UDK 004.91 http://doi.org/10.15407/pp2023.03.22

V.I. Shynkarenko, I.M. Demydovych

METHODS AND SOFTWARE FOR SIGNIFICANT INDICATORS DETERMINATION OF THE NATURAL LANGUAGE TEXTS AUTHOR PROFILE

Methods for the formation and optimization of author profiles are presented. The author profile is an image – a vector in a multidimensional space, which components are author's texts measurements by a number of methods based on 4-grams, stemming, recurrence analysis and formal stochastic grammar. The author's profile is a model of his language, including vocabulary, sentence syntax features. A comparative analysis of each of the methods effectiveness is carried out. By means of the genetic algorithm, a reduced profile of the author is formed. Insignificant indicators are excluded, which allows to reduce their number by 20%. The reduced author's profile contains attributes that are significant for this author and is an effective attribution of a particular author.
Keywords: natural language texts, authorship determination, genetic algorithm, recurrent analysis, statistical analysis, text classification, pattern recognition, formal grammars

Introduction

Attribution of authorship is the problem of identifying an anonymous text author or a text whose authorship is in doubt [1]. There are many examples in the literature of different countries, when doubts arose in the work authorship and authorship was not reliably established.

To resolve such controversial issues, an analysis of the other authors works is carried out, during which it is required to determine the significant characteristics of the text and the author's style as a whole. Subsequently, the belonging of the text to one or another author's pen will be determined by the closeness of the text under study writing style to one of them. In most cases, such a task of determining the text authorship refers to classification tasks.

There are various subtasks in text classification, and they can be divided into thematic and non-thematic. The traditional classification of texts is based on their subject matter.

However, over the past 20 years, areas of non-thematic classification have also been actively used, for example, in such subtasks as genre classification [2,5], sentiment classification, spam identification, language identification, authorship identification, and plagiarism detection [3].

Many algorithms have been developed to evaluate text authorship. These algorithms rely on the fact that the authors are characterized by the linguistic features of their own language at all levels – semantic, syntactic, lexicographic, spelling and morphological [4], which manifest themselves in the writing of texts.

As a rule, these features appear unconsciously in the authors works and thus provide a useful basis for determining authorship. The most common approach to determining authorship is to use stylistic analysis, which takes place in two stages: first, certain style markers are extracted, then, some classification procedures are applied to the resulting model.

These methods are usually based on the calculation of lexical measures representing the author's vocabulary richness and the commonly used words appear frequency [5].

The extraction of style markers is usually done using some form NLP analysis, such as tagging, parsing, and morphological analysis.

However, this standard approach has several drawbacks. First, the methods used to extract style markers are language specific. For example, the English parser is not applicable to texts in German, Ukrainian, or Chinese.

Second, feature selection is not a trivial process and usually involves setting thresholds to exclude non-informative features [6].

These decisions can be extremely subtle because although rare features contribute less signal than common features, they can still have an important cumulative effect [7].

Thirdly, modern authorship attribution systems – determining the author of a text – invariably analyze by words. However, although word-level analysis seems intuitive, it ignores the fact that morphological features can also play an important role, and in addition, many Asian languages such as Chinese and Japanese do not have well-defined word boundaries in text.

When working with a small number of authors and their works, the number of measures for comparison will also be small. However, if the number of authors or classes is much larger, it is necessary to set a limit on the amount of information about the author, i.e. create an author profile that will include only the most informative indicators from a large list of them.

At present, approaches starting with the theory of pattern recognition, mathematical statistics and probability theory, algorithms of neural networks and cluster analysis, and many others are used for text attribution.

This article solves the problem of determining the text authorship various attributions effectiveness – from the set of text attributes obtained by different methods, their subset is distinguished, which is sufficient to identify a specific author of the text. We will consider these subsets as effective attribution of a particular author.

The work is carried out on Ukrainian literary texts and explores the features of speech constructions and sentence construction that are specific to the Ukrainian language.

The allocation of effective attribution of the author is carried out on the basis of experiments with texts of different Ukrainian authors by means of a genetic algorithm.

Methods

Several methods are used to analyze the texts of different authors, form their profiles, highlight the most significant indicators, and then reduce the data of each profile to reduce the time and computational resources required during the experiment.

Below is a general scheme for highlighting the effective attribution of authors (Fig. 1).

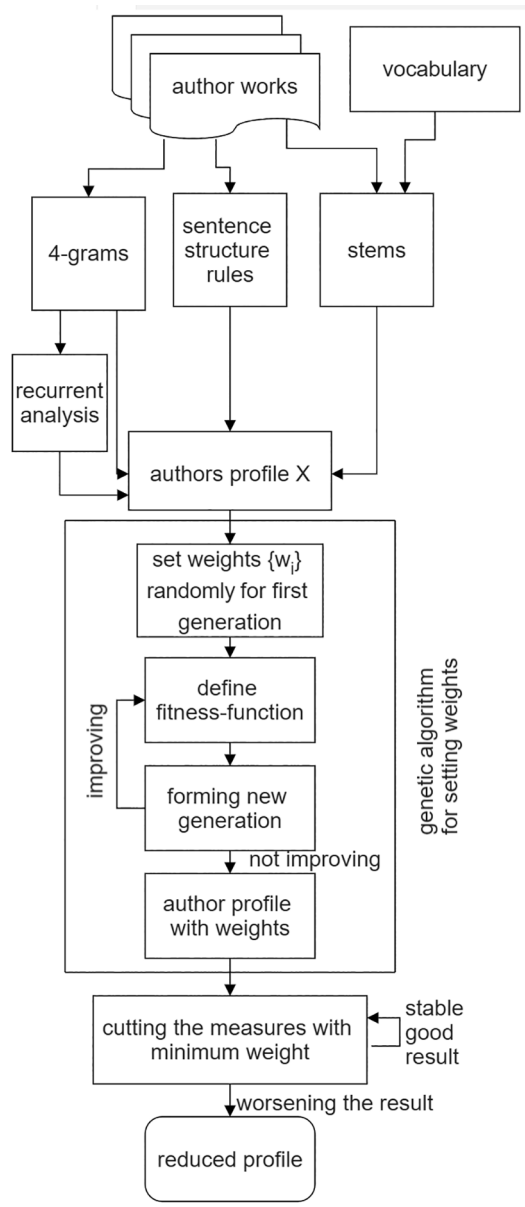

Figure 1 – General experiment scheme

In the selecting weights process for each of the indicators using a genetic algorithm, the following is performed: the initial weight vector Wk of the first generation is randomly formed, the fitness function is determined, and the best ones are selected with a crossover and mutation to form a new generation Wk.

Fitness function $\sum_{k=1}^{40} \rho(W_k' \cdot X_k)$, where X_k – is the profile of the k-th work author, W_k' – are the measurement weights corresponding to this author, ρ – is a function that experimentally determines whether the authorship of the k-th work is established correctly.

The last two steps are repeated until the improvement of the function result stops, af-

ter which the process is considered completed, and the weights are determined.

The last step is to reduce the number of indicators. xj and wj are successively eliminated such that $w_j = \min_k(w_k)$. If the result remains the same or slightly deteriorates, the profile reduction continues. As soon as the result begins to deteriorate significantly, the contraction stops and is considered complete.

Frequency analysis in creating an author profile

Frequency analysis is one of the most common text analysis methods. For many languages and a large number of authors, linguists compiled an author's language frequency dictionary or for the individual author's texts [8, 9]. The basis of such text processing is the calculation of a single character occurrence frequency for a particular text. Based on the data obtained, it can be concluded that each text will be characterized by its own individual frequency structure.

This method is based on the fact that there is a non-standard statistical distribution of characters within the text.

Practical application of this approach can be very different. A large number of works have been devoted to this problem. Also, the problems of frequency analysis occur when the process of decoding is necessary, the necessary set of data selection in large arrays, the analysis of texts that were written in ancient languages, and the conduct of categorization processes. The implementation of frequency analysis can be used in expert systems. At the same time, the frequency component underlines the measure of texts proximity.

The method of text analysis using N-grams is a relatively new method and in most cases is used to search for plagiarism in various text sources [10, 12]. This method also shows the best results in determining the authorship of texts using frequency analysis [12, 13].

In the current work, 4-grams are used due to their greatest efficiency in determining authorship in previous works [12, 13].

Based on the obtained frequencies of 4-grams, a recurrent analysis adapted for working with texts is carried out – a time series is built based on the frequency of occurrence of each 4-gram in order (advance to the next corresponding element is taken as a unit of time), on the basis of which a recursive diagram is formed. According to the resulting diagram, the following indicators are calculated: for repeating statistically similar symbols, DIV – is a value, reverse maximum length of diagonal structures; ENT – indicate the frequency distribution of the statistically similar characters repetition, LAM – indicates the repetition of statistically similar characters, TT – indicates the average frequency of statistically similar characters repetition. [12, 13].

An example of 4-grams from the work "Доля" by T. Shevchenko:

Ти не лукавила зо мною,

Ти другом, братом і сестрою…

Obtained 4-grams: тине, инел, нелу, елук, лука, укав, кави, авил, вила, илаз, лазо, азом, зомн, омно, мною, ноют, оюти, ютид,…

Using stems to form an author profile

Stemming is the process of shortening a word to its base by cutting off parts, such as an ending or a suffix. The basic concept of stemming is words with the same stem or root that refer to the same concept.

The results of stemming are sometimes very similar to determining the root of a word, but its algorithms are based on other principles. Therefore, the word after processing by the stemming algorithm may differ from the morphological root of the word. Stemming is used in linguistic morphology and information retrieval [16]. Many search systems use stemming to establish synonymous relationships if they have the same forms after stemming.

Martin Porter's stemming algorithm has become widespread and has become the de facto standard stemming algorithm for the English language.

In this work, Porter's stemmer adapted to the Ukrainian language is also used and studied from its effectiveness point of view for determining authorship [14, 15]. It is used to work directly with the texts of various authors and also to build a various stems frequency profile, specific to each author.

An example of the same passage from the work "Доля" by T. Shevchenko after stemming: т, лукав, мн, друг, брат, сестр.

Using dictionaries to create author profile

To conduct an experiment in this paper, we studied the effectiveness of using a dictionary. In general, the dictionary was developed on the basis of two approaches. The first, dictionary was the public dictionary the Large Electronic Dictionary of Ukrainian (VESUM) [17]. And the second, one was formed on the basis of Ukrainian text bank, including literacy texts, messages, posts, etc.

Based on it, a complex dictionary was built containing unique word stems, their endings and prefixes. To reduce its size, a preliminary selection of unique endings lists was carried out and only an index from it was assigned to the stem of the word. Maintaining a list of vowel alternations in words is also supported.

To create lists of prefixes for the bases, the formed dictionary was analyzed for the presence of bases that differ only in the presence of a prefix by simple enumeration. As a result, the original dictionary of bases has decreased – all key bases have been assigned the corresponding index from the list of prefixes, and the extra bases with prefixes have been removed.

The advantage of the resulting dictionary is its support for taking into account all word forms for stems, each of them will be assigned a unique index. Thus, all cases, different forms of words, as well as words obtained by adding a prefix, will unmistakably lead to a single stem.

The process of dictionary formation and its form is described in more detail in the previous works of the authors [18].

Using formal stochastic grammar to model sentence structure

Stochastic grammar is used to create rules that describe the structure of sentences in a text. For each rule, the probability of application in a particular work is determined. The probability of inferring the entire sentence is defined as the probabilities of the speech parts sequences product used in it. The resulting rules will generate a language characteristic of the processed and structurally similar a certain author works [19].

To describe the structure of the text under study, speech parts are used as a characteristic of the word. Thus, each word in the sentence is replaced by the part of speech that it is. For more information about the structure of sentences and the rules for their construction, characteristic of a particular author, reading not only parts of speech, but also forms, numbers, gender, etc. for the word under study [19].

For each speech part, its occurrence probability in a certain place of the sentence in the given text is calculated. The certain speech part appearance probability in the studied sequence will more accurately capture the each of the authors under study individual writing style characteristic. After receiving the text in the form a speech parts sequences set in sentences with the probability of their occurrence in a particular place, rules are formed. The process is described in more detail in the previous work of the authors [19].

An example of the same passage from the work "Доля" by T. Shevchenko in terms of rules:

$$\sigma \overset{\text{p1}}{\rightarrow} pronA_{1,1};\ A_{1,1} \overset{\text{p2}}{\rightarrow} partA_{2,1};\ A_{2,1} \overset{\text{p3}}{\rightarrow} vA_{3,1};$$
$$A_{3,1} \overset{\text{p4}}{\rightarrow} prepA_{4,1};\ A_{4,1} \overset{\text{p5}}{\rightarrow} pronA_{5,1};\ldots$$

where σ – is the initial nonterminal, $A_{i,j}$ the j -th nonterminal in the rule of the i -th level, p_i – is the probability of applying the corresponding rule when parsing this work.

More details are given in the work of the authors. [12].

Forming an author profile

To obtain the profile of a specific author, calculations are carried out to determine each of the studied indicators groups for all the works of the author in the training sample. Further, they are all collected in one vector X – the profile of the author.

For example, when working with 4-grams, based on the obtained indicators, a vector is formed that contains the frequency of each such 4-gram occurrence in the text. To compile the author's profile, such vectors are taken into account for each texts in the train-

ing sample and the average value for each of them is found. A similar procedure is repeated to form vectors based on the remaining groups of indicators.

An example of a vector image of T. Shevchenko based on 4-grams::

Y' = [АБАЗ, АБАЙ, АБАР, АБАС, АБАТ, АБАУ, АБЕР, АБІК, АБІЛ, АБЛА, АБОГ, АБОТ, АБОЮ, АБОЯ, АБУД, АБУД, АБУЛ, АБУС, АБУТ, АВАБ, АВАВ, АВАЛ, …].

In total, there are 8748 4-grams used in the text in the vector. And their frequencies:

$X' =$[0.0001249, 0.0001565, 0.0001249, 0.0001565, 0.0001249, 0.0001249, 0.0001565, 0.0001565, 0.0001249, 0.0001249, 0.0004998, 0.0001249, 0.0001249, 0.0001249, 0.0004381, 0.0001249, 0.0001249, 0.0001249, 0.0001565, 0.0002499, 0.0001565, 0.0004696, …].

As can be seen, there are a large number of obtained 4-grams and their frequencies, which is time-consuming and computationally expensive to work with. However, since each author has his own style of writing, different 4-grams may be most informative for different authors. In addition, often the least common letter combinations can be of the greatest importance, as they will be a characteristic feature of the author's language. Thus, the list of received frequencies requires additional analysis of their informativeness and subsequent data reduction to work with only the most significant indicators.

To optimize performance and obtain best result, when working with different indicators in the vectors, a genetic algorithm was applied to determine the weights of each of them in each group.

In this work, on the basis of all the above indicators and further determination of their weight, profiles of the authors were compiled. In total, the author's profile included four main groups according to the methods studied. Each of the groups includes a list of indicators with individual weights for each. Thus, for each author, a list of indicators was determined that most accurately reflect his author's style and allow you to identify similar elements in the texts of the control sample.

An example of a T. Shevchenko profile vector based on stems, created on the basis of the Large Electronic Dictionary of Ukrainian (VESUM):

$X' =$ […а, аа, аб, абатів, абатівськ, абатств, абатськ, абет, абетк, аби, аби-аби, абиде, абиколи, абикуди, аби-но, абискільки, абись, аби-то, абич, …].

In total, there are 7239 stems used in the text in the vector. As can be seen from the data obtained, the number of topics for analysis is as large as previous, which will also require subsequent reduction and selection the most informative of them.

Their weights for the profile T. Shevchenko:

Y' = […0.91, 0.12, 0.55, 0.08, 0.18, 0.82, 0.9, 0.85, 0.99, 0.89, 0.17, 0.86, 0.38, 0.99, 0.42, 0.58, 0.98, 0.62, 0.43, 0.34, …].

And working with the rules when creating a profile, all the rules obtained in the process of analyzing the texts in the training sample were collected in a single database, and for each of them was also found a weight. The total number of rules was 6946, the following is an example of a vector with weights for them:

$X' =$ […0.35, 0.88, 0.25, 0.44, 0.21, 0.6, 0.41, 1, 0.08, 0.2, 0.72, 0.21, 0.86, 0.49, 0.62, 0.12, 0.54, 0.14, 0.12, 0.24, …].

The number of rules is somewhat less, but still requires the selection of the most important and informative ones for the correct determination of authorship with the least expenditure of resources.

For a repeat experiment, the profile of each author was reduced for each group of indicators. The indicators with the smallest weights for each of the groups were discarded in order to reduce the time and computing power of the computer.

During the experiment, the authorship of natural language texts was determined by two samples. The sample included works of art due to the presence of the author style characteristic in them and confirmed information about their authorship, which is not subject to doubt.

For the first experiment, 40 texts of fiction by 10 Ukrainian authors were selected in the training sample. The control sample consisted of 60 texts by the same authors.

The works of the following authors are presented: IB – I. Bahrianyi, AV – A. Vyshnia,

MV – M. Vovchok, AD – A. Dovzhenko, HK – H. Kvitka-Osnovianenko, PM – P. Myrnyi, VN – V. Nestaiko, VP – V. Pidmohylnyi, IF – I. Franko, MK – M. Khvylovyi.

Attribution results

In working with a control sample, when determining the authorship of a text based on the author's profile, the following results were obtained.

Based on the data presented, working with the author's profile, the number of works with correctly identified authorship in the control sample was 54 works out of 60. The method under study made it possible to determine the authorship of most texts correctly, with some exceptions. While when comparing the profile of the following authors – Bahrianyi, Vovchok, Kvitka-Osnovianenko, Franko and Khvylovyi – one of the works was not correctly identified and showed a great similarity with the profile of another author in the sample.

During analyzing the result obtained, some similarity of styles in the two works was shown by Bahrianyi and Franko, and it can also be argued that Khvylovyi's style most often echoes the styles of other authors: in 3 cases out of 6.

Table 1 – Authorship establishing result with the full profiles

real	defined	real	defined
IB	IB	MV	MV
IB	IB	MV	MK
IB	IB	MV	MV
IB	IB	AD	AD
IB	IB	AD	AD
IB	IF	AD	AD
AV	AV	AD	AD
AV	AV	AD	AD
AV	AV	AD	AD
AV	AV	HK	HK
AV	AV	HK	HK
AV	AV	HK	MK
MV	MV	HK	HK
MV	MV	HK	HK
MV	MV	HK	MK
PM	PM	VP	VP
PM	PM	VP	VP
PM	PM	VP	VP
PM	PM	IF	IF
PM	PM	IF	IF
PM	PM	IF	IF
VN	VN	IF	IB
VN	VN	IF	IF
VN	VN	IF	IF
VN	VN	MK	MK
VN	VN	MK	MK
VN	VN	MK	MK
VP	VP	MK	MK
VP	VP	MK	MK
VP	VP	MK	IF

When excluding from the list of indicators the least significant for each author. Thus, the number of 4-grams in the profile decreased by 1750, stem by 1448 and rules by 1390, which amounted to 20% in each of the classes. When working with optimized vectors, the following results were obtained.

As a result of the experiment with a reduced author profile, the result was 53 works with correctly established authorship out of 60.

Results and discussion

As a result of the experiment using a genetic algorithm and obtaining the best solution, the following results were obtained: out of 60 texts in the control sample, the authorship of 54 works was established correctly, which amounted to a total 90%.

Table 2 – Authorship establishing result with the reduced profiles

real	elimi	real	elimin
IB	IB	MV	MV
IB	IB	MV	MHK
IB	IB	MV	MV
IB	IB	AD	AD
IB	IB	AD	AD
IB	IF	AD	AD
AV	AV	AD	AD
AV	AV	AD	AD
AV	AV	AD	AD
AV	AV	HK	HK
AV	AV	HK	HK
AV	AV	HK	MKH
MV	MV	HK	HK
MV	MV	HK	HK

MV	MV	HK	MKH
PM	PM	VP	VP
PM	PM	VP	VP
PM	PM	VP	VP
PM	PM	IF	IF
PM	PM	IF	IF
PM	PM	IF	IF
VN	VN	IF	IB
VN	VN	IF	IF
VN	VN	IF	IF
VN	VN	MKH	MKH
VN	VN	MKH	MKH
VN	VN	MKH	MKH
VP	VP	MKH	MKH
VP	VP	MKH	MKH
VP	PM	MKH	IF

For comparison in previous works and the application of these methods separately, the following results were obtained. The best indicator – 91% coincidence of the texts authorship – was obtained when working with 4-grams. Working with the basics of words using dictionaries and stemming gave a result of 88%.

As you can see, the combination of different approaches and methods did not significantly improve the result, however, it made it possible to take into account additional features of the text due to working with grammars.

Based on the data obtained, the most successful methods of working with text are 4-grams – working with them is average in terms of resources and time, relative to other methods, and gives the best result. As well as work with stochastic grammars, due to the display the features of the phrases and sentences construction by the author, however, this method requires significant computational and time resources.

The result of working with stems and dictionaries shows that they are less informative. Taking into account the high cost of these methods in calculations and time, the methods are the most expensive and the least informative among all those used.

With the exception of the least significant indicators and, as a result, a reduction in their number, the result obtained was 52 works with correctly established authorship, which is a good result – 87% the accuracy of the definition.

This approach made it possible to significantly reduce the complexity and time of calculation, while the result did not decrease significantly.

Conclusions

In the work, various approaches were explored for the formation of the general author profile: work with 4-grams, stems, recurrent analysis and sentence structure formalized by means of a formal stochastic grammar.

This approach made it possible to obtain an effective profile of the author, taking into account the various features of his personal language, from the use of individual words to the peculiarities of constructing sentences. The results obtained demonstrate the effectiveness of an integrated approach that provides better results compared to approaches that take into account individual aspects of the author's style.

References

1. H. Love. 2002. Attributing Authorship: An Introduction. Cambridge University Press.
2. Aidan Finn and Nicholas Kushmerick. 2003. Learning to classify documents according to genre. In IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis.
3. D. Khmelev and W. Teahan. 2003. A repetition based measure for verification of text collections and for text categorization. In SIGIR'2003, Toronto, Canada.
4. M. Ephratt. 1997. Authorship attribution – the case of lexical innovations. In Proc. ACH-ALLC-97.
5. E. Stamatatos, N. Fakotakis, and G. Kokkinakis. 2001. Computer-based authorship attribution without lexical measures. Computers and the Humanities, 35:193–214.
6. S. Scott and S. Matwin. 1999. Feature engineering for text classification. In Proceedings ICML-99.
7. A. Aizawa. 2001. Linguistic techniques to improve the performance of automatic text categorization. In Proceedings 6th NLP Pac. Rim Symp. NLPRS-01.
8. Darchuk N. 2023. Automatic frequency dictionary of connectivity by Lina Kostenko and Mykola Vingranovskyi. Linguistic and concep-

tual pictures of the world, 73 (1), 10.17721/2520-6397.2023.1.01.
9. Danyliuk, I., Zagnitko, A. and Sytar, G., 2019. Text corpus of Yury Shevelyov: structure, functions, navigation. APPLIED LINGUISTICS. LINGUISTICS. 10.18523/1p.2522-9281.2019.5.158-169.
10. Kuzma, K.T., 2020. Information technology for estimating the level of simslarity of strings based on the N-gram method. Academic notes of TNU named after V.I. Vernadskyi. Series: technical sciences. 31 (7), p. 96-98. 10.32838/TNU-2663-5941/2020.6-1/16.
11. H. Gómez-Adorno, JP. Posadas-Durán, G. Sidorov, Document embeddings learned on various types of n-grams for cross-topic authorship attribution. Computing 100 (2018) 741–756. doi: 10.1007/s00607-018-0587-8.
12. V.I. Shynkarenko, I.M. Demidovich Determination of the attributes of authorship of natural texts. Artificial Intelligence 3 (2018) 27-35.
13. V.I. Shynkarenko, I.M. Demidovich Authorship Determination of Natural Language Texts by Several Classes of Indicators with Customizable Weights, in: Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021). Volume I: Main Conference. Lviv, Ukraine, April 22-23, 2021, pp. 832-844.
14. T. V. Golub, M. Yu. Tyagunova, Method of steaming Ukrainian-language texts for classification of documents based on Porter's algorithm. Scientific works of Donetsk National Technical University. Series: Informatics, cybernetics and computer engineering No. 1(24) (2017) 59–63.
15. Dukhnovska KK, Strashok YaA, Shilo PV. Information technology for performing lemmatization and steming in Ukrainian-language texts. Applied systems and technologies in the information society. Pp.. 119-127.
16. S. Memon, K. Memon, F. Dehraj and others. 2020. Comparative Study of Truncating and Statistical Stemming Algorithms. International Journal of Advanced Computer Science and Applications.
17. Great electronic dictionary of the Ukrainian language (VESUM). URL: https://github.com/brown-uk/dict_uk.
18. I. Demidovich, V. Shynkarenko, O. Kuropiatnyk, O. Kirichenko, Processing Words Effectiveness Analysis in Solving the Natural Language Texts Authorship Determination Task, XVI International Scientific and Technical Conference (CSIT'2021). September 22-25, 2021, Lviv, Ukraine.
19. V. I. Shynkarenko, I. M. Demidovich Natural Language Texts Authorship Establishing Based on the Sentences Structure, in: Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2022), Volume I: Main Conference, Gliwice, Poland, May 22- 23, 2022, pp. 328-337

Received: 07.09.2023

About the authors:

Viktor Shynkarenko,
Doctor of Science, Professor,
Number of scientific publications
in Ukrainian publications – more than 200
Number of scientific publications
in foreign publications – more than 30
Index Girsh – 6
https://orcid.org/0000-0001-8738-7225
Scopus Author ID: 26635896100

Inna Demydovych,
PhD student,
Number of scientific publications
in Ukrainian publications – 4
Number of scientific publications
in foreign publications – 1
Index Girsh – 2
https://orcid.org/0000-0002-3644-184X
Scopus Author ID: 57224201949

Place of work:

Ukrainian State University
of Science and Technologies,
49010, Ukraine, Dnipro, str. Lazaryana, 2
E-mail:office@ust.edu.ua