УДК 004.632.5

V.V. Voitsekhovskyi
I.V.Bogach
A.Y. Baranovska

THE SYSTEM OF INDEXING DOCUMENTS AND SEARCHING FOR THEIR INDEXES STORED IN DATABASE WRITTEN IN JAVA USING GRADLE BUILD AUTOMATION TOOL

Вінницький національний технічний університет

Анотація:

У *даній доповіді розглянуто систему індексування документів та їх пошуку по збереженим в базу даних індексам написану на мові Java використовуючи інструмент автоматичного збирання проектів Gradle та особливості її використання.*

Ключові слова: *програмування, програмне забезпечення, індексація, бази даних, автоматична збірка проєктів, Java, Gradle.*

Abstract:

This article examines the system of indexing documents and searching for their indexes stored in database written in Java language using Gradle build automation tool and peculiarity of its use.

Keywords: *programming, software engineering, indexing, databases, build automation tool, Java, Gradle.*

Introduction

Indexing is the process of viewing content on your computer and classifying information about them, such as the words and metadata they contain.

This is helps to search for files fast; in addition, it takes up little space – less than 10 percent of the size of indexed files. This process used on many various operating systems and platforms in order to optimize a process of search [1].

Methods of Indexing

There are two types of indexes – forward and inverted. Forward – consists in comparing the document with a list of words encountered in it. Inverted (named to contrast a forward type) compares the word is matched with a list of documents in which it is. It is logical to assume that an inverted index is best for fast searches.

Typically, search engines rank a list of documents containing queries after using an inverted index to list documents from the query. An inverted index is the most popular data structure used in information retrieval.

There are also two variants of the inverted index:

1. Index that contains only a list of documents for each word;
2. Index, additionally includes the position of the word in each document [6].

MySQL and formats of files

MySQL is highly effective DBMS that is fast and scalable. It has a number of advantages such as multithreading, support for multiple simultaneous requests; optimization of connections with joining many data in one pass; fixed and variable length records; flexible support for number formats, variable length strings and timestamps. All of these aspects is very important to create a quality indexing system so that is why it’s used here [3].

Doc, docx and pdf formats are used because it considered as a documents standard (using widely) and it is easy to extract and read data from this types of files.

Java Language and Gradle Tool

A Java programming language was used to create this system of indexing. Java has many advantages over other programming languages, which allows you to solve almost any problem with it, for example:

Java is an object-oriented language. This allows you to create modular programs whose source code that can be used repeatedly;

Java is easy to learn comparing to other OOP languages;

One of the main advantages of the Java language is the ability to transfer programs from one system to another. This language is platform independent

The disadvantage is that compared to other languages, Java is rather slow in execution [4].

Gradle is an open-source tool that helps us to create software with mechanization. This tool is widely used for the creation of different kinds of software due to its high performance. It works on Java and a Groovy-based Domain-Specific Language (DSL) for developing the project structure. Gradle supports the creation of mobile and web applications with testing and deploying on various platforms.

Main benefits of Gradle [2]:

This tool is highly customizable as it supports a variety of IDE’s. It avoids compilation.

Performance is better than in Maven and other build tools.

More flexible than other tools. You even can write some code for configuring Gradle in a script. A lot of different plugins and platforms which support Gradle

Usability in different version control systems such as GitHub

Open-source tool, so it is free.

Some disadvantages:

More difficult and complex than other build automation tool

Has own script language, not like XML using in Maven. So it can be difficult to understand it without learning documentation

Maven has more examples and more dependencies than Gradle [6].

Graddle script listing:

```
plugins {
    id 'java'
    id 'maven-publish'
}

repositories {
    mavenLocal()
    maven {
        url = uri('https://repo.maven.apache.org/maven2/')
    }
}

dependencies {
    implementation 'org.apache.poi:poi-ooxml:3.16'
    implementation 'org.apache.poi:poi-scratchpad:3.16'
    implementation 'org.apache.pdfbox:pdfbox:2.0.19'
    implementation 'mysql:mysql-connector-java:8.0.18'
}

group = 'org.example'
version = '1.0-SNAPSHOT'
description = 'TestTask'
java.sourceCompatibility = JavaVersion.VERSION_1_8
```

```
publishing {
    publications {
        maven(MavenPublication) {
            from(components.java)
        }
    }
}

tasks.withType(JavaCompile) {
    options.encoding = 'UTF-8'
}
```

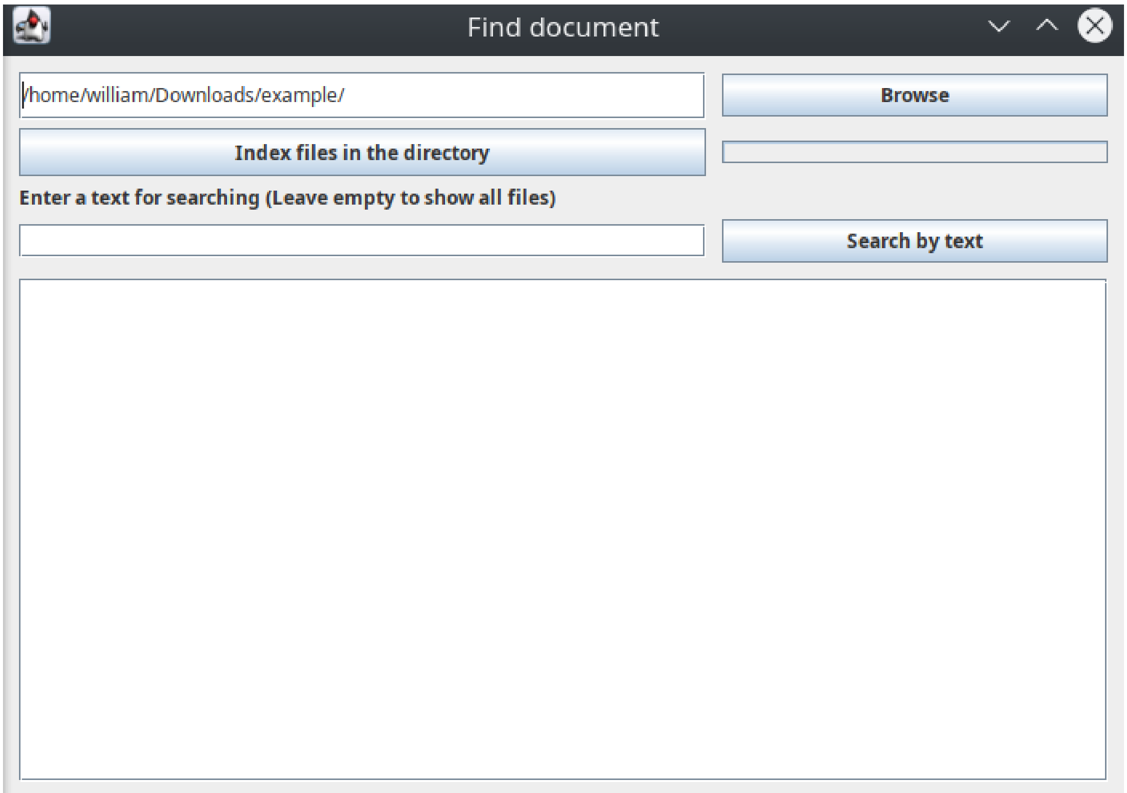
File reader method for reading pdf files example code Java:

```
public static String readPdfFile(String filePath){
        try {
            PDDocument document = PDDocument.load(new File(filePath));
            if (!document.isEncrypted()) {
                PDFTextStripper stripper = new PDFTextStripper();
                return stripper.getText(document);
            }
            document.close();
        } catch (IOException e) {
            e.printStackTrace();
        }
        return "";
    }
```

Example method of searching file by searched word Java:

```
public DefaultListModel<String> getFileListIncludesSearchingWord(String 
wordForSearch ) {
        DefaultListModel<String> fileList = new DefaultListModel<>();

        try {
            Connection connection = DriverManager.getConnection(connectionUrl, 
userName, password);
            Statement statement = connection.createStatement();

            String query = "select file_name, file_type from Files where 
file_content like " + "\"%" + wordForSearch + "%\"";
            ResultSet rs = statement.executeQuery(query);

            while (rs.next())
                fileList.addElement(rs.getString("file_name"));

            connection.close();

            return fileList;
        }
        catch (SQLException throwables) {
            throwables.printStackTrace();
            return null;
        }
    }
```

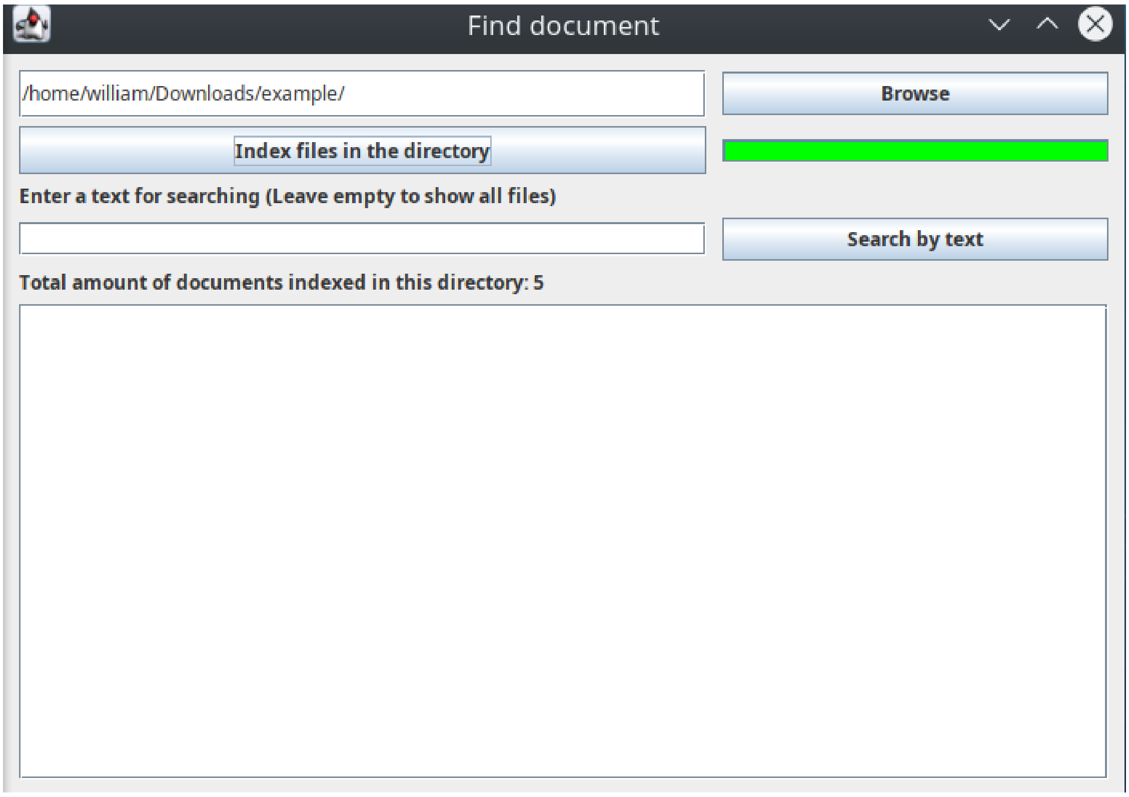
Using the system of indexing



/home/william/Downloads/example/

Enter a text for searching (Leave empty to show all files)

Figure 1 – General view of the program

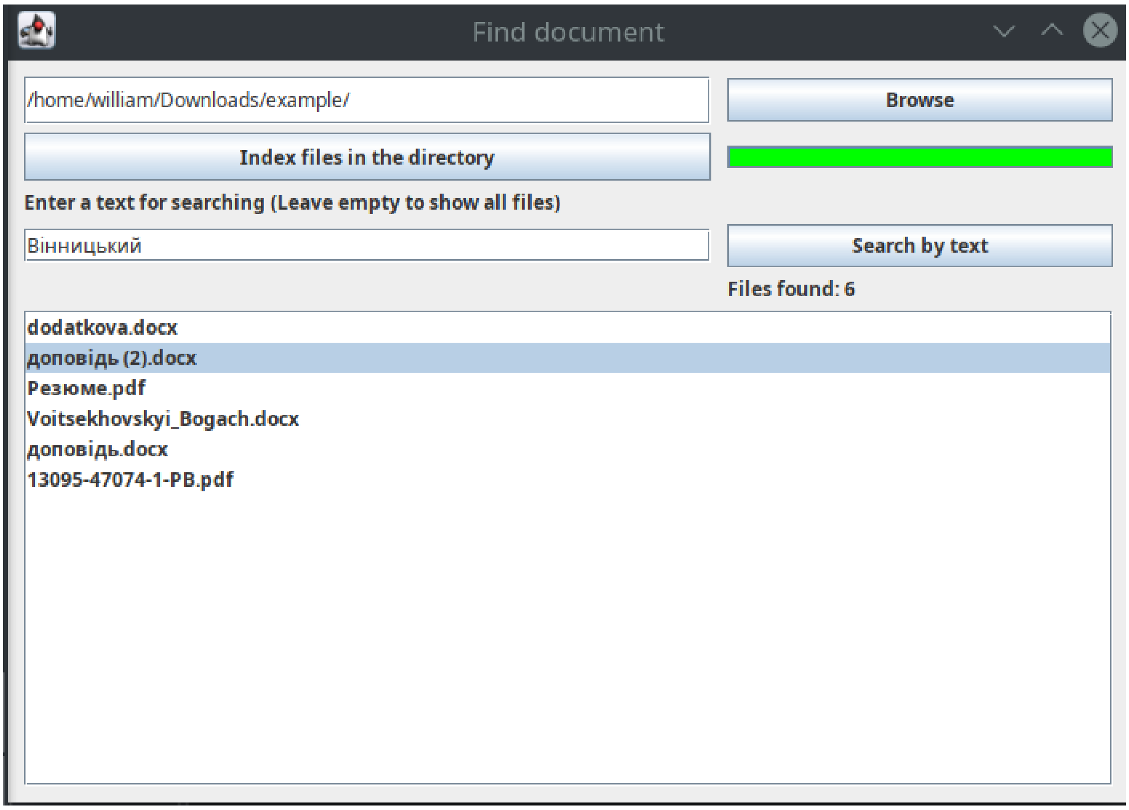


/home/william/Downloads/example/

Enter a text for searching (Leave empty to show all files)

Total amount of documents indexed in this directory: 5

Figure 2 – A specific folder is selected and a certain number
of documents that fit the file type are indexed



/home/william/Downloads/example/

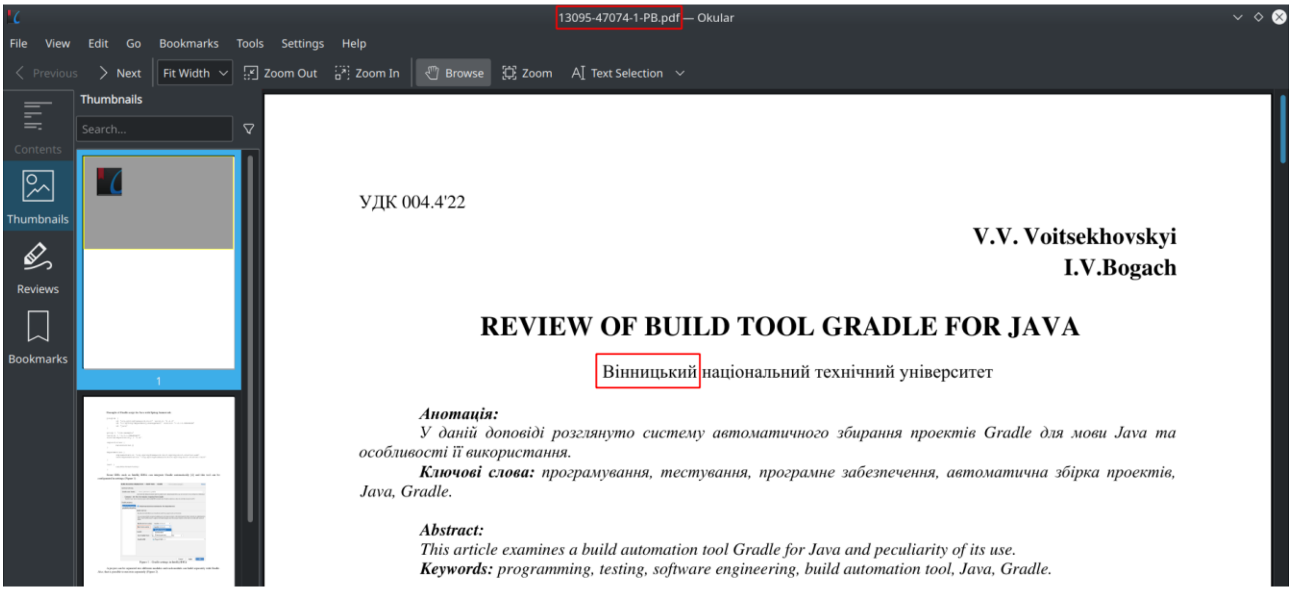
Enter a text for searching (Leave empty to show all files)

Вінницький

Files found: 6

dodatkova.docx
доповідь (2).docx
Резюме.pdf
Voitsekhovskyi_Bogach.docx
доповідь.docx
13095-47074-1-PB.pdf

Figure 3 – Showing files that match the search query.
Open the file selected in the screenshot



УДК 004.4'22

V.V. Voitsekhovskyi
I.V.Bogach

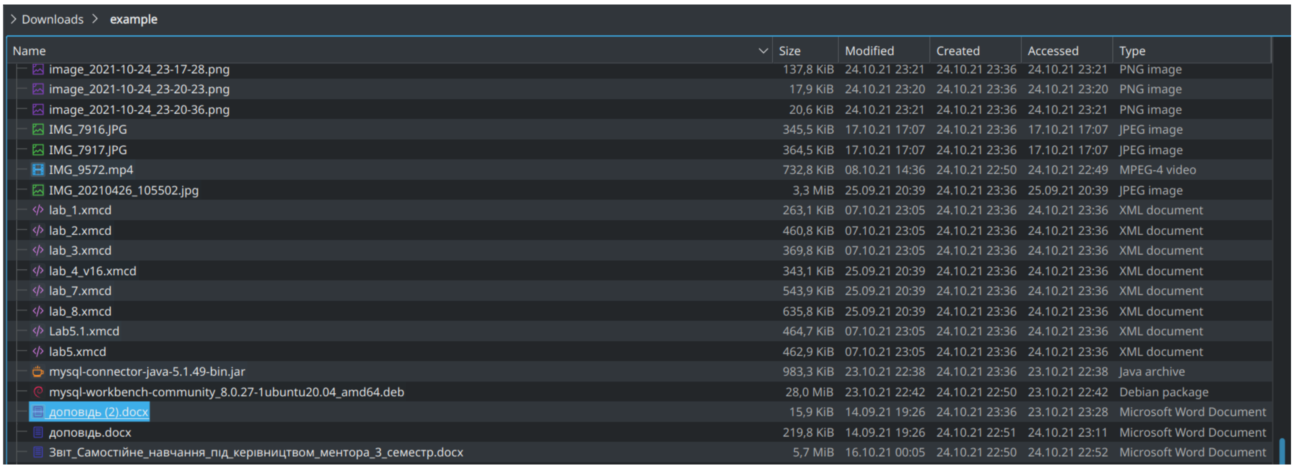
REVIEW OF BUILD TOOL GRADLE FOR JAVA

Вінницький національний технічний університет

Анотація:
У даній доповіді розглянуто систему автоматичного збирання проектів Gradle для мови Java та особливості її використання.
Ключові слова: *програмування, тестування, програмне забезпечення, автоматична збірка проектів, Java, Gradle.*

Abstract:
This article examines a build automation tool Gradle for Java and peculiarity of its use.
Keywords: *programming, testing, software engineering, build automation tool, Java, Gradle.*

Figure 4 – The specified file contains the word from the search entry



Name	Size	Modified	Created	Accessed	Type
image_2021-10-24_23-17-28.png	137,8 KiB	24.10.21 23:21	24.10.21 23:36	24.10.21 23:21	PNG image
image_2021-10-24_23-20-23.png	17,9 KiB	24.10.21 23:20	24.10.21 23:36	24.10.21 23:20	PNG image
image_2021-10-24_23-20-36.png	20,6 KiB	24.10.21 23:21	24.10.21 23:36	24.10.21 23:21	PNG image
IMG_7916.JPG	345,5 KiB	17.10.21 17:07	24.10.21 23:36	17.10.21 17:07	JPEG image
IMG_7917.JPG	364,5 KiB	17.10.21 17:07	24.10.21 23:36	17.10.21 17:07	JPEG image
IMG_9572.mp4	732,8 KiB	08.10.21 14:36	24.10.21 22:50	24.10.21 22:49	MPEG-4 video
IMG_20210426_105502.jpg	3,3 MiB	25.09.21 20:39	24.10.21 23:36	25.09.21 20:39	JPEG image
lab_1.xmcd	263,1 KiB	07.10.21 23:05	24.10.21 23:36	24.10.21 23:36	XML document
lab_2.xmcd	460,8 KiB	07.10.21 23:05	24.10.21 23:36	24.10.21 23:36	XML document
lab_3.xmcd	369,8 KiB	07.10.21 23:05	24.10.21 23:36	24.10.21 23:36	XML document
lab_4_v16.xmcd	343,1 KiB	25.09.21 20:39	24.10.21 23:36	24.10.21 23:36	XML document
lab_7.xmcd	543,9 KiB	25.09.21 20:39	24.10.21 23:36	24.10.21 23:36	XML document
lab_8.xmcd	635,8 KiB	25.09.21 20:39	24.10.21 23:36	24.10.21 23:36	XML document
Lab5.1.xmcd	464,7 KiB	07.10.21 23:05	24.10.21 23:36	24.10.21 23:36	XML document
lab5.xmcd	462,9 KiB	07.10.21 23:05	24.10.21 23:36	24.10.21 23:36	XML document
mysql-connector-java-5.1.49-bin.jar	983,3 KiB	23.10.21 22:38	24.10.21 23:36	23.10.21 22:38	Java archive
mysql-workbench-community_8.0.27-1ubuntu20.04_amd64.deb	28,0 MiB	23.10.21 22:42	24.10.21 22:50	23.10.21 22:42	Debian package
доповідь (2).docx	15,9 KiB	14.09.21 19:26	24.10.21 23:36	23.10.21 23:28	Microsoft Word Document
доповідь.docx	219,8 KiB	14.09.21 19:26	24.10.21 22:51	24.10.21 23:11	Microsoft Word Document
Звіт_Самостійне_навчання_під_керівництвом_ментора_3_семестр.docx	5,7 MiB	16.10.21 00:05	24.10.21 22:50	24.10.21 22:52	Microsoft Word Document

Figure 4 – As we can see the program ignored files with the incorrect data type

Conclusion

In this article, we reviewed and described the system of indexing documents and searching for their indexes stored in database written in Java language using Gradle build automation tool. Also summed up pros and cons of every technology. We added some code examples, configurations, and use. In conclusion, we can say that we used most proper tools and algorithms to create a fast and accurate system for both indexing and search.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Search indexing in Windows [Електронний ресурс] : [Веб-сайт]. – Режим доступу: https://support.microsoft.com/en-us/windows/search-indexing-in-windows-10-faq-da061c83-af6b-095c-0f7a-4dfecda4d15a . – Назва з екрана.
2. Gradle. Wikipedia [Електронний ресурс] : [Веб-сайт]. – Режим доступу: https://en.wikipedia.org/wiki/Gradle. – Назва з екрана.
3. 8 Advantages using MySQL [Електронний ресурс] : [Веб-сайт]. – https://devops.com/8-advantages-using-mysql/. – Назва з екрана.
4. Advantages of Java [Електронний ресурс] : [Веб-сайт]. – Режим доступу: https://www.ibm.com/docs/en/aix/7.1?topic=monitoring-advantages-java .– Назва з екрана.
5. Gradle User Manual [Електронний ресурс] : [Веб-сайт]. – Режим доступу: https://docs.gradle.org/current/userguide/userguide.html. – Назва з екрана.
6. Inverted index, Wikipedia [Електронний ресурс] : [Веб-сайт]. – Режим доступу: https://en.wikipedia.org/wiki/Inverted_index . – Назва з екрана.

Войцеховський Вільям Вільямович *– студент групи 1АКІТ-18Б, кафедра автоматизації та інтелектуальних інформаційних технологій, Факультет комп'ютерних систем і автоматики, Вінницький національний технічний університет, м.Вінниця, e-mail: fkca.1akit18.VVV@gmail.com*

Богач Ілона Віталіївна *– к.т.н., доцент кафедри Автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, м.Вінниця, e-mail: ilona.bogach@gmail.com*

Барановська Анастасія Юріївна *– студентка групи 1ІСТ-19Б, кафедра автоматизації та інтелектуальних інформаційних технологій, Факультет комп'ютерних систем і автоматики, Вінницький національний технічний університет, м.Вінниця, e-mail: 01-19-051.stud@vntu.edu.ua*

Voitsekhovskyi Viliam Viliyamovych *– student of 1AKIT-18B group, Department of Automatization and Intellectual Informational Technologies, Faculty of Computer Systems and Automatics, Vinnytsia National Technical University, Vinnytsia, e-mail: fkca.1akit18.VVV@gmail.com*

Bogach Ilona Vitaliivna *- Associate Professor of Automation and Intelligent Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: ilona.bogach@gmail.com*

Baranovska Anastasiia Yuriivna *– student of 1IST-19B group, Department of Automatization and Intellectual Informational Technologies, Faculty of Computer Systems and Automatics, Vinnytsia National Technical University, Vinnytsia, e-mail: 01-19-051.stud@vntu.edu.ua*