ИНФОРМАТИКА

UDC 004.93
MSC 93B03

Extending the applicability of the Zipf's laws to the sequences of byte data

S. L. Sergeev, I. S. Blekanov, F. V. Ezhov, N. A. Tarasov

St. Petersburg State University, 7–9, Universitetskaya nab., St. Petersburg, 199034, Russian Federation

For citation: Sergeev S. L., Blekanov I. S., Ezhov F. V., Tarasov N. A. Extending the applicability of the Zipf's laws to the sequences of byte data. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2024, vol. 20, iss. 3, pp. 391–403. https://doi.org/10.21638/spbu10.2024.307

Zipf's law have been shown to hold true in many places. From it's first idea of a statistical phenomenon related to natural language to it's later adaptations for economical, social and many other fields, it has been shown to work almost universally. In all of these cases authors discuss the applicability of the Zipf's law in terms of semantically complex structures. We take this notion a step further and show how this law can work for data analysis, in particular for the sequences of byte data, obtained from various sources. We show that, using the basic chunking methodology, the Zipf's law can be shown to hold true for many different types of raw sequences of byte data. In particular, the law holds true in all caes for the "middle point" of data, where it is present with a degree of certainty of more than 90 %. We conclude by discussing the implications and potential use cases of these findings.

Keywords: Zipf's laws, byte data, chunking, frequency analysis.

1. Introduction. Frequency dictionaries of natural and even artificial languages obey Zipf's law. This effect has been shown to hold true in a variety of studies, and for natural languages, follows a fairly simple process. To compile the dictionary the text is divided into a sequence of words, the words are normalized in a certain way and are put the dictionary. Each word is included in the dictionary only once, with an indication of how many times it appears in the text. The dictionary is then sorted in descending order by the number of occurrences. Zipf [1, 2] discovered this inverse proportion of frequencies and rank. Later this observation was confirmed by numerous studies, related to the frequency analysis of natural languages. We use the following mathematical representation of the Zipf's law:

$$Rk \times N \approx B. \tag{1}$$

In (1) Rk is the rank of the word in the dictionary (ordinal number in the ordered list), N is the number of occurrences of the word in the text. It is common, however, for several words to have the same frequency. This is especially true for large texts. As such Zipf's law has another interpretation called the Zipf — Mandelbrot law or generalized Zipf's law [3, 4]:

$$Rk_N \times N^a = B. \tag{2}$$

1.1. Chunking. Data representation is the first step necessary for frequency analysis. Chunking is one the methods commonly used in the tasks related to data compression due to it's universal, data-independent nature [5]. We consider chunking to be a procedure of splitting the data set into a sequence of segments of equal length. At the same time, a dictionary of unique chunks is compiled. Each word in the dictionary is matched with a list of positions occupied by a chunk in the data. The dictionary together with the list of positions takes up less space in memory resulting in a lossless compression. The original text is easily restored using the chunk dictionary — a with R rows, where R is the number of unique chunks in the set. Each row consists:

- Ch is a chunk binary number, which consists of L bytes;
- N is the number of occurrences of a chunk in the source set;
- $A_1, A_2, ..., A_N$ is a list of addresses — the positions occupied by the chunk in the source set.

According to the dictionary of chunks, a dictionary of groups is built. The group dictionary consists of R_{gr} rows. The table is formatted as follows:

- Rk is the rank of the group;
- N number of occurrences of chunk;
- M is the number of chunks in the group.

The table is sorted in descending order of N.

On some data sets, we have found that the dictionary of chunks, at least in part, obeys Zipf's law (more on that in the later section). In addition, we have found that there is another form of dependence that complements Zipf's law for chunks with low frequencies, later referred as the addition to the Zipf's law:

$$M_N \times N^c = D, \tag{3}$$

where M_N is the number of chunks with the same N is the number of occurrences. Previously, similar relation was found to hold true for low ranking words in statistical studies of literary dictionaries [6].

There is a fundamental difference between a dictionary of natural language words and a dictionary of chunks. In the first case, words are text elements of various lengths and, most importantly, with certain semantics. In the second — formal units of the same length with undefined semantics.

1.2. Research goals. To study the distribution laws for arbitrary data a more convenient form of Zipf's law was chosen. This definition can be obtained from Equations (2) and (3) by taking their logarithm and a applying a slight transformation:

$$\log_2 N = a_1 \log_2 Rk_n + b_1, \tag{4}$$

$$\log_2 M_N = a_2 \log_2 N + b_2. \tag{5}$$

The study pursued the following goals:

- examine the applicability of the Zipf's law in the form of Equation (4) and it's addition as in Equation (5) on data sets of different nature;
- establish the features of the implementation of Zipf's law for different types of data;

• find the dependence of the parameters a_1, b_1, a_2, b_2 on the set size — K and on the chunk size — L.

To the applicability of Zipf's law, as well as it's generalization, 9 datasets were selected containing data from 4 different areas: in-memory data, text data, sound data and image data.

The remainder of this paper is organised as follows. Section 2 provides an overview on the existing testing methods and potential applications. Section 3 discusses the applicability of the Zipf — Mandelbrot law for different types of data. Section 4 provides the results of testing of Equation (5) — the addition to the Zipf's law. Section 5 concludes the paper, outlines it's contributions and discusses the potential for future research and applications.

2. Previous work.

2.1. General applicability. Zipf's law have been heavily studied since it's inception. Although initially proposed for natural languages it has since found a variety of other applications. Authors in [7] provide an overview on the usability of Zipf's law in regards to natural language, comparing distributions with different semanitcs, languages, etc. Authors in [8] show the modeling effectiveness of the Zipf's model across 50 languages. The authors present the evidence of a 3 stage division of data, with only the middle stage being close to the theoretical definition in the case of non-ininite amount of data. This coincides with our findings, present in the next section. Another widely discussed application of Zipf's law relates to city size distribution. In the work [9] an overview of 114 studies on the topic is provided. The authors conclude that the law, while applicable in some cases, does not provide accurate representations and thus is not universal and theorize that this effect is at least partially related to the question of defining the subject of research — city. Limitations of empirical methods is another common problem for this task. Once again, authors emphasize the prevalence of the idea that different levels of size (from lower-tail small cities to upper-tail metropolitan cities) are different in their characteristics in regard to the Zipf's law.

2.2. Data analysis. Zipf's law have been found to accurately describe other types of data and have been used to improve and optimize storage, compression and data access speed. For text research, in work [10] authors present a novel language encoding model, utilizing the Zipf's law to efficiently build coocurence matrix. Authors [10] show, that the resulting model is capable of encoding the text with quality, comparable to that of a traditional neural language models.

For sound data researchers have found, that Zipf's law holds true for speech [11], for music [12] and even animal vocal communications [13].

For image data authors [14] present an image encoding procedure, and state that image data, encoded in a way conforming the Zipf's law, can be used to improve the quality of image recognition and other related tasks, and introduce these features and demonstrate that some of them have characteristics that can be closely defined by the Zipf's law.

3. Applicability of the Zipf's law to the sequences of byte data.

3.1. On importance of segmentation. Preliminary studies have established that Zipf's law is not valid for all groups of chunks in the set. Like many other researches [15, 16] we have found that to for Zipf's law to hold true it is necessary to segment the data. That is, a group with $Rk = 1$, and sometimes with $Rk = 2$, somewhat falls out of the pattern described by Equation (4). In addition, they do not fit well into the law of the group of chunks with low occurrence, for small N.

We define a good fitting chunk group as $3 \leqslant Rk_N \leqslant Rk_{end}$. To assess the quality of the pattern matching for groups of chunks to the linear Zipf's model (4), we used the coefficient of determination, denoted us — R_z^2. It is important to note, that not all groups of chunks are subject to the additon law, described by Equation (5). It was previously established that the addition law described by formula (5) is satisfied only for groups of chunks with high values of Rk, when $3 \leqslant Rk_{beg} \leqslant Rk_N \leqslant Rk_{gr}$.

3.2. Datasets. Four types of data were tested:

• RAM memory dumps of mobile devices on the Android OS — 3 datasets with the combined size of 12Gb: d1ram, d2ram, d3ram;

• text files — 2 datasets — text files in various formats and corresponding raw text data;

• Image data:

1. VisualQA — 20 000 files, 2.7Gb total, images1,
2. Indoor images dataset — 15 620 files, 2.4Gb total, images2;

• Sound data:

1. UrbanSound — 8733 files, 6.6Gb total, sounds1,
2. Golos — 79 801 files, 7.6Gb total, sounds2.

3.3. Experimental methodology and results on the first dataset. This section describes the studies related to Equation (4) using the RAM data from the third device set with chunks of $L = 8$ bytes, as an example. Number of chunks $K = 469\,718\,016$ (original set f3, size $V_0 = 3\,757\,744\,128$ bytes). For the set, a table was built — a dictionary of $R_{gr} = 5697$ chunk groups. Table elements were plotted and can be seen in Figure 1, *a*.

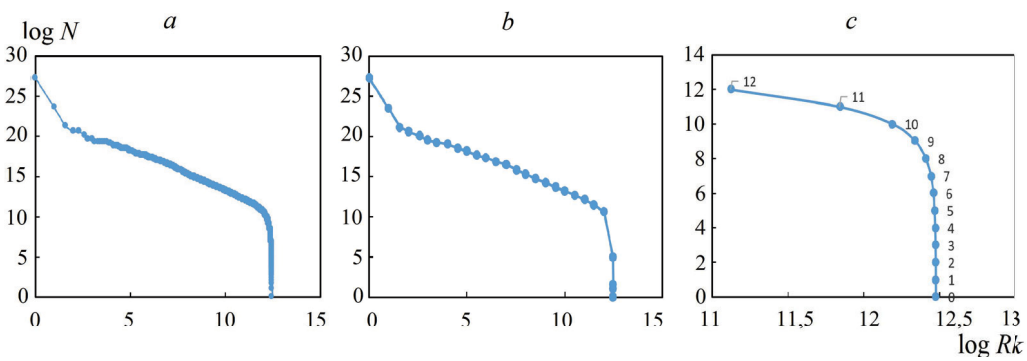

Figure 1. Dependence charts
a — dependence of the frequency of occurrence of chunks ($\log_2 N$) on the rank ($\log_2 Rk$) according to the Zipf — Mandelbrot law; *b* — sparse version frequency dependence chart; *c* — detailed tail of the dependence chart.

For greater clarity, part of the points were plotted separately. When plotting graphs, the overall table of chunk groups was thinned out. The $\log_2 Rk$ column was identified. 26 rows were selected, based on this column. Row with $\log_2 Rk = 0$ and $\log_2 Rk = 1$. And then with a step of about 0.5 along the column of the logarithms. That is, $\log_2 Rk \approx 1.5$, $Rk = 2$, $\log_2 Rk \approx 2.5$, etc., up to $\log_2 Rk \approx 12.47$ ($12.47 \approx \log_2 R_{gr}$). The graph constructed according to these lines is shown in Figure 1, *b*.

Each point on the plots corresponds to one group. The first two points correspond to the groups with the highest rank. They each consist of a single chunk. The last point corresponds to the group with rank $Rk = 5.697$. We are interested in the middle part of the plot, which looks linear — indicating it's Zipf property. It is important to define

these cutoff parameters and boundaries. In all further examples, it turned out that on the left (upper-cutoff) the linear part can be considered to take an effect from the third or even from the second point. Determining the right boundary (lower-cutoff) presents some problems. At first glance, the last point of the linear part is $\log_2 N = 12$, this is a group with a rank of 4096. But the rank of the last point is much higher — 5697. To clearly define this problem, we present a more detailed plot of the right side of the curve (Figure 1, *c*).

It can be seen from the Figure 1, *c* that the 12, as well as the 11, 10, etc. Can all be considered to be the last point of this segment, depending on the criterion of correspondence of the experimental points to the linear model. As a measure of compliance, it is natural to use the coefficient of determination. We decided to select the right boundary (lower cut-off) of the linear part so that the $R^2 < R^*$ condition is satisfied. The following algorithm was used. We take a subsample of points. The first point in the subsample will coincide with the upper bound, the last point of the subsample will coincide with the last point of the entire sample. We construct a regression line on the subsample using the least squares method and calculate R^2. If $R^2 < R^*$, we discard the last point from the subsample and plot the regression line again. We continue to drop points from the subsample until $R^2 < R^*$. The last point of the final subsample will be the lower bound — Rk_{end}.

Next, we built a straight line $\log_2 N = a_1 \log_2 Rk + b_1$ approximating the points in the middle part of the table, and excluded the points at the beginning and the end. The parameters a_1 and b_1 were calculated using the least squares method: a_1 and b_1 were found, by minimizing the function:

$$\Delta = \sqrt{\frac{\sum_{i=2}^{Rk_{end}} (\log_2 N_i - a_1 \log_2 i - b_1)^2}{Rk_{end} - 1}}. \tag{6}$$

Since there is no clear answer to the question of choosing the optimal value of R^*, calculations were performed for several values: $R^* = 0.80, 0.85, 0.90, 0.95, 0.99$. The following parameters were chosen as the set of parameters characterizing the dictionary:

- R_{gr} is the number of groups in the dictionary;
- Rk_{end} is the rank of the group at which Zipf's law stops working;
- a_1 and b_1 are parameters of Equation (4);
- R^2 is a linear model correlation coefficient;
- $\sigma = \frac{Rk_{end}-1}{R_{gr}}$ proportion of groups conforming to the Zipf's law.

Table 1 shows the parameter values for $K = 469\ 718\ 016$ and $L = 8$ bytes. In this table, the last 2 columns are of primary interest. It can be deduced that R^2 values equal to or greater than 0.85, and σ of more than 90 % can be considered acceptable. Following this the three middle rows of the table can be considered evidence of the fulfillment of the generalized Zipf's law for this dictionary.

Table 1. **Endpoint approximation parameters**

R_{gr}	Rk_{end}	a_1	b_1	R^2	σ, %
5697	5695	−1.35	26.6	0.80	99.96
	5678	−1.34	26.42	0.85	99.67
	5574	−1.29	25.9	0.90	97.84
	5287	−1.21	25.2	0.95	92.80
	4378	−1.10	24.2	0.99	76.85

Similar calculations were performed for the same K value and $L = 4, 16, 32, 64$, 128 bytes. In all cases, at $R^2 \approx 0.85$, the proportion of groups corresponding to the generalized Zipf's law — σ is more than 99 %, at $R^2 \approx 0.95$ — over 90 %.

Table 2 shows the values of σ for various values of L and portions of the original $K = 469\,718\,016$ at $R^2 \approx 0.9$. It was found that for the purposes of our study, the specific values of the parameters a and b are of no interest and are only needed to calculate R^2 and σ. Therefore, they are not shown in further tables.

Table 2. **Parameter testing results for subsets of data**

L, bytes	V_0, %	$V_0/2$, %	$V_0/4$, %	$V_0/8$, %	$V_0/16$, %
4	97	97	97	97	97
8	98	98	98	97	97
16	98	98	97	97	97
32	97	97	97	97	97
64	97	97	97	97	97
128	98	98	98	98	97

Thus, for all considered sizes of sets and chunks, Zipf's law is fulfilled for 97–98 % of chunk groups with a certainty of $R^2 \approx 0.9$.

3.4. Results on the other datasets. As already mentioned, 8 more datasets were studied. The results for their testing are presented in Table 3. For each set, only the results with $R^2 = 0.85$ or slightly more are presented.

Based on the tables, the following conclusions can be drawn:

a. For all data sets, regardless of the nature of the data itself, generalized Zipf's law is satisfied for most of the chunk groups (from 89 to 99 %) with chunks of size $L = 4$ bytes. The chunk with the highest occurrence does not follow the law — it occurs more often than the law predicts, and from 1 to 11 % of the groups of chunks with the lowest occurrence. It must be emphasized that here, we are talking about the groups. Groups with low occurrence are the most numerous.

b. For some of the sets, generalized Zipf's law also holds true for L greater than four — testing showed this to be the case for up to $L = 128$ bytes.

c. The degree of the fulfillment of the law (R^2 and σ) does not see an increase with the growth of L.

Table 3. **Parameter testing results for other datasets**

Dataset	$V_0 \cdot 10^6$	L, bytes					
		4	8	16	32	64	128
d1ram	4.27	99	99.6	99.6	99.8	99.8	99.7
d2ram	4.11	99	99.7	99.5	99.5	99.7	99.6
texts1	3.20	99	98	97	98	99	99
texts2	3.11	99	98	98	98	98	91
sounds1	7.09	89	94	99	98	96	92
sounds2	7.63	99	98	99	98	92	15
images1	2.91	95	79	68	61	62	69
images2	2.58	95	99	99.6	99	—	12

One might suspect that the fulfilment of the generalized Zipf's law for chunk groups is related to how the data is formatted in the source sets. For example, ASCII encoding is done performed for bytes; 4, 8 and so on bytes are combinations of any given characters. To dispel these suspicions, we compiled three dictionaries based on the f3 dataset with chunks of 61, 64 and 67 bytes as it is unlikely that there are standard combinations of 61 or 67 bytes. The results of this experiment are presented in Figure 2.

This experiment shows that fulfilment of the generalized Zipf's law is not related to the data representation formats.

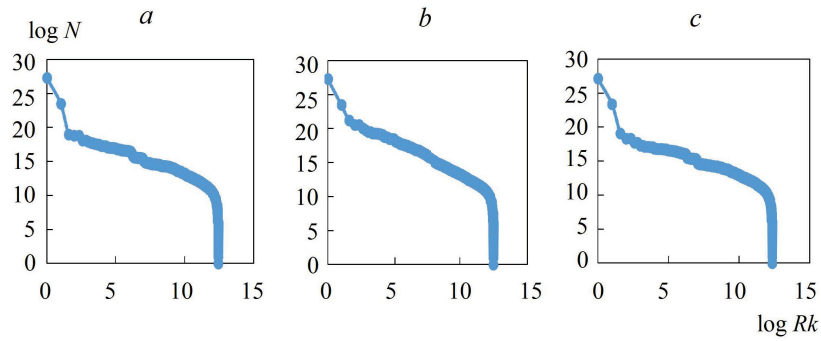

Figure 2. Dependence of $\log_2 N$ on $\log_2 Rk$ according to generalized Zipf's law for chunk length $L = 61$ (a), 64 (b) and 67 (c) bytes, respectively

3.5. Results on the randomised data. Another important experiment was designed to check the methodology used in this study, dispelling the potential suspicion that any data set created as a random sequence of bits obeys the Zipf—Mandelbrot law.

For this purpose, we have built a random dataset with the size $V_0 = 3,\ 757,\ 744,\ 128$ bytes, equal to the size of the f3 dataset. To obtain bit sequences, the following random number generators were used:

- MT19937 — Mersenne Twister 19937 generator [17];
- PCG64 — Permuted congruential generator [18];
- PCG64DXSM — 128-bit implementation of permutation congruential generator [19];
- Philox — Counter-based random number generator [20];
- SFC64 — Small Fast Chaotic PRNG [21].

The random bit sequences generated by each method were combined into datasets with the size V_0. As a result, 5 randomized datasets were built. The datasets were processed according to the standard scheme, as described in the previous section:

- the data was split into chunks;
- chunk dictionaries were constructed;
- lists of chunk groups were constructed.

The result for chunks of length 8, 16, 32, 64 and 128 bytes was the same for all 5 randomized datasets.

The results for chunks with the length of 4 bytes, generated by the MT19937 algorithm are presented in the Table 4.

Table 4. **Parameter testing results for other datasets**

Rk	N	M
1	8	2
2	7	12
3	6	519
4	5	14 381
5	4	329 920
6	3	6021 784
7	2	82 557 866
8	1	754 860 149

Figure 3 shows that neither the Zipf—Mandelbrot law (a) nor its addition (b) are satisfied for this set. Similar results, with slight variations in column M, were also shown by sets obtained using other random number generators: PCG64, PCG64DXSM, Philox,

SFC64. Using our tetsing methodology it was determined that all chunks belong to the same group and have the same frequency of occurrence $N = 1$. This result shows that the fulfillment of the Zipf — Mandelbrot law is related to the nature of the information encoded in the dataset files.

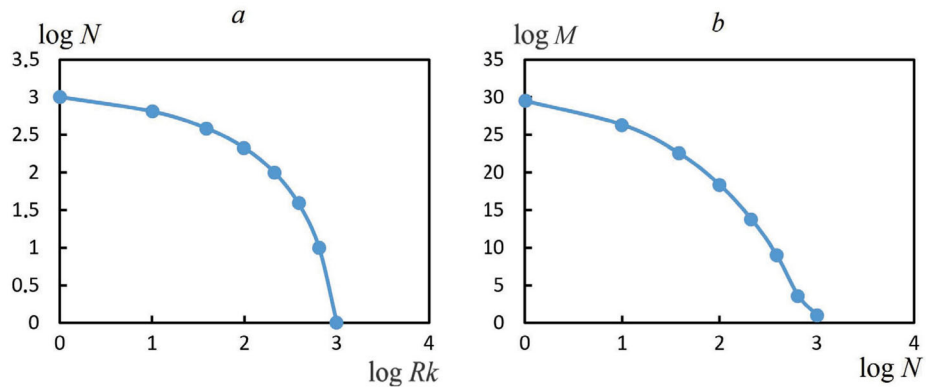

Figure 3. Applicability of the Zipf's law to random data
a — dependence of the frequency of occurrence of chunks ($\log_2 N$) on the rank ($\log_2 Rk$) according to the Zipf for MT19937 data; *b* — dependence of the number of chunks with the same number of occurrences ($\log_2 M$) on the chunk frequency ($\log_2 N$) according to the addition to the Zipf's law for MT19937 data.

4. The addition to the Zipf's law.

4.1. Experimental results. Studies related to Equation (5) were carried out according to the same scheme. Figure 4, *a* shows the results for the d3ram dataset at $L = 8$ bytes. Similarly to the previous section, a sparse list was made (with a constant step along the $\log_2 N$ column). The results are shown in the Figure 4, *b*.

It can be seen from the plots that a large number of points corresponding to smaller N, starting from $N = 1$, lie on a straight line, although there are some noticeable outliers. At the same time, a large number of points are concentrated on or are near the horizontal axis. Most of these points correspond to groups that consist of a single chunk ($M = 1$). There is also a transition region, that makes the boundaries between the left and right regions less determinate. And here, as a measure of compliance of the fixed part of the points to the linear model, the coefficient of determination R^2_{ad} was used (ad subscript denotes the difference in the parameter related to the generalized Zipf's law from the one related to the addition). We decided to divide all the points into two parts: the left one corresponding to the linear model, and the right one not corresponding to the linear model. To separate the parts and find the last point of the left part — N_0, the same algorithm from the previous section was used.

Next, a straight line $\log_2 M = a_2 \log_2 N + b_2$ was constructed, approximating the points on the left side of the table. The parameters a_2 and b_2 were calculated using the least squares method, minimizing the function

$$\Delta = \sqrt{\frac{\sum_{N=1}^{N_0} (\log_2 M_N - a_2 \log_2 N - b_2)^2}{N_0}}. \tag{7}$$

The following parameters were chosen as the main parameters characterizing the dictionary:

• R_gr, Rk_N are the number of the group from which the generalized form of the law begins to appear;

• a_2 and b_2 are the parameters of Equation (7);

• $\sigma_M = \frac{R_{gr} - Rk_N}{R_{gr}}$ is the proportion of groups that obey the addition to the Zipf's law;

• R^2_{ad} is the correlation coefficient of the linear model.

Table 5 shows the parameter values for $K = 469\,718\,016$ and $L = 8$ bytes

Table 5. **Approximation parameters for the addition to the Zipf's law**

R_{gr}	Rk_N	a_2	b_2	R^2_{ad}	σ_M, %
5697	249	−1.35	17.3	0.80	96
	505	−1.51	18.9	0.85	91
	970	−1.71	20.9	0.90	83
	2717	−2.07	24.3	0.95	83

Table 6 shows the values of σ_M for parts of the set d3ram for various L. For the whole table $R^2_{ad} = 0.85$. The tables show that the addition to the Zipf's law holds true for the greater amount of the chunks, with larger the set size and the smaller size chunks increasing that number.

Table 6. **Applicability of the the addition to the Zipf's law**

L, bytes	V_0, %	$V_0/2$, %	$V_0/4$, %	$V_0/8$, %	$V_0/16$, %
4	92	92	91	91	90
8	91	90	90	90	90
16	90	89	89	88	87
32	88	87	86	86	85
64	83	83	83	83	81
128	84	80	80	78	77

Table 7 contains the results of testing of the addition to the Zipf's law for 8 other datasets with various values of L.

Table 7. **Results of the addtion to the Zipf's law testing on other sets of data**

Parameters		L, bytes					
		4	8	16	32	64	128
f1	R^2_{ad}	0.85	0.85	0.85	0.85	0.85	0.85
$V_0 = 4.27106$	σ_M	94 %	92 %	92 %	91 %	89 %	86 %
f2	R^2_{ad}	0.85	0.85	0.85	0.85	0.85	0.85
$V_0 = 4.11106$	σ_M	94 %	92 %	92 %	90 %	86 %	83 %
texts1	R^2_{ad}	0.85	0.85	0.85	0.85	0.85	0.86
$V_0 = 3.20106$	σ_M	83 %	87 %	94 %	91 %	85 %	64 %
texts2	R^2_{ad}	0.85	0.85	0.85	0.85	0.85	0.85
$V_0 = 3.11106$	σ_M	81 %	88 %	93 %	92 %	91 %	90 %
sounds1	R^2_{ad}	0.86	0.85	0.86	0.80	0.80	0.85
$V_0 = 7.09106$	σ_M	99.97 %	99.1 %	99.1 %	56 %	43 %	38 %
sounds2	R^2_{ad}	0.85	0.85	0.85	0.85	0.85	69 %
$V_0 = 7.63106$	σ_M	87 %	98 %	93 %	98 %	—	—
images1	R^2_{ad}	0.85	0.85	0.85	0.89	0.89	0.86
$V_0 = 2.91106$	σ_M	93 %	78 %	99 %	98 %	90 %	85 %
images2	R^2_{ad}	0.85	0.87	0.85	0.86	0.86	0.85
$V_0 = 2.58106$	σ_M	99,8 %	13 %	9 %	6 %	64 %	53 %

Conclusions for the addition to the Zipf's law:

• with $L = 4$ bytes, the addition is satisfied for more than 80 % of groups with an accuracy of R^2 or more;

• for most sets, the addition is satisfied with high accuracy even for $L > 40$ bytes.

4.2. On importance of segmentation. Both charts can be presented in one figure (Figure 5). At first glance, it seems that the $\log_2 Rt$ plot consists of two lines (excluding the last two points), however, a closer look (Figure 6, *a*) shows that there is only one line (Figure 6, *b*).

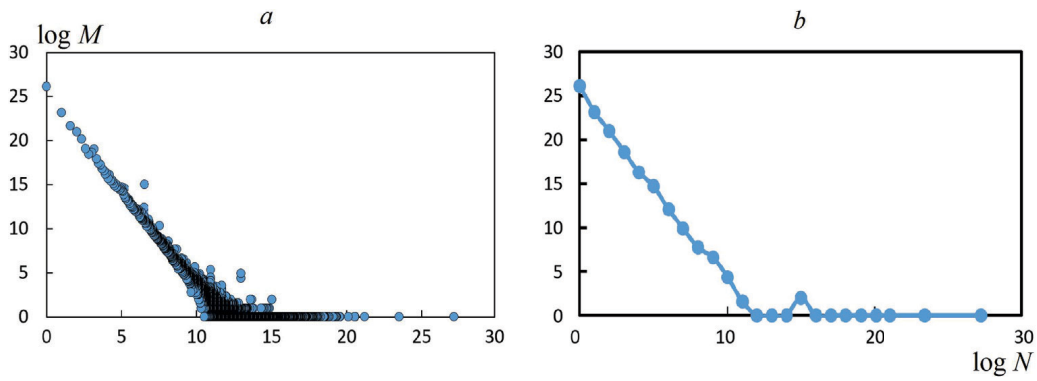

Figure 4. Dependence of the number of chunks with the same number of occurrences ($\log_2 M$) on the chunk frequency ($\log_2 N$) according to the addition to the Zipf's law
a — full form, $L = 61$ bytes; *b* — sparse frequency chart, $L = 64$ bytes.

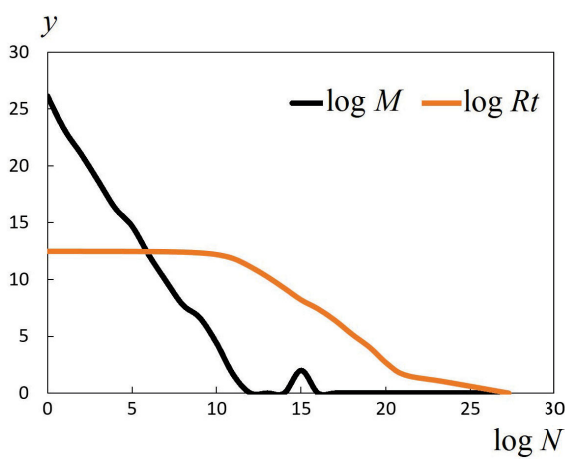

Figure 5. Charts of the addition to the Zipf's law and its inverse correlation with the generalized formulation of the Zipf's law

From Figure 5 it can be seen that the entire space of chunk groups is divided into three zones. The one on the right (lower cutoff) — splits into 2 additional groups for some cases. The middle one, where Zipf—Mandelbrot law is fulfilled, and the left one (higher cutoff), where the addition to the Zipf's law is fulfilled. The left and middle zones overlap.

5. Conclusions. In this study we presented the idea of general applicability of the Zipf's law using the ideas and methods of analysing raw byte data. We have shown the

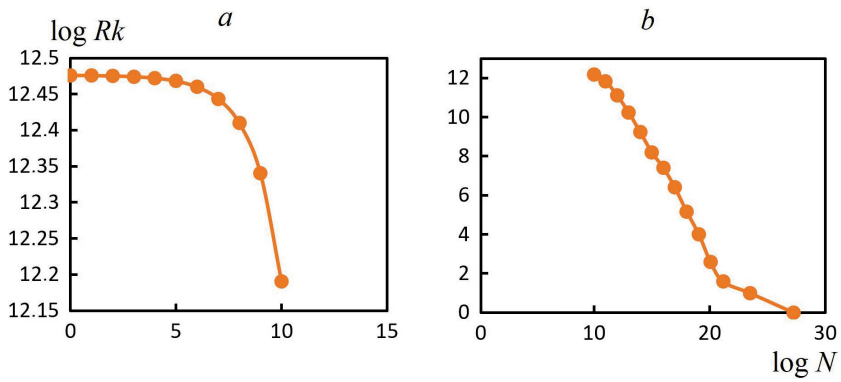

Figure 6. Inverse dependence of the Zipf — Mandelbrot law
a — for $\log_2 N$ from 0 to 10, $L = 61$ bytes; b — for $\log_2 N$ from 10 to the maximum value of N, $L = 64$ bytes.

evidence of the Zipf-like structure of byte-level data on 9 varied datasets. For each set, several dictionaries of chunk groups were built (different dictionaries for one set arise for different L) and for all of them, a significant part of the dictionary was found to correspond to Zipf — Mandelbrot law and to addition to the Zipf's law. One of the initial sets (d3ram) was sequentially divided into 2, 4, 8 and 16 parts, and dictionaries for different values of L were compiled for each. Thus, partial compliance to the Zipf's law was established for about 80 different dictionaries.

Our findings can prove to be useful in a variety of topics, related to data analysis, be it efficient model training, signal processing and data storage. Data storage and compression, in particular, are of the greatest importance to our further research. Previously it has been found that the usage of Zipf's law can allow for faster more efficient compression of text data [22, 23]. Using similar methodologies we expect to create a general compression and byte data encoding approach, using the data structuring in accordance with the Zipf's law.

References

1. Zipf G. K. *The psycho-biology of language: An introduction to dynamic philology.* London, Routledge Publ., 1999, 356 p.

2. Zipf G. K. *Human behavior and the principle of least effort.* Cambridge, Mass., 1965, 573 p.

3. Mandelbrot B. An informational theory of the statistical structure of language. *Communication Theory*, 1953, vol. 84, pp. 486–502.

4. Mandelbrot B. *The fractal geometry of nature.* New York, W. H. Freeman & Co. Publ., 1982, 468 p.

5. Lu G., Jin Y., Du D. H. C. Frequency based chunking for data de-duplication. *2010 IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, IEEE, 2010, pp. 287–296.

6. Baayen R. H. *Word frequency distributions.* Dordrecht, Springer Science & Business Media, 2001, 335 p.

7. Piantadosi S. T. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 2014, vol. 21, no. 5, pp. 1112–1130.

8. Yu S., Xu C., Liu H. *Zipf's law in 50 languages: its structural pattern, linguistic interpretation, and cognitive motivation.* arXiv preprint, arXiv: 1807.01855, 2018.

9. Arshad S., Hu S., Ashraf B. N. Zipf's law and city size distribution: A survey of the literature and future research agenda. *Physica A: Statistical Mechanics and its Applications*, 2018, vol. 492, pp. 75–92.

10. Gao L., Zhou G., Luo J., Huang Y. Word embedding with Zipf's context. *IEEE Access*, 2019, vol. 7, pp. 168934–168943.

11. Baumann A., Każmierski K., Matzinger T. Scaling laws for phonotactic complexity in spoken english language data. *Language and Speech*, 2021, vol. 64, no. 3, pp. 693–704.

12. Perotti J. I., Billoni O. V. On the emergence of Zipf's law in music. *Physica A: Statistical Mechanics and its Applications*, 2020, vol. 549, art. no. 124309.

13. Kershenbaum A., Demartsev V., Gammon D. E., Geffen E., Gustison M. L., Ilany A., Lameira A. R. Shannon entropy as a robust estimator of Zipf's law in animal vocal communication repertoires. *Methods in Ecology and Evolution*, 2021, vol. 12, no. 3, pp. 553–564.

14. Crosier M., Griffin L. D. Zipf's law in image coding schemes. *BMVC 2007 — Proceedings of the British Machine Vision Conference*, 2007, pp. 1–10.

15. Kornai A. Zipf's law outside the middle range. *Sixth Meeting on Mathematics of Language*, 1999, pp. 347–356.

16. Corral Á., Boleda G., Ferrer-i-Cancho R. Zipf's law for word frequencies: Word forms versus lemmas in long texts. *PloS One 10*, 2015, vol. 549, no. 7, pp. 1–23.

17. Matsumoto M., Nishimura T. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 1998, vol. 8, no. 1, pp. 3–30.

18. O'Neill M. E. *PCG: A family of simple fast space-efficient statistically good algorithms for random number generation*. Available at: https://www.pcg-random.org/ (accessed: May 1, 2024).

19. Upgrading PCG64 with PCG64DXSM — NumPy v1.24 Manual. Available at: URL: https://numpy.org/doc/stable/reference/random/upgrading-pcg64.html, (accessed: May 01, 2024).

20. Salmon J. K., Moraes M. A., Dror R. O., Shaw D. E. Parallel random numbers: as easy as 1, 2, 3. *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, 2011, pp. 1–12.

21. SFC64. *Small Fast Chaotic PRNG*. Available at: https://numpy.org/doc/stable/reference/random/bit_generators/sfc64.html (accessed: May 1, 2024).

22. Bakulina M. P. Application of the Zipf law to text compression. *Journal of Applied and Industrial Mathematics*, 2008, vol. 2, no. 4, pp. 477–483.

23. Mahmood M. A., Hasan K. A. Efficient compression scheme for large natural text using Zipf distribution. *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, 2019, pp. 1–6.

Received: May 19, 2024.
Accepted: June 25, 2024.

Authors' information:

Sergey L. Sergeev — PhD in Technics, Associate Professor; slsergeev@yandex.ru

Ivan S. Blekanov — PhD in Technics, Associate Professor; https://orcid.org/0000-0002-7305-1429, i.blekanov@spbu.ru

Fedor V. Ezhov — Postgraduate Student; https://orcid.org/0009-0007-1468-0042, st056053@student.spbu.ru

Nikita A. Tarasov — Postgraduate Student; https://orcid.org/0000-0002-9473-6179, nkt.tarasov@yandex.ru

Расширение применимости закона Ципфа для произвольных последовательностей битовых данных

С. Л. Сергеев, И. С. Блеканов, Ф. В. Ежов, Н. А. Тарасов

Санкт-Петербургский государственный университет,
Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9

Для цитирования: *Sergeev S. L., Blekanov I. S., Ezhov F. V., Tarasov N. A.* Extending the applicability of the Zipf's laws to the sequences of byte data // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2024. Т. 20. Вып. 3. С. 391–403. https://doi.org/10.21638/spbu10.2024.307

Доказано, что закон Ципфа справедлив для самых разнообразных статистических распределений, начиная с его первоначальной идеи о статистической закономерности, связанной с его применением для обработки естественных языков, и заканчивая его более поздними адаптациями для экономической, социальной и многих других предметных областей, в которых было установлено, что он работает практически повсеместно. Во всех этих случаях авторы различных исследований обсуждают применимость закона Ципфа в терминах семантически сложных структур. Сделан следующий шаг в этом вопросе и показано, как такой закон может работать для анализа данных, в том числе для последовательностей байтовых данных, полученных из разных источников. Используя базовую методологию разбиения на блоки, можно доказать, что закон Ципфа справедлив для многих типов необработанных последовательностей байтовых данных, в частности во всех случаях для «средней точки» данных, где они присутствуют со степенью достоверности более 90 %. В заключение приводятся рассуждения о последствиях и возможных вариантах использования полученных результатов.

Ключевые слова: законы Ципфа, битовые данные, фрагментация данных, частотный анализ.

Контактная информация:

Сергеев Сергей Львович — канд. техн. наук, доц.; slsergeev@yandex.ru

Блеканов Иван Станиславович — канд. техн. наук, доц.; https://orcid.org/0000-0002-7305-1429, i.blekonov@spbu.ru

Ежов Федор Валерьевич — аспирант; https://orcid.org/0009-0007-1468-0042, st056053@student.spbu.ru

Тарасов Никита Андреевич — аспирант; https://orcid.org/0000-0002-9473-6179, nkt.tarasov@yandex.ru