http://doi.org/10.35784/iapgos.5423 *received: 22.10.2023 | revised: 11.02.2024 | accepted: 03.03.2024 | available online: 31.03.2024*

INTELLIGENT DATA ANALYSIS ON AN ANALYTICAL PLATFORM

Dauren Darkenbayev[1], Arshyn Altybay[1], Zhaidargul Darkenbayeva[2], Nurbapa Mekebayev[3]
[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan, [2]Kazakh Ablai Khan University of International Relations and World Languages, Almaty, Kazakhstan,
[3]Kazakh National Women's Teacher Training University, Almaty, Kazakhstan

Abstract. *The article discusses methods for processing unstructured data using an analytical platform. The authors analyze existing methods and technologies used to implement data processing and propose new approaches to solving this problem. The possibilities of using analytical platforms to solve the problem of processing source data are considered. The purpose of the article is to explore the possibilities of data import, partial preprocessing, missing data recovery, anomaly removal, spectral processing and noise removal. The authors explored how analytics platforms can function without a data warehouse, obtaining information from any other sources, but the most optimal way is to use them together, and how big data and unstructured data can be processed using an analytics platform. The authors solved a specific problem related to processing problems and proposed ways to solve them using an analytical platform. Particular attention is paid to a complete set of mechanisms that allows you to obtain information from any data source, carry out the entire processing cycle and display the results. Overall, the paper represents an important contribution to the development of raw data processing technologies. The authors plan to continue research in the field of processing big unstructured data.*

Keywords: raw data, processing, analytical platform, technology, analysis

INTELIGENTNA ANALIZA DANYCH NA PLATFORMIE ANALITYCZNEJ

Streszczenie. *W artykule omówiono metody przetwarzania surowych danych z wykorzystaniem platformy analitycznej. Autorzy analizują istniejące metody i technologie stosowane do realizacji przetwarzania danych i proponują nowe podejścia do rozwiązania tego problemu. Rozważane są możliwości wykorzystania platform analitycznych do rozwiązania problemu przetwarzania surowych danych. Celem artykułu jest zbadanie możliwości importu danych, częściowego przetwarzania wstępnego, przywracania brakujących danych, usuwania anomalii, przetwarzania spektralnego i usuwania szumu. Autorzy sprawdzili, jak platformy analityczne mogą funkcjonować bez hurtowni danych, otrzymując informacje z innych źródeł, jednak najbardziej optymalnym sposobem jest ich wspólne wykorzystanie oraz jak duże zbiory danych można przetwarzać za pomocą platformy analitycznej. Autorzy omawiają możliwe problemy związane z problemami przetwarzania i sugerują sposoby ich rozwiązania. Szczególną uwagę zwrócono na kompletny zestaw mechanizmów, który pozwala na pozyskanie informacji z dowolnego źródła danych, przeprowadzenie całego cyklu przetwarzania i wyświetlenie wyników. Ogólnie rzecz biorąc, artykuł stanowi ważny wkład w rozwój technologii przetwarzania surowych danych. Artykuł kończy się przyszłościowym planem dalszych badań w tym obszarze.*

Słowa kluczowe: surowe dane, przetwarzanie, platforma analityczna, technologia, analiza

Introduction

The modern world and information systems are data-driven, generated in huge quantities every day [1]. Therefore, to obtain truly valuable results in modern realities in industrial, business or scientific problems, it is very important to be able to effectively process the big data available to us, using the tools available for this, and some scientists use parallel data processing on many computing nodes [2].

This article contains the theoretical material necessary to understand the basic principles of how the analytical platform works, as well as solving practical problems that allow you to start working with raw data and processing it. The review material covers issues of Big Data processing.

Big Data is one of the key tools of digitalization. Their use in government administration and business began at the turn of 2010. But the relevance and possibilities of using this technology are only increasing over time [5].

The classic tool for working with large volumes of information – structured databases – cannot process such volumes of information. This was the main reason for the emergence of Big Data technology. This term refers to working with a large volume of loosely structured information stored in different formats and frequently updated. "Big Data" may include text documents, video and audio recordings, program code, etc. The main problem here is adequate analysis tools that make it possible to compare these data with each other and ensure their useful use [6].

Unlike DBMS with a set of Data Mining algorithms, analytical platforms are initially focused on data analysis and are designed to create ready-made analytical solutions.

An analytical platform is an information and analytical system, as well as a specialized software solution that contains all the tools for carrying out the process of extracting patterns from "raw" data; the process of extracting some patterns from the entire data array is carried out: a means of consolidating information in a single source (storage data), extraction, transformation, transformation of data, Data Mining algorithms, visualization, shifting simple and complex methods and models [3].

The main purpose of a scientific article is data import, partial preprocessing, restoration of missing data, removal of anomalies, spectral processing, noise removal. For this purpose, the Deductor analytical platform was used.

Deductor is a program that implements the functions of importing, processing, visualizing and exporting data. Deductor can function without a data warehouse, receiving information from any other sources, but the most optimal way is to use them together.

In recent years, there has been a lot of thought about the role of technology in the analysis of search data. When we create technologies of all types, we must always think carefully about their consequences [8]. Typically, new technologies are created to solve specific problems or meet specific needs, so we try to determine how well they will succeed in doing so. But this is not enough. We also need to consider possible downsides-the ways in which these technologies can cause harm. This is especially true for information technology, particularly technologies for creating meaning from data, but this is rarely done by the companies that produce it [4].

1. Literature review

The relevance of the selected study is studied in the works of such scientists as Franks B. [9], Lubanovic B. [11], Rastorguev V. [12], Rimmer J. [13], Saar-Tsechansky M., Provost F. [14] etc.

Today, there are many developments in data processing, such as SAS [13], Statistica Data Miner [10] etc.

However, there are problems in processing large raw data [18]. Big Data processing technologies are developing every day and many soft-ware systems are being developed for processing big unstructured data [17].

2. Formulation of the problem

Big Data is everywhere and everywhere [7]. Every day we are faced with big data and problems of processing it. There are many analytical platforms, but their mathematical model is a trade secret, so the authors decided to use a ready-made analytical platform. Using the analytical platform, the authors wanted to show the process of processing large unstructured data [16].

Deductor Studio is a program that implements the functions of importing, processing, visualizing and exporting data. Deductor Studio can function without a data warehouse, receiving information from any other sources, but the most optimal way is to use them together [15]. Deductor Studio includes a full set of mechanisms that allows you to obtain information from an arbitrary data source, carry out the entire processing cycle (cleaning, transforming data, building models), display the results in the most convenient way (OLAP, charts, trees) and export the results externally. This is entirely consistent with the concept of Knowledge Discovery from Databases (KDD) [19].

Given a data set containing columns such as "Argument", "Sinus", "Anomalies", "More noise", "Medium noise", "Small noise". The separator between columns is a tab character. The Argument column is assigned values from 0 to 2.96 in increments of 0.02. For any twenty argument values, skip entering data in sine values. The values in the Anomalies column are equal to the values in the Sine column, but have no missing data, but 10 values deviate sharply from the true value of the sine of the argument. The values for the columns "More noise", "Medium noise", "Small noise" have values close to the value of the sine of the argument, but have some deviation (dispersion and are selected from the range -1.5 to 1.5) (Fig. 1). Import data, create a file, process data, restore missing sinus values, perform partial processing, remove anomalies and noise.

Argument	Sinus	Anomalies	Big noises	Medium noises	Small noises
0.14	0.139543115	0.139543115	-0.006298541	0.126864534	0.174892379
0.16	0.159318207	0.159318207	0.33294777	0.085238625	0.145575649
0.18	0.179029573	0.179029573	0.279505602	0.085723008	0.134508346
0.2	0.198669331	0.198669331	0.258280061	0.277157184	0.176827573
0.22	0.218229623	0.5	0.139744277	0.28969863	0.231458805
0.24	0.237702626	0.237702626	0.317646146	0.177087062	0.282627785
0.26	0.257080552	0.257080552	0.266913669	0.19389017	0.288192043
0.28	0.276355649	0.276355649	0.131932825	0.325878695	0.243445006
0.3		0.2955200207	0.403496144	0.30362014	0.30123014
0.32		0.314566561	0.430780976	0.306273759	0.312952364

Fig. 1. Data set

Often the source data is not complete enough or has various noises and is not suitable for analysis, and the quality of the data affects the quality of the results. So the issue of preparing data for subsequent analysis is very important. Typically, "raw" data contains various noises, behind which it is difficult to see the overall picture, as well as anomalies – the influence of random or rare events. Obviously, the influence of these factors on the general model must be minimized, because a model that takes them into account will be inadequate.

3. Partial preprocessing

Partial preprocessing is used to recover missing data, edit anomalous values, and spectral data processing (for example, data smoothing). This step is often carried out first. In data processing, in the case of missing data, a repeating number is often written in their place, so blanks are filled, and it is better to monitor the filling of blanks. In many cases, those gaps affect the results of data processing and it can be poor quality data. Poor quality data directly leads to poor results.

4. Recovering missed data

It often happens that some data in a column is missing for some reason (the data is unknown, or they forgot to enter it, etc.). Normally, this would cause all rows that contain missing data to be removed from processing. But Deductor Studio mechanisms allow you to solve this problem. One of the partial processing steps is responsible for restoring missing values. If the data is ordered (for example, by time), then it is recommended to use approximation to recover missing values. The algorithm itself will select a value that should replace the missing value based on nearby data. If the data is not ordered, then you should use the maximum likelihood mode, when the algorithm substitutes the most probable values for the missing data, based on the entire sample.

5. Removing anomalies

Anomalies are deviations from the normal behaviour of something. This could be, for example, a sharp deviation of a value from its expected value. Automatic editing of anomalous values is carried out using robust filtering methods, which are based on the use of robust statistical estimates, such as the median. In this case, it is possible to set an empirically selected criterion for what is considered an anomaly. For example, setting the degree of suppression of anomalous data to "weak" means the most tolerant attitude towards the value of permissible emissions. Essentially, anomalies should not have any effect on the result at all. If they are present during the construction of the model, then they have a very large influence on it. They must first be eliminated. They also spoil the statistical picture of the data distribution. Data with anomalies, as well as a histogram of their distribution are presented in Fig. 2.

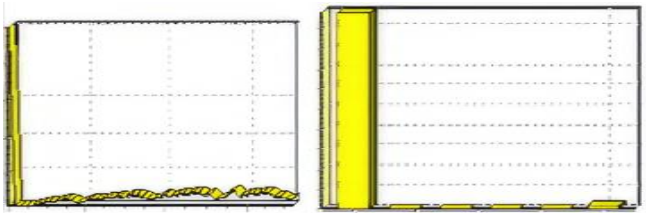

Fig. 2. Anomaly data and histogram distribution

Obviously, anomalies do not allow us to determine both the nature of the data themselves and the statistical picture. The data after eliminating the anomalies is presented in Fig. 3.

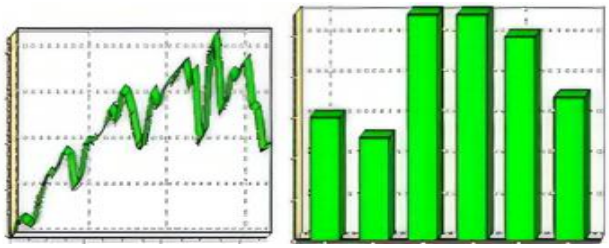

Fig. 3. Results of anomaly removal

6. Spectral processing

Data smoothing is used to remove noise from the original data set. The Deductor Studio platform offers several types of spectral processing: data smoothing by specifying the bandwidth, noise subtraction by specifying the degree of noise subtraction, and wavelet transform by specifying the depth of decomposition and wavelet order.

Spectral analysis or spectral analysis is analysis in terms of the spectrum of frequencies or associated quantities, such as energies, eigenvalues, etc. In specific fields, it may refer to:

Spectroscopy in chemistry and physics, a method for analyzing the properties of substances by their electromagnetic interactions.

Spectral estimation – in statistics and signal processing, an algorithm that estimates the strength of various frequency components (power spectrum) of a signal in the time domain. This can also be called frequency domain analysis.

A spectrum analyzer is a hardware device that measures the magnitude of the input signal as a function of frequency over the full frequency range of the instrument.

Spectral theory in mathematics is a theory that extends eigenvalues and eigenvectors to linear operators in Hilbert space and, more generally, to elements of Banach algebra.

In nuclear and particle physics, gamma-ray spectroscopy, and high energy astronomy, the analysis of pulse amplitude analyzer output for characteristic features such as spectral lines, edges, and various physical processes that create continuous shapes.

7. Noise removal

Noise in the data not only hides the overall trend, but also manifests itself when building a forecast model. Because of them, the model may turn out to have poor generalizing qualities. Spectral processing allows you to do this by specifying these fields as the processing type "Noise Subtraction" The settings have some flexibility. Thus, there is a large, medium and small degree of noise subtraction. The analyst can choose a degree that suits him. In some cases, wavelet transform gives good results for removing noise.

8. Processing results

After launching the import wizard, we will specify the import type "Text file with delimiters" and proceed to setting up the import. Let's indicate the name of the file from which we need to get data. In the viewing window of the selected file, you can see the contents of this file; we will also process the data from the "TestForPPP.txt" file.

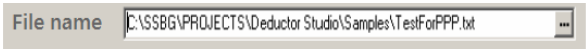

Fig. 4. Uploading a file for processing

It contains a table with the following fields: "Argument" – argument, "Sinus" – sinus values of the argument (some values are empty), "Anomalies" – sinus with outliers, "Big noises" – sinus values with large noises, "Medium noises" – sine values with medium noise, "Small noises" – sinus values with small noise. All data can be seen on the chart after importing from a text file.

After importing the file, you can see that in the "Sinus" column contains empty values. In the diagram above you can see that some sine values are missing. For further processing it is necessary to restore them. To do this, run the Partial Processing Wizard.

Since the data in the source set is ordered, at the next step of the processing wizard, select the "Sinus" field and specify the "Approximation" processing type for it. Since in this case nothing else is required, we leave the remaining processing parameters disabled. Having gone to the page for launching the processing process, we execute it by clicking on start, and then select the type of visualization of the processed data. After completing the processing process, as can be seen from Fig. 10, the gaps in the data in the diagram have disappeared, which is what needed to be done.

Next, remove anomalies from the "Anomalies" field imported table. In the partial preprocessing wizard at the third step select the "Anomalies" field and indicate to it the type of processing "Removal of anomalous phenomena", the degree of suppression "Large".

After completing the processing process, the diagram shows that the outliers have disappeared, leaving only small disturbances that can be easily smoothed out using spectral processing.

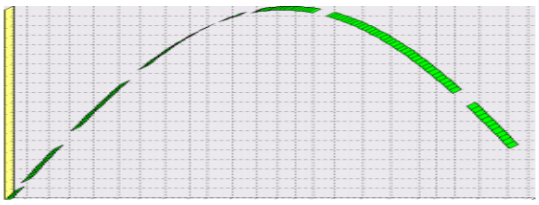

Fig. 5. Column with missing data

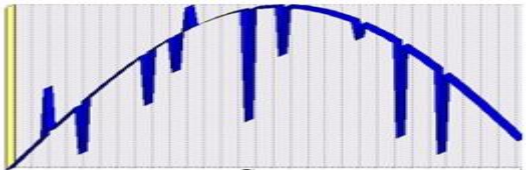

Fig. 6. Column with anomalies

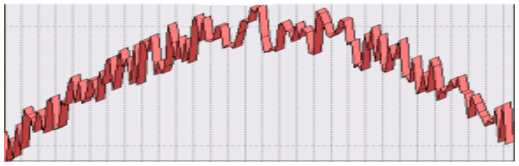

Fig. 7. Column with big ears

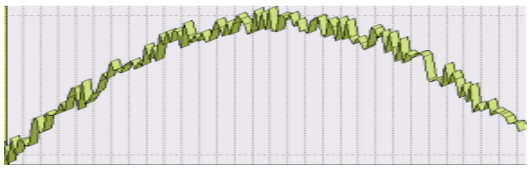

Fig. 8. Column with medium spikes

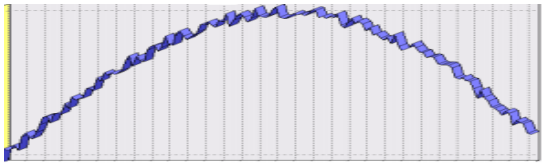

Fig. 9. Column with low noise

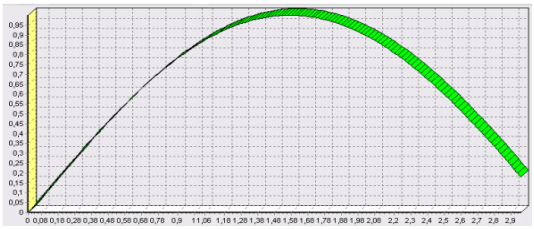

Fig. 10. Diagram after processing process

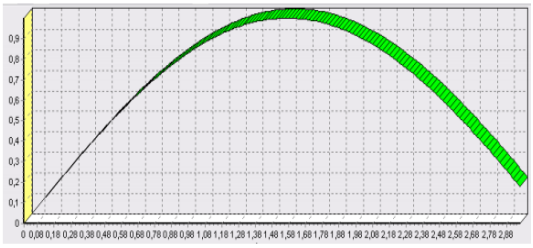

Fig. 11. Results after removing the anomaly

As can be seen in Fig. 11, the anomalies were eliminated, however minor disturbances remained. Let's smooth them out using partial processing. To do this, after removing the anomalies, we will run the partial processing wizard again. In it, in the fourth step, we select the "Anomalies" field and indicate to it the processing type "Wavelet transform" with default parameters (decomposition depth 3, wavelet order 6).

After processing, you can see in the diagram that there is no emissions and compare the result with the reference sine value (column "Sinus"). In Fig. 12 green (light) graph – sine values, blue (dark) – smoothed sine values after eliminating anomalies.

As was given in the partial processing dataset, as shown earlier, there are 3 columns of noise: "Big noise", "Medium noise", and "Small noise" – respectively a sine with large, medium and small noise. It is clear that for further work with the data, these noises must be eliminated. Thus, in the fourth step of the partial processing wizard, select the fields "Big noise", "Medium noise" and "Small noise" in turn, set the type of processing to "Subtraction of noise" and indicate the degree of suppression – "Large", "Medium" and " Small" respectively.

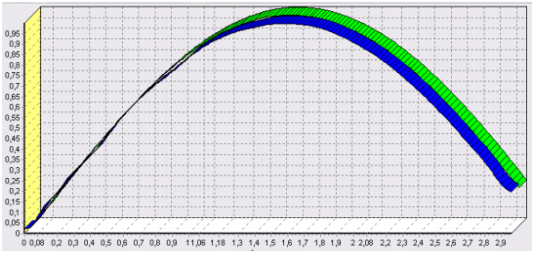

Fig. 12. Results after processing

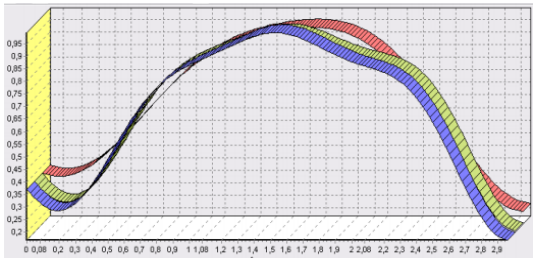

Fig. 13. Results of partial processing

Now let's remove noise using wavelet transform. In the partial processing wizard, select the "Large" Fields noise", "Medium noise" and "Small noise", indicate the type "Wavelet transform" processing, leaving the default processing parameters (decomposition depth – 3, wavelet order – 6). In the diagram you can see that the data has been smoothed (Fig. 14). The quality of noise smoothing can be improved in this way by selecting satisfactory processing parameters.

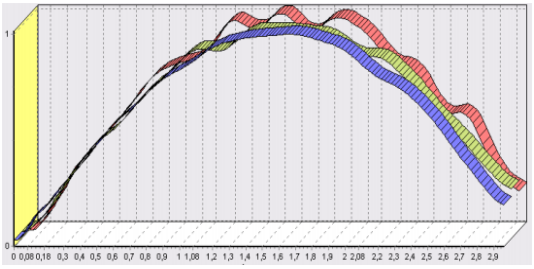

Fig. 14. Results of noise removal

9. Conclusion

The rate of data growth has become significant in the last decade. We have to work with data in different formats. Servers monitor incoming emails and transactions. It can be said that the Internet has become a major source of data storing large amounts of data. Processing data using computers is one of the main tasks of many information systems.

The technologies implemented in Deductor allow you to go through all the stages of building an analytical system on the basis of a single architecture: from data consolidation to building models and visualizing the results obtained.

Before the advent of analytics platforms, data analysis was carried out mainly in statistical packages. Their use required high user qualifications. Most algorithms implemented in statistical packages did not allow for efficient processing of large volumes of information. To automate routine operations, it was necessary to use built-in programming languages.

The article solves the problems of processing raw data. The authors concluded that such analytics platforms are needed to process unstructured data, eliminate processing errors, and obtain accurate results. The results of the study are used by the authors of this article for further modeling of big data processing and for the development of analytical platforms.

References

[1] Abdiakhmetova Z. M.: Wavelet data processing in the problems of allocation in recovery well logging. Journal of Theoretical and Applied Information Technology 95(5), 2017, 1041–1047.
[2] Altybay A. et al: Numerical Simulation and Parallel Computing of the Acoustic Wave Equation. AIP Conference Proceedings 3085(1), 2024, 020006.
[3] Balakayeva G. et al: Development of an application for the thermal processing of oil slime in the industrial oil and gas sector. Informatics, Control, Measurement in Economy and Environmental Protection 13(2), 2023, 20–26.
[4] Balakayeva G. et al: Digitalization of enterprise with ensuring stability and reliability. Informatics, Control, Measurement in Economy and Environmental Protection 13(1), 2023, 54–57 [http://doi.org/10.35784/iapgos.3295].
[5] Balakayeva G., Darkenbayev D.: The solution to the problem of processing Big Data using the example of assessing the solvency of borrowers. Journal of Theoretical and Applied Information Technology 98(13), 2020, 2659–2670.
[6] Balakayeva G. T. et al: Using NoSQL for processing unstructured Big Data. News of the NAS of the Republic of Kazakhstan 6(438), 2019, 12–21.
[7] Big Data Big Opportunity [http://www.oracle.com] (28.01.2012).
[8] Darkenbayev D. K.: Numerical solution of the regression model for analysis and processing of Big Data. Vestnik KazNRTU 6(130), 2018, 132–139.
[9] Franks B.: The Taming of Big Data: How to Extract Knowledge from Arrays of Information Using Deep Analytics. Mann, Ivanov and Ferber, 2014, 180.
[10] Highlights: Unique Features of Statistica Data Miner [http://www.statsoft.com] (01.02.2014).
[11] Lubanovic B.: Introducing Python: Modern Computing in Simple Packages 2nd Edition. O'Reilly Media, 2019.
[12] Rastorguev V.: DataMining technology for data analysis in credit scoring methods. Banking Technologies (11), 2003, 14–18.
[13] Rimmer J.: Contemporary changes in credit scoring. Credit Control 26 (4), 2005, 56–60.
[14] Saar-Tsechansky M., Provost F.: Active sampling for class probability estimation and ranking. Machine Learning 54(2), 2004, 153–178.
[15] Semenov Yu. A.: Large amounts of data (big data) [http://book.itep.ru] (21.04.2013).
[16] Usachev S.: Credit scoring: desktop or enterprise solutions. Banks and technologies (4), 2008, 50–54.
[17] [http: //www.basegroup.ru].
[18] [http://www.nosql-database.org].
[19] [https://basegroup.ru/deductor/components/studio].

Ph.D. Dauren Darkenbayev
e-mail: dauren.kadyrovich@gmail.com

Ph.D., associate professor of the Department of Computer Science, Faculty of Information Technologies, Al-Farabi Kazakh National University, Almaty, Kazakhstan.
Research interests: Big Data processing, mathematical and computer modeling, development of computer systems for the educational process, machine learning, deep learning in inverse problems.

https://orcid.org/0000-0002-6491-8043

Ph.D. Arshyn Altybay
e-mail: arshyn.altybay@gmail.com

Ph.D., associate professor of the Department of Computer Science, Faculty of Information Technologies, Al-Farabi Kazakh National University and researcher at the Institute of Mathematics and Mathematical Modeling, , Almaty, Kazakhstan.
Research interests: numerical simulation of PDE, numerical analysis, high-performance computing, numerical methods, parallel programming, machine learning, deep learning in inverse problems.

https://orcid.org/0000-0003-4939-8876

Ph.D. Zhaidargul Darkenbayeva
e-mail: zhaidargul.d@mail.ru

Candidate of Philological Sciences, associate professor at Kazakh Ablai Khan University of International Relations and World Languages.
Research interests: information technology, processing of voice and text data, semantic features of phraseological units and computational linguistics. Speech recognition using Big Data technology, machine learning, deep learning.

https://orcid.org/0000-0003-3756-0581

Ph.D. Nurbapa Mekebayev
e-mail: nurbapa@mail.ru

Ph.D. in Computer Science from Al-Farabi Kazakh National University, Almaty, Kazakhstan, in 2020. Currently, he is a senior researcher at the Institute of Information and Computing Technologies in Almaty, Kazakhstan, and an associate professor at the Department of Computer Science of the Kazakh National Women's Teacher Training University in Almaty, Kazakhstan.
Research interests: machine learning, deep learning in inverse problems and computational linguistics.
https://orcid.org/0000-0002-9117-4369