ДОДАТОК А

Графічний матеріал кваліфікаційної роботи

Харківський національний університет радіоелектроніки
Кафедра ЕОМ

МЕТОД ТА ЗАСОБИ БЕНЧМАРКІНГА МУЛЬТИМОДЕЛЬНОЇ БАЗИ ДАНИХ

Кваліфікаційна робота
Другий (магістерський рівень)

Автор:
Хомич В. М.,
студ. гр. СПм-20-2

Керівник:
Можаєв О. О.,
проф. каф. ЕОМ

МЕТА І НАУКОВА НОВИЗНА РОБОТИ

Мета: розробка методу і засобів бенчмаркінга мультимодельної бази даних.

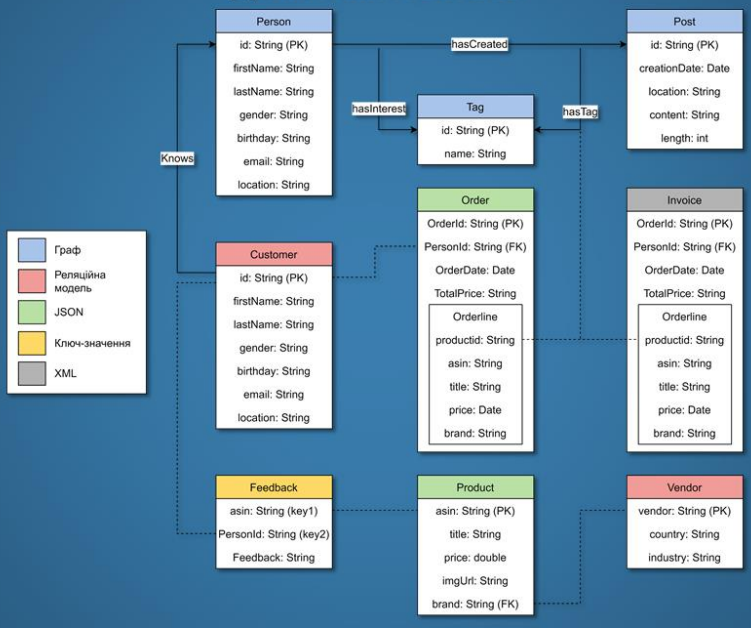
Наукова новизна: розробка наскрізного бенчмарка для мультимодельних баз даних.

ЗАДАЧІ РОБОТИ

- розробка нового генератора даних, який надає корельовані дані в різних моделях;
- розробка набору робочих навантажень із кількома моделями;
- визначення проблеми під назвою «мультимодельне курування параметрів»;
- проведення комплексної оцінки чотирьох мультимодельних баз даних.

3

ДІАГРАМА КЛАСІВ



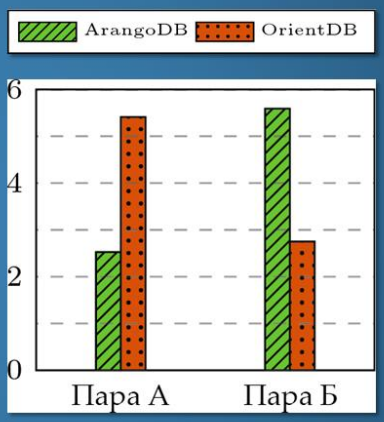
4

МОТИВ ДЛЯ КУРУВАННЯ ПАРАМЕТРІВ

Приклад запиту з двома парами параметрами підстановки:

Пара А: *@PersonId*=33, *@BrandName*="Adidas"

Пара Б: *@PersonId*=56, *@BrandName*="Nike"



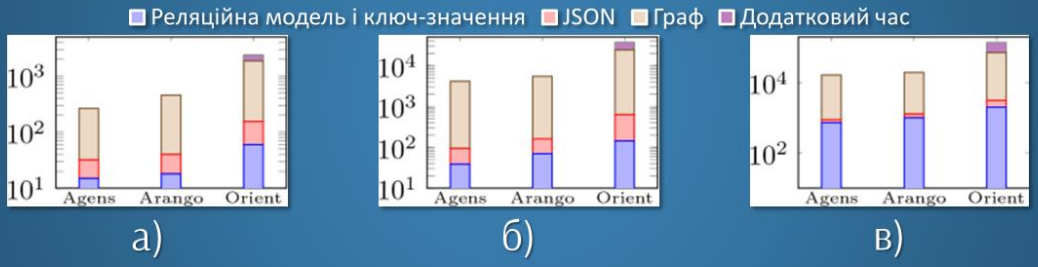
5

МУЛЬТИМОДЕЛЬНИЙ РОЗПОДІЛ СТВОРЕНОГО НАБОРУ ДАНИХ

КУРУВАННЯ ПАРАМЕТРІВ ЕФЕКТИВНОСТІ, РІЗНОМАНІТНОСТІ ТА СТАБІЛЬНОСТІ

6

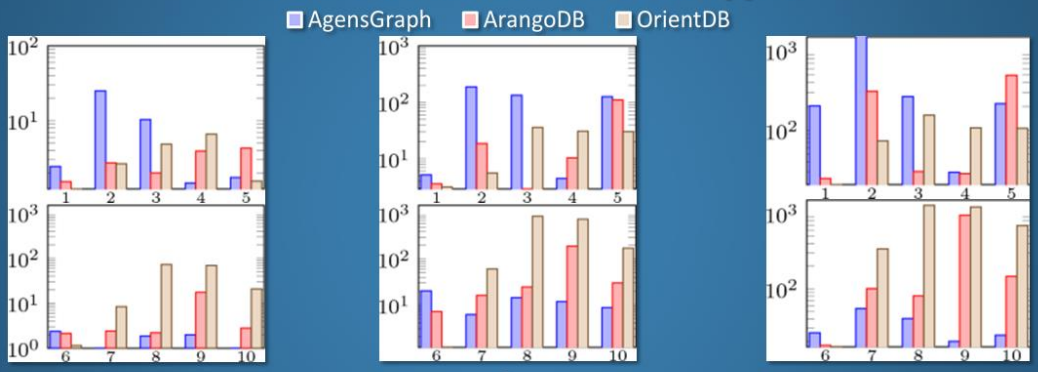
ЧАС ОБРОБКИ ДЛЯ ІМПОРТУ МУЛЬТИМОДЕЛЬНИХ НАБОРІВ ДАНИХ З ОДНИМ ПОТОКОМ В СЕКУНДАХ



а) б) в)

7

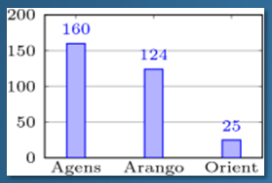
ЧАС ОБРОБКИ ЗАПИТІВ В СЕКУНДАХ



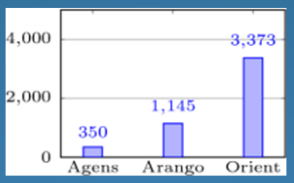
а) б) в)

8

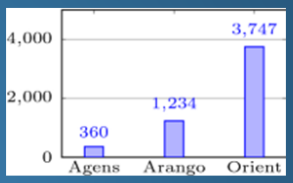
ПРОДУКТИВНІСТЬ ТРАНЗАКЦІЇ В МІЛІСЕКУНДАХ



а)

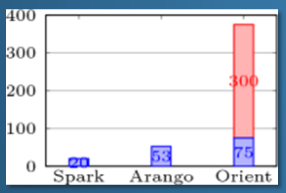


б)

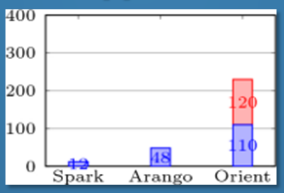


в)

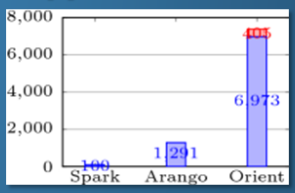
ЧАС ОБРОБКИ ІМПОРТУ МУЛЬТИМОДЕЛЬНИХ НАБОРІВ ДАНИХ У СЕКУНДАХ



а)



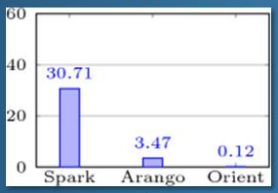
б)



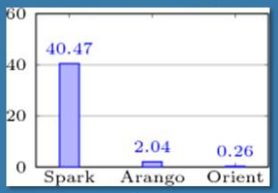
в)

9

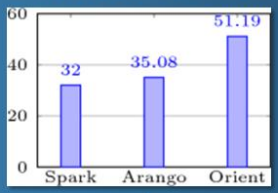
РОЗПОДІЛЕНИЙ ЧАС ОБРОБКИ МУЛЬТИМОДЕЛЬНОГО ЗАПИТУ В СЕКУНДАХ



а)



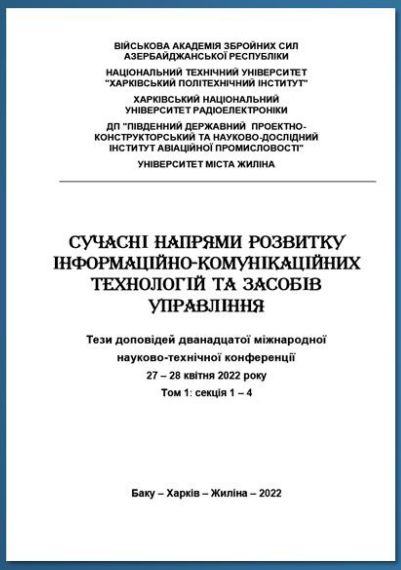
б)



в)

10

АПРОБАЦІЯ РЕЗУЛЬТАТІВ



ВІЙСЬКОВА АКАДЕМІЯ ЗБРОЙНИХ СИЛ АЗЕРБАЙДЖАНСЬКОЇ РЕСПУБЛІКИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ "ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ"
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ РАДІОЕЛЕКТРОНІКИ
ДП "ПІВДЕННИЙ ДЕРЖАВНИЙ ПРОЕКТНО-КОНСТРУКТОРСЬКИЙ ТА НАУКОВО-ДОСЛІДНИЙ ІНСТИТУТ АВІАЦІЙНОЇ ПРОМИСЛОВОСТІ"
УНІВЕРСИТЕТ МІСТА ЖИЛІНА

СУЧАСНІ НАПРЯМИ РОЗВИТКУ ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ ТА ЗАСОБІВ УПРАВЛІННЯ

Тези доповідей дванадцятої міжнародної науково-технічної конференції
27 – 28 квітня 2022 року
Том 1: секція 1 – 4

Баку – Харків – Жиліна – 2022

11

ВИСНОВКИ

Результатом виконання кваліфікаційної роботи став бенчмарк мультимодельної бази даних, який складається зі змішаної моделі даних, масштабованого мультимодельного генератора даних і набору робочих навантажень, включаючи мультимодельну агрегацію, об'єднання та транзакцію.

12

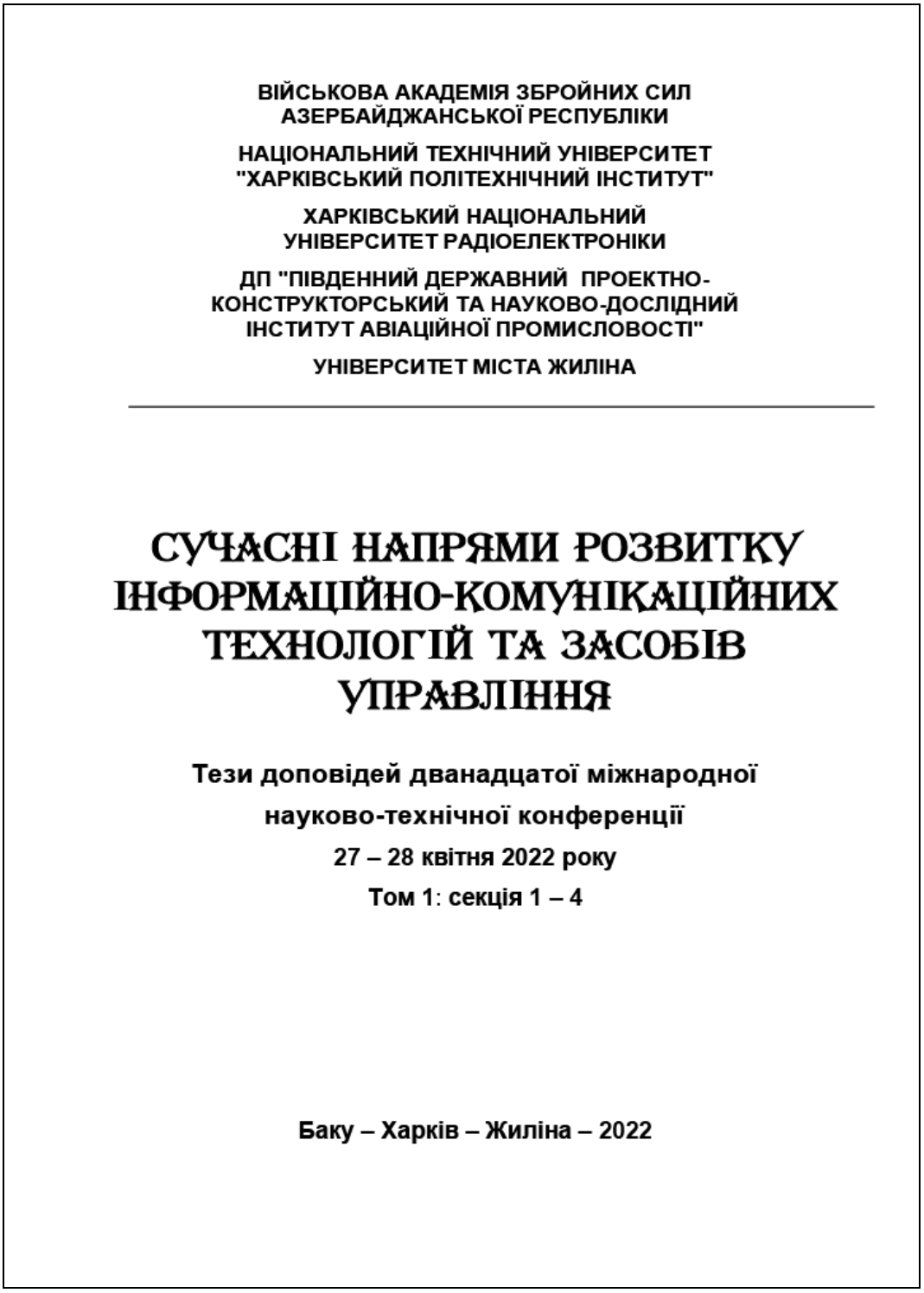
ПЕРСПЕКТИВИ РОЗВИТКУ

Подальші дослідження полягають у впровадженні гнучкості в генерацію даних, в оцінці продуктивності мультимодельних баз даних щодо різних стратегій розподілу, а також в оцінці продуктивності мультимодельних баз даних щодо оптимізації запитів різних систем.

13

ДОДАТОК Б

Публікації за темою кваліфікаційної роботи



ВІЙСЬКОВА АКАДЕМІЯ ЗБРОЙНИХ СИЛ
АЗЕРБАЙДЖАНСЬКОЇ РЕСПУБЛІКИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
"ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ"

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ
УНІВЕРСИТЕТ РАДІОЕЛЕКТРОНІКИ

ДП "ПІВДЕННИЙ ДЕРЖАВНИЙ ПРОЕКТНО-
КОНСТРУКТОРСЬКИЙ ТА НАУКОВО-ДОСЛІДНИЙ
ІНСТИТУТ АВІАЦІЙНОЇ ПРОМИСЛОВОСТІ"

УНІВЕРСИТЕТ МІСТА ЖИЛІНА

СУЧАСНІ НАПРЯМИ РОЗВИТКУ ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ ТА ЗАСОБІВ УПРАВЛІННЯ

Тези доповідей дванадцатої міжнародної
науково-технічної конференції
27 – 28 квітня 2022 року
Том 1: секція 1 – 4

Баку – Харків – Жиліна – 2022

Рисунок Б.1 – Тези доповіді конференції «Сучасні напрями розвитку інформаційно-комунікаційних технологій та засобів управління»

МЕТОД ТА ЗАСОБИ БЕНЧМАРКІНГА МУЛЬТИМОДЕЛЬНОЇ БАЗИ ДАНИХ

Хомич В. М., Можаєв О. О.
Харківський національний університет радіоелектроніки, Харків, Україна

Оскільки все більше компаній усвідомлюють, що дані в усіх формах і розмірах мають вирішальне значення для прийняття найкращих можливих рішень, ми бачимо постійне зростання систем, які підтримують величезний обсяг реляційних або нереляційних форм даних. На відміну від традиційних систем керування базами даних, які організовані навколо єдиної моделі даних, яка визначає, як дані можуть бути організовані, збережені й маніпульовані, мультимодельна база даних розроблена для підтримки кількох моделей даних на одному інтегрованому сервері [1]. Наявність єдиної платформи даних для керування як добре структурованими даними, так і даними NoSQL є вигідним для користувачів, позаяк такий підхід значно зменшує проблеми інтеграції, міграції, розробки, обслуговування та експлуатації.

Бенчмаркінг є загальноприйнятою практикою для оцінки систем баз даних, позаяк все більше і більше платформ пропонуються для роботи з мультимодельними даними. Тому стає важливим мати бенчмарки, які можна використовувати для оцінки продуктивності та зручності використання наступного покоління мультимодельних систем баз даних.

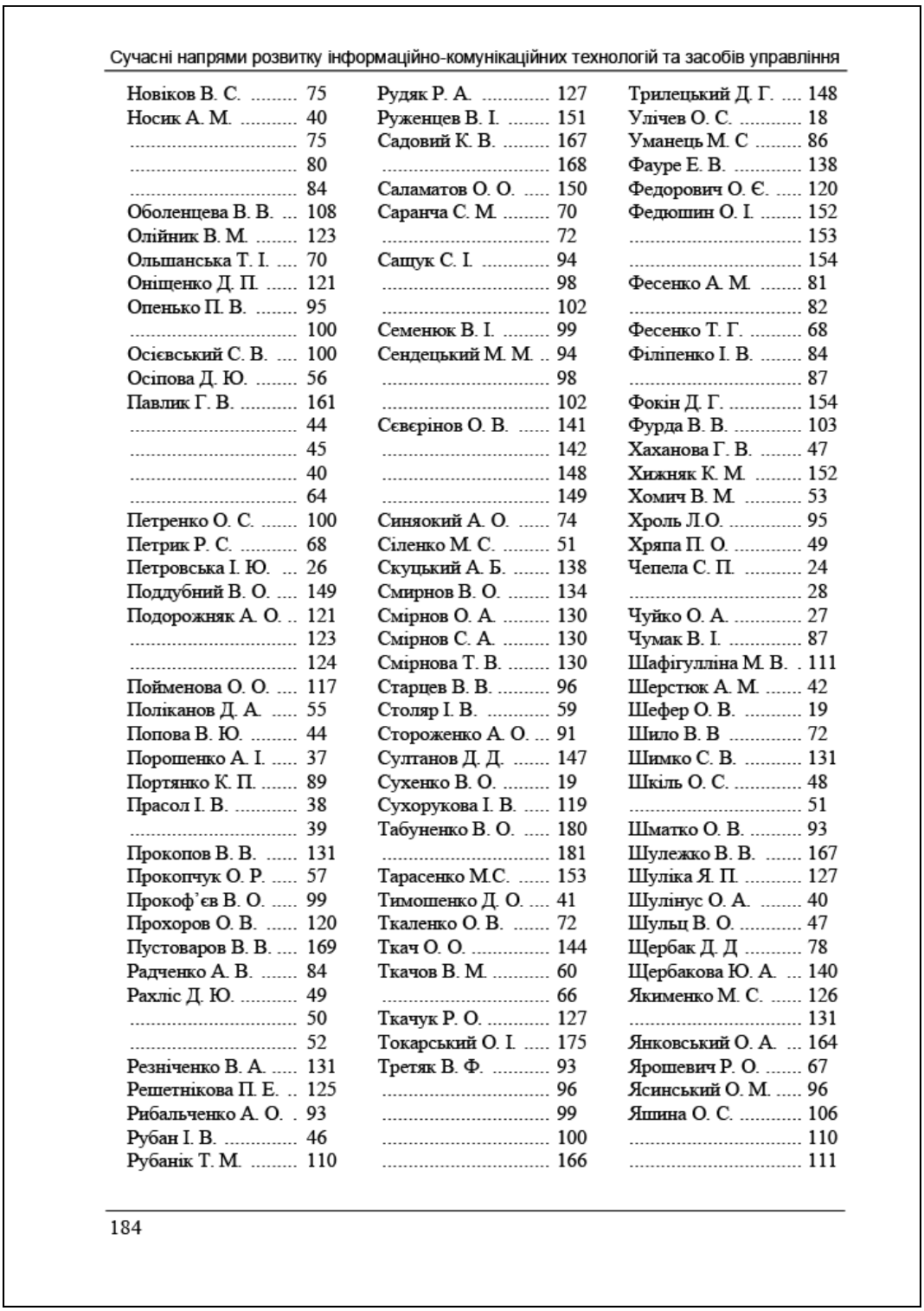
Метою доповіді є аналіз методу та засобів бенчмаркінга мультимодельної бази даних.

В доповіді наводиться, що ретельна оцінка мультимодельних систем баз даних ставить перед собою кілька нових проблем, які необхідно подолати. По-перше, оскільки стандартної мультимодельної мови запитів зараз немає, загальнодоступні реалізації даних бенчмаркінга та запитів для різних систем слід розробляти, спільно використовувати, уніфікувати та оптимізувати. По-друге, на відміну від реляційного світу, системи NoSQL дотримуються парадигми «спочатку дані, схема пізніше або ніколи». Для ретельної оцінки має бути можливість контролювати вхідну схему та складність еволюції схеми для мультимодельних даних. Бенчмарк повинен підвищувати продуктивність, дозволяючи створювати багато мультимодельних даних із різноманітною схемою, використовуючи невеликі ручні зусилля. Нарешті, мультимодельні бази даних повинні підтримувати міжмодельну транзакцію та узгодженість. Тому нові метрики узгодженості, які описують поведінку узгодженості для різних моделей даних, повинні бути запропоновані точним чином.

Список літератури

1. Спасітєлєва С. О., Жданова Ю. Д., Чичкань І. В. Проблеми безпеки універсальних платформ управління даними. *Кібербезпека: освіта, наука, техніка*. 2019. Т. 2, № 6. С. 122–133. DOI: https://doi.org/10.28925/2663-4023.2019.6.122133

Рисунок Б.2 – Тези доповіді конференції «Сучасні напрями розвитку інформаційно-комунікаційних технологій та засобів управління»



Сучасні напрями розвитку інформаційно-комунікаційних технологій та засобів управління

Новіков В. С. 75
Носик А. М. 40
..... 75
..... 80
..... 84
Оболенцева В. В. ... 108
Олійник В. М. 123
Ольшанська Т. І. 70
Оніщенко Д. П. 121
Опенько П. В. 95
..... 100
Осієвський С. В. 100
Осіпова Д. Ю. 56
Павлик Г. В. 161
..... 44
..... 45
..... 40
..... 64
Петренко О. С. 100
Петрик Р. С. 68
Петровська І. Ю. ... 26
Поддубний В. О. 149
Подорожняк А. О. .. 121
..... 123
..... 124
Пойменова О. О. 117
Поліканов Д. А. 55
Попова В. Ю. 44
Порошенко А. І. 37
Портянко К. П. 89
Прасол І. В. 38
..... 39
Прокопов В. В. 131
Прокопчук О. Р. 57
Прокоф'єв В. О. 99
Прохоров О. В. 120
Пустоваров В. В. 169
Радченко А. В. 84
Рахліс Д. Ю. 49
..... 50
..... 52
Резніченко В. А. 131
Решетнікова П. Е. .. 125
Рибальченко А. О. . 93
Рубан І. В. 46
Рубанік Т. М. 110
Рудяк Р. А. 127
Руженцев В. І. 151
Садовий К. В. 167
..... 168
Саламатов О. О. 150
Саранча С. М. 70
..... 72
Сащук С. І. 94
..... 98
..... 102
Семенюк В. І. 99
Сендецький М. М. .. 94
..... 98
..... 102
Сєвєрінов О. В. 141
..... 142
..... 148
..... 149
Синяокий А. О. 74
Сіленко М. С. 51
Скуцький А. Б. 138
Смирнов В. О. 134
Смірнов О. А. 130
Смірнов С. А. 130
Смірнова Т. В. 130
Старцев В. В. 96
Столяр І. В. 59
Стороженко А. О. ... 91
Султанов Д. Д. 147
Сухенко В. О. 19
Сухорукова І. В. 119
Табуненко В. О. 180
..... 181
Тарасенко М.С. 153
Тимошенко Д. О. 41
Ткаленко О. В. 72
Ткач О. О. 144
Ткачов В. М. 60
..... 66
Ткачук Р. О. 127
Токарський О. І. 175
Третяк В. Ф. 93
..... 96
..... 99
..... 100
..... 166
Трилецький Д. Г. 148
Улічев О. С. 18
Уманець М. С 86
Фауре Е. В. 138
Федорович О. Є. 120
Федюшин О. І. 152
..... 153
..... 154
Фесенко А. М. 81
..... 82
Фесенко Т. Г. 68
Філіпенко І. В. 84
..... 87
Фокін Д. Г. 154
Фурда В. В. 103
Хаханова Г. В. 47
Хижняк К. М. 152
Хомич В. М. 53
Хроль Л.О. 95
Хряпа П. О. 49
Чепела С. П. 24
..... 28
Чуйко О. А. 27
Чумак В. І. 87
Шафігулліна М. В. . 111
Шерстюк А. М. 42
Шефер О. В. 19
Шило В. В 72
Шимко С. В. 131
Шкіль О. С. 48
..... 51
Шматко О. В. 93
Шулежко В. В. 167
Шуліка Я. П. 127
Шулінус О. А. 40
Шульц В. О. 47
Щербак Д. Д 78
Щербакова Ю. А. ... 140
Якименко М. С. 126
..... 131
Янковський О. А. ... 164
Ярошевич Р. О. 67
Ясинський О. М. 96
Яшина О. С. 106
..... 110
..... 111

184

Рисунок Б.3 – Тези доповіді конференції «Сучасні напрями розвитку інформаційно-комунікаційних технологій та засобів управління»

Рисунок Б.4 – Стаття в журналі «Сучасний стан наукових досліджень та технологій в промисловості»

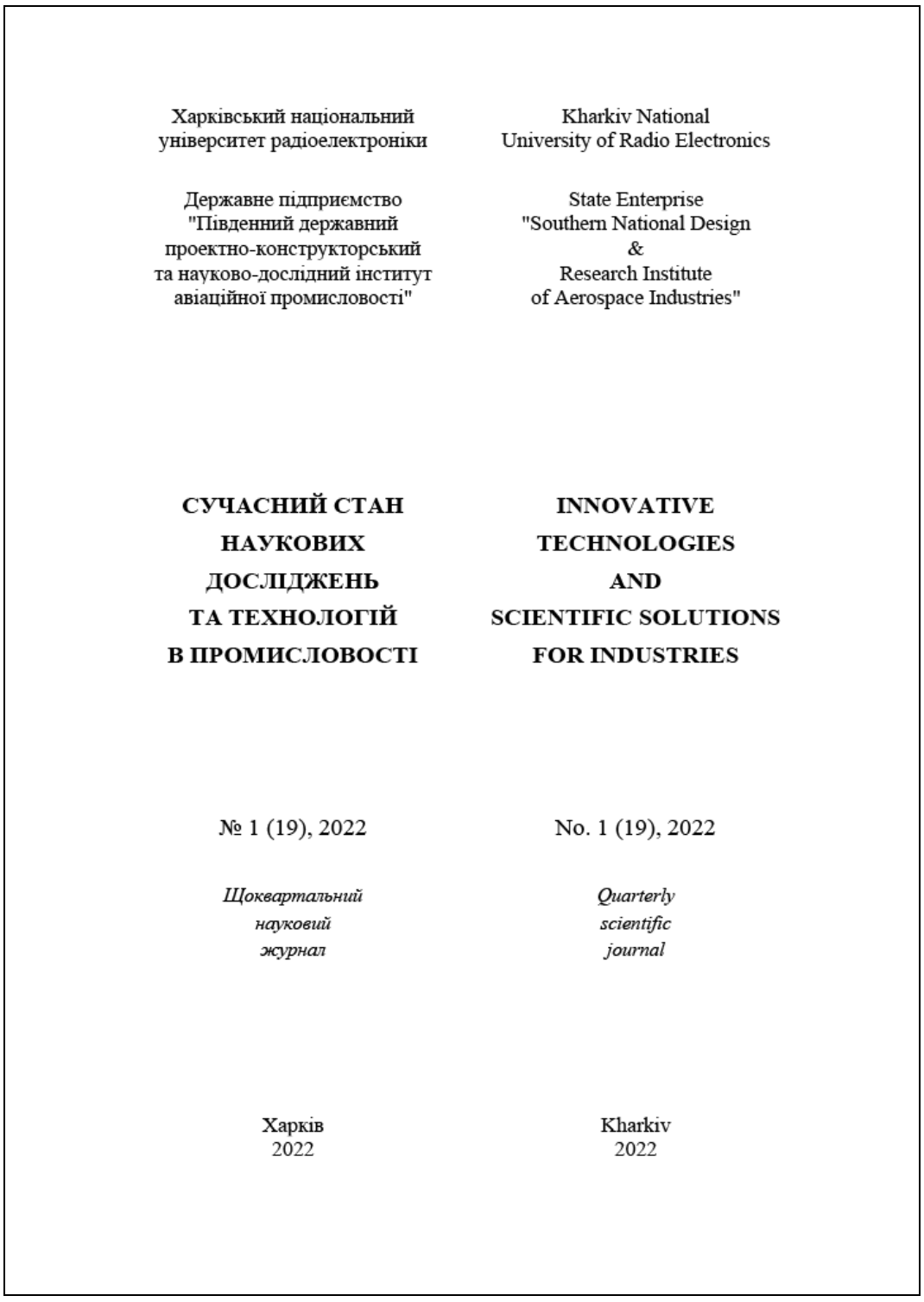
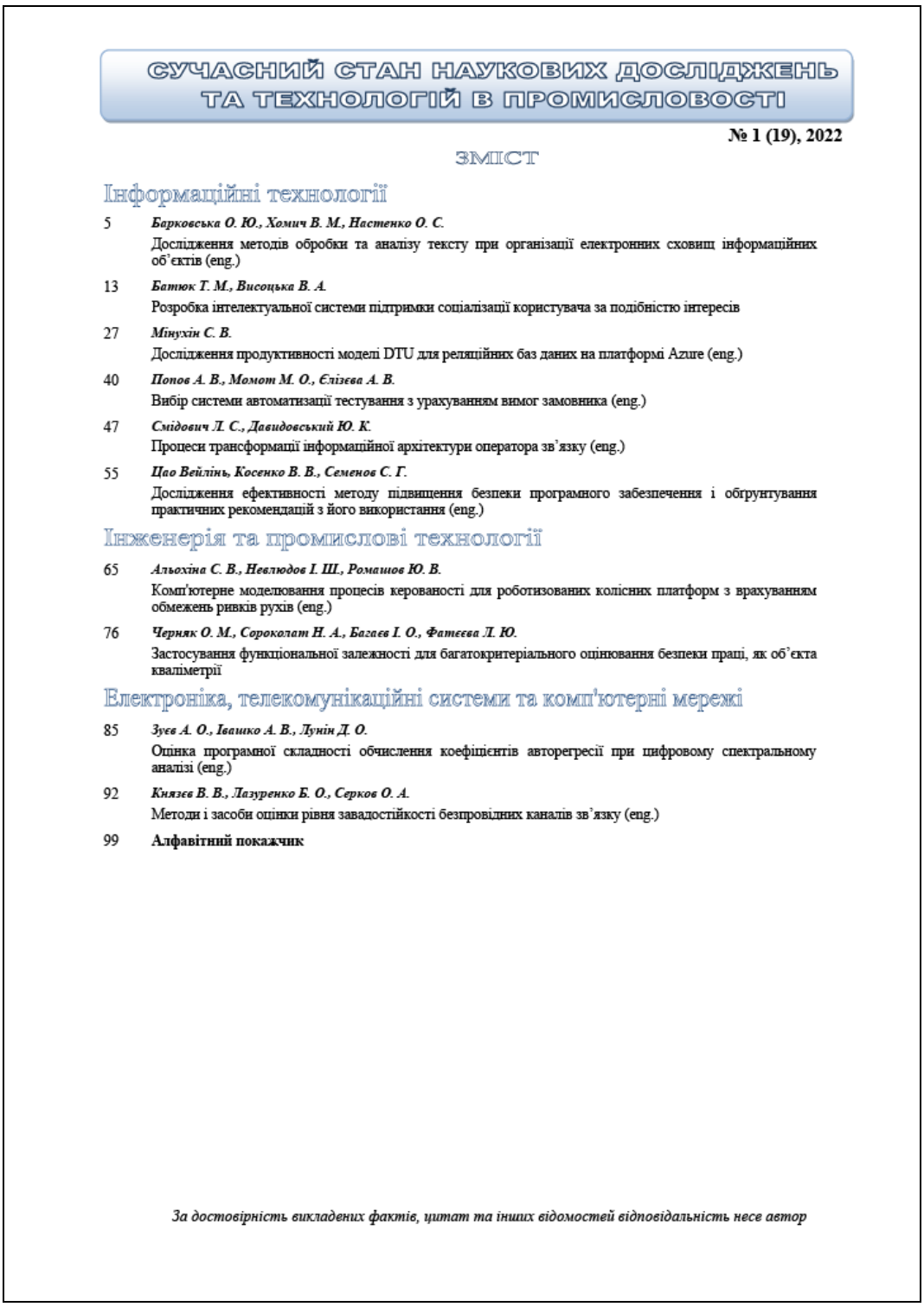


Рисунок Б.5 – Стаття в журналі «Сучасний стан наукових досліджень та технологій в промисловості»



СУЧАСНИЙ СТАН НАУКОВИХ ДОСЛІДЖЕНЬ ТА ТЕХНОЛОГІЙ В ПРОМИСЛОВОСТІ

№ 1 (19), 2022

ЗМІСТ

Інформаційні технології

5 *Барковська О. Ю., Хомич В. М., Настенко О. С.*
Дослідження методів обробки та аналізу тексту при організації електронних сховищ інформаційних об'єктів (eng.)

13 *Батюк Т. М., Висоцька В. А.*
Розробка інтелектуальної системи підтримки соціалізації користувача за подібністю інтересів

27 *Мінухін С. В.*
Дослідження продуктивності моделі DTU для реляційних баз даних на платформі Azure (eng.)

40 *Попов А. В., Момот М. О., Єлізєва А. В.*
Вибір системи автоматизації тестування з урахуванням вимог замовника (eng.)

47 *Смідович Л. С., Давидовський Ю. К.*
Процеси трансформації інформаційної архітектури оператора зв'язку (eng.)

55 *Цао Вейлінь, Косенко В. В., Семенов С. Г.*
Дослідження ефективності методу підвищення безпеки програмного забезпечення і обгрунтування практичних рекомендацій з його використання (eng.)

Інженерія та промислові технології

65 *Альохіна С. В., Невлюдов І. Ш., Романов Ю. В.*
Комп'ютерне моделювання процесів керованості для роботизованих колісних платформ з врахуванням обмежень ривків рухів (eng.)

76 *Черняк О. М., Сороколат Н. А., Багаєв І. О., Фатєєва Л. Ю.*
Застосування функціональної залежності для багатокритеріального оцінювання безпеки праці, як об'єкта кваліметрії

Електроніка, телекомунікаційні системи та комп'ютерні мережі

85 *Зуєв А. О., Івашко А. В., Лунін Д. О.*
Оцінка програмної складності обчислення коефіцієнтів авторегресії при цифровому спектральному аналізі (eng.)

92 *Князєв В. В., Лазуренко Б. О., Серков О. А.*
Методи і засоби оцінки рівня завадостійкості безпровідних каналів зв'язку (eng.)

99 **Алфавітний покажчик**

За достовірність викладених фактів, цитат та інших відомостей відповідальність несе автор

Рисунок Б.6 – Стаття в журналі «Сучасний стан наукових досліджень та технологій в промисловості»

Сучасний стан наукових досліджень та технологій в промисловості. 2022. № 1 (19)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

UDC 004.9

DOI: https://doi.org/10.30837/ITSSI.2022.19.005

O. BARKOVSKA, V. KHOMYCH, O. NASTENKO

RESEARCH OF THE TEXT PROCESSING METHODS IN ORGANIZATION OF ELECTRONIC STORAGES OF INFORMATION OBJECTS

The **subject** matter of the article is electronic storage of information objects (IO) ordered by specified rules at the stage of accumulation of qualification thesis and scientific work of the contributors of the offered knowledge exchange system provided to the system in different formats (text, graphic, audio). Classified works of contributors of the system are the ground for organization of thematic rooms for discussion to spread scientific achievements, to adopt new ideas, to exchange knowledge and to look for employers or mentors in different countries. The **goal** of the work is to study the libraries of text processing and analysis to speed-up and increase accuracy of the scanned text documents classification in the process of serialized electronic storage of information objects organization. The following **tasks** are: to study the text processing methods on the basis of the proposed generalized model of the system of classification of scanned documents with the specified location of the block of text processing and analysis; to investigate the statistics of change in the execution time of the developed parallel modification of the methods of the word processing module for the system with shared memory for collections of text documents of different sizes; analyze the results. The **methods** used are the following: parallel digital sorting methods, methods of mathematical statistics, linguistic methods of text analysis. The following **results** were obtained: in the course of the research fulfillment the generalized model of the scanned documents classification system that consist of image processing unit and text processing unit that include unit of the scanned image previous processing; text detection unit; previous text processing; compiling of the frequency dictionary; text proximity detection was offered. **Conclusions:** the proposed parallel modification of the previous text processing unit gives acceleration up to 3,998 times. But, at a very high computational load (collection of 18144 files, about 1100 MB), the resources of an ordinary multiprocessor-based computer with the shared memory obviously is not enough to solve such problems in the mode close to real time.

Keywords: information system; parallelism; word processing; linguistic programming; library; acceleration; method.

Introduction

The paper [1] offers the system of young scientists from different countries knowledge exchange, that provides for communication of the specialists and interested people on the one of available topics. The topic is defined on the basis of qualification thesis and scientific work of the system participants, that prescribes the previous stage of the data collection in database, their analysis to classify and identify topical and scientific and practical directions and authors who are interested in and look into specified directions. Classified works of contributors of the system are the ground for organization of the thematic rooms for discussion to spread scientific achievements, to adopt new ideas, to exchange knowledge and to look for an employers or mentors in different countries.

In the thesis [2] it was offered the organization model of the storage system [3-6] and access to the scientific thesis of the researchers, lecturers and university students within electronic storage IO. Both stages are working with different types of input documents: graphic, text and audio-files (the last format is used only on the data access stage). That is why applicability of the research in the scanned documents classification field for the classification and structured storing in electronic storage of the scientific knowledge exchange system is reasonable.

Analysis of last achievements and publications

Scanned documents processing is a complex process as it represents the combination of the methods of the work with the image and text processing, that was noted in thesis [7-9]. Further, a lot of works are addressed to the text processing methods [10-14], and realization of methods is performed in numerous linguistic programming libraries – CoreNLP, NLTK, TextBlob, Spacy, Spark NLP etc. The specific feature of a library is a possibility to use different hardware (massive parallel system, general memory systems etc). Processing libraries analysis and text analysis is the goal of the given work. It will provide fast and accurate detected text processing for the further classification.

In the sphere of natural language processing there are many libraries, but the libraries with the limited functionality and operate for the specific cases will not be productive for solving the given task. For instance, NLP gensim library was originally created for thematic modeling and can not be used for complete NLP pipeline construction as a collection of basic and additional functions (fig. 1).

The analysis of the NLP libraries expended features about text processing revealed that the following libraries with the open code meet the necessary requirements: Spark NLP; spaCy; OpenNLP; Stanford CoreNLP., NLTK [15-17]. The comparison of the NLP libraries features is given in table 1.

Although most NLP libraries support learning new models by users it is important that any NLP library was able to provide available pre-trained high-quality models. However, most libraries only support general pre-trained models (POS, NER etc.). Some do not allow the use of their pre-trained models for commercial purposes because of the way they are licensed (table 2).

Рисунок Б.7 – Стаття в журналі «Сучасний стан наукових досліджень та технологій в промисловості»

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2022. No. 1 (19)

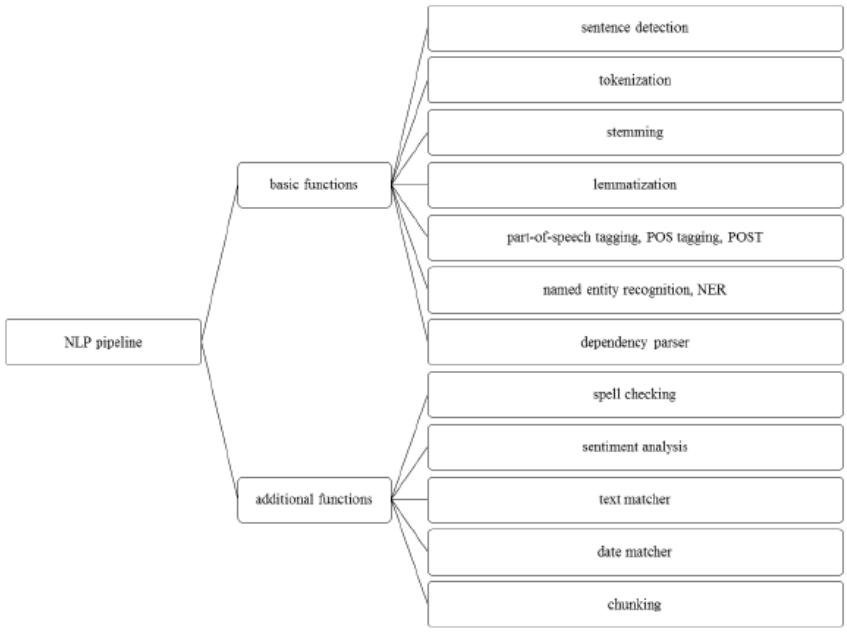


Fig.1. Definition of basic and additional functions of NLP pipeline

Table 1. *The comparison of the NLP libraries features*

Name of the text processing method	NLTK	Spark NLP	spaCy	CoreNLP
Sentence detection	+	+	+	+
Tokenization	+	+	+	+
Stemming	+	+	+	+
Lemmatization	+	+	+	+
POS tagging	+	+	+	+
NER	+	+	+	+
Dependency parser	+	+	+	+
Text matcher	–	+	+	+
Date matcher	–	+	–	+
Chunking	+	+	+	+
Spell checking	–	+	–	–
Sentiment analysis	–	+	–	+
Pre-trained models	+	+	+	+
Trained models	+	+	+	+

Table 2. *Comparison of NLP libraries with the support of pre-trained models*

Name	Pre-trained models (general)	Pre-trained models (subject-oriented)	Free license to use pre-trained models commercially
NLTK	+	–	+
Spark NLP	+	+	+ (partially)
spaCy	+	+	+ (partially)
CoreNLP	+	–	–
OpenNLP	+	–	+

The conducted analysis of the linguistic programming libraries, the functionality of which is focused on natural language processing with ready-made pre-trained models of neural networks, conveyors and

Рисунок Б.8 – Стаття в журналі «Сучасний стан наукових досліджень та технологій в промисловості»

Сучасний стан наукових досліджень та технологій в промисловості. 2022. № 1 (19) *ISSN 2522-9818 (print) ISSN 2524-2296 (online)*

vector representations of words, as well as supporting the learning of their own models, showed that Spark NLP and spaCy are the leaders, therefore, these libraries can be successfully used for research and experimentation in order to reduce the operating time of the text unit in the proposed generalized model of the scanned documents classification system.

Comparing Spark NLP and spaCy, it should be noted that spaCy is the most documented library and possesses an extended training course, as well as the wide industrial application. In the research [18] the above mentioned libraries were analyzed on the accuracy of training, model size, time forecast, F-measures on the same data sets for all libraries in the process of training and testing. The spaCy library provided the most efficient performance and the highest results of the accuracy of training in comparison with the other models. Therefore, the experimental part of this work was performed using the spaCy linguistic library.

Aim of the study is to study the text processing and analysing libraries with the aim to speed up and increase the accuracy of the scanned text documents classifications when organizing an orderly electronic storage of the information objects.

Achieving the goal is possible by solving the following tasks:

- to perform the analysis of the modern linguistic programming libraries for sequential and parallel implementations of the most common methods of text processing and analysis;
- to offer the generalized model of the scanned documents classification system with the specified place of the block of processing and the analysis of the text;
- to provide a modification of the methods of the pre-processing module for the system with shared memory;
- to obtain statistics of the time spent on the work of the pre-processing module for the collections of the text documents of different sizes;
- to analyze the obtained results.

Materials and methods

The topicality of the research lies in reducing the time of organizing the new information resources entering the storage by increasing the speed of processing methods and analysis of the recognized text in the original image.

Figure 2 shows the place of the methods of text processing and analysis on the generalized model of the scanned documents classification system, which consists of the following units – image and text processing units, which in their turn include the modules of the scanned image pre-processing; text detection; text pre-processing; creating the frequency dictionary; defining the textual proximity. The mentioned units work in the sequence shown in fig. 2, that is the final result (the accuracy and time of classification) depends on all methods used.

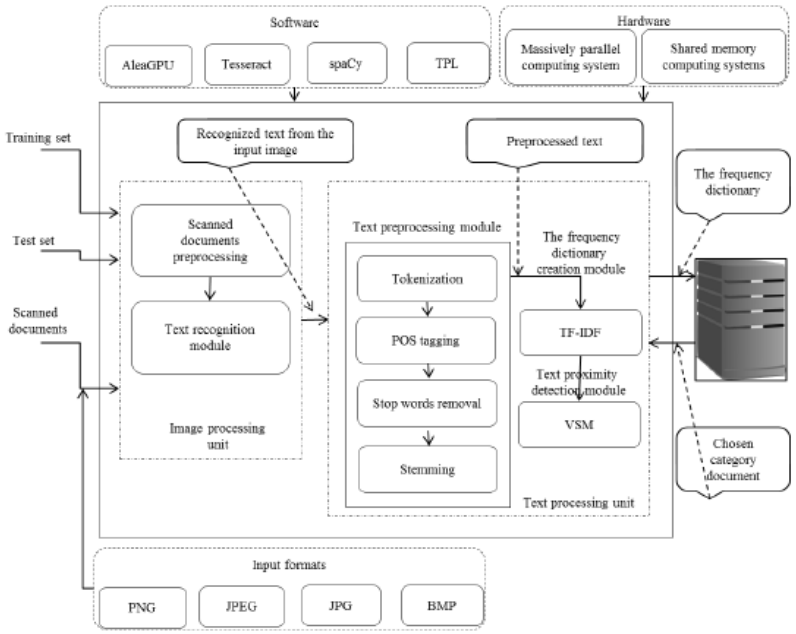
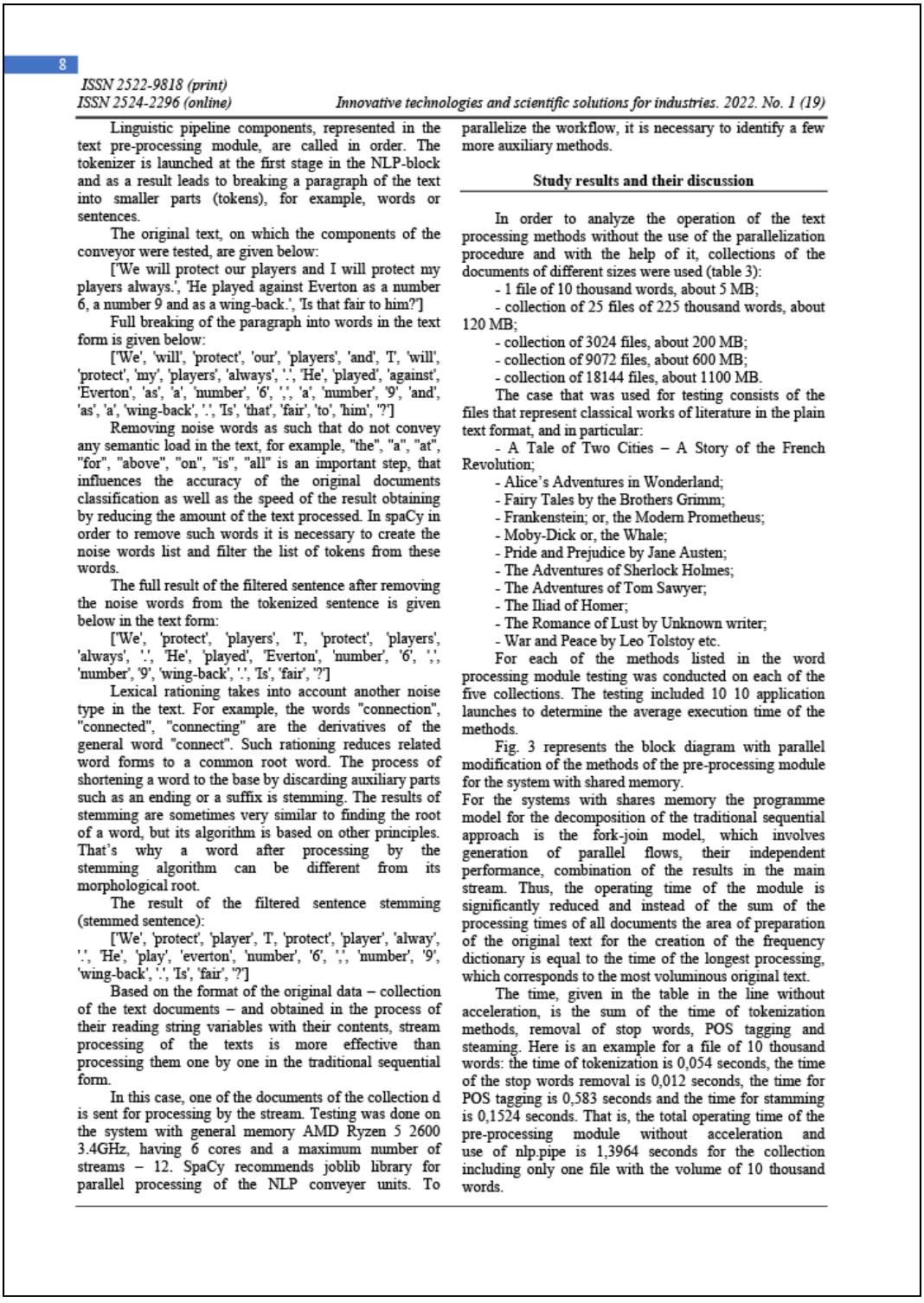


Fig. 2. The scheme of the text processing and analysis methods on the generalized model of the scanned documents classification system

Рисунок Б.9 – Стаття в журналі «Сучасний стан наукових досліджень та технологій в промисловості»



ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2022. No. 1 (19)

Linguistic pipeline components, represented in the text pre-processing module, are called in order. The tokenizer is launched at the first stage in the NLP-block and as a result leads to breaking a paragraph of the text into smaller parts (tokens), for example, words or sentences.

The original text, on which the components of the conveyor were tested, are given below:

['We will protect our players and I will protect my players always.', 'He played against Everton as a number 6, a number 9 and as a wing-back.', 'Is that fair to him?']

Full breaking of the paragraph into words in the text form is given below:

['We', 'will', 'protect', 'our', 'players', 'and', 'I', 'will', 'protect', 'my', 'players', 'always', '.', 'He', 'played', 'against', 'Everton', 'as', 'a', 'number', '6', ',', 'a', 'number', '9', 'and', 'as', 'a', 'wing-back', '.', 'Is', 'that', 'fair', 'to', 'him', '?']

Removing noise words as such that do not convey any semantic load in the text, for example, "the", "a", "at", "for", "above", "on", "is", "all" is an important step, that influences the accuracy of the original documents classification as well as the speed of the result obtaining by reducing the amount of the text processed. In spaCy in order to remove such words it is necessary to create the noise words list and filter the list of tokens from these words.

The full result of the filtered sentence after removing the noise words from the tokenized sentence is given below in the text form:

['We', 'protect', 'players', 'I', 'protect', 'players', 'always', '.', 'He', 'played', 'Everton', 'number', '6', ',', 'number', '9', 'wing-back', '.', 'Is', 'fair', '?']

Lexical rationing takes into account another noise type in the text. For example, the words "connection", "connected", "connecting" are the derivatives of the general word "connect". Such rationing reduces related word forms to a common root word. The process of shortening a word to the base by discarding auxiliary parts such as an ending or a suffix is stemming. The results of stemming are sometimes very similar to finding the root of a word, but its algorithm is based on other principles. That's why a word after processing by the stemming algorithm can be different from its morphological root.

The result of the filtered sentence stemming (stemmed sentence):

['We', 'protect', 'player', 'I', 'protect', 'player', 'alway', '.', 'He', 'play', 'everton', 'number', '6', ',', 'number', '9', 'wing-back', '.', 'Is', 'fair', '?']

Based on the format of the original data – collection of the text documents – and obtained in the process of their reading string variables with their contents, stream processing of the texts is more effective than processing them one by one in the traditional sequential form.

In this case, one of the documents of the collection d is sent for processing by the stream. Testing was done on the system with general memory AMD Ryzen 5 2600 3.4GHz, having 6 cores and a maximum number of streams – 12. SpaCy recommends joblib library for parallel processing of the NLP conveyer units. To parallelize the workflow, it is necessary to identify a few more auxiliary methods.

Study results and their discussion

In order to analyze the operation of the text processing methods without the use of the parallelization procedure and with the help of it, collections of the documents of different sizes were used (table 3):

- 1 file of 10 thousand words, about 5 MB;
- collection of 25 files of 225 thousand words, about 120 MB;
- collection of 3024 files, about 200 MB;
- collection of 9072 files, about 600 MB;
- collection of 18144 files, about 1100 MB.

The case that was used for testing consists of the files that represent classical works of literature in the plain text format, and in particular:

- A Tale of Two Cities – A Story of the French Revolution;
- Alice's Adventures in Wonderland;
- Fairy Tales by the Brothers Grimm;
- Frankenstein; or, the Modern Prometheus;
- Moby-Dick or, the Whale;
- Pride and Prejudice by Jane Austen;
- The Adventures of Sherlock Holmes;
- The Adventures of Tom Sawyer;
- The Iliad of Homer;
- The Romance of Lust by Unknown writer;
- War and Peace by Leo Tolstoy etc.

For each of the methods listed in the word processing module testing was conducted on each of the five collections. The testing included 10 10 application launches to determine the average execution time of the methods.

Fig. 3 represents the block diagram with parallel modification of the methods of the pre-processing module for the system with shared memory.

For the systems with shares memory the programme model for the decomposition of the traditional sequential approach is the fork-join model, which involves generation of parallel flows, their independent performance, combination of the results in the main stream. Thus, the operating time of the module is significantly reduced and instead of the sum of the processing times of all documents the area of preparation of the original text for the creation of the frequency dictionary is equal to the time of the longest processing, which corresponds to the most voluminous original text.

The time, given in the table in the line without acceleration, is the sum of the time of tokenization methods, removal of stop words, POS tagging and steaming. Here is an example for a file of 10 thousand words: the time of tokenization is 0,054 seconds, the time of the stop words removal is 0,012 seconds, the time for POS tagging is 0,583 seconds and the time for stamming is 0,1524 seconds. That is, the total operating time of the pre-processing module without acceleration and use of nlp.pipe is 1,3964 seconds for the collection including only one file with the volume of 10 thousand words.

Рисунок Б.10 – Стаття в журналі «Сучасний стан наукових досліджень та технологій в промисловості»

Сучасний стан наукових досліджень та технологій в промисловості. 2022. № 1 (19) *ISSN 2522-9818 (print) ISSN 2524-2296 (online)*

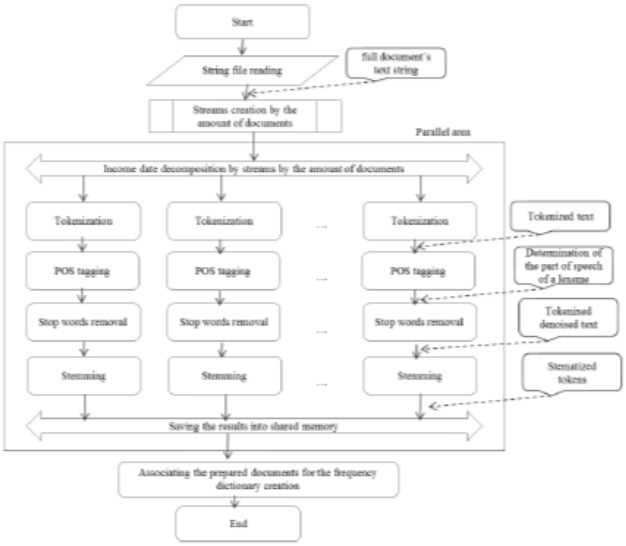


Fig.3. Block diagram with parallel modification of the preprocessing methods for the shared memory systems

Table 3. *Statistics of the time spent on the operation of the text pre-processing module*

Method		Number of files				
		1	**25**	**3024**	**9072**	**18144**
Without acceleration		1,3964 sec	8,42 sec	758,7 sec	1956,78 sec	5894,6 sec
nlp.pipe		1,781 c	6,354c	620,35c	1604,9c	5691,4c
Paralleling	2 streams	1,413 c	3,247c	350,4c	821,3c	3854,8м
	4 streams	1,348 c	2,87c	192,7c	443,46c	2310,7c

According to the results depicted on the graph (fig. 4), we can observe the tendency of acceleration but the resources of an ordinary multiprocessor-based computer with the shared memory obviously is not enough to solve such problems in the mode close to real time. Parallel algorithm for several streams has an advantage only in case of presence of a large number of documents, and the higher the number of streams the greater the acceleration. If it is necessary to analyze a small amount of text it is better to divide the data into parts and transfer to nlp.pipe.

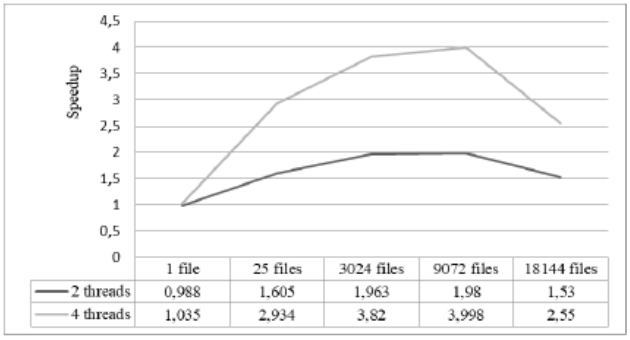


Fig.4. Graph of the text preliminary processing methods acceleration

Рисунок Б.11 – Стаття в журналі «Сучасний стан наукових досліджень та технологій в промисловості»

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2022. No. 1 (19)

The next step of the functioning of the text processing unit is the work on the creation of the frequency dictionary module and the module of determining the textual proximity. As the compilation and organization of the frequency dictionary is given in the article of the authors of this survey [2], further research will be related exactly to the definition of the textual proximity which is an important step for the efficient storage of documents and its organization. Determining the percentage of similarity between the texts in the given generalized classification system of the scanned documents is proposed to avoid duplication of documents, that is resaving of an existing document with a different name.

Conclusion

In the course of the research it was proposed a generalized classification system of the scanned documents, which consists of the image and text processing units, which in their turn include the modules of the scanned image pre-processing; text detection; text pre-processing; creating the frequency dictionary; defining the textual proximity. As the given units operate in the sequence, that is the final result (the accuracy and time of classification) depends on all methods used, the reduction of the time for classification and arrangement of the new information resources entering the system is possible by increasing the speed of processing methods and analysis of the recognized text in the original image.

Therefore, in the research it was proposed the parallel modification of the methods of the pre-processing module for the system with the shared memory and conducted its software testing with the definition of the obtained acceleration depending on the different number of the computational streams used.

Analysis of the libraries of linguistic programming, the functionality of which is focused on the processing of natural language with ready-made pre-trained models of neural networks, conveyors and vector representations of the words as well as the support of learning their own models, showed that Spark NLP and spaCy are the leaders, although, such advantages as the best documentation, industrial application, high productivity and accuracy of learning compared to other libraries determined the experimental research using the spaCy linguistic programming library.

The results showed a tendency to accelerate up to 3,998 times but, at a very high computational load (collection of 18144 files, about 1100 MB), the resources of an ordinary multiprocessor-based computer with the shared memory obviously is not enough to solve such problems in the mode close to real time. If it is necessary to analyze a small amount of text it is recommended to divide the data into parts and transfer to nlp.pipe.

Perspectives of further development

Future research will be further concentrated on reducing the time required to prepare text documents for classification, focusing the attention on large collections (more than 200 thousand words). Besides, significant is the time of the method of determining the textual proximity, which is an important step for efficient storage and organization of documents.

To achieve greater acceleration, it is planned to conduct the experiments with the application of the mass parallelism computer systems and various approaches to the original data decomposing – ordinal distribution, sentence distribution, adaptive distribution.

References

1. Barkovska, O., Kholiev, V., Pyvovarova, D., Ivaschenko, G., Rosinskiy, D. (2021), "International system of knowledge exchange for young scientists", *Advanced Information Systems*, No. 5 (1), P. 69 – 74. DOI: https://doi.org/10.20998/2522-9052.2021.1.09
2. Barkovska, O., Pyvovarova, D., Kholiev, V., Ivashchenko, H, Rosinskyi, D. (2021), "Information Object Storage Model with Accelerated Text Processing Methods", *Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021)*, No. 2870, P. 286 – 299.
3. Koroteev, M. (2020), "On the Usage of Semantic Text-Similarity Metrics for Natural Language Processing in Russian", *13th International Conference "Management of large-scale system development"* (MLSD), P. 1 – 4. DOI: https://doi.org/10.1109/MLSD49919.2020.9247691
4. Liu, Y. Sheng, Wei, Z., Yang, Y. (2018), "Research of Text Classification Based on Improved TF-IDF Algorithm", *IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE)*, P. 218 – 222. DOI: https://doi.org/10.1109/IRCE.2018.8492945
5. Zhang, Y. (2021), "Research on Text Classification Method Based on LSTM Neural Network Model", *IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, P. 1019 – 1022. DOI: https://doi.org/10.1109/IPEC51340.2021.9421225
6. Jindal, R., Shweta, (2018), "A Novel Method for Efficient Multi-Label Text Categorization of research articles", International Conference on Computing, Power and Communication Technologies (GUCON), P. 333 – 336. DOI: https://doi.org/10.1109/GUCON.2018.8674985
7. Martinek, J., Lenc, L., Král, P. (2020), "Building an efficient OCR system for historical documents with little training data", *Neural Computing and Applications*, No. 32, P. 17209 – 17227. DOI: https://doi.org/10.1007/s00521-020-04910-x
8. Pawar, N., Shaikh, Z., Shinde, P., Warke Y. (2019), "Image to Text Conversion Using Tesseract", *International Research Journal of Engineering and Technology (IRJET)*, No. 6 (2), P. 516– 519.
9. Revathi, A., Modi, N. A. (2021), "Comparative Analysis of Text Extraction from Color Images using Tesseract and OpenCV", *8th International Conference on Computing for Sustainable Global Development (INDIACom)*, P. 931 – 936. DOI: https://doi.org/10.1109/INDIACom51348.2021.00167

Рисунок Б.12 – Стаття в журналі «Сучасний стан наукових досліджень та технологій в промисловості»

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2022. № 1 (19) *ISSN 2524-2296 (online)*

10. Burns, S. (2019), *Natural Language Processing: A Quick Introduction to NLP with Python and NLTK (Step-by-Step Tutorial for Beginners)*, Amazon KDP Printing and Publishing C, 123 p.
11. Lane, H., Hapke, H., Howard, C. (2019), *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python*, Manning; 1st edition, 544 p.
12. Jurafsky, D., Martin, J.H., "Speech and Language Processing", available at: https://web.stanford.edu/~jurafsky/slp3/ (last accessed: 16.02.2022)
13. Kim, J., Hur, S., Lee, E., Lee, S. (2021), "NLP-Fast: A Fast, Scalable, and Flexible System to Accelerate Large-Scale Heterogeneous NLP Models," *30th International Conference on Parallel Architectures and Compilation Techniques (PACT)*, P. 75– 89. DOI: https://doi.org/10.1109/PACT52795.2021.00013
14. Berko, A., Matseliukh, Y., Ivaniv, Y., Chyrun, L., Schuchmann, V. (2021), "The Text Classification Based on Big Data Analysis for Keyword Definition Using Stemming," *IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT)*, P. 184– 188. DOI: https://doi.org/10.1109/CSIT52700.2021.9648764
15. Sakthi vel, S. (2021), "Pre-Processing techniques of Text Mining using Computational Linguistics and Python Libraries," *International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, P. 879–884. DOI: https://doi.org/10.1109/ICAIS50930.2021.9395924
16. Al Omran, F. N. A., Treude, C. (2017), "Choosing an NLP Library for Analyzing Software Documentation: A Systematic Literature Review and a Series of Experiments," *IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*, P. 187– 197. DOI: https://doi.org/10.1109/MSR.2017.42.
17. Vasiliev, Y. (2020), *Natural Language Processing with Python and SpaCy: A Practical Introduction*, No Starch Press, 217 p
18. Naseer, S., Mudasar Ghafoor, M., Alvi, S. bin K., Kiran, A., Shafique Ur Rahmand, Ghulam Murtazae, & Murtaza, G. (2022), "Named Entity Recognition (NER) in NLP Techniques, Tools Accuracy and Performance", *Pakistan Journal of Multidisciplinary Research*, No. 2 (2), P. 293– 308.

Received 25.02.2022

Відомості про авторів / Сведения об авторах / About the Authors

Барковська Олеся Юріївна – кандидат технічних наук, доцент, доцент кафедри Електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна; e-mail: olesia.barkovska@nure.ua; ORCID ID: https://orcid.org/0000-0001-7496-4353.

Барковская Олеся Юрьевна – кандидат технических наук, доцент, доцент кафедры Электронных вычислительных машин, Харьковский национальный университет радиоэлектроники, Харьков, Украина.

Barkovska Olesia – Ph.D (Engineering Sciences), Docent, Associate Professor Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Хомич Віктор Михайлович – студент кафедри Електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна; e-mail: viktor.khomych@nure.ua.

Хомич Виктор Михайлович – студент кафедры Электронных вычислительных машин, Харьковский национальный университет радиоэлектроники, Харьков, Украина.

Khomych Viktor – students of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Настенко Олександр Сергійович – студент кафедри Електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна; e-mail: oleksandr.nastenko2@nure.ua.

Настенко Александр Сергеевич – студент кафедры Электронных вычислительных машин, Харьковский национальный университет радиоэлектроники, Харьков, Украина.

Nastenko Oleksandr – students of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

ДОСЛІДЖЕННЯ МЕТОДІВ ОБРОБКИ ТА АНАЛІЗУ ТЕКСТУ ПРИ ОРГАНІЗАЦІЇ ЕЛЕКТРОННИХ СХОВИЩ ІНФОРМАЦІЙНИХ ОБ'ЄКТІВ

Предметом дослідження в статті є електронне сховище інформаційних об'єктів, впорядковане за визначеними правилами на етапі накопичення кваліфікаційних та наукових робіт учасників запропонованої системи обміну знаннями, наданими до системи у різних форматах (текстові, графічні, аудіо). Класифіковані роботи учасників системи є підставою для організації тематичних кімнат для обговорення із метою розповсюдження наукових досягнень, запозичення нових ідей, обміну знаннями та пошуку роботодавців або менторів у різних країнах світу. **Мета** роботи – дослідження бібліотек обробки та аналізу тексту із метою прискорення та збільшення точності класифікації сканованих текстових документів при організації впорядкованого електронного сховища інформаційних об'єктів. В статті вирішуються наступні **завдання**: дослідити методи обробки та аналізу тексту на базі запропонованої узагальненої моделі системи класифікації сканованих документів із зазначеним місцем блоку обробки та аналізу тексту; дослідити статистику зміну часу виконання розробленої паралельної модифікації методів модулю попередньої обробки тексту для системи із загальною пам'яттю для колекцій текстових документів різного розміру; проаналізувати отримані результати. Використовуються такі **методи**: паралельні чисельні методи сортування, методи

Рисунок Б.13 – Стаття в журналі «Сучасний стан наукових досліджень та технологій в промисловості»

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2022. No. 1 (19)

математичної статистики, лінгвістичні методи аналізу тексту. Отримано наступні **результати**: в ході виконання досліджень, було запропоновано узагальнену модель системи класифікації сканованих документів, яка складається з блоку роботи із зображенням та блоку роботи із текстом, які, в свою чергу, включають модулі попередньої обробки сканованого зображення; модуль розпізнавання тексту; попередньої обробки тексту; побудови частотного словнику; визначення текстової близькості. **Висновки**: запропонована паралельна модифікація модулю попередньої обробки тексту дає прискорення до 3,998 разів. Але, при дуже високому обчислювальному навантаженні (колекція з 18144 файлів, близько 1100Мб), ресурсів обчислювача на базі багатопроцесорного ЦПУ із загальною памяттю не достатньо для вирішення подібних задач у режимі, наближеному до реального часу.

Ключові слова: інформаційна система; паралелізм; обробка тексту; лінгвістичне програмування; бібліотека; прискорення; метод.

ИССЛЕДОВАНИЕ МЕТОДОВ ОБРАБОТКИ И АНАЛИЗА ТЕКСТА ПРИ ОРГАНИЗАЦИИ ЭЛЕКТРОННЫХ ХРАНИЛИЩ ИНФОРМАЦИОННЫХ ОБЪЕКТОВ

Предметом исследования в статье является электронное хранилище информационных объектов, упорядоченное по определенным правилам на этапе накопления квалификационных и научных работ участников предлагаемой системы обмена знаниями, поступающими в систему в различных форматах (текстовые, графические, аудио). Классифицированные работы участников системы являются основанием для организации тематических комнат для обсуждения с целью распространения научных достижений, заимствования новых идей, обмена знаниями и поиска работодателей или менторов в разных странах мира. **Цель** работы – исследование библиотек обработки и анализа текста с целью ускорения и увеличения точности классификации сканированных текстовых документов при организации упорядоченного электронного хранилища информационных объектов. В статье решаются следующие **задачи**: исследовать методы обработки и анализа текста на основе предложенной обобщенной модели системы классификации сканированных документов с указанным местом блока обработки и анализа текста; исследовать статистику изменение времени выполнения разработанной параллельной модификации методов модуля предварительной обработки текста для системы с общей памятью для коллекций текстовых документов разного размера; проанализировать полученные результаты. Используются следующие **методы**: параллельные численные методы сортировки, методы математической статистики, лингвистические методы анализа текста. Получены следующие **результаты**: в ходе выполнения исследований была предложена обобщенная модель системы классификации сканированных документов, состоящая из блока работы с изображением и блока работы с текстом, которые, в свою очередь, включают модули предварительной обработки сканируемого изображения; модуль распознавания текста; предварительной обработки текста; построения частотного словаря; определение текстовой близости. **Выводы**: предложенная параллельная модификация модуля предварительной обработки текста дает ускорение в 3,998 раза. Но, при очень высокой вычислительной нагрузке (коллекция из 18144 файлов, около 1100Мб), ресурсов вычислителя на базе многопроцессорного ЦПУ с общей памятью недостаточно для решения подобных задач в режиме, приближенном к реальному времени.

Ключевые слова: информационная система; параллелизм; обработка текста; лингвистическое программирование; библиотека; ускорение; метод.

Бібліографічні описи / Bibliographic descriptions

Барковська О. Ю., Хомич В. М., Настенко О. С. Дослідження методів обробки та аналізу тексту при організації електронних сховищ інформаційних об'єктів. *Сучасний стан наукових досліджень та технологій в промисловості*. 2022. № 1 (19). С. 5–12. DOI: https://doi.org/10.30837/ITSSI.2022.19.005

Barkovska, O., Khomych, V., Nastenko, O. (2022), "Research of the text processing methods in organization of electronic storages of information objects", *Innovative Technologies and Scientific Solutions for Industries*, No. 1 (19), P. 5–12. DOI: https://doi.org/10.30837/ITSSI.2022.19.005

Рисунок Б.14 – Стаття в журналі «Сучасний стан наукових досліджень та технологій в промисловості»