ISSN No:-2456-2165

# Copy Checker

[1]Ajith A P
Student, Computer Science and Engineering Department
Sahrdaya College of Engineering & Technology Thrissur, India

[2]Ajmal Noorudheen Arakkal
Student, Computer Science and Engineering Department
Sahrdaya College of Engineering & Technology Thrissur, India

[3]Ashish Xavier
Student, Computer Science and Engineering Department
Sahrdaya College of Engineering & Technology Thrissur, India

[4]Ebin Joseph
Student, Computer Science and Engineering Department
Sahrdaya College of Engineering & Technology Thrissur, India

[5]Dr.Arun Thomas
Asso. Professor, Computer Science and Engineering Department Sahrdaya College of Engineering & Technology Thrissur, India

**Abstract**:- **As technology has a great impact in our current lifestyle compared to old times, plagiarism is a phenomenon that is increasing day by day. The effect of the pandemic Covid-19 even moved into online space.. Students complete their assignments independently, at the same time another part of students either copy from others or download from the internet. In case of textual works detection process is easy and some of the system already exists.. But students mask plagiarism by changing the order, shuffling contents, changing the structure and other similar things. We are trying to find a solution for this type of problem and make it very low cost. Here we propose a system that finds the similarity between the given document images which can either be a text document or a handwritten text image, find the similarity score between them using the cosine similarity method, and determine whether the submitted documents are plagiarized or not or is it similar to the submitted text images. In this way, teachers can easily determine whether a particular text document with images has been plagiarized.. To identify the text from the image we use the OCR technique.**

***Keywords:-*** *Plagiarism System, Text mining, Data Mining.*

## I. INTRODUCTION

With the advancement of the internet all over the world it is easy to get someone's works or paper documents. So someone's steals some other's work and by doing minute changes they will submit it as their own. Plagiarism is the act of using someone's work without his permission. The author of the real paper even doesn't know about these practises.

Since we are in the computer era the usage of computers is very vast. It extends from school, institutes to industries. Even large scale industries are fully automated with computer. So the use of computer is inevitable in current scenario. Since todays teaching is e-learning the plagiarism is increased many times. Students find easy to copy from internet and friends documents. Data can be taken from different sources including the internet, papers, books over the internet, newspapers etc... These actions lead to a lack of learning in students. So a proper system is needed to detect plagiarism and the learning of students.

The scope of plagiarism detection projects is very high since most of the students submit assignments as scanned documents. It is very difficult to find plagiarism manually since there are so many documents. One steals someone else's documents and presents them as their own. This will affect their potential as well as they became. So it is inevitable for a proper plagiarism technique. The key mechanism behind the plagiarism detection is by using K Means algorithm and the cosine distance method. The results from these two are capable of solving the problems. In addition, the combination of K-means and cosine method requires a specific design of parts to achieve the goal of the system.. Plagiarism detection using automated mechanisms is very much needed considering the huge number of documents as assignments and as paper works.

ISSN No:-2456-2165

*A. Motivation*

The entire world has broken in covid 19. The world has moved into online space. Students complete their assignments and submit them into the online space for valuation. In this situation, students write their assignments, and at the same time, another part of students either copy from the internet or copy from the other's work. This practice is not suitable for students' future. We have found a better solution for that.

*B. Proposed System*

The effect of the current pandemic of covid-19 moved us further into the online space. Students complete their assignments independently, at the same time another part of students either copy from others or download from the internet. By using our system it is able to find similarities between different documents. In the case of students, students have uploaded their assignments in this portal system compared to other content. System detected above forty percent of similarity, that document is moved into another folder. The system rejects the document because of high plagiarism. Students can not upload that file as their assignment.

To find document similarity a combined process consisting of K Means algorithm and the Cosine distance method are used. The results from these two are capable of solving all the problems. Moreover, the combination between the K-means algorithm and cosine distance method requires a specific process schema to fulfill the objective of the research. The design of this system consists of various modules or parts that have to be integrated together to complete the system. As technology is improving day by day the tendency of plagiarism is also increasing day by day.

In the case of textual data, various mechanisms already exist.. But students mask plagiarism by changing the order, shuffling contents, changing the structure and other similar things. We are trying to find a solution for this type of problem and make it very low cost. Here we propose a system that finds the similarity between the given document images which can either be a text document or a handwritten text image, find the similarity score between them using the cosine similarity method, and determine whether the submitted documents are plagiarized or not or is it similar to the submitted text images. This will help the teachers to easily find whether the given textual image documents are plagiarised or not. To identify the text from the image we use the OCR technique.

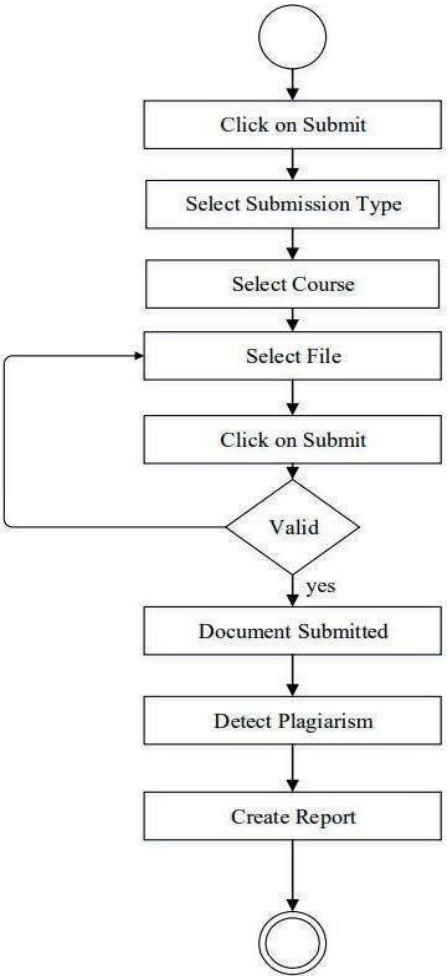


FIG.1 FLOW CHART OF THE SYSTEM

## II. METHODOLOGY

Finding plagiarism in a large amount of data is not an easy task, it is a complicated process. Students attach their assignments into a particular folder. The system divides that document into several small documents and calculates the similarity score of each document. Then calculate the average of similarity. The clustering of comparable vectors is taking place because of better seek efficiency.. To cluster similar vectors, we use the known and widely used K Means algorithm, which we modify for our purpose. In this section there can be facts that are to be saved in a database. It is in a form that allows them to be easily looked up. Clustering is done for a logical division of the data into related units.

The final section of the system deals with obtaining character suits from the database and their assessment. Earlier than the actual generation of the record, a filter out of non-extensive matches can be included which serves to clarify the reviews. As an example, commands for importing programs generated commands, and many others. After we have processed the source code, getting a list of vectors that represents the source code. There are many of these vectors and that they need to be processed.. The new designed system won't explore plagiarism within the whole knowledge set. however this may permit getting

ISSN No:-2456-2165

a listing of matches for one specific work. The two source code fragments are marked as identical given that their characteristic vectors square measure identical. The plagiarism search rule primarily consists of 3 components. The primary one is to get similarities from the information, the other is to match and filter these similarities, and therefore the third half is, the degree of similarity calculated bythe detected pairs of works.

The system tries to detect plagiarized content from handwritten copies as well as a word document. For using many packets in the programming language the data are converted to vector form. Then take the dot product of pairs of documents. Inorder to visualise the similarity plot a similarity. To identify the text from the image we use the OCR technique. This will help teachers to evaluate assignments.

## III. RESULT

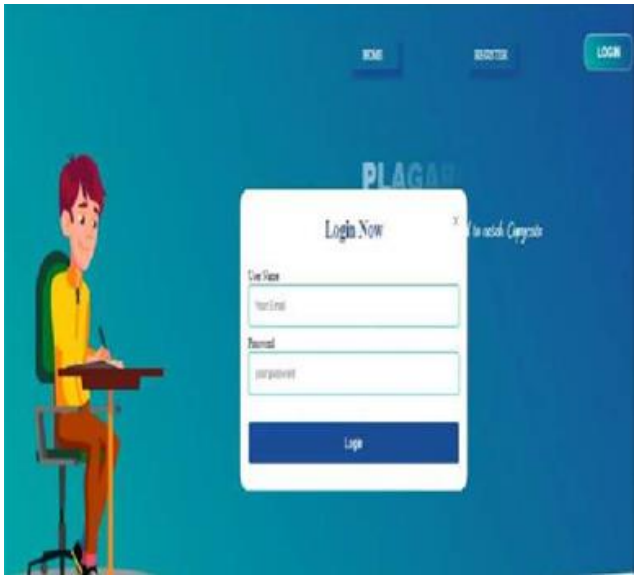


FIG.2 USER INTERFACE OF THIS MODEL


FIG.3 RESULT AFTER THE CHECKING

## IV. CONCLUSION

Manually finding plagiarism especially among a lot of documents like assignment submission are really difficult. If it is automated it can be really helpful. Plagiarism tools should be clear and effective. Plagiarism detection is a process that requires automation to effectively improve the plagiarism detection process. Inorder to decrease the overhead of the process preprocessing and clustering techniques can be used. Similarity scores are assigned to each document based on their plagiarism nature. It will display how much the document is similar to other documents.. Plagiarism checker is used to detect the similarities between 2 or more scanned documents. The practice of plagiarism leads to a lack of learning among students.. As teachers are concerned with plagiarism detection is having very much importance in students future

## FUTUREWORK

Our project aims at rectifying the difficulties of teachers who are searching for whether documents are unique or not. Traditional way of plagiarism detection is by checking each and every document manually. It is difficult and time consuming if the number and size of documents increases . Since our system is automatic it is easy and efficient. We will describe problems and detect the same or similar parts and describe existing similar systems and tools. Subsequently, we will present our designed and implemented system that we will verify and compare with two currently most used tools. Increase loading speed of image documents in future.

## REFERENCES

[1]. AntiCutAndPaste—Copied and Pasted Source Code Detector. Accessed: Jul. 16, 2020. http://www.plagiarism-report.com/anticutandpaste/
[2]. F. B. Allyson, M. L. Danilo, S. M. Jose, and B. C. Giovanni, ‘‘Sherlock Noverlap: Invasive normalization and overlap coefficient for the similarity analysis between source code,’’ IEEE Trans. Comput., vol. 68, no. 5, pp. 740–751, May 2019.
[3]. T. S. Adiningrum, “Reviewing plagiarism: an input for indonesian higher education,” Journal of Academic Ethics, pp. 107-120, 2015
[4]. D. Moeljadi, I. Kamajaya, and D. Amalia, “Building the kamus besar bahasa indonesia (kbbi) database and its applications,” in Proc. of The 11th International Conference of the Asian Association for Lexicography, pp. 64-80, 2017.
[5]. U. Rani, and S. Sahu, “Comparison of clustering techniques for measuring similarity in articles,” in Proc. of The 3rd International Technology (CICT), pp. 1-7, IEEE, 2017.
[6]. F. Rozi, and F. Sukmana, “Document grouping by using meronyms and type-2 fuzzy association rule mining,” Journal of ICT Research and Applications, 11(3), pp. 268-283, 2017.
[7]. https://en.wikipedia.org/wiki/K-means_clustering
[8]. https://www.w3schools.com/css/default.asp
[9]. https://developer.mozilla.org/enUS/docs/Learn/Getting_started_with_the_web/HTML_basics