Plagiarism Detection Systems

Ugo Chidera Chinedu,Dr. Charles Ikerionwu, Engr. Obi Nwokonkwo

Information Management Technology, Federal University of Technology Owerri

DOI: 10.29322/IJSRP.10.03.2020.p9969
http://dx.doi.org/10.29322/IJSRP.10.03.2020.p9969

Abstract- Plagiarism is an inappropriate practice in which a person makes use of the words and ideas of another person without fully acknowledging the person. Presently, most tertiary institutions' student's project work in Nigeria is stored on shelves which can make it impossible to compare one or more projects to determine if plagiarism has been committed. When manually done, this is inefficient and tedious for any individual to be picking files one by one and comparing against other files to determine the occurrence of plagiarism. The developed system when subjected to a test of plagiarism indicates a 40% successful match of words in the body of the thesis. JAVA Programming Language is used for software development. To accomplish this task using this language the following tools was used, the JAVA development kit 1.7, Netbeans source code editor 7.4, MySQL database management 5.0 for back end support, fireworks graphics editor for the graphics. The benefits of implementing this system are that it could save time for academic staff trying to detect plagiarism in student's project and mitigate plagiarism within the institution.

I. Introduction

It is hard to define plagiarism in a way that is simple and has a wide acceptance this is because different fields and organizations may have diverse ethics as well as what might be considered ‘widely accepted’, thus does not need referencing by an expert in a given subject. On the other hand, there is a mutual understanding that plagiarism happens when someone else’s thoughts ,ideas and work is passed off as that of another person’s effort whether deliberately or unintentionally, without germane acknowledgement. It is noteworthy to recognize that plagiarism does not just apply to written work – be it essays, reports, dissertations or laboratory results- but can also apply to plans, projects, designs, music, presentations or other work presented for assessment (Ranal,2009).To detect textual plagiarism, there are some standalone tools that are valuable they are as follows:

II. Literature Review

A conceptual framework speaks to the scientist's amalgamation of writing on the best way to clarify a wonder. It maps out the activities required throughout the investigation-given his past learning of other scientists' perspective and his perceptions regarding the matter of research.

As McGaghie et al. (2001) put it: The conceptual framework "sets the stage" for the presentation of the particular research question that drives the investigation being reported based on the problem statement. The issue proclamation of a proposition shows the specific situation and the issues that made the specialist lead the examination.

Plagiarism as a word to most people may sound strange. Everyone has committed plagiarism in one way or the other during their life time without attaching any meaning to it as it is understood today. As indicated by Shiva (2006), the word plagiarism is gotten from the Latin word plagiare, which means to grab or steal. The word started being used in the English language at some point during the 1600s. While it initially meant to kidnap somebody, it bit by bit came to mean to pass off, in part of entire, another person's work as your own. Further, Merriam-Webster dictionary is of the supposition that plagiarism could mean any of the following: taking others' subjects, innovation, thoughts or words and report either verbally or recorded as a hard copy as one's own. Expansion of idea or product from an established source with credibility; theft in writing and expressions; and without giving required credits or acquiring consent before the utilization of others' creation.

For example a larger piece of religious writings have no author and was duplicated and integrated into later works. The word scholarship means to substantiate mastery of the ancient greats. There was a detectable change in this pattern precisely during the renaissance when novelty in scholarship began commanding respect and individual accomplishments are compensated in many more disciplines than it has ever been previously. This began when painters started marking their works. By the mid-1600s, allegations of copyright infringement and taking thoughts were regular in each imaginative field including technical studies. The main English copyright law was passed in the year 1709 and It has more to do with protecting the privileges of authors against book theft as it did with securing the writer's rights against corrupt printers, however, the advancement of the writer's privilege is speedy. James Boswell, otherwise called Samuel Johnson's biographer, was a legal advisor who contended one of the significant cases over to what extent copyrights went on for a creator and his or her beneficiaries (Hygeia, 2011). By the start of the nineteenth century, the idea and the law were fundamentally the same as what they are today. Indeed, even references were being utilized in a structure fundamentally the same as what they are today. What has changed from that point forward has been the issue of authorizing copyrights crosswise over fringes. Most European nations finished up understandings to forestall book robbery. The United States was the odd man out and would not give any security to foreign authors and publishers until 1891 and didn't sign on to the Berne Convention until 1988 Vinod et al. (2011).

Some outstanding characters have in one way or the other committed plagiarism. As indicated by Stephen (2005), Shakespeare stole the vast majority of his chronicled plots

straightforwardly from Holinshed. Laurence Sterne and Samuel Taylor Coleridge were both blamed for written falsification. The degree of Coleridge's copyright infringement has been battered by researchers since Thomas de Quincey, himself a practiced borrower, distributed a report in Tait's Magazine two or three weeks after Coleridge's passing. Oscar Wilde was blamed for written falsification: consequently the commended trade with Whistler: "I wish I'd said that, James." Furthermore, Stephen (2005) is of the view that students who are lazy and mendacious are not by any means the only individuals who plagiarize. For example, Martin Luther King counterfeited some portion of a part of his doctoral thesis. Expressed that George Harrison was effectively sued for stealing the Chiffons' He's So Fine for My Sweet Lord. Stephen (2005) indicates that Alex Haley duplicated huge entries of his novel Roots from The African by Harold Courlander. Princess Michael was blamed for literary theft over her book on illustrious ladies. Jayson Blair, at that point a journalist for the New York Times, plagiarized numerous articles and faked quotes. In 1997, under a half year after winning the Booker prize, Graham Swift's Last Orders was at the focal point of allegations that the writer had gone too far among motivation and copyright infringement by "legitimately mirroring" a prior work, the 1930 novel As I Lay Dying by William Faulkner. Gone up against the allegations, Swift said his book was a "reverberation" of Faulkner's.

Innovation has made a difference a lot over the most recent 200 years, however, the significance we append to it might decay. As observed by McKay (2009) this" started to change during the renaissance when unique grant turned out to be increasingly regarded and singular achievement was perceived in a lot more fields than it had been already (for instance, this is when painters started marking their works). The point here is that, by the mid-1600s, allegations of stealing and plagiarism were common in every innovative field including the science. An allegation of taking another person's words or thoughts and passing them off as your very own was one of the most exceedingly awful affronts believable and reason for claims and legal actions".

Object-oriented Programming (OOP) : As per Kwanzul-natal (2007), Object-oriented Programming speaks to an endeavor to make programs all the more closely model how individuals consider and manage the world. In the more established styles of programming, a software engineer who is confronted with some issue must distinguish a processing task that should be performed to tackle the issue. It is a lot of instruments and strategies that empower programming specialists to manufacture dependable, easy to use, and viable, very much archived, reusable programming frameworks that satisfy the necessities of its clients. It is asserted that object-direction gives programming engineers new personality apparatuses to use in taking care of a wide assortment of issues. Item direction gives another perspective on calculation. A product framework is viewed as a network of items that collaborate by passing messages in taking care of an issue. Instances of Object-oriented Programming Languages are Java, Python, Ruby, C++, and Smalltalk.

III. Related Works

Skyline. Inc. developed standalone plagiarism-detector anti plagiarism software, which detects plagiarized text. It is an autonomous Microsoft Windows-based computer desktop application made with Visual C#.Net. It follows the exact substring detection method and it is used in the academic environment. At any rate, it has its shortcomings which is that it runs just on windows operating system, it raises a huge amount of false positives (it flags a sentence as plagiarized even though it is not). It also lacks plagiarism prevention mechanism because there is no module or subsystem in place to deter or discourage plagiarism.

Sherlock: As demonstrated by Shahabi (2012), Sherlock, is a program used to recognize copyright encroachment for essays, computer source codes files and other kinds of textual documents in digital form. Sherlock works by converting text it receives into digital signatures to measure the similarity between the documents. A digital signature is a number formed by changing several words (3 by default) in the input into a series of bits and joining those bits into a number. Sherlock is developed with C programming language and it requires compilation before being installed either on Unix/Linux or Windows. It doesn't have a GUI as it is a command-line program. When a "sherlock *.txt" command is issued sherlock will compare all the text files in the current directory and produce a list of file pairs together with a similarity percentage. Research has shown that 100% similarity index does not suggest that the files are identical because Sherlock actually throws away some data randomly in the process in order to simplify and speed up the match (Shahabi 2012). In addition, sherlock requires recompilation each time it is to be used in other platforms apart from windows. The nature of sherlock as a command line tool makes it not to be user friendly as users have to remember all the relevant commands. As observed by Martins, *et al.* (2010), It allows for control over the threshold which conceals the percentage with lower values, the number of words per digital signature and the granularity of the comparison by use of the respective arguments: - t, - n and - z. It supports natural language. Sherlock is a fascinating case as its outcomes are a rundown of sentences in a "Document 1 and File 2: Match%" group. This configuration does not enable the client to locate the most noteworthy matches, it essentially makes the outcomes simpler to post-process. Results were created with every one of the blends of the settings from - n 1 to 4 and - z 0 to 5, Sherlock was created utilizing the C language.

Ferret: It was developed at the University of Hertfordshire it is a freely available standalone plagiarism detection software. It runs on windows environment and very easy to install as well. It can process files with the following extensions .txt, .rtf, .doc and .pdf. Ferret's algorithm was written in C++. Ferret takes a set of documents, converts each text into reference numbers and set of characteristic trigrams. When documents are fed to Ferret it converts those documents into reference numbers-set of characteristic trigrams. In order evaluate the writings it got it checks the quantity of one of a kind trigrams and afterward depends on its findings to create a rundown as record sets with its accompanying likeness score that generally ranks from the most comparable pair to the least comparable pair. This count is used to count the resemblance measure, as the number of similar trigrams in a pair of documents, divided by the total number of different trigrams in the pair. Ferret manifests the scores of similarity precisely, such as 0.90991. The system allows user to select any pair of texts and do further investigation as they will be displayed

side by side with similar paragraphs highlighted for instance similar parts in blue and different parts in black. Since it is developed with C++ language it implies that it can only run on windows platform and not on any other platform. C++ being a poor string manipulator makes Ferrets detection ability slow (Shahabi, 2012).

Extant literature suggest that, there is no significant work done to address these shortcomings.

iThenticate: According to Asim *et al.* (2013) iThenticate compares a given document against the document sources available on the World Wide Web. It also compares the given document against proprietary databases of published works (including ABI/Inform, Periodical Abstracts, Business Dateline), as well as numerous electronic books and produces originality reports. The originality reports provide the amounts of materials copied (in percentages) to determine the extent of plagiarism, was developed using PHP and supported by an MSQL back end database.

IV. Methodology Adopted

Structured System Analysis and Design Methodology (SSADM) is a collection well thought out standards used for analyzing systems and application design. It uses a formal methodical approach to the analysis and design of information system. It was explicitly adopted because there is the need to interact with the supervisors and other stake holders who will be using it as to the require features needed to be present on it and also the need to inquire about the current system of storing students projects and how it is utilized to combat plagiarism. The advantage of utilizing this technique is that firstly when used it's many stages form a beneficial set of real-world prescriptions that consider real challenges and conditions. This empowers the creators to practically consider the suggested data framework. We must recall that Key phases of SSADM incorporate feasibility study, where the designer should establish if the recommended system is attainable and the study current system demands that systems are reviewed. This guarantees all suggested changes are valuable and essential.

Source Code

```
import java.awt.Component;
import java.awt.FlowLayout;
import java.awt.PopupMenu;
import java.awt.event.ActionEvent;
import java.io.FileInputStream;
/**
 *
 * @author franc
 */
public class ProofRead extends javax.swing.JFrame {
   private BufferedImage img2;
   private JButton b;
   private JTextFieldtx;
   private JLabel jp1;
   private JLabeltxl;
   private JLabeljpl;
private void jButton12ActionPerformed(java.awt.event.ActionEventevt) {
      File file = null;
      File oldfile = null;
ResultSetrs = null;


      Connection con = null;
FileInputStreamoldfis = null;
PDDocument extractor = null;
PDDocumentoldextractor = null;

PDDocumentolddocument = null;
      try{
         file = new File(jTextField7.getText());
         Properties prop = new Properties();
prop.setProperty("user","root");

prop.setProperty("password","");
con = DriverManager.getConnection("jdbc:mysql://localhost/proofread",prop);
      if(con != null )
System.out.print("CONNECTION SUCCESSFUL");
PreparedStatementst = null;
FileInputStreamfis = new FileInputStream(file.getAbsolutePath());
      // document = new XWPFDocument(fis);
      // extractor = new XWPFWordExtractor(document);
       //String fileData = extractor.getText();
      // String[] filed = fileData.split("\\n");

          document  =  PDDocument.load(fis);
            if (!document.isEncrypted()) {
PDFTextStripper stripper = new PDFTextStripper();
                String fileData = stripper.getText(document);
                filed = fileData.split("\\n");
             }
st = con.prepareStatement("select project from registration");
rs = st.executeQuery();
      List<String> result = new ArrayList<String>();

      while(rs.next()){
result.add(rs.getString("project"));
      }

          while(ii <result.size()){
System.out.print(ii);
oldfile = new File(result.get(ii));
oldfis = new FileInputStream(oldfile.getAbsolutePath());
olddocument = PDDocument.load(oldfis);
PDFTextStripper stripper2 = new PDFTextStripper();
                String fileDataold = stripper2.getText(olddocument);
                String[] filedo = fileDataold.split("\\n");

                for(int i = 0; i<= filed.length; i++)
            try{
                if(filedo[i] != null && filed[i] != null)
    if(filed[i].equals(filedo[i]) && filed[i] != ""){
```

```
    intern = intern + 1;
    list1.add("File " + file.getName() + " equals " +
oldfile.getName()+ " on "+ filed[i]   );
                }
            }
                catch(IndexOutOfBoundsExceptionie){
Logger.getLogger(ProofRead.class.getName()).log(Level.SEVE
RE, null, ie);
                continue;
                }
                ii++;
            perce = (intern * 100)/filedo.length;
            }
```

System design

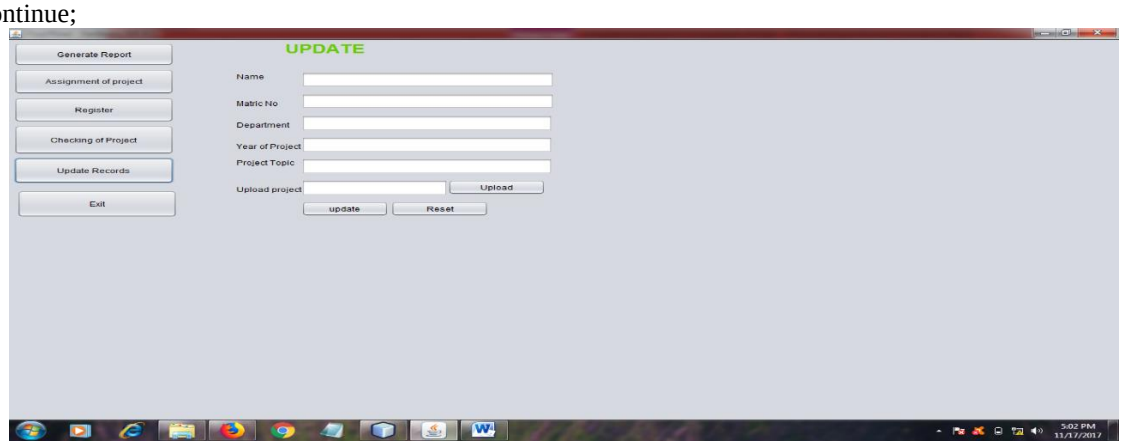

Figure 2:Update screen capture of the system

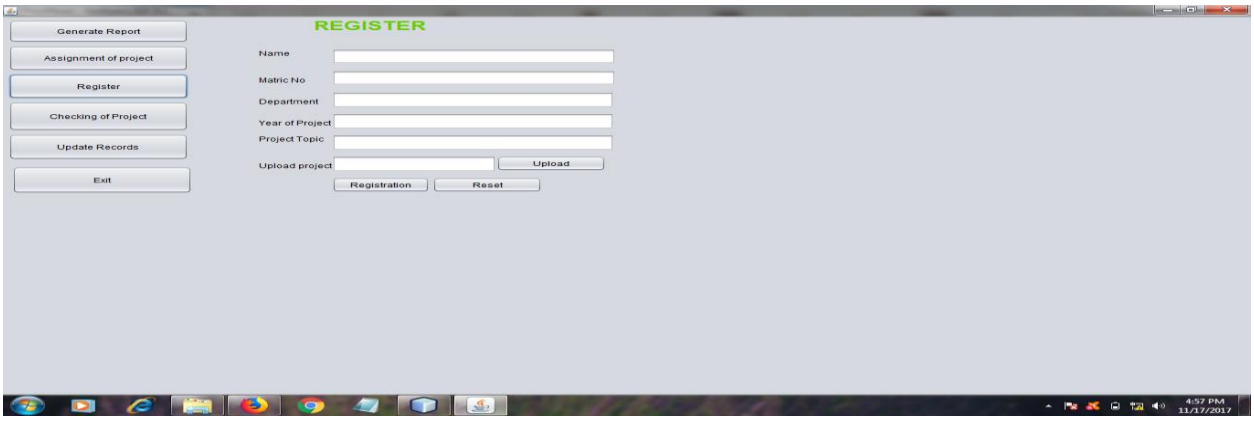

Figure 3: Registration screen capture of the system

V. RESULTS

The data presented below is the performance indicator (speed) of each of the aforementioned plagiarism detection system including the manual system being practiced in FUTO and the proposed system when presented with a large volume of the file to check for plagiarism.

Table 4.1 as shown below presents the performance details of the existing system using the speed performance indicator with file size as yard stick.

Table 4.1: Data presentation of the existing system

S/N	Plagiarism detection tool	Max file size accepted	Speed of completion
1	Ferret	15MB	3 Mins
2	Plagiarism-detectorby Skyline inc	10MB	4Mins
3	Sherlock	30MB	4Mins

4	Proposed System	Tested with 100MB file	10seconds
5	Manual System	120 page booklet	4hours

The Manual System: The manual system of detecting plagiarism is tedious as it requires the user to flip through multiple hard copy files and reading through page by page to determine if plagiarism has occurred.

As a result of the tediousness of the manual system, users are reluctant to indulge in it thereby losing interest in plagiarism detection.

There is virtually no plagiarism deterrence in this system as project topics that have been treated before can still be reassign to students because there is no electronic means of storing project topics treated before so as to avoid reassignment or awarded to a student if he promises improvement on the existing one.

Benefits of the proposed system

It is user friendly as it made use of graphical user interfaces and very intuitive.

It runs on every platform as it is developed with the Java program.

It does not raise false positives as Java is an efficient string manipulation language in other words it is accurate and precise.
Since it is a standalone desktop application it is reliable in the context of its purpose as it does not rely on internet service availability to function.

It features a plagiarism deterrence system as it spots a module to assist supervisors to check for already taken project topics so as to assign project topics that will yield the needed impact and encourage originality of ideas thereby reducing the tendency to plagiarize.

VI. CONCLUSION

Based on the findings of the study the following conclusions are drawn.

1. Plagiarism detection systems developed with PHP, C, C# or C++ lack speed due to the poor string manipulation capabilities of these languages.
2. Plagiarism detection tools developed with C or PHP has a very high possibility of throwing false positive alarm that is flagging a sentence or a sequence of words as plagiarized whereas in the real sense it is not.
3. Plagiarism tools developed with C, C++, C# run only on the Windows operating system denying users the ease to use the tools across all platforms be it Linux, Ubuntu, Macintosh and so on.
4. Virtually all standalone plagiarism detection system lacks a plagiarism deterrence module to try and discourage plagiarism.

REFERENCES

[1] Asim, M., El Tahir Ali, Hussam, M., Dahwa Abdulla, and Vaclav Snasel, (2011)."Overview and Comparison of Plagiarism Detection Tools", V. Snasel, J. Pokorny, K. Richta (2nd Eds.): Dateso, pp. 161-172.

[2] Hygeia, J. D., Med Vinod, K. R. (2011). Plagiarism- history, detection and prevention. Retrieved March 20, 2017 from http://www.hygeiajournal.com/downloads/Editorial/1597787464plagiarism.pdf.

[3] Martins, V. T., Fonte, D., Henriques, P. R and Cruz, D. D. (2014).Plagiarism Detection: A Tool Survey and Comparison. Retrieved March 20, 2017 from https://pdfs.semanticscholar.org/70b7/7f68d493021b10608eefd050e8491ac65cfa.pdf.

[4] McKay, J. J. (2009, March 02). A brief history of plagiarism. Retrieved April 5, 2017 from http://johnmckay.blogspot.com.au/2009/03/very-brief-history-of-plagiarism.html

[5] McGaghie, W. C.; Bordage, G.; and J. A. Shea (2001). Problem Statement, Conceptual Framework, and Research Question. Retrieved from http://goo.gl/qLIUFg

[6] Ranald , M (2009, April 4). UK Physical Sciences Centre - Briefing Paper, Retrieved 16th May 2018 from https://www.heacademy.ac.uk/system /files/new_ plagiarism.pdf

[7] Shahabi, M. (2012, February 10). International Journal of Computational Linguistics (IJCL), Volume (3):Issue (1). Retrieved April 23, 2018 from http://www.cscjourals.org/manuscript/Journals/IJCL/Volume3/Issue1/IJCL-33.pdf.

[8] Shiva, S. (2006, May 07). What's the origin of the word plagiarism?. Retrieved March 8, 2017 fromhttps://timesofindia.indiatimes.com/home/sunday-times/Whats-the-origin-of-the-wordplagiarism/articleshow/1519035. cms

AUTHORS

First Author – Ugo Chidera Chinedu, M.Sc Information Manage, Federal University of Technology Owerri,chidera_chinedu@yahoo.com
Second Author – Dr. Charles Ikerionwu, PhD Software Engineer, Glasgow University, charles.ikerionwu@futo.edu.ng
Third Author – Engr. Obi Nwokonkwo(PhD), PhD Information Management Technology, Federal University of Technology Owerri, obinwokonkwo@gmail.com