NUCET NUCLEAR ENERGY AND TECHNOLOGY

Nuclear Energy and Technology 9(2): 115–120
DOI 10.3897/nucet.9.106759

Research Article

Application of machine learning methods for filling and updating nuclear knowledge bases*

Victor P. Telnov[1], Yury A. Korovin[1]

1 *IATE MEPhI, 1 Studgorodok, 249039 Obninsk, Kaluga Reg., Russia*

Corresponding author: Victor P. Telnov (telnov@bk.ru)

Academic editor: Georgy Tikhomirov ♦ **Received** 06 July 2022 ♦ **Accepted** 1 March 2023 ♦ **Published** 20 June 2023

Citation: Telnov VP, Korovin YuA (2023) Application of machine learning methods for filling and updating nuclear knowledge bases. Nuclear Energy and Technology 9(2): 115–120. https://doi.org/10.3897/nucet.9.106759

Abstract

The paper deals with issues of designing and creating knowledge bases in the field of nuclear science and technology. The authors present the results of searching for and testing optimal classification and semantic annotation algorithms applied to the textual network content for the convenience of computer-aided filling and updating of scalable semantic repositories (knowledge bases) in the field of nuclear physics and nuclear power engineering and, in the future, for other subject areas, both in Russian and English. The proposed algorithms will provide a methodological and technological basis for creating problem-oriented knowledge bases as artificial intelligence systems, as well as prerequisites for the development of semantic technologies for acquiring new knowledge on the Internet without direct human participation. Testing of the studied machine learning algorithms is carried out by the cross-validation method using corpora of specialized texts. The novelty of the presented study lies in the application of the Pareto optimality principle for multi-criteria evaluation and ranking of the studied algorithms in the absence of a priori information about the comparative significance of the criteria. The project is implemented in accordance with the Semantic Web standards (RDF, OWL, SPARQL, etc.). There are no technological restrictions for integrating the created knowledge bases with third-party data repositories as well as metasearch, library, reference or information and question-answer systems. The proposed software solutions are based on cloud computing using DBaaS and PaaS service models to ensure the scalability of data warehouses and network services. The created software is in the public domain and can be freely replicated.

Keywords

semantic web, knowledge base, machine learning, classification, semantic annotation, cloud computing

Introduction

Nuclear science and technology are among the areas with a high intensity of information exchange and knowledge generation. Research carried out at elementary particle accelerators annually produces hundreds of terabytes of new experimental results (CERN Document Server). World nuclear data centers accumulate and systematize information on thousands of nuclear reactions and nuclear constants (Centre for Photonuclear Experiments Data). The IAEA (IAEA Nuclear Knowledge Management) and respective national agencies (Rosatom State Corporation. Knowledge Management System) create and maintain databases and knowledge bases on nuclear technology and radiation safety.

The practical contribution of the authors of this paper to the development of knowledge bases consists in the creation of working prototypes, and then scalable seman-

* Russian text published: Izvestiya vuzov. Yadernaya Energetika (ISSN 0204-3327), 2022, n. 4, pp. 122–133.

tic web portals, which are deployed on cloud platforms and are intended for use in the educational activities of universities (Telnov 2017, Telnov and Korovin 2019a, 2019b, 2019c, 2020a, 2020b). The first project (Semantic Educational Portal. Nuclear Knowledge Graphs. Intelligent Search Agents) is related to education in the field of nuclear physics and nuclear power engineering. The second project (Knowledge Graphs on Computer Science. Intelligent Search Agents) involves the study of computer sciences and programming. Both projects deal with models and methods for representing and processing problem-oriented knowledge for specific subject areas. They create and test technologies for accumulating and integrating knowledge as well as for increasing the level of competence of knowledge bases as artificial intelligence systems.

The relevance of the first project is explained by the fact that it is aimed at the creation and computer-aided filling of semantic repositories (knowledge bases) on nuclear physics and nuclear power engineering. These are areas in which Russia is able to achieve competitive advantages and world leadership. As of 2022, the educational web portals of universities, nuclear data centers, and nuclear knowledge management systems of the IAEA and Rosatom State Corporation do not use sufficiently the capabilities of the semantic web and machine learning methods.

This study is aimed at searching for and testing optimal classification and semantic annotation algorithms applied to the textual network content for computer-aided filling and updating of nuclear knowledge graphs, both in Russian and English. The corresponding optimization problem is formulated and solved below in the section on the results of computational experiments. The proposed algorithms will provide a methodological and technological basis for continuously filling and updating problem-oriented knowledge bases as artificial intelligence systems, as well as prerequisites for the development of semantic technologies for acquiring new knowledge on the Internet without direct human participation.

From a practical standpoint, the software implementation of effective classification and semantic annotation algorithms is carried out as part of a scalable semantic web portal hosted on a cloud platform. Fig. 1 shows the control panel that is used to set the parameters for the semantic annotation (classification) of an arbitrary text on the Internet. Any knowledge graphs that form a knowledge base can be used in any quantity and in any combination. Semantic annotation and classification can be performed using only ontology classes (TBox), only ontology objects (ABox), or both. The choice of preferred data processing technologies (traditional text analysis or machine learning methods) is the prerogative of the knowledge engineer who manages the process. The search for and extraction of initial text data on the Internet and their primary clustering are carried out by the “Context-sensitive search” agent, which is an integral part of the semantic portal (Semantic Educational Portal. Nuclear Knowledge Graphs. Intelligent Search Agents).

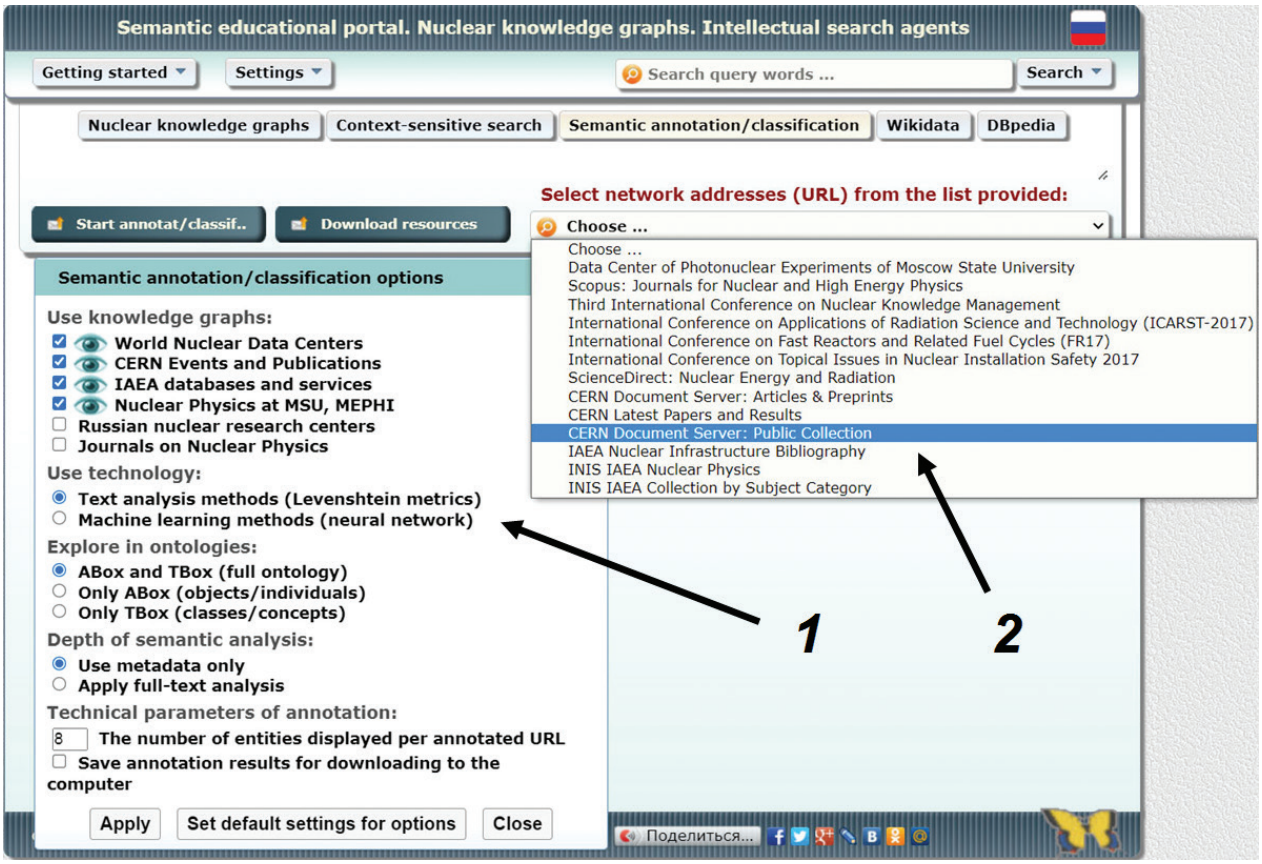

Figure 1. Setting the parameters of the semantic annotation/classification process: 1) choice of semantic annotation (classification) technology; 2) network addresses (URL) of documents to be annotated (classified).

The created online solutions are in the public domain (excluding confidential information) and can be freely replicated. The project is implemented in accordance with the Semantic Web standards (RDF, OWL, SPARQL, etc.) (W3C Semantic Web, W3C RDF Schema 1.1, W3C OWL 2 Web Ontology Language). For this reason, there are no technological restrictions for integrating the created knowledge bases with third-party data repositories as well as metasearch, library, reference or information and question-answer systems.

The scalability of semantic repositories (knowledge bases) is carried out directly by means of the cloud platform used. The scientific novelty of the approaches used in this project is defined by the use of the Pareto optimality principle, which allows for multi-criteria evaluation and ranking of the studied machine learning algorithms in the absence of a priori information about the comparative significance of the criteria.

Text data classification methods

Classifying text data refers to the tasks of Machine Learning (ML) in the field of Natural Language Processing (NLP). By 2022, at least a dozen machine learning methods had been created that were potentially suitable for solving problems related to text classification and semantic annotation Geron A (2019). There are now dozens of software implementations of these methods (Scikit-learn. Machine Learning in Python).

Naive Bayes Classifier

This classification algorithm (Naive Bayes Classifier) is considered to be one of the simplest classification algorithms. Bayes' theorem is invariant with respect to the causes and effects of events. If we know the probability with which a particular cause leads to a certain effect, Bayes' theorem allows us to calculate the probability that this particular cause has led to the observed event. This idea underlies the Bayes classifier, while the principle of maximum likelihood is used to determine the most probable class.

For natural languages, the probability of the next word or phrase appearing in the text is highly dependent on the current context. The Bayes classifier ignores this circumstance and represents the document as a set of words, the probabilities of which are conditionally independent of each other. This approach is sometimes referred to as the bag-of-words model. Despite strong simplifying assumptions, the Naive Bayes classifier performs well in many real world problems. It does not require a large amount of training data and, with moderate text corpora, is often not inferior to more sophisticated algorithms.

MaxEnt Classifier (Softmax)

If the Naive Bayes classifier ignores correlations between words, then the MaxEnt stochastic classifier allows and takes these correlations into account. From the logistic regression models corresponding to the training data, the one is selected that contains the least number of assumptions about the true probability distribution of the text data. In other words, an empirical probability distribution with the maximum information entropy is chosen. This approach is especially productive exactly for solving text classification problems, when the words in the text are obviously not independent.

The Softmax function, or normalized exponential function, is a generalization of the logistic function for the multivariate case. In multiclass classification problems, the Softmax function is built in such a way that the number of neurons on the last layer of the neural network is equal to the number of the desired classes. In this case, each neuron should give the value of the probability that the object belongs to the class, and the values of all the neurons in the sum should give unity.

The Maxent classifier usually takes more time to train compared to the Naive Bayes classifier due to the optimization that needs to be done to estimate the model parameters. After calculating these parameters, the method yields very reliable results and is competitive in terms of computing resource and memory consumption.

SVM Classifier with SGD

Support vector machines (SVM) are a set of linear binary classification methods with a simple and clear interpretation. For example, the task is to find in a multidimensional space such a surface, a hyperplane in the simplest case, which divides objects into two classes with the largest gap. The SVM classifier is equivalent to a two-layer neural network, where the number of neurons in the hidden layer is defined as the number of support vectors.

Stochastic gradient descent (SGD) is an iterative method for optimizing an objective function with suitable smoothness properties, which is widely practiced in deep learning models. Here, the gradient of the function being optimized is calculated not as the sum of the gradients from each sample element, but as the gradient from one randomly selected subset of elements. The slower convergence of the algorithm can be compensated by the high speed of iterations on large data sets.

Assessing the quality of classification algorithms

The following three proven binary (Classification Metrics) are used here as quality functionals for the machine learning algorithms.

1. Precision (classification accuracy). It is calculated as the proportion of objects that really belong to some positive class and are classified correctly.
2. Recall (completeness of classification). It is calculated as the proportion of objects that are assigned by the algorithm to some positive class and are classified correctly.
3. F1-score. This aggregated metric is calculated as the harmonic mean of the accuracy and completeness of the classification.

The first two metrics do not depend on the filling of classes with objects and therefore are applicable in conditions of unbalanced samples. The Precision metric characterizes the algorithm's ability to distinguish among classes, and the Recall metric shows the algorithm's ability to detect a particular class in general. The third metric, F1-score, is the most informative in cases where the values of the first two metrics differ significantly from each other. To assess the quality of the multiclass classification algorithms, the so-called macro-averages are used, when the metric values are averaged over all the classes, regardless of the number of objects in these classes.

Cross-validation is used to improve the reliability of classification algorithm testing results. The initial training set is randomly divided N times into N samples of approximately the same length. Each of the N samples is in turn declared a control sample, the remaining $N-1$ samples are combined into a training sample. The algorithm is tuned to the training sample and then classifies the control sample objects. The described procedure is repeated N times, and the value of N varies from 3 to 10.

Calculation results

To identify the most effective methods for classifying text data for the purpose of computer-aided filling and updating of nuclear knowledge graphs, a series of tests was carried out using corpora of specialized texts on nuclear physics and nuclear power engineering. In total, seven nuclear knowledge graphs (Semantic Educational Portal. Nuclear Knowledge Graphs. Intelligent Search Agents) were used (see Table 1). For each studied classification method and each knowledge graph, three metrics were calculated: Precision, Recall and F1-score.

Table 1. Metrics for three text data classification methods calculated using seven nuclear knowledge graphs

Nuclear knowledge graphs as training and control samples	Method								
	Naive Bayes Classifier, %			Maxent Classifier (Softmax), %			SVM Classifier with SGD, %		
	P	R	F	P	R	F	P	R	F
World nuclear data centers	55	33	42	46	51	48	87	83	85
Events and publications (CERN)	94	72	82	72	71	71	42	43	43
IAEA databases and services	97	46	62	46	51	48	56	57	56
Nuclear physicsa at MSU, MEPhI	96	57	71	78	59	67	87	82	84
Russian nuclear research centers	75	13	21	82	68	74	95	94	95
Magazines in nuclear physics	86	57	69	83	100	91	17	25	20
Combined nuclear knowledge graph	99	25	39	63	37	46	88	85	86

Note: P = Precision metric, R = Recall metric, F = F1-score metric.

The results of calculations from Table 1 in the form of three-dimensional graphs are visually presented in Fig. 2. As can be seen from the data in the table and the figure, each of the three studied text data classification methods is characterized by 14 independent quality indicators of

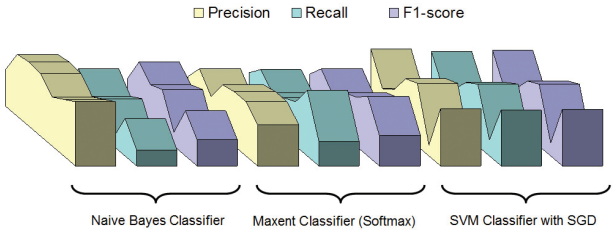

Figure 2. Visual representation of data from Table 1 in the form of three-dimensional graphs.

the Precision and Recall metrics and 7 derived quality indicators of the F1-score metric. The optimization problem is defined as follows. It is required to choose the best classification method, taking into account all the calculated quality indicators, without making any assumptions about the comparative significance of these indicators. To do this, in the class of transitive antireflexive binary relations, we shall consider the Pareto ratio in a Euclidean space. For any two elements x and y from the set W, the Pareto ratio P is determined as follows:

$$(\forall x,y\in\Omega)[xPy]\Leftrightarrow\left\{(\forall j=1,\ldots,m)\left[x_j\geq y_j\right]\&(\exists j_0\in\{1,\ldots,m\})\left[x_{j_0}>y_{j_0}\right]\right\} \quad (1)$$

The set of P-optimal elements on W is the Pareto set W^P:

$$|\ \Omega^P=\{x\in\Omega:(\forall y\in\Omega)[y\overline{P}x]\} \quad (2)$$

The Pareto ratio provides a universal mathematical model for the multicriteria context-independent choice in a Euclidean space. If we denote by d(y, x) the number of the criteria by which the element y is superior to the element x, then the value

$$D_\Omega(x)=\max_{y\in\Omega}d(y,x) \quad (3)$$

is called the dominance index of the element x when the set W is presented. Roughly speaking, the dominance index is equal to the number of the criteria by which the element x does not exceed all other elements from the set W. We shall determine the function $C^D(W)$ for choosing the best elements as follows:

$$C^D(\Omega)=\{x\in\Omega:D_\Omega(x)=\min_{x\in\Omega}D_\Omega(x)\} \quad (4)$$

The value D_W is the dominance index of the entire set W. The elements with the minimum value of the dominance index form the so-called Pareto set. The Pareto set includes elements that are best in terms of the totality of all the criteria taken into account, without any a priori assumptions about the comparative significance of these criteria. In conditions of a real choice, the Pareto set often contains more than one element.

Returning to the original task of finding the most effective method for classifying text data, we shall turn to the data in Table 2. There, for each method studied, three dominance indicators were calculated for 7, 14 and 21 metrics, respectively. The initial data for the calculations were taken from Table 1.

Table 2. Dominance indices for the three text data classification methods calculated using seven nuclear knowledge graphs

Text data classification method	Dominance index		
	Calculated by F1-score	Calculated by Precision and Recall	Calculated by F1-score, Precision and Recall
Naive Bayes Classifier	4	7	11
MaxEnt Classifier (Softmax)	5	10	15
SVM Classifier with SGD	3	7	10

As can be seen from the data in Table 2, the leader is the SVM Classifier with SGD with dominance indices of 3, 7, and 10. The Naive Bayes Classifier method is only slightly inferior to it. However, the Maxent Classifier (Softmax) method looks like an outsider as compared to the other two methods. This conclusion was obtained by carrying out computational experiments on the corpora of specific texts on nuclear physics and nuclear power engineering. In this case, moderate volumes of initial data were used. Each of the seven knowledge graphs involved contained no more than one thousand objects and no more than one hundred classes. It should be noted that the SVM Classifier with SGD is binary, i.e., it allows us to distribute elements by only two classes. This technical limitation is overcome, for example, by multiple classification according to the "one against all" or "one against one" principle. The other two methods – the MaxEnt Classifier (Softmax) and the Naive Bayes Classifier – are inherently multiclass ones, which makes them more convenient to use.

Project software architecture

The software solutions implemented in the project are based on cloud computing using DBaaS and PaaS service models to ensure the scalability of data warehouses and network services. The server scripts of the working software prototype run on the Jelastic cloud platform in a Java and Python runtime environment.

Fig. 3 shows a component diagram made according to the UML 2 standard (ISO/IEC 19505–2:2012(E) (2012)), which shows the features of the software implementation of the Semantic Annotation agent and the Semantic Classification agent as part of a scalable semantic web portal (Semantic Educational Portal. Nuclear Knowledge Graphs. Intelligent Search Agents). The created software is tested on the corpora of specialized texts on nuclear physics and nuclear power engineering, including relevant texts from the IAEA, CERN, MEPhI, the Faculty of Physics of MSU, as well as texts from specialized journals, publications of nuclear research centers and nuclear data centers.

Related works and conclusions

Research groups from Stanford University (Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D (2014)), the Massachusetts Institute of Technology, the University of Bari, the University of Leipzig, and the University of Manchester are focusing on the development of the Semantic Web, related issues of machine learning and natural language processing. The global IT giants are actively developing knowledge representation models and machine learning technologies, including IBM Watson Studio, Google AI and Machine Learning, Amazon Comprehend NLP, AWS Machine Learning, Yandex DataSphere (Jupyter Notebook), etc. Software tools for research in the field of artificial intelligence and natural language processing are

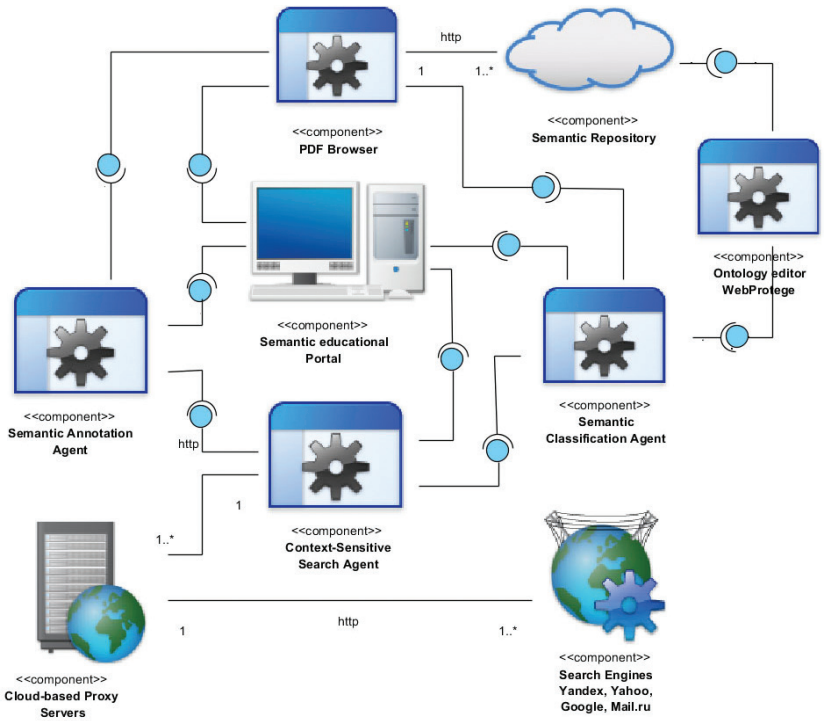

Figure 3. Diagram of the components of a scalable semantic web portal.

provided by Matlab (Machine Learning with MATLAB & Simulink), Stanford NLP (Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D (2014)), Scikit-learn (Scikit-learn. Machine Learning in Python) etc. In Russia, specialized research is carried out at the Competence Center of the National Technological Initiative of Moscow Institute of Physics and Technology (MIPT), Institute of Precision Mechanics and Optics (ITMO), the Faculty of Computational Mathematics and Cybernetics of Moscow State University (MSU), Ivannikov Institute for System Programming (ISP RAS) (Stupnikov S, Kalinichenko A (2019)), and Russian divisions of Hyawei.

In this study, seven corpora of specialized texts on nuclear physics and nuclear power engineering were used to show the effectiveness of relatively simple, intuitive machine learning methods for solving the problem of continuous filling and updating of nuclear knowledge bases without direct human participation. The SVMs and the Naive Bayes classifier ensure the competence of semantized knowledge bases as artificial intelligence systems. Some of the results that were obtained by the authors in the study of such classifiers as the "*K* nearest neighbors (*kNN*)" algorithm and terminological decision trees, which are built based on the results of text parsing, remained outside the scope of this article.

Acknowledgments

The study was supported by the Russian Science Foundation grant No. 22-21-00182, https://rscf.ru/project/22-21-00182/.

References

- Centre for Photonuclear Experiments Data (2022) CDFE Website. http://cdfe.sinp.msu.ru/index.en.html [accessed Jun. 26, 2022]
- CERN Document Server (2022) Access articles, reports and multimedia content in HEP. https://cds.cern.ch [accessed Jun. 26, 2022]
- Classification Metrics (2022) Classification Metrics. https://github.com/turi-code/userguide/blob/master/evaluation/classification.md [accessed Jun. 26, 2022]
- IAEA Nuclear Knowledge Management (2022) Nuclear Knowledge Management. https://www.iaea.org/topics/nuclear-knowledge-management [accessed Jun. 26, 2022]
- Geron A (2019) Hands-on ML with Scikit-Learn, Keras & TensorFlow. 2nd edn. O'Reilly Media, Inc. Boston.
- ISO/IEC 19505–2:2012(E) (2012) Information Technology – Object Management Group Unified Modeling Language (OMG UML) – Part 2: Superstructure. ISO/IEC, Geneva.
- Knowledge Graphs on Computer Science (2022) Intelligent Search Agents. http://vt.obninsk.ru/s/ [accessed Jun. 26, 2022]
- Machine Learning with MATLAB & Simulink (2022) MATLAB for Machine Learning. https://www.mathworks.com/solutions/machine-learning.html [accessed Jun. 26, 2022]
- Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D (2014) The Stanford Core NLP Natural Language Processing Toolkit. Proceedings of the LIInd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics, 55–60. https://doi.org/10.3115/v1/P14-5010
- Naive Bayes Classifier (2022) Naive Bayes. https://scikit-learn.org/stable/modules/naive_bayes.html [accessed Jun. 26, 2022]
- Rosatom State Corporation (2022) Knowledge management system. http://www.innov-rosatom.ru/suz-rosatoma/ [accessed Jun. 26, 2022]
- Scikit-learn. Machine Learning in Python (2022) Machine Learning in Python. https://scikit-learn.org/stable/ [accessed Jun. 26, 2022]
- Semantic Educational Portal (2022) Nuclear Knowledge Graphs. Intelligent Search Agents. http://vt.obninsk.ru/x/ [accessed Jun. 26, 2022]
- Stupnikov S, Kalinichenko A (2019) Extensible unifying data model design for data integration in FAIR data infrastructures. Proceedings of the 20th International conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2018), Springer, 17–39. https://doi.org/10.1007/978-3-030-23584-0_2
- Telnov V (2017) Semantic educational web portal. Selected papers of the XIX International conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2017), Moscow, Russia, October 9–13. http://ceur-ws.org/Vol-2022/paper11.pdf [accessed Jun. 26, 2022]
- Telnov V, Korovin Yu (2019a) Semantic web and knowledge graphs as an educational technology of personnel training for nuclear power engineering. Izvestiya vuzov. Yadernaya energetika [News of Higher Educational Institutions. Nuclear Power Engineering] 2: 219–229. https://doi.org/10.26583/npe.2019.2.19
- Telnov V, Korovin Yu (2019b) Semantic web and knowledge graphs as an educational technology of personnel training for nuclear power engineering. Nuclear Energy and Technology 5 (3): 273–280. https://doi.org/10.3897/nucet.5.39226
- Telnov V, Korovin Yu (2019c) Programming knowledge graphs, Reasoning on Graphs. Software Engineering 2: 59–68. https://doi.org/10.17587/prin.10.59-68
- Telnov V, Korovin Yu (2020a) Machine learning and text analysis in the tasks of knowledge graphs refinement and enrichment. CEUR Workshop Proceeding 2790: 48–62; Supplementary Proceedings of the XXII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2020), Voronezh, Russia, October 13–16, EID: 2-s2.0-85098723055, ISBN: 16130073. http://ceur-ws.org/Vol-2790/paper06.pdf [accessed Jun. 26, 2022]
- Telnov V, Korovin Yu (2020b) Semantic web and interactive knowledge graphs as educational technology. In: Harkut DG (Ed.) Cloud Computing Security. IntechOpen, London, ISBN: 978-1-83880-703-0. https://doi.org/10.5772/intechopen.83221
- W3C OWL 2 Web Ontology Language (2022) Document Overview (2nd edn). https://www.w3.org/TR/owl2-overview/ [accessed Jun. 26, 2022)]
- W3C RDF Schema 1.1 (2022) W3C First Public Working Draft. https://www.w3.org/TR/rdf-schema/ [accessed Jun. 26, 2022]
- W3C Semantic Web (2022) Semantic web. https://www.w3.org/standards/semanticweb/ [accessed Jun. 26, 2022]