Количество информации} = 5360 \text{ символов} * 1 \text{ байт} = 5360 \text{ байт}$$

$$\text{Переводим в Кб: } 5360 \text{ байт} : 1024 \text{ байт} = 5,23 \text{ Кб}$$

Если бы необходимо было получить ответ в битах, то количество информации = 5360 символов * 8 бит = 42880 бит.

Практическая часть

1. Вероятность первого события составляет 0,5, а второго и третьего 0,25. Какое количество информации мы получим после реализации одного из них?
2. Какое количество информации будет получено при игре в рулетку с 32-мя секторами?
3. Какое количество символов содержится на странице энциклопедического словаря, если в памяти компьютера эта страница занимает 13 Кб?
4. Сколько символов содержится в алфавите, при помощи которого написана книга из 10 страниц, на каждой из которых содержится 32 строк по 40 символов и занимает в памяти компьютера вся книга 12,5 Кб.

Решения задач жду по адресу: yaro4kin72@yandex.ru (в письме указать Ф.И. и класс).