Conserve the Originality: Detection of Plagiarism in Documents

Aakanksha Shinde[1], Mrunmayee Dixit[2], Himani Solanki[3], Prof. Sheetal Thakare[4]
Department of Computer Engineering
[1,2,3,4] Bharati Vidyapeeth College of Engineering, Navi Mumbai, Maharashtra, India

Abstract- *Plagiarism has repeatedly occurred in India, resulting from focusing on such academic misbehavior as a "major issue" in Indian higher education. Plagiarism is stealing and publication of another person's language, thoughts, or ideas and the representing as one's own original work. It defiles ethical work and degrade the quality of the Education and Research in any university across the globe. Plagiarism is not in itself a crime, but it constitutes copyright infringement. In academia and industry, it is a serious ethical offense. Many new models have emerged over the years for strategies and systems for detection, penalties and mitigation, based on deeper understanding of the underlying reasons behind student plagiarism. This paper presents the review of the technology implemented that checks plagiarism of any document. It uses Naïve algorithm for detection of plagiarism. The plagiarism detection tool has been developed ins Node.js. It suggests that faculty members and research scholars can use this application for checking their thesis or research papers before submitting to universities or conferences and also this tool can help to analyze student's true capabilities and help the teachers tremendously in plagiarism detection.*

Keywords- plagiarism detection, comparison, word counter, naïve algorithm, database.

I. INTRODUCTION

Nowadays, the Internet is the biggest source of information. People can easily search, get access and browse the web to get the information they need. It would be difficult to do scientific research without the Internet and web space. Furthermore, due to the size and digital structure of the internet, it is easy to illegally use someone else's work now. The most common plagiarism is written text document which is formed by copying some or all parts of the original document, sometimes with some modifications. Identification of documents which were copied is stressful and time-consuming process to humans due to the large number of documents which have to be analyzed. The documents in digital format make the process of plagiarism quite simple, it means that such cases of plagiarism can be traced automatically.

Plagiarism Detection tool is a perfect platform to check paper for plagiarism, in order to verify the integrity of its written content. It actually identifies fragments of identical text. It compares the submitted text against database and identifying identical, or near-identical passages and over the world wide web. Plagiarism Checker API checks content sentence wise and allow full article detection with word limit 6500 words per month. So, the application includes word counter tool which counts the number of words in the dissertation. The implementation of the Plagiarism Detector tool also works for comparing the submitted files and highlight their differences in result.

Plagiarism detection software benefits academicians, research scholars and students interested in safeguarding their writing. Through plagiarism detection tools research community can benefit by having their research paper/thesis and dissertation checked for any plagiarism done unintentionally and ensure that the text is unique.

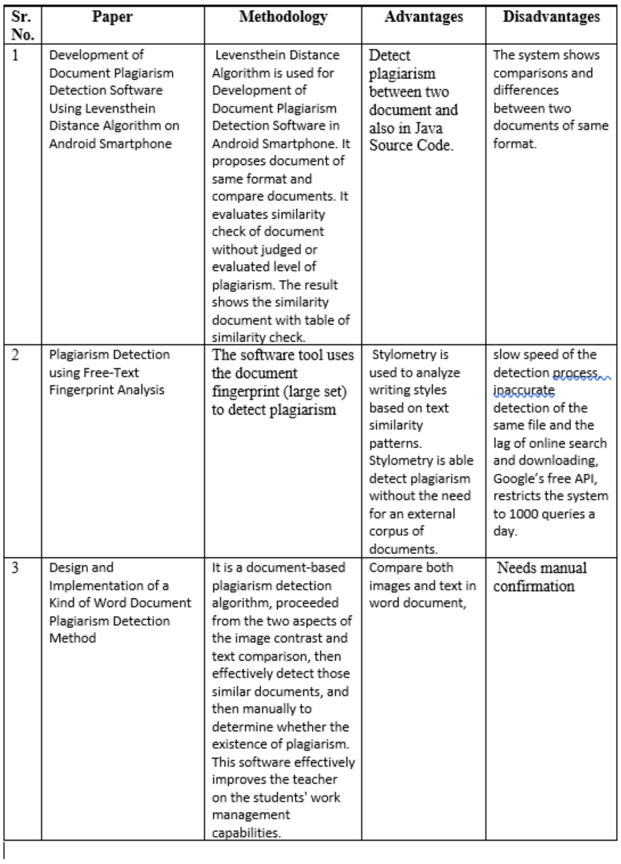
II. LITERATURE SURVEY

The Plagiarism Detection Tool has to be effective and efficient in its implementation. Various proposals have been put forward and some of them already implemented. So, a survey was done among different proposals and this paper includes survey about different methods for checking plagiarism.

The paper [1] Development of Plagiarism Detection Software is based on Levenshtein Distance Algorithm. The system shows the comparisons and differences between two documents of same format. In information theory and computer science, the Levenshtein distance is a string metric, which is one way to measure edit distance. The Levenshtein distance between two strings is given by the minimum number of operations, and that needed to transform one string into the other, where an operation is an insertion, deletion, or substitution of a single character. In this paper, it uses a diagonal line that is derived from the Levenshtein distance in the plagiarism detection to decrease the scarcity of the dynamic programming array.

Research conducted by [2] uses Fingertip Analysis and Google search API for plagiarism detection. The system is able to detect plagiarism between two given documents, given document and group of local documents, and between given document and online available documents. Agile software methodology was used to develop the software and some open source libraries were manipulated and used to search the internet. It splits a document into a (large) set of ‘fingerprints. A set of fingerprint contains pieces of text that may overlaps with one another. A fingerprint is then used as a query to search the web or a database, in order to estimate the degree of plagiarism.

Based on the needs of document plagiarism detection the author [3] put forward a document-based detection method, which compare the images and text of the documents, so that to effectively identify similar documents. The paper [3] is implemented using Visual Basic Language.



Sr. No.	Paper	Methodology	Advantages	Disadvantages
1	Development of Document Plagiarism Detection Software Using Levensthein Distance Algorithm on Android Smartphone	Levensthein Distance Algorithm is used for Development of Document Plagiarism Detection Software in Android Smartphone. It proposes document of same format and compare documents. It evaluates similarity check of document without judged or evaluated level of plagiarism. The result shows the similarity document with table of similarity check.	Detect plagiarism between two document and also in Java Source Code.	The system shows comparisons and differences between two documents of same format.
2	Plagiarism Detection using Free-Text Fingerprint Analysis	The software tool uses the document fingerprint (large set) to detect plagiarism	Stylometry is used to analyze writing styles based on text similarity patterns. Stylometry is able detect plagiarism without the need for an external corpus of documents.	slow speed of the detection process, inaccurate detection of the same file and the lag of online search and downloading, Google’s free API, restricts the system to 1000 queries a day.
3	Design and Implementation of a Kind of Word Document Plagiarism Detection Method	It is a document-based plagiarism detection algorithm, proceeded from the two aspects of the image contrast and text comparison, then effectively detect those similar documents, and then manually to determine whether the existence of plagiarism. This software effectively improves the teacher on the students' work management capabilities.	Compare both images and text in word document,	Needs manual confirmation

Table 1: Literature Survey

III. METHODOLOGY

The overall working is divided into four sections:

i. **Plagiarism check using Database -** It processes the text to find matching sections of words between the input text it is processing and the ones it has indexed in its databases. The database consists of numerous downloaded research papers and thesis of different branches in the field. It gives the percentage of the plagiarized text after clicking ‘Check Plagiarism’.

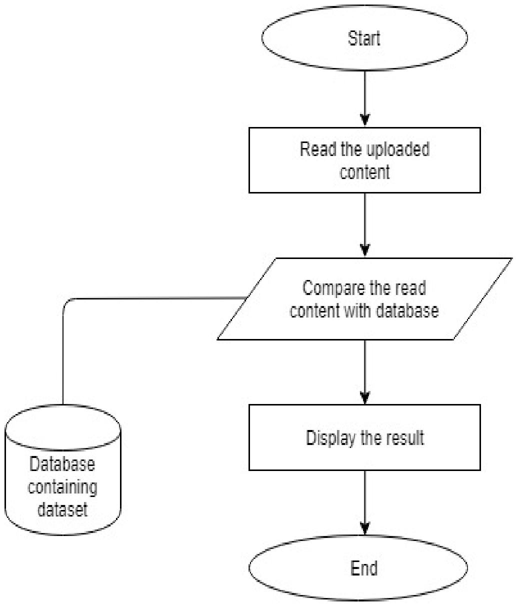

Start
Read the uploaded content
Compare the read content with database
Database containing dataset
Display the result
End

Fig.1 Flow Chart of plagiarism check over DB

ii. **Check Duplicate content in uploaded files-** In this section plagiarism is checked by comparing text files from dataset and the data files uploaded by the users to check similar content between two files. It will not match content all over the internet. Comparing them with each other does not mean that content is 100% plagiarism free. It means that text is not matched or matched with another specific document or website. It could be useful when teachers want to check work of their students on same topic, etc. By using plagiarism detection tool, you can easily compare two documents for duplicate content. It finds out the similarities between text documents, and can compare two text files for plagiarism.

The text file uploaded on the browser and department chosen by the users will be fetched and stored by the server. By using the in-built module of node i.e. fs module the server reads the fetched text file and stores the text file into the string and convert the string into the tokens. Once the conversion is done then the code will check the field chosen i.e. computer, mechanical, chemical, etc. and accordingly fetch the document from the dataset for comparison. The comparison will be done by using Naïve algorithm. After comparing and displaying the result, the uploaded file will get deleted from the server itself.

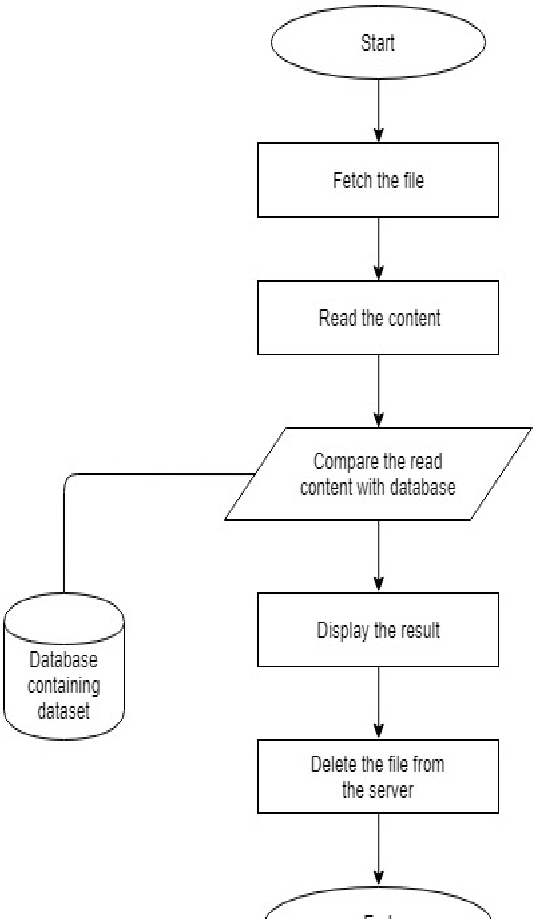

Start
Fetch the file
Read the content
Compare the read content with database
Database containing dataset
Display the result
Delete the file from the server

Fig.2 Flowchart of Plagiarism check in uploaded file

iii. **Plagiarism over world wide web-** It is home-grown plagiarism detection method built on top of Google's search API and has produced superior results as compared to leading software packages like Turnitin and Mydropbox. This is mainly due to Google's indexing of many more web sites as compared to these tools for plagiarism detection. The disadvantages of using Google's free API have restricted their system to 6500 words per month. Instead of license their Search engines, Google maintain the exclusive control to numerous potential applications. This application works on the same plagiarism test principle but the function is very much the same as Google or any other search engine works to find the matching words or phrases in other sources and provides the best results, along with the plagiarism checker percentage and provide all the link from which the fragment is copied.

The process of file detection is as follows:

1) Firstly, the text is fetched on the server and the process for directing to API gets initiated. This process directs the user to the Copyleaks API.
2) API then compares the input text with all the data over internet and a status is provided. The status gets displayed according to the progress of the comparison.
3) Once the status has reached the value of 100% then it redirects to the new page which displays the result.
4) This new page displays the output of the plagiarism in the json format. JSON stands for JavaScript Object Notation. It is a lightweight format for storing and transporting data.

iv. **Word Counter-** In this section, it checks the number of words of the input text. To check the word count, simply place the cursor in the box and start typing or simply copy paste the text from another file into the editor box. Word Counter will help to make sure its word count reaches a specific requirement or stays within a certain limit and then one can accordingly use this data to check plagiarism over World Wide Web.

The HTML DOM is a standard **object** model and **programming interface** used for word counter. The Document Object Model (DOM) is a platform and language-neutral interface that allows programs and scripts to dynamically access and update the content, structure, and style of a document on browser. The text-miner packages are also installed to interact with the document. Here, it dynamically calculates the number of words and displays the result on the browser itself.

The implementation of the Plagiarism Detection application is in Node.js. Node.js is a runtime open source server-side environment. It also facilitates handling multiple client requests. The requests of the users are stored into the buffer and executed properly. Thus, unnecessary use of threads is avoided and working is done smoothly.

Nodemon is used for the automatic starting of the server. Nodemon is a tool that helps develop node.js based applications by automatically restarting the node application when file changes in the directory are detected. Different npm modules like express, etc. are also implemented. Express helps in routing which refers to how an application's endpoints (URIs) respond to client requests. In-built modules of node like fs is used for file access and to perform different operations on the uploaded file. While, HyperTerminal is used to set up the connections to other computers using Telnet.

IV. CONCLUSION

In this technological era, plagiarism detection is essential for protecting the written work. Plagiarism happens for a number of reasons; some students try to gain credits for the work of others. However, most incidents of plagiarism are not the product of deliberate cheating, but of underdeveloped academic skills. Students try to plagiarism because of intellectual insecurity i.e. use of own words paradox, poor time management, lack of clear argument, inadequate research and poor note making. So, faculty members or teachers and research scholars have to use one of this anti plagiarism software in checking their thesis or research paper or articles before submitting to universities. This will definitely improve the quality of the thesis work of the research scholars, authors of journal article and Students submitting their work and will also help teachers to check and improve their student's assignment and work.

ACKNOWLEDGEMENT

After having endured the experience, there were many who helped us in our project and we would like to thank everyone. We are deeply indebted to our Principal, Dr.M.Z. Shaikh, and our Head of Department, Dr. D. R. Ingle, for giving us this opportunity without which it wouldn't have been possible to successfully complete our survey. We also thank our guide, Professor Sheetal Thakare, for her kind help and guidance during the work, and also for providing us with the opportunity to work on "Conserve the originality: Detection of Plagiarism in Documents".

REFERENCES

[1] Nurhayati, Busman, Development of Document Plagiarism Detection Software Using Levensthein Distance Algorithm on Android Smartphone.

[2] Mohamed Elkhidir, Mohannad M. Ibrahim, Tarig A. Khalid, Shawgi Ibrahim, Mohamed Awadalla, Plagiarism Detection using Free-Text Fingerprint Analysis.

[3] Feng Jian, Design and Implementation of a Kind of Word Document Plagiarism Detection Method.