Entropy in information theory is a quantity used to describe the information density of a sequence of characters (or words). Although it is related to entropy, which is known from physics, chemistry and thermodynamics, the latter differs from the entropy of information theory due to the Boltzmann constant and the natural logarithm. Further consideration of this relationship would go too far here. The formulas for the entropy of information theory go back to C.E. Shannon (C. E. Shannon: *A Mathematical Theory of Communication*. In: *Bell System Technical Journal*. Vol. 27, No. 3, 1948, pp. 379-423, doi:10.1002/j.1538-7305.1948.tb01338.x)

A derivation for the entropy of a discrete character string can be found at https://de.wikipedia.org/wiki/Entropie:

There you will find, among other things, the basic formula for calculating the entropy of a character string of length N, where each character has the proportion of occurrence (=relative frequency) p_i , namely:

Entropy = $\sum [p(i) \times \log2(1/p(i))]$ or

Entropy= $- \sum [p(i) \times \log2p(i)]$

For our application with the usernames, N is the number of characteristic values of the username.

The logarithm to the base 2 ($\log_2$ or log2) is - as can be seen above - chosen arbitrarily and is based on the application with binary numbers.

If Log2 is used, entropy values greater than 1 occur for longer character strings (Length>2). Therefore, a normalization must take place (log2/log2(N)).

A practical and application-oriented explanation of entropy in statistics can be found at https://welt-der-bwl.de/Entropie?ssp=1&setlang=de&cc=DE&safesearch=moderate, which forms the basis for the following plausibility considerations:

Entropy definition

Statistical entropy as a measure of statistical distribution can be applied to nominally scaled data, as it does not refer to the characteristic values themselves, but only to the relative frequencies of the characteristic values.

The entropy is usually converted into a normalized entropy, which can only assume values between 0 and 1. Values close to 1 then represent a broad distribution of the data; with a normalized entropy of 1, the data is uniformly distributed.

Example: Calculate normalized entropy

The following 3 nominally scaled characteristic values (Merkmalsausprägungen) are available with their respective absolute and relative frequencies:

Qualification	Absolute frequency	Relative frequency
Apprenticeship	30	0,3
Bachelor	10	0,1
Master	60	0,6

For example, 30 out of 100 employees have completed an apprenticeship; this corresponds to a proportion or relative frequency of 0.3 (or 30 %), etc.

The formula for entropy is

Entropy= $\sum [p(i) \times \log2(1/p(i))]$ or (1st version)

Entropy= $- \sum [p(i) \times \log2 p(i)]$ (2nd version)

The sum runs over i = 1 to n with n as the number of characteristic values (number of different qualifications). Here, p(i) is the relative frequency of characteristic value i, whereby p(i) (probability) is used here for simplicity instead of f(i) for relative frequency. The term "log2" is the logarithm to the base 2; this can be calculated using a calculator, e.g. $\log2\ 0.1 = \ln 0.1 / \ln 2 = -3.3219$. ln is the natural logarithm (key: LN).

Using the 2nd version of the formula results in the following for the example data:

Entropy = $- (0.3 \times \log2(0.3) + 0.1 \times \log2(0.1) + 0.6 \times \log2(0.6))$

$= - (0{,}3 \times -1{,}7370 + 0{,}1 \times -3{,}3219 + 0{,}6 \times -0{,}7370)$

$= - (- 0{,}5211 - 0{,}33219 - 0{,}4422) = 1{,}29549$.

If the calculated entropy value is divided by log2(N) (where N is the number of different characteristic values, in this case 3), the normalized entropy is obtained.

Normalized entropy = entropy / log2(3) = 1.29549 / 1.5850 = 0.8173

<u>If the above is transferred to the USERNAMES, the following relationship results:</u>

The number of characteristic values N is the number of different characters.

The calculation of entropy is based on the relative frequency of each character.

<u>Here are a few examples:</u>

Username	Length (number of characters)	N (charact. value)	Abs. frequent.	Rel. frequency f(i) Abs.frq./length
aaaaaaaaaa	10	1	a: 10	a:1

ABC	3	3	A:1, B:1, C:1	A:1/3, B:1/3, C:1/3
llzaaayy	8	5	l:2, z:1,a:3,y:2	l:2/8, z:1/8,a:3/8,y:2/8

Entropy (aaaaaaaaaaaa)= -(1x log2(1) /log2(1))=0 (must be avoided, if N==1)

Entropy(ABC)= -(1/3xlog2(1/3)+ 1/3xlog2(1/3)+ 1/3xlog2(1/3))/log2(3)=XXXXX

Entropy(llzaaayy)=-(2/8xlog2(2/8)+1/8xlog2(1/8)+3/8xlog2(3/8)+ 2/8xlog2(2/8))/log2(8)=XXXXX

If all characters have the same value, the relative frequency has the value 1 (i.e. 100%). This means that the logarithm has the value 0. The entropy is therefore also 0, regardless of the number of characters, i.e. the length of the username and regardless of the type of character.

The number of different characters within a username is the characteristic value.

As the relative frequencies from which we calculate the logarithm are always less than 1, the values of the logarithm are all negative. Therefore, the entropy must still be multiplied by -1 (see also the 1st version and 2nd version of the formula above).