National Research University Higher School of Economics

as a manuscript

Malashina Anastasia Gennadievna

Development of methods and algorithms for analyzing the characteristics of natural language texts based on the information theory approach for application to information security tasks

DISSERTATION SUMMARY

for the purpose of obtaining academic degree

Doctor of Philosophy in Engineering

Academic supervisor:

Doctor of Technical Sciences

Associate professor

Los Alexey Borisovich

Moscow – 2024

Contents

Introduction

Problem statement and relevance

One of the aspects of information security ensuring is the application of the information-theoretic approach to the analysis of algorithmic methods for information protection. In particular, within the framework of this approach, the information characteristics and parameters of the message source are calculated. On these characteristics, among other things, the issues of source protection depend. This task is very relevant, but it is quite complex, and, despite the existing theoretical base, little applied research has been carried out in this subject area.

Considering the relevance of this research area, the thesis develops methods and algorithms for analyzing the characteristics of various texts in natural languages, such as the entropy of short s-grams, power characteristics and coverage of s-gram dictionaries, based on the information-theoretic approach.

The results of the text characteristics analysis are applicable to solving the problems of information security and the development of methods for analyzing and synthesizing modern information security algorithms. In particular, they are applicable for building methods for restoring message parts based on the information-theoretic approach, as well as for studying the effectiveness of authentication algorithms based on password protection.

For example, in several cases, there are situations when information about the possible values of the original characters appears on the communication channel output the relative to the characters of an unknown text message.

With some information about the characters of an unknown text and limiting the number of input messages in advance (for example, with closed dictionaries), it is possible to restore the original message or its discrete s-grams [12, 15].

The object of research in this thesis is texts in natural language. **The subject of this research** is the methods and algorithms for analyzing the information characteristics of these texts.

The relevance is due to the need to develop theoretical-informational and algorithmic approaches to several information security problems and the improvement of methods for analyzing and synthesizing modern information security algorithms.

The relevance and significance of the study is confirmed by clause 1.4 "Development of Theoretical and Information Methods for the Analysis of Information Transmission, Storage and Protection Systems" of the science field passport "Engineering Sciences and Applied Mathematics" [11], approved by the Academic Council of the Higher School of Economics on February 2, 2018, as well as recent research conducted in this area [1].

Goals and objectives

The main goal of this dissertation is to develop methods and algorithms for analyzing the characteristics of natural language texts based on the information-theoretic approach.

To achieve this goal, the following tasks are set:

1. To develop algorithms for normalizing text corpora and building s-gram dictionaries. To form various text corpora and build s-gram dictionaries.
2. To develop a mathematical model of s-gram dictionary coverage. Get numerical estimates of dictionary coverage.
3. To develop a method for estimating the entropy of s-grams. To obtain numerical estimates of the entropy of short s-grams.
4. To investigate the limits of the permissible polysemy which appears in the process of the s-grams restoration.
5. To develop a method for restoring the discrete parts of input messages using the information-theoretic approach. Obtain numerical estimates of the share of the recovered message s-grams.

Literature review

Theoretical and informational aspects of information protection are described in detail in the monographs of many world scientists, including K. Shannon [7].

Some cases of information recovery at the output of a communication channel are currently being studied by A. V. Babash [1], and studied by V. M. Deundyak and E. A. Pashkova [10]. In these works, the method of "reading in columns" is used to fully restore the output message in the presence of some output information, which significantly limits the number of unknown message characters (with a probability of their occurrence close to 1).

According to previous studies [15], a complete restoration of a message is possible if the number of assumed values of the original characters is limited (for example, for the Russian language, it should be no more than 16 with an alphabet of 32 characters), while the probability of the true character appearance of among them is close to 1. This approach can lead to the loss of the true recovery option, the probability of which is estimated based on the introduced process probabilistic model under study. With an increase in the number of values at the output of the communication channel, the restoration of the original text becomes difficult due to the significant number of possible recovery options and the uncertainty of choice that arises in the enumeration process. Thus, when the number of characters at the output of the channel is small, the recovery option from the set of input messages can be constructed with a valid polysemy, since all other combinations will turn out to be a text of a random structure. However, as the number of values increases, this approach leads to finding a set of variants that exceed the valid polysemy. In this case, it is impossible to determine which of the found texts is the original message, without having additional information. That is, it becomes impossible to fully restore the message. However, the problem of the restoring discrete s-grams of the message remains open.

The concept of entropy is the basis of the information-theoretic approach to information security. Text and language are also systems with information entropy. Moreover, the entropy of natural language texts is significantly lower than the maximum entropy of the alphabet [17]. Since the number of valid s-grams in the language is significantly less than the forbidden ones, this approach significantly

reduces the complexity of recovery compared to a brute force method. A similar approach, for example, can be used when guessing passwords [4].

There are various methods for estimating the entropy of s-grams. The most popular among them is the Shannon method [8]. Using the representation of the text using a Markov chain of depth *s*, it is possible to approximate the probabilities of s-grams.

In this dissertation, the author uses a method for determining the entropy of s-grams, based on the compilation of dictionaries, the ideas of which go back to the combinatorial approach of Kolmogorov [5] and proposes methods for asymptotic assessment of the coverage of the created s-gram dictionaries.

Scientific novelty

1. Algorithms for normalizing text corpora and building s-gram dictionaries have been developed, which make it possible to form an applied base for analyzing the characteristics of texts in natural language based on the information-theoretic approach.
2. Mathematical models of s-gram dictionaries coverage based on the information-theoretic properties of natural language texts are proposed and investigated.
3. A method for estimating the short s-grams entropy based on the power characteristics of s-gram dictionaries has been developed.
4. Studies of the permissible polysemy boundaries which appears un the process of the s-grams restoration are carried out.
5. A method for restoring the discrete parts of input messages using the information-theoretic approach has been developed.

Theoretical and practical significance

The theoretical significance of this dissertation research lies in the development of theoretical-informational and algorithmic approaches to solving the problems of information security and the development of methods for analyzing and synthesizing modern algorithms of information security.

The practical significance of the dissertation results is as follows:

1. The developed methods and algorithms make it possible to obtain numerical estimates of the information characteristics of natural languages.
2. The developed method for reconstructing the discrete parts of input messages makes it possible to estimate the share of information that can be recovered using the information-theoretic approach at the given power characteristics.
3. Methods for constructing dictionaries and estimating the entropy of short s-grams can be used to analyze the strength of authentication algorithms built based on password systems.

Research methods

In the thesis, we use the methods of statistical analysis, limit theorems of probability theory, including the central limit theorem and theorems on the distribution (including conditional and multidimensional). We also use fundamental concepts from information theory (discrete communication channel, entropy of a discrete stationary source, Shannon's second theorem, etc.). When creating and conducting research on text corpora and dictionaries of s-grams, we apply theoretical aspects of corpus linguistics, including the principles of completeness and representativeness of compiled corpora.

To study the possibility of recovering discrete message s-grams, we implement a software of the recovery algorithm using object-oriented programming methods in C++. To obtain several numerical estimates and extrapolate empirical results, we use computer algebra tools (numerical methods), unclouding the Wolfram Mathematica software package.

Personal contribution

All the results were obtained by the author personally. During the dissertation research, the author has developed an approach that allows evaluate the coverage of dictionaries. The author has empirically studied several informational properties of the Russian and English languages (assessment of the *volumes of s-gram*

dictionaries for corpora of different styles and sizes, estimating the coverage of dictionaries, estimating the entropy of s-grams).

In addition, the author has investigated the problem of reconstructing discrete s-grams of a message from a pre-formed dictionary in the case when full recovery is impossible. The author has obtained the boundaries of a valid polysemy of the s-gram message recovering and has experimentally studied the restoration of *suitable* s-gram within the framework of certain theoretical-probabilistic models. Moreover, the author has conducted theoretically studies on the evaluation of probability distributions that arise in the problem of reconstructing discrete s-grams of a message.

Content of the thesis

The thesis consists of a glossary, an introduction, 3 chapters, a conclusion, a list of references, and an appendix.

The **first chapter** analyzes well-known approaches to the study of information characteristics of languages, including foreign standards for estimating the entropy of texts. An overview of language models that are most widely used in practice to describe texts in natural language is given. The s-gram language models, problems of estimating the probabilities of s-grams in the text corpus and the problem of out-of-vocabulary (OOV) s-grams are considered. The methods of increasing the s-gram dictionaries coverage are analyzed. Estimates of the English corpora sizes necessary to achieve optimal coverage of word bigrams are considered. Markov models are studied as approximate models of natural language, the shortcomings of these models for describing texts in natural language are described.

A review of methods for estimating the entropy of a random sequence from the American standard NIST SP 800-90B and the Belarusian standard MI.10127.10.02 is carried out. It has been established that the analyzed methods have several limitations for their use to assess the entropy of short s-grams of natural language texts, since language, as a rule, has a high structural predictability and contains a lot of repetitions and generally accepted phrases. In addition, entropy estimation

methods from standards are typically designed to analyze large amounts of data and require a large sample for statistical significance, making them difficult to apply for analyzing the entropy of short texts. The need to develop a method for estimating the entropy of short s-grams is revealed.

It also provides an overview of known approaches to estimating the marginal entropy for natural language texts. The approaches of Shannon and Kolmogorov and their experiments are considered. Several other studies on the estimation of marginal entropy for the English and Russian languages are also analyzed, which correct and clarify the previously obtained data. In *Table 1* An analysis of known estimates is given.

Table 1 – Entropy limit values H_0 (bit/symbol)

English language			
Author	*Corpus*	*Alphabet*	*Entropy, H_0*
Shannon [8], 1951	One literary text (Dumas Malone "Jefferson the Virginian")	26 letters and a space	0.6 – 1.3
Brown [2], 1992	General Language (Brown Corpus)	all printed ASCII characters (95 characters)	1.75
Teahan & Cleary [9], 1996	One literary text (Dumas Malone "Jefferson the Virginian")	26 letters and a space	1.46
Kontoyiannis [6], 1997	Literary texts (4 novels by J. Austen)	all printed ASCII characters (95 characters)	1.77
Calin [3], 2020	Poetry of D. G. Byron	26 letters, 10 digits, space and other characters (63 characters in total)	1.37
Project No338 MIEM [18], 2021	Literary texts	26 letters, space, period, comma	1.32-1.72
Russian			
Author	*Corpus*	*Alphabet*	*Entropy, H_0*
Piotrovsky [14], 1968	Language in general	32 letters, space	0.87-1.37
	Literary texts	32 letters, space	0.87-1.19
	Journalism and Scientific and Business Speech	32 letters, space	0.59-0.83
Kolmogorov [5], 1968	One literary text (Aksakov S.T. "Childhood of Bagrov the Grandson")	31 letters and space	1.0-1.2
Project No338 MIEM [18], 2021	Literary texts	32 letters, space, period, comma	1.49-1.89

Source: results of research by Shannon (1951), Brown (1951), Teahan (1996), R. G. Piotrovsky (1968), Kolmogorov (1968) and others.

In the **second chapter**, methods and algorithms for analyzing the information characteristics of texts in Russian and English are developed: the construction of corpora and dictionaries of s-grams, the development of methods for estimating the entropy of short s-grams and long texts, the development of models for the coverage of s-gram dictionaries and methods for estimating the permissible polysemy that appears during the restoration of individual sections of messages. An algorithm for the formation (normalization) of corpora of texts and its software implementation in C++ are developed. Four corpora of texts for the study are created: the journalistic corpus of the Russian language, the literary corpus of the Russian language, the general corpus of the Russian language and the general corpus of the English language. Within the framework of the s-gram language model, an algorithm for the formation of s-gram dictionaries from the created corpora and its software implementation in the C++ language are developed. Dictionaries of 10-grams, 15-grams, 20-grams and 25-grams are formed for corpora of various styles and languages. A comparative experimental study of the volume of s-gram dictionaries for 4 corpora is carried out. The dependence of the dictionary volume on the size of the text corpus is analyzed. Using the Wolfram Mathematica system, we interpolate this relationship for different corpora.

The first theoretical coverage model is based on the number of s-grams that occur once. The second theoretical coverage model is based on finding the distribution of the number of "empty boxes" in the classical allocation problem. For the second coverage model, an asymptotic analysis of the formula in the case of $s \to \infty$. It is established that within the framework of this model, the number of valid s-grams of the language that are not present in the dictionary is distributed asymptotically normally. Numerical estimates of the s-gram dictionaries coverage for 4 corpora are given. An experimental method for estimating the coverage of s-gram dictionaries is implemented to test the proposed theoretical coverage models.

A method for estimating the entropy of short s-grams based on the power characteristics of s-gram dictionaries and the coverage value of these dictionaries is

developed. The ideas of the method go back to A. Kolmogorov's combinatorial approach to estimating the amount of information based on the statistical characteristics of the text, without using a probabilistic model. Extrapolations of the entropy of short s-grams are carried out. Numerical estimates of the marginal entropy for different corpora are obtained.

The boundaries of the polysemy of s-gram reconstruction are investigated. An approach to estimation based on the calculation of the average number of word fragments in the s-gram is proposed. A parametric form of theoretical assessment is derived. An experimental study of the boundaries of the s-gram reconstruction polysemy is carried out to verify theoretical assumptions.

Conclusions on Chapter 2

1. Based on the results of an experimental study of the power characteristics of s-gram dictionaries on different textual material, the following conclusions can be drawn:
   - The sizes of s-gram dictionaries for Russian and English are different: English has a smaller dictionary. However, with an increase in *s*, this difference is smoothed out (for 10 grams it is about 25%, for 25 grams it is less than 6%). This fact is explained by the decrease in the saturation of the corpus within the framework of the s-gram model of the language with an increase in *s*.
   - For different styles of texts within the same language, there are some differences in the size of dictionaries. For example, for the Russian language, journalistic texts (news and articles on political topics), literary texts (Russian literature of the XIX-XX centuries) and texts of mixed styles were studied. For all s-gram lengths, journalistic texts have the smallest vocabulary, and literary texts have the largest vocabulary (for 10-grams, the difference in the size of dictionaries was more than 15%). However, with an increase in *s*, the difference is smoothed out, which is also explained by a decrease in the saturation of the corpus within the s-gram model of the language with an increase in *s*.

2. The study of the relationship between the size of the original corpus and the volume of the corresponding dictionaries allows us to conclude the following:
    - For the English 10-gram dictionary, it is possible to observe a slight decrease in the growth rate of the dictionary with a linear increase in the volume of the original corpus. For other s-gram lengths, there is an almost linear relationship between the size of the dictionary and the volume of the corpus.
    - For the Russian language, for all s-gram lengths, there is a close to linear relationship between the size of the dictionary and the volume of the corpus.

This conclusion means that the compiled corpora are not saturated within the framework of certain s-gram models under consideration. That is, an increase in the volume of source material leads to a proportional replenishment of dictionaries with new elements. The issue of corpora saturation is closely related to an important information characteristic: coverage of corpora or dictionaries.

3. The study results of coverage models and experimental verification allow us to conclude the following:
    - The coverage of dictionaries from the journalistic corpus turned out to be higher than the coverage of dictionaries from the literary corpus of the same volume (according to the results of the experiments, the coverage of 10 grams is 2 times more). This fact is explained by the lower lexical variety of journalistic texts and the presence of frequently repeated phrases in comparison with literary texts.
    - An increase in material, as expected, leads to an increase in coverage. For example, according to both models for English, a 10-fold increase in the corpus (from 10^7 to 10^8 characters) leads to an almost twofold increase in coverage.
    - With the same volume of source corpora for the English language, it is possible to achieve greater coverage of s-gram dictionaries. For example, when studying dictionaries from a corpus of 10^7 characters, the coverage of 10-grams for the Russian language was 20%, and for English was 33%.

- The accuracy of the estimates of the proposed coverage models increases with increasing *s*.

2. According to the results of the study, the marginal entropy level was 0.78-0.91 bits per character for the Russian language and was 0.55-0.87 bits per character for the English language. It should be noted that in the experiments on the study of the entropy (Russian prose) conducted by Kolmogorov, an estimate of the marginal level of entropy was also obtained starting from H_{50} [17]. A comparison of the estimates of the marginal level of entropy in this dissertation study and the previously obtained results of Piotrovsky and Shannon is given in Table *2*.

Table 2 – Known estimates of the entropy limit H_0

Evaluation	Russian Language: Journalism and Scientific and Business Speech	English language
known results (Piotrovsky and Shannon, respectively)	0.59-0.83 bits/character	0.6-1.3 bits/ character
obtained in this study	0.59-0.89 bits/ character	0.55-0.87 bits/character

Source: results obtained by the author, as well as the results of R. G. Piotrovsky (1968) and K. Shannon (1951)

In the **third chapter**, a method for restoring discrete parts of incoming message using the information-theoretic approach is developed and investigated, including theoretical and experimental methods for estimating the share of recovered message s-grams. An analysis of methods for the complete restoration of the original message is carried out, the conditions for the impossibility of complete restoration of the message are clarified. A model of a communication channel is introduced, at the output of which the theoretical-probabilistic distributions of the appearance of information about the symbols of the input message are set. Within the framework of the communication channel model, a method for restoring discrete s-grams of a message is developed, based on the search for suitable s-grams and the recovery of s-grams from a pre-formed dictionary within the boundaries of permissible polysemy. The labor intensity of this method is estimated.

To conduct an experimental study of the method, a software implementation of the algorithm for restoring individual sections of a message in the C++ language is

developed. Theoretical-probabilistic models of the appearance of symbol values at the output of a communication channel are introduced: discrete uniform distribution, polynomial distribution, and distribution arising in the case of multiple use of a key in stream transformations are simulated. The study is carried out within the framework of various probability-theoretical models of the appearance of output sets. An experimental study is conducted to restore suitable s-grams and the average share of s-grams that can be restored at the output of the communication channel is estimated. The experimental study has the following structure:

1. Investigation of s-gram recovery within the given probability distributions of the appearance of symbol values at the output of the communication channel.
    a. Restoring the s-grams of a message that is part of the original corpus of dictionaries (i.e., the true s-gram is always present in dictionaries).
    b. Restoring of s-grams of a message that **is not** part of the original corpus of dictionaries (i.e., the true s-gram may not be present in dictionaries due to incomplete coverage).
2. Investigation of s-gram recovery in cases of repeated use of the key in stream transformations.
    a. The probability of symbols coincidence of the unknown message with one of the symbols of the known message.
    b. Restoring of the s-gram of a message that is part of the original corpus of dictionaries (i.e., the true s-gram is always present in dictionaries).

In addition to the experimental study, theoretical estimates of the share of reconstructed s-grams are obtained, which depends on the probability of the appearance of a suitable s-gram in the message at a fixed critical sampling boundary and the probability that the recovery of a suitable s-gram does not exceed the limit of permissible polysemy at fixed parameters. A theoretical estimate of the probability of the suitable s-grams appearance in a message within the framework of a given theoretical-probabilistic model at the output of a communication channel

is obtained. An evaluation model is built using a normal distribution. A theoretical estimate of the probability of the suitable s-grams appearance in case of repeated use of the key in stream transformations is obtained. It is shown that such a probability can be estimated from above and below using a binomial distribution, the parameter estimates of which are obtained in the experimental study in step 2a. An estimation model is constructed using a hypergeometric distribution, and an asymptotic analysis of the obtained formulas is carried out. Experimental verification of the proposed models for approximation of probability distributions is carried out. Numerical calculations of these models are performed.

Conclusions on Chapter 3

Based on the results of the study described in this chapter, the following conclusions can be drawn.

1. Research within the framework of a given theoretical-probabilistic model without considering the dictionary coverage:
    - For the Russian language, the examination of 10-grams allows us to restore no more than 0.05 share of the s-gram message. When considering 15-grams, it is possible to reconstruct about 0.50 share of the s-grams of the original message, increasing the critical boundary for the selection of suitable s-grams (16 or more characters). When considering 20- and 25-grams, an increase in the critical boundary leads to a corresponding increase in the share of recovered s-grams to 1. It should be borne in mind that the coverage of experimental dictionaries of 20 and 25-grams does not exceed 4% and 3%, respectively.
    - For English, considering 10-grams resulted in the recovery of no more than 0.05 share of s-grams. When looking at 15-grams, it is possible to recover just under 0.25 share of s-grams. When considering 20- and 25-grams, the average share rises to 1 (with an increase in the critical boundary).

2. Research within the framework of a given theoretical-probabilistic model, considering the possible incompleteness of the dictionary coverage:
    - For the Russian language, the examination of 15-grams made it possible to reconstruct the largest share (0.17) of s-grams of the message compared to other s-gram lengths. When considering other s-gram lengths, the recovered share does not exceed 0.05.
    - For English, looking at 15-grams and 20-grams allows us to reconstruct the largest share (0.16 and 0.17 respectively) of the s-gram message compared to other s-gram lengths. When considering other s-gram lengths, the recovered share does not exceed 0.03.
3. Based on the results of the experiments, the following conclusions can be drawn:
    - The probability of matching the characters of message texts for English is slightly higher than for Russian.
    - With the double use of the key, it is possible to recover less than 0.0001 share of 20-grams and 25-grams in Russian (excluding the coverage of dictionaries); with three uses of the key, it is possible to recover about 0.05 of 25-grams; with a quadruple is about 0.22, with a five-fold is about 0.66 (excluding the coverage of dictionaries).

Conclusions on the thesis

1. The share of recovery of discrete s-grams of a message under certain parameters can reach more than 0.9, but this result does not take into account the coverage of the dictionaries used. Nevertheless, in several practical cases, it is possible to limit a set of input messages to a narrow class in which the texts are standard in their structure and are limited to a few repetitive topics (for example, the information-theoretic properties of the ground control language were previously investigated). In this case, it is possible to achieve coverage of dictionaries that is close in full under the conditions under consideration.

2. In more general cases, where there are no significant restrictions on the content and structure of the text, it is difficult to achieve practically acceptable coverage, especially for 20- and 25-grams. The imposition of some restrictions on the style and subject matter of texts makes it possible to increase the coverage of dictionaries (for example, the coverage of 10-grams of literary texts is 2 times less than the coverage of journalistic texts limited to political topics).
3. The highest coverage is achieved when considering 10-grams (40% for Russian and 54% for English), even without imposing stylistic restrictions on the texts of the original corpus. But from a practical point of view, as this study has shown, considering 10-grams allows us to recover only a small share of s-grams due to high polysemy.
4. Considering the exponential decrease in the coverage of the s-gram dictionary with an increase in *s*, the largest share of reconstructed s-grams is obtained by considering 15-grams for the Russian language (up to 0.17) and 20- and 25-grams for English (0.16 and 0.17, respectively).
5. The choice of the critical boundary for the selection of suitable s-grams depends on the language of the source text (for example, the strength of its alphabet). For example, within the framework of a given theoretical-probabilistic model of the appearance of sign variants, the choice of the critical boundary for the selection of s-grams, equal to 14 characters, made it possible to restore about 0.5 share of the message s-grams for the Russian language and more than 0.8 for the English language (without considering the coverage of dictionaries).
6. In
7. Table 3 the results of the 15-grams restoration of the massage that are not part of the original corpus within the framework of a given theoretical-

probabilistic model (that is, taking into account the influence of incomplete coverage of dictionaries) are presented.

Table 3 – Share of reconstructed 15-grams, considering possible incomplete dictionaries

Critical boundary	*Share of 15-grams*	
	Russian	English language
12 characters	0.03	0.13
16 characters	0.17	0.16

Source: calculated by the author.

8. The difference in the information-theoretical properties of the languages and the power characteristics of the original alphabets determines a different share of reconstructed s-grams with identical parameters.

Theses for approval

1. Algorithms for normalizing text corpora and building s-gram dictionaries, which make it possible to form an applied base and prepare empirical material for analyzing the characteristics of texts in natural language.
2. Dictionary coverage models that provide numerical estimates of the information characteristics of s-grams in natural language.
3. Method for estimating the entropy of short s-grams based on the power characteristics of s-gram dictionaries. This method implies an approach based on the statistical characteristics of texts, without the use of a probabilistic model for building estimates.
4. Estimates of the permissible polysemy of s-grams arising during the restoration of texts in Russian and English. This result allows us to consider the influence of language variability on text recovery algorithms.
9. Method for restoring discrete parts of input messages based on the information-theoretic approach. This method allows us to search for and restore suitable s-grams of the message and is applicable if it is impossible to fully restore the message.

Conclusion

As part of the thesis:

1. Algorithms for normalizing text corpora and building s-gram dictionaries were developed, which make it possible to form an applied base for analyzing the characteristics of texts in natural language based on the information-theoretic approach.
2. Mathematical models of s-gram dictionary coverage based on the information characteristics of natural language texts were proposed and investigated.
3. A method for estimating the entropy of short s-grams based on the power characteristics of s-gram dictionaries was developed.
4. Studies of the boundaries of the permissible polysemy of the s-grams restoration were carried out.
5. A method for restoring the discrete parts of input messages using the information-theoretic approach were developed.
6. Estimates of the share of reconstructed s-grams using dictionaries for Russian and English were obtained within the framework of a certain theoretical-probabilistic model of the appearance of symbol values at the output of the channel and in cases of repeated use of the stream transformation key.
7. Theoretical estimates of the share of reconstructed s-grams of the message within the framework of various probabilistic models were obtained.

Publications and approbation

Author's publications in Scopus/WoS:

1. Malashina A. Possibility of Recovering Message Segments Based on Side Information about Original Characters // Doklady Mathematics. 2024. Vol. 108. No. Suppl 2. (**Scopus Q2** / WoS SCIE) *(translated from Russian)*
2. Malashina A. G. O vozmozhnosti vosstanovleniya otrezkov soobshcheniya po informacii o znacheniyah iskhodnyh simvolov [On the possibility of recovering message based side on information about the original characters]. Doklady Rossijskoj Akademii Nauk. Mathematika, Informatika, Processy Upravlenia. 2023. T. 514. № 2. Pp. 138-149. (**HSE List B**)
3. Malashina A. The Combinatorial Analysis of n-Gram Dictionaries, Coverage and Information Entropy based on the Web Corpus of English // Baltic Journal of Modern Computing, 2021, vol.9, No3, pp. 363-376. (**Scopus Q4**)
4. Malashina A. G., Los A. B. Postroenie i analiz modelej russkogo yazyka v svyazi s issledovaniyami kriptograficheskih algoritmov [The construction and analysis of the Russian language models for a cryptographic algorithm research] // Chebyshevskii Sbornik, 2022, Vol. 23, No2, pp. 151-160. (**Scopus Q3**)

Publications in other publications:

5. Malashina A. G. Razrabotka instrumental'nykh sredstva dlya issledovaniia informatsionnykh kharakteristik estestvennogo yazyka [Software development for the study of natural language characteristics]. Industrial Automatic Control Systems and Controllers, 2021. № 2, pp. 9-15. *(Higher Attestation Commission K2)*
6. Malashina A. G. Modifikaciya odnogo algoritma vosstanovleniya tekstovyh soobshchenij i matematicheskaya model' raspredeleniya chisla osmyslennyh tekstov [Modification of one algorithm for the restoration of text messages and a mathematical model for the distribution of the number of meaningful texts] // Electronic means and control systems: materials of reports of the XVI

International Scientific and Practical Conference (November 18–20, 2020): in part 2. Tomsk: V-spectrum, 2020. Pp. 85-88. *(RSCI)*

7. Malashina A., Los A. The construction and analysis of the Russian language models for a cryptographic algorithm research // Algebra, number theory and discrete geometry: modern problems, applications and problems of history. Tula: TSPU, 2020. P. 177-181. *(RSCI)*

Reports at conferences:

1. Interuniversity Scientific and Technical Conference of Students, Postgraduates and Young Specialists named after E.V. Armensky (Moscow, February 2023) with the report "Study of the entropy of texts in natural language by the method of compiling n-gram dictionaries".
2. Entropy 2021: The Scientific Tool of the 21st Century (Portugal, May 2021) with a poster presentation "Entropy analysis of n-grams and estimation of the number of meaningful language texts".
3. XXI All-Russian Competition-Conference of Students and Postgraduates on Information Security "SIBINFO-2021" (Tomsk, April 2021) with the report "Mathematical Model of the Algorithm for Restoring Discrete Parts of a Text Message".
4. RusCrypto'2021 (Solnechnogorsk, March 2021) with the report "Algorithm for restoring certain parts of text messages based on information about possible variants of its characters".
5. Interuniversity Scientific and Technical Conference of Students, Postgraduates and Young Specialists named after E.V. Armensky (Moscow, March 2021) with the report "Limit Distributions Arising in the Problem of Restoring Discrete Segments of a Text Message".
6. XVI International Scientific and Practical Conference "Electronic Means and Control Systems" (Tomsk, November 2020) with the report "Modification of one algorithm for restoring text messages and a mathematical model for the distribution of the number of meaningful texts".

7. XVIII International Conference "Algebra, Number Theory and Discrete Geometry: Modern Problems, Applications and Problems of History" (Tula, September 2020) with the report "Construction and Analysis of Russian Language Models in Connection with the Research of Cryptographic Algorithms".
8. All-Russian Competition-Conference of Students and Postgraduates on Information Security "SIBINFO-2020" (Tomsk, April 2020) with the report "Modification of one algorithm for recovering text messages and a mathematical model for the distribution of the number of meaningful texts".
9. Interuniversity Scientific and Technical Conference of Students, Postgraduates and Young Specialists named after E.V. Armensky (Moscow, February 2020) with the report "Statistical Analysis of Language Models of the Russian Language Based on the Textual News Corpus".

Presentations at seminars by the HSE MIEM Academic Supervisor and the HSE Department of Computer Security:

1. Seminar dated April 23, 2024, report: "Study of information characteristics of natural languages in connection with the development of methods for evaluating secure information systems".
2. Seminar dated April 18, 2023, report: "Study of information characteristics of natural languages in connection with the development of methods for evaluating secure information systems".
3. Seminar dated November 25, 2021, report: "Study of information characteristics of natural languages in connection with the development of methods for evaluating secure information systems".

Results of intellectual activity:

1. Certificate No RU2022662474 on the registration of the computer program "Program for the restoration of certain sections of the message based on information about possible symbols of its signs", 2022.

2. Certificate No RU2020665906 on the registration of the computer program "Program for creating n-gram dictionaries and calculating their information characteristics", 2020.

Projects:

The results of the dissertation research were also used in the implementation of projects under the supervision of the author:

1. MIEM Student Project No338, "Research on the Information Characteristics of Natural Languages[1]", 2020-2021.
2. Project "Study of the Information Characteristics of Natural Languages" within the framework of the "Project Fair", 2020.

[1] The results of the project are available in the publication *by Nagayev I. E., Savchenkova D. M. Study of Information Characteristics of Literary Texts and Their Translations // Interuniversity Scientific and Technical Conference of Students, Graduate Students and Young Specialists named after E. V. Armensky. – 2021. – P. 241-244.*

References used in the summary

1. Babash A. V. Attacks on the Random Gamming Code // Mathematics and Mathematical Modeling, 2020, no. 6, pp. 35–38.
2. Brown P., Della Pietra V., Della Pietra S., Lai J., Mercer R. An estimate of an upper bound for the entropy of English // Computational Linguistics, 1992, vol. 18(1), pp. 31–40.
3. Calin O. Statistics and machine learning experiments in English and Romanian poetry // Sci 2(4), 2020, https://doi.org/10.3390/sci2040092
4. Florencio D., Herley C. A large-scale study of web password habits // Proceedings of the 16th international conference on World Wide Web, Association for Computing Machinery: New York, USA, 2007, pp. 657–666
5. Kolmogorov A. Three approaches to the definition of the notion of amount of information // Shiryayev, A.N. (eds) Selected Works of A. N. Kolmogorov, 1993, pp. 184-193.
6. Kontoyiannis I., Algoet P. H., Suhov Y. M., Wyner, A. J. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text // IEEE Trans. Inf. Theory 1998, 44, pp. 1319–1327.
7. Shannon C. E. A mathematical theory of communication // The Bell system technical journal, 1948, 27(3), pp. 379–423.
8. Shannon C. E. Prediction and entropy of printed English // Bell system technical journal, 1951, 30(1), pp. 50–64.
9. Teahan W., Cleary J. The entropy of English using ppm-based models // Proceedings of Data Compression Conference - DCC'96, IEEE, Snowbird, UT, USA, 1996, pp. 53–62.
10. Deundyak V. M., Pashkova E. A. Matematicheskie voprosy kriptoanaliza shifra gammirovaniya v sluchae serii oshibok operatora tipa propusk-povtor [Mathematical issues of cryptanalysis of the gamma cipher in the case of a series of errors of the skip-repeat operator]. URL: http://scholar.googleusercontent.com/scholar?q=cache:KJEuFdgLDG0J:scholar.google.com/+Math+Questions+Cryptanalysis+Cipher+Gamma+In+Case+Ser

ies+Errors+Operator+Type+Skip-Repeat+&hl=ru&as_sdt=0.5 (Accessed: 13.02.2024)

11. Dissertation Council on Engineering Sciences and Applied Mathematics. Passport of the Science Field "Engineering Sciences and Applied Mathematics", approved by the Academic Council of the Higher School of Economics on February 2, 2018. URL: https://www.hse.ru/mirror/pubs/share/335373657
12. Los A. B., Nesterenko A. Yu., Rozhkov M. I. Kriptograficheskie metody zashchity informacii [Cryptographic methods of information protection] // Urait Publishing House, 2016, pp. 56-65.
13. Malashina A. G. O vozmozhnosti vosstanovleniya otrezkov soobshcheniya po informacii o znacheniyah iskhodnyh simvolov [On the possibility of restoring the segments of the message based on the information about the values of the initial symbols]. Mathematics, Informatics, Control Processes, 2023, vol. 514. No 2, pp. 138-149.
14. Piotrovsky R. G. Informatsionnye izmereniya yazyka [Information measurements of language] // Moscow: Nauka Publishing House, 1968.
15. Proskurin G. V. Printsipy i metody zashchita informatsii [Principles and methods of information protection] // MIEM: Moscow, 1997.
16. Shannon K. Raboty po teorii informatsii i cybernetike [Works on the theory of information and cybernetics] // Moscow: Izdatelstvo inostrannoi literatury, 1963, pp. 669-687.
17. Yaglom A. M., Yaglom I. M. Veroyatnost' i informaciya [Probability and information]. Third edition revised and supplemented. Moscow: Nauka. 1973, pp. 236-312.
18. Nagaeva I. E., Savchenkova D. M. Issledovanie informatsionnykh kharakteristik khudozhestvennykh tekstov i ikh perevodov [Study of information characteristics of literary texts and their translations]. 2021, pp. 241-244.