Information systems research

UDC 621.396 **doi: 10.20998/2522-9052.2021.1.09**

Olesia Barkovska, Vladyslav Kholiev, Daria Pyvovarova, Heorhii Ivashchenko, Dmytro Rosinskiy

Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

INTERNATIONAL SYSTEM OF KNOWLEDGE EXCHANGE FOR YOUNG SCIENTISTS

Abstract. The paper proposes a system which is electronic data storage (of qualification works of students from different countries) and provides the capability to identify and connect young scientists conducting research on a related problem area. The purpose of developing this system is to provide opportunities for knowledge exchange, research in a team on a common problem, as well as to identify scientific trends in different countries. In this paper, the preprocessing **methods** influence on the work of classifiers such as Logistic Regression, LSTM, BERT, LightGBM was researched. A study was conducted on the speed of classification and F1 assessment. **Conclusions.** Lemmatization showed to require a shorter operating time compared to steaming by almost twice and a better score by an average of 5 percent, so it was decided to use the Logistic Regression classifier with lemmatization at the stage of text preparation in the subsequent operation of the proposed ISKE.

Keywords: system; NLP; text, processing; acceleration; shingles; proximity; likeness; classification; preprocessing; lemmatization; stemming.

Introduction

It is known that the accumulation of information has been going on since ancient times. From the first years of its existence, mankind has used such a natural information technology as language.

Later, along with language, people began to use imagery and writing to store and transmit information. With the development of language and general culture of peoples, different types of writing began to appear as a co called "hard copy". The main purpose of writing is storing information. Thus, the main task of writing is to record information on media and transmit it to other people. The phenomenon of the global information revolution of the late XX century led to the introduction of information technology using computers, data servers, computer servers, various types of digital media for information storage. Stages of information technology development depending on the types of tools are shown in Fig. 1.

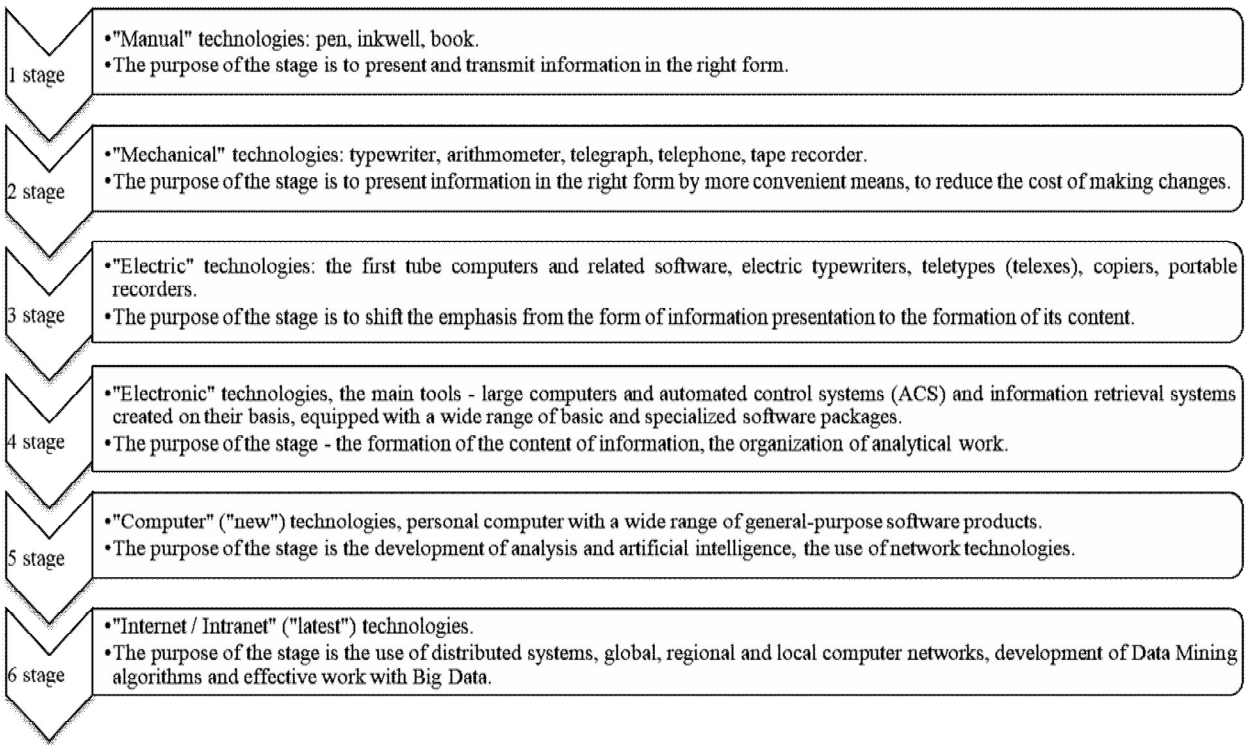

Fig. 1. Stages of information technology development

The importance of storing and accumulating information in a structured form is also evidenced by the fact that for the first time in library practice to classify hundreds of thousands of books by field of knowledge and compile catalogs with the author and the name of each book began as early as in the Library of

© Barkovska O., Kholiev V., Pyvovarova D., Ivashchenko H., Rosinskiy D., 2021

Alexandria. Today, the development of the information society has changed the library world in the direction of remote access to libraries through the creation of electronic catalogs. The main type of storage media is electronic media, which has advantages such as small size, large storage capacity, easy mobility and access to the library collection from different corners of the world. Additional features include the relative ease of structuring information as well as searching [1-4].

The problem is the continuous increase in the amount of information, which leads to such requirements as increasing the speed of search engines and systems for categorizing information.

Among the types of libraries that can be distinguished (as shown in Fig. 2) there are thematic libraries (juridical, medical, military, musical, transport, art libraries, philosophical) and specialized-corporate, i.e. those that are relevant and in demand by a group of readers with a certain status, for example – student, PhD student, researcher, young scientist. In the XXI century, a prominent feature of library development is the work of international library associations in the direction of:

– Integration;
– Development of electronic library science;
– The establishment and strengthening of international ties between libraries;
– Conducting research in the field of library science and bibliography.

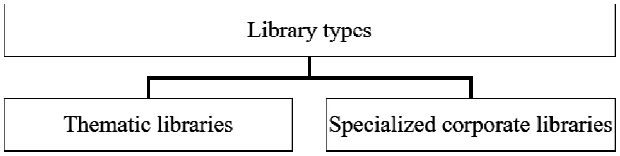

Fig. 2. Typing of libraries

The following typing is also relevant for electronic repositories, as it provides easy access to targeted data – namely, scientific and research works of young scientists for specialized corporate libraries.

Research task rationale

Libraries of educational institutions have undergone the greatest reform in recent years due to the introduction of quarantine conditions and modernization of the educational process. [4, 6]. They are actively implementing the latest technologies, increasing the volume of storage and server capacity, which provide storage and access to electronic versions of qualification works of said educational institutions' graduates. This, in turn, leads to improved search algorithms and classification of textual information [5].

The aforementioned proves the relevance of the proposed project to create an electronic library of young scientists' works with the capability for them to communicate and exchange knowledge, to perform further joint research and analyze the state of the most relevant scientific fields as well as to attract young professionals directly to industrial and practical experience corresponding to the fields their of knowledge and interests, providing the possibility of implementing the obtained results into the practice.

Aims and tasks of the work

The aim of the study is to develop a project of an International System of Knowledge Exchange of young scientists (ISKE) through the organization of a community of people with certain scientific interests, their further communication and integration into international research projects.

To achieve this goal, the system must provide the following functionality:

- scalability of the data storage solution;
- guarantee of the integrity, reliability, confidentiality and fault tolerance of the system;
- implementation of processing of various types of search queries (voice, text, scanned and photo media);
- implementation of accelerated data search in the system, based on the specified area of user interest;
- a high degree of accuracy and completeness of the division of records into classes with determining the most pressing tasks and problems;
- in-depth analysis of uploaded work to group possible research teams;
- creation of a virtual room for research teams, sharing results, discussing research prospects with anticipation of the possibility of communication of different languages native speakers.

ISKE provides the opportunity to scale and internationalize the research.

Task fulfillment

Schematic conceptual representation of the proposed ISKE is shown in Fig. 3. The system is able to operate in two main modes – the data accumulation and processing mode and access mode. The constituent elements of the system are:

– scalable cloud storage;
– high-performance computing module for text documents vectorization;
– high-performance computing module for determining the text proximity of documents;
– graphical user web-interface;
– additional utility services.

The data collection and processing mode provides the formation of a repository of qualification and research papers in a structured form. The operation of this mode is to perform the following steps:

- uploading qualification works in docx or pdf formats to the cloud storage, namely to the buffer that stores unprocessed works;
- allocation in the main repository a separate cell-record for each of them containing the document itself, its author, as well as its metadata – frequency dictionary, etc.;
- text preprocessing (tokenization, removal of punctuation and stop words, lemmatization of text, etc.);
- creation of the frequency dictionary of the sample using the basis of the processed texts (Radix sorting is chosen as sorting method for terms) and finally placement of the received data in the main storage; moreover, if the system had already processed documents, the frequency dictionary gets rebuilt again from the start.

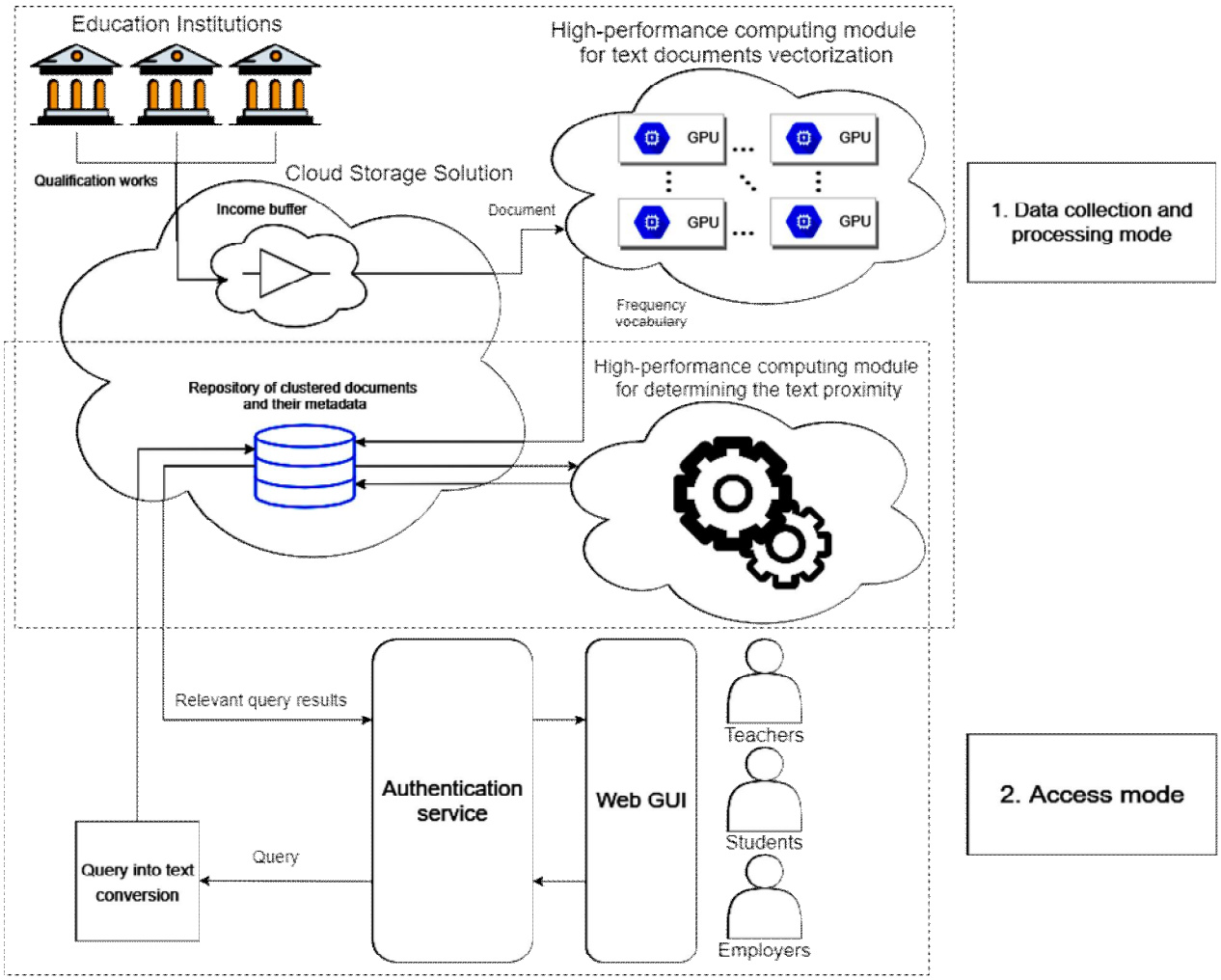

Fig. 3. The general scheme of proposed ISKE

- comparative analysis of the frequency dictionary of a sample;
- clustering of documents, i.e. assigning to all of them a specific value of proximity to other documents, on the basis of which clusters of semantically similar texts are formed, which are combined into groups.

Meanwhile, the second mode (access) consists of setting up and applying an authentication system and access rights, which affects the ability to view and search texts, as the system is designed for several groups of users depending on their status (student, teacher or employer). The system also takes into account the boundaries of different educational institutions to ensure the integrity and confidentiality of information, i.e. a user from one institution will not have access to non-public documents of another institution. The access itself is available through a web interface and has the search by keywords functionality, using the formed text relations when determining the relevant results for the query. The available query formats range from both text and voice to image format (scan), as all formats are converted to text using appropriate detection and data processing methods.

The functioning of the proposed ISKE is impossible without the accumulation of data and cataloging of new works that come regularly several times a year in large quantities (depending on the educational institution). Therefore, the solution to the problem of documents classification is one of the main tasks that affects the efficiency and speed of ISKE.

Research of influence of datasets statistical properties on classification result

The methods of pre-processing and text preparation have a great influence on the results of classification, which necessitates the study of source data pre-processing methods such as stop words removal, stemming and lemmatization. A training dataset of 4,725,865 sentences was prepared for the study. Of these, 153254 sentences describe the problem of natural language processing with the most commonly used words shown in Figure 4.

The rest (4572611 sentences) do not relate to the given problem; The most commonly used words are the words in Figure 5. The sample shows that there are a large number of noise words (get, one, can't, still), which can worsen the learning outcomes of classifiers. The validation date set is 92,000 sentences.

The paper investigates the influence of preprocessing methods (cleaning the text from words and symbols based on established rules; lemmatization and stemming) on the work of text classification methods such as Logistic Regression, LSTM, BERT, LightGBM.

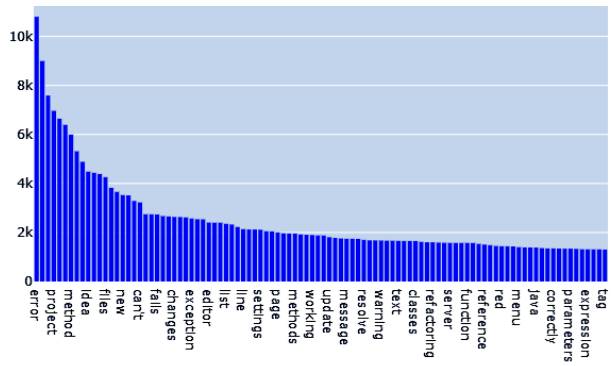

Fig. 4. The number of occurrences of words in the dataset from the problem area

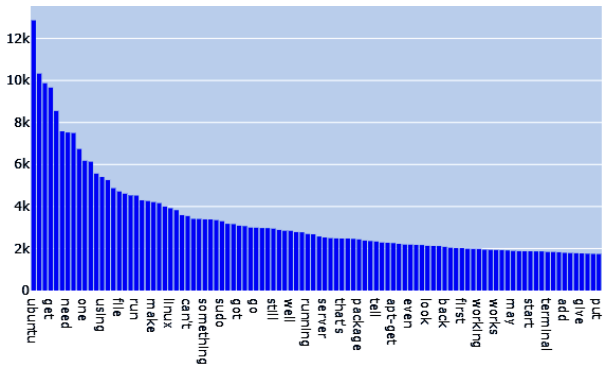

Fig. 5. The number of occurrences of words in the dataset from the non-problem area

A study was conducted on the speed of classification and F1 assessment. All the above classifiers worked on a validation sample of 92,000 sentences; the average number of characters in a sentence was 54.

The results of deleting stop words and symbols, frequently or little used punctuation words, and deleting POS tagging based on rule-based algorithms are shown in Table 1.

The efficiency of Logistic Regression, LSTM, BERT, LightGBM classifiers after text cleaning is set on the basis of F1-score, which allows to combine accuracy and completeness estimates and is given in table 2. Based on the experiments using different methods of clearing the text from the characters, the drawn conclusions is shown in table 3.

Table 1 – **Research of the text clearing runtime**

Method name	Execution speed in seconds
Stop words removal	1.17
Frequently or little used words deletion	1.62
Punctuation deletion	2.87
POS tagging deletion	4.70

Table 2 – **Study of F1-assessment of classification based on Logistic Regression, LSTM, BERT, LightGBM classifiers after performing data cleaning in the text**

Preprocessing stage	Classification method	F1 assessment
Baseline	Logistic Regression	0.71
	LightGBM	0.63
	LSTM	0.73
	BERT	0.9
Stop words removal	Logistic Regression	0.63
	LightGBM	0.50
	LSTM	0.65
	BERT	0.75
Frequently or little used words deletion	Logistic Regression	0.65
	LightGBM	0.55
	LSTM	0.7
	BERT	0.73
Punctuation deletion	Logistic Regression	0.72
	LightGBM	0.71
	LSTM	0.75
	BERT	0.8
POS tagging deletion	Logistic Regression	0.75
	LightGBM	0.7
	LSTM	0.73
	BERT	0.74

Table 3 – **Conclusions on the influence of the stages of text cleaning on the result of the classifier**

Classification method	Conslucion
Logistic Regression	Removing the POS tagging is the most influential and effective step in clearing text for this classification method.
LightGBM	Removing punctuation has the best effect on the performance of this classification method.
LSTM	Removing POS tagging is the most influential and effective method of clearing text for this classification method.
BERT	Cleaning the text negatively affects the quality of the classification. To improve the result of the model, it was decided not to perform text clearing of the dataset for this classification method.

The results of the experiments on the runtime of lemmatization and stemming for the validation date set of 92,000 sentences with an average number of words in a sentence of 54, are shown in table 4.

Table 4 – **Research of runtime of lemmatization and stemming**

Method name	Execution time in seconds
Stemming	9.8
Lemmatization	4.46

The obtained results showed that the speed of the lemmatization operation is twice as high as the time of the stemming.

Based on the experiments, the F1-assesement was measured on the basis of a validation dataset to determine the effectiveness of the Logistic Regression, LSTM, BERT, LightGBM classifiers after performing lemmatization and stemming operations. The results are shown in table 5.

Lematization and steaming improved the evaluation of Logistic Regression, LSTM and LightGBM classification methods on the validation dataset. For BERT, the score deteriorated slightly after performing lemmatization, which confirms the absence of the need to process the text when submitting it to the BERT classifier. Taking into account the results of table 4, the execution of lemmatization at the stage of preprocessing gives the result with high speed and efficiency for all these classifiers, except BERT.

Lemmatization showed a shorter execution time compared to stemming by almost twice and a better score by an average of 5 percent, so it was decided to use the Logistic Regression classifier with lemmatization at the stage of text preparation in the subsequent operation of the proposed ISKE.

Table 5 – **Study of F1-assessment of classification based on classifiers Logistic Regression, LSTM, BERT, LightGBM after lemmatization and text stemming**

Method name	Model name	F1 assessment
Baseline	Logistic Regression	0.71
	LightGBM	0.63
	LSTM	0.73
	BERT	0.9
Stemming	Logistic Regression	0.8
	LightGBM	0.75
	LSTM	0.8
	BERT	-
Lemmatization	Logistic Regression	0.81
	LightGBM	0.75
	LSTM	0.8
	BERT	0.72

Conclusion

The paper proposes a system which is electronic data storage (of qualification works of students from different countries) and provides the capability to identify and connect young scientists conducting research on a related problem area.

The purpose of developing this system is to provide opportunities for knowledge exchange, research in a team on a common problem, as well as to identify scientific trends in different countries. The system is able to operate in two main modes – the data accumulation and processing mode and access mode. During access mode available query formats range from both text and voice to image format (scans or photo), as well as the possibility of accelerated data search throughout the system.

The paper investigates the influence of preprocessing methods (cleaning the text from words and symbols based on established rules; lemmatization and stemming) on the work of text classification methods such as Logistic Regression, LSTM, BERT and LightGBM. A study was conducted on the speed of classification and F1 assessment. Lemmatization showed a shorter execution time compared to stemming by almost twice and a better score by an average of 5 percent, so it was decided to use the Logistic Regression classifier with lemmatization at the stage of text preparation in the subsequent operation of the proposed ISKE.

REFERENCES

1. Zaiceva, S. and Barkovska, O. (2020), “Analysis of Accelerated Problem Solutions of Word Search in Texts”, *The Fourth International Scientific and Technical Conference «COMPUTER AND INFORMATION SYSTEMS AND TECHNOLOGIES»*, KhNURE, Kharkiv, p.66, DOI: https://doi.org/10.30837/IVcsitic2020201445.
2. Barkovska, Olesia, Mikhal, Oleg, Pyvovarova, Daria, Liashenko, Oleksii, Diachenko, Vladyslav and Volk, Maxim (2020), “Local Concurrency in Text Block Search Tasks”, *International Journal of Emerging Trends in Engineering Research*, Vol. 8. No. 3, March, pp. 690-694, DOI: https://doi.org/10.30534/ijeter/2020/13832020.
3. Barkovska, O., Pyvovarova, D. and Serdechnyi, V. (2019), “Accelerated word-image search algorithm in text with adaptive decomposition of input data”, *Systemy upravlinnja, navighaciji ta zv'jazku*, Vol. 4 (56), pp. 28-34, DOI: https://doi.org/10.26906/SUNZ.2019.4.028 (in Ukrainian).
4. Vuksanović, Petrijevčanin and Sudarević, B. (2010), “Migration process and data modeling in National and University Library in creating ILS”, *Proceedings ELMAR-2010*, Zadar, Croatia, pp. 155-158.

5. Puritat, K. and Intawong, K. (2020), "Development of an Open Source Automated Library System with Book Recommedation System for Small Libraries", *2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering* (ECTI DAMT & NCON), Pattaya, Thailand, pp. 128-132, DOI: https://doi.org/10.1109/ECTIDAMTNCON48261.2020.9090753.
6. Bhushan, S. and Weingroff, M. (2005), "Tools for managing collaboration, communication, and website content development in a distributed digital library community", *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL '05), Denver, CO, USA, pp. 401-401, DOI: https://doi.org/10.1145/1065385.1065504.

Received (Надійшла) 25.11.2020
Accepted for publication (Прийнята до друку) 17.02.2021

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Барковська Олеся Юріївна – кандидат технічних наук, доцент, доцент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Olesia Barkovska – Candidate of Technical Sciences, Associate Professor, Associate Professor of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: olesia.barkovska@nure.ua; ORCID ID: http://orcid.org/0000-0001-7496-4353.

Холєв Владислав Олександрович – студент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Vladyslav Kholiev – student of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail**:** vladyslav.kholiev@nure.ua; ORCID ID: http://orcid.org/0000-0002-9148-1561.

Пивоварова Дар'я Ігорівна – асистент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Daria Pyvovarova – Assistant lecturer of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: daria.pyvovarova@nure.ua; ORCID ID: http://orcid.org/0000-0002-7251-994X.

Іващенко Георгій Станіславович – кандидат технічних наук, доцент, доцент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Georgiy Ivaschenko – Candidate of Technical Sciences, Associate Professor, Associate Professor of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: heorhii.ivashchenko@nure.ua; ORCID ID: http://orcid.org/0000-0003-1027-5262.

Росінський Дмитро Миколайович – старший викладач кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Dmytro Rosinskiy – Senior Lecturer of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: dmytro.rosinskyi@nure.ua; ORCID ID: http://orcid.org/0000-0002-0725-392X.

Система обміну знаннями молодих науковців із різних країн

О. Ю. Барковська, В. О. Холєв, Д. І. Пивоварова, Г. С. Іващенко, Д. М. Росінський

Анотація. У роботі запропонована система, яка являє електронне сховище даних (кваліфікаційних робіт студентів із різних країн) та забезпечує можливість виявити та зв'язати між собою молодих вчених, що ведуть дослідження над єдиною проблемною областю. **Метою** розробки даної системи є забезпечення можливості обміну знаннями, виконання досліджень у команді над спільною проблемою, а також визначення наукових тенденцій у різних країнах світу. У роботі досліджено вплив методів препроцесінгу на роботу таких класифікаторів, як Logistic Regression, LSTM, BERT, LightGBM. Проведено дослідження щодо швидкості класифікації та F1 оцінки. **Висновки.** Лематизація показала коротший час роботи у порівнянні зі стемінгом майже в два рази та кращу оцінку в середньому на 5 відсотків, тому було прийнято рішення використовувати класифікатор Logistic Regression із лематизацією на етапі підготовки тексту у подальшій роботі запропонованої системи обміну знаннями молодих науковців.

Ключові слова: система; NLP; текст; обробка; прискорення; шингли; близькість; подібність; класифікація; попередня обробка; лематизація; стемінг.

Система обмена знаниями молодых ученых из разных стран

О. Ю. Барковская, В. А. Холев, Д. И. Пивоварова, Г. С. Иващенко, Д. М. Росинский

Аннотация. В работе предложена система, которая представляет электронное хранилище данных (квалификационных работ студентов из разных стран) и обеспечивает возможность выявить и связать между собой молодых ученых, ведущих исследования над единой проблемной областью. **Целью** разработки данной системы является обеспечение возможности обмена знаниями, проведения исследований в команде над общей проблемой, а также определение научных тенденций в разных странах мира. В работе исследовано влияние методов препроцесинга на работу таких классификаторов, как Logistic Regression, LSTM, BERT, LightGBM. Проведено исследование скорости классификации и F1 оценки. **Выводы**. Лемматизация показала меньшее время работы по сравнению со стеммингом почти в два раза и лучшую оценку на 5%, поэтому было принято решение использовать классификатор Logistic Regression с лемматизацией на этапе подготовки текста в дальнейшей работе предложенной системы обмена знаниями ученых.

Ключевые слова: система; NLP; текст; обработка; ускорение; шинглы; близость; сходство; классификация; предварительная обработка; лемматизация; стемминг.