Text and Image Plagiarism Detection Using Deep Learning

[1.]Mrs.G. Swathi, Assistant professor, Computer Science and Engineering (Artificial Intelligence), Gates Institute of technology, Gooty, Andhra Pradesh, Email:swathi.ganta20gmail.com
[2.]D. Supraja, Computer Science and Engineering (Artificial Intelligence), Gates Institute of Technology, Gooty, Andhra Pradesh Email:dasarisuprajaroyal@gmail.com
[3.]J.Yashwanth, Computer Science and Engineering (Artificial Intelligence), Gates Institute of technology, Gooty, Andhra Pradesh Email:yashwanthyushu2003@gmail.com
[4.]N. Zuberia, Computer Science and Engineering (Artificial Intelligence), Gates Institute of technology, Gooty, Andhra Pradesh Email: nzuberia636@gmail.com
[5.]B. Sreehari, Computer Science and Engineering (Artificial Intelligence), Gates Institute of technology, Gooty, Andhra Pradesh, Email: Sreehari1official@gmail.com
[6.]O. Uday Kiran, Computer Science and Engineering (Artificial Intelligence), Gates Institute of technology, Gooty, Andhra Pradesh , Email: Kiranreddy18451@gmail.com

Abstract- Plagiarism in research is being debated more than ever before. There have been considerable harms to research as a consequence of web conditions and the ability to do complicated and intelligent searches in a short period of time. Text-focused plagiarism detection tools disregard visuals. Images, on the other hand, are a vital component of the process of transmitting the massive amounts of data included inside a research paper or other piece of scholarly writing. It's possible that plagiarism might occur because of the vast variety of pictures and the huge number of images present in computer-generated texts, and since flowcharts hold a lot of information. Using the Histogram Model, we hope to determine how many photographs in a paper have been plagiarised.

Keywords— Plagiarism, Detection, Research paper, Histogram model.

I. INTRODUCTION

The problem of plagiarism is often debated in the academic community. It refers to the practise of passing off someone else's work or ideas as your own without attribution. In essence, it's a repackaging of already existent data. By "is the act of copying or exploiting someone else's invention or idea without permission and presenting it as one's own," S. Hannabuss defines plagiarism [5]. So many materials are now publicly available because to the enormous popularity of the internet. The internet has grown to be a vast repository for information. There is no need for people to write their own text documents since they can quickly get the information they need from the internet. Plagiarism detection is becoming more relevant in light of the ease with which a plagiarist might locate an acceptable text fragment to copy. On the other hand, as the number of alternative sources grows, it becomes more difficult to accurately detect plagiarised sections[7]. Plagiarism is a common occurrence in a variety of fields, including academia, media, science, and even politics. In cases when there is no reference collection

accessible or not all the probable copy sources are provided, this technique to plagiarism detection is particularly beneficial since documentto-document comparison algorithms cannot be applied. Text manipulation and other forms of plagiarism are also forms of plagiarism [3]. Similarly, a variety of methods for detecting plagiarism are available. System implementations relying on the text manipulation approach are currently insufficient for practical use. Therefore, we have developed a novel and simple method that employs a machine learning methodology to identify plagiarism across text sets. According to our threshold value for plagiarism detection, we generate a percentage value based on the number of words that are similar between the two files, and then we can identify the plagiarised text series.

II. RELATEDWORKS

Text-based, citation-based and shape-based plagiarism detection systems have been compared to each other in various cases. Compared to citation-based plagiarism detection approaches, text-based plagiarism detection methods have proven over 70 percent effective. Text-based approaches for detecting plagiarism in translated materials have been effectively implemented. Fewer than 5%, while in citation-based technique, this figure is approximately 80%. The comparison of photos has not yet been conducted in the existing system. Table 1 shows literature review of existing works. Disadvantages are there is a far lower level of accuracy in identifying information sources for plagiarism using photos than there is with text-based approaches.

230

Table 1 Literature survey

Reference and year	Approach and Method	Performance
Imam Much IbnuSubroto and Ali Selamat, 2014	Plagiarism Detection through Internet using Hybrid Artificial Neural Network and Support Vectors Machine	most of the plagiarism detections are using similarity measurement techniques. Basically, a pair of similar sentences describes the same idea
UpulBandara and GaminiWijayrathna , 2012	Detection of Source Code Plagiarism Using Machine Learning Approach	Source code plagiarism is currently a severe problem in academia. In academia' s programming assignments are used to evaluate students in programming courses.
SalhaAlzahrani, Naomie Salim, Ajith Abraham, and Vasile Palade, 2011	iPlag: Intelligent Plagiarism Reasoner in Scientific Publications	Texts that are acceptable to be redundant and texts that are cited properly are all highlighted as plagiarism, and the real decision of plagiarism is left up to the user.

A Selamat, IMI Subroto and Choon-Ching Ng, 2009	Arabic Script Web Page Language Identification Using HybridKNN Method	One of the crucial tasks in the text-based language identification that utilizes the same script is how to produce reliable features and how to deal with the huge number of languages in the world
Ahmad Gull Liaqat and Aijaz Ahmad, 2011	Advanced Supervised Learning in Multi-layer Perceptrons - From Backpropagation to Adaptive Learning Algorithms,	Since the presentation of the backpropagation algorithm [1] a vast variety of improvements of the technique for training the weights in a feed-forward neural network have been proposed.

III. PROPOSED SYSTEM ARCHITECTURE

Training and testing are the two main components of the system as it is currently envisioned. They are seen as using the Histogram in the learning phase and the modelling done by this network in the testing phase for the recognition stage in the train phase. Based on correlation rates between query photos and images in database, the data analysis approach selects the images with the most comparable correlations to the query image. Correlation levels at this step are used to report on the tested picture plagiarism, and the expert is responsible for the ultimate interpretation of the results. The architecture of proposed system is shown in Fig. 2.

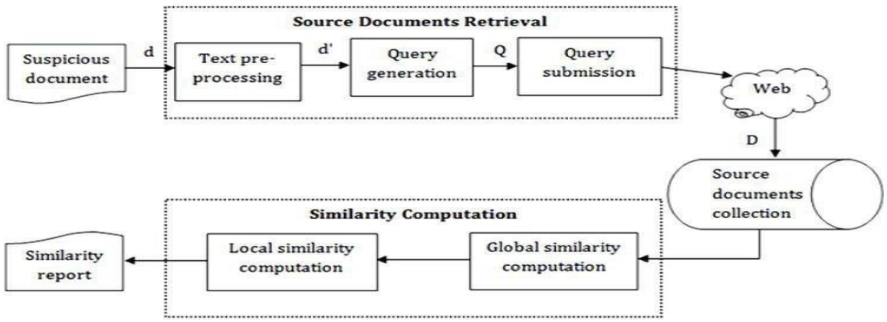

Fig. 1 Proposed System architecture

IV. RESULTS AND DISCUSSION

The results obtained after executing the implementation code is shown from Fig.2 to Fig.20.

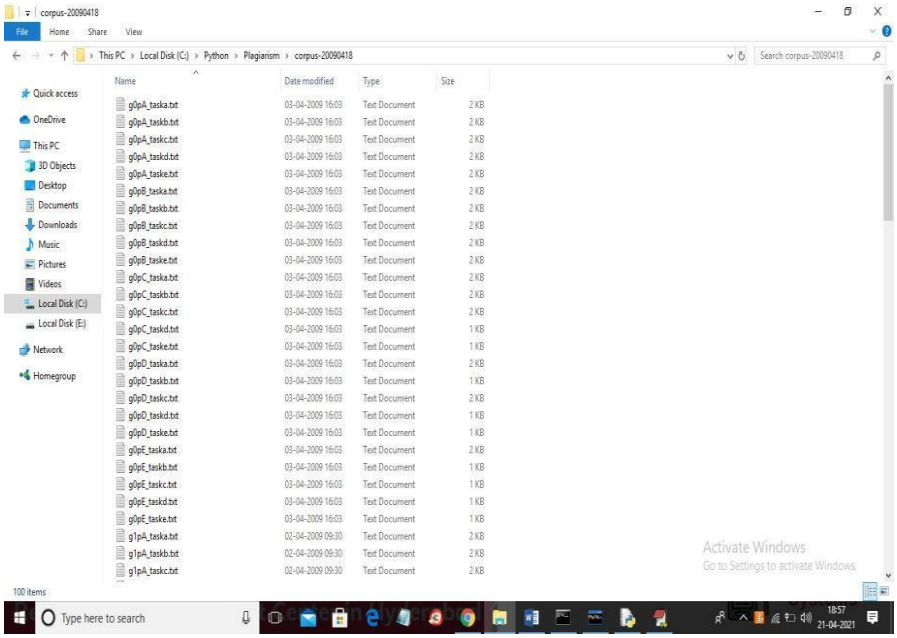

Fig.2 Text files used to build histogram

We are using below images to build histogram model and if any suspicious image similarity finds with this histogram then plagiarism will be detected. See below images used to build histogram model

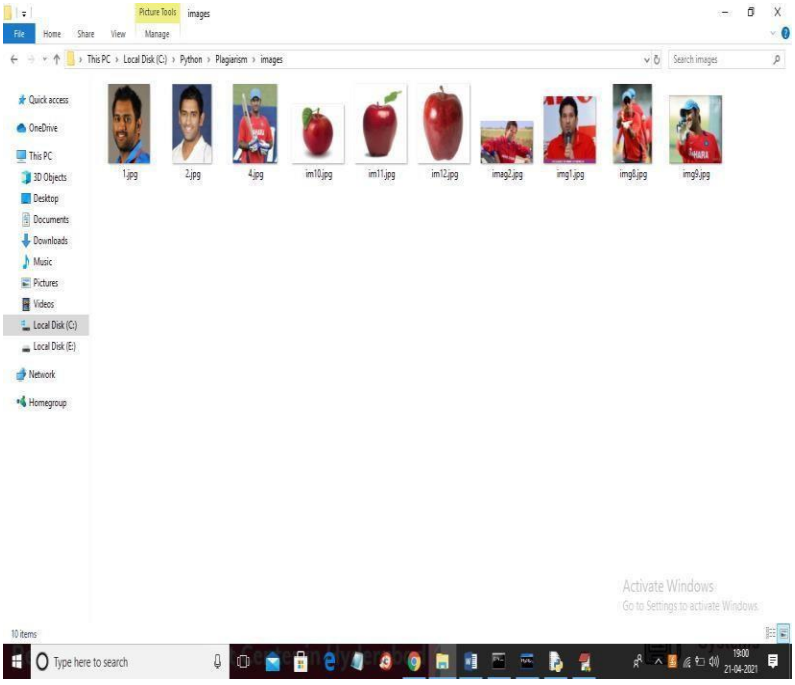

Fig.3 Images used to build histogram

Above images are available inside " images" folder

To run project install python 3.7 and then install DJANGO server and deploy code on that server and run from browser to get below screen

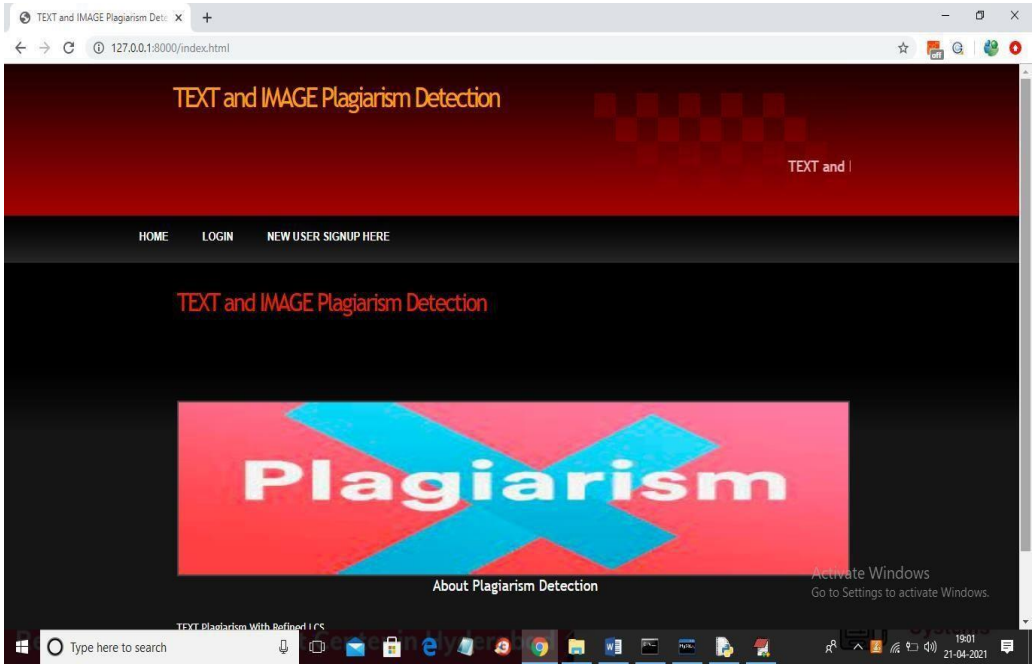

Fig.4 Home Page

In above screen click on ‘ New User Signup Here’ link to get below screen

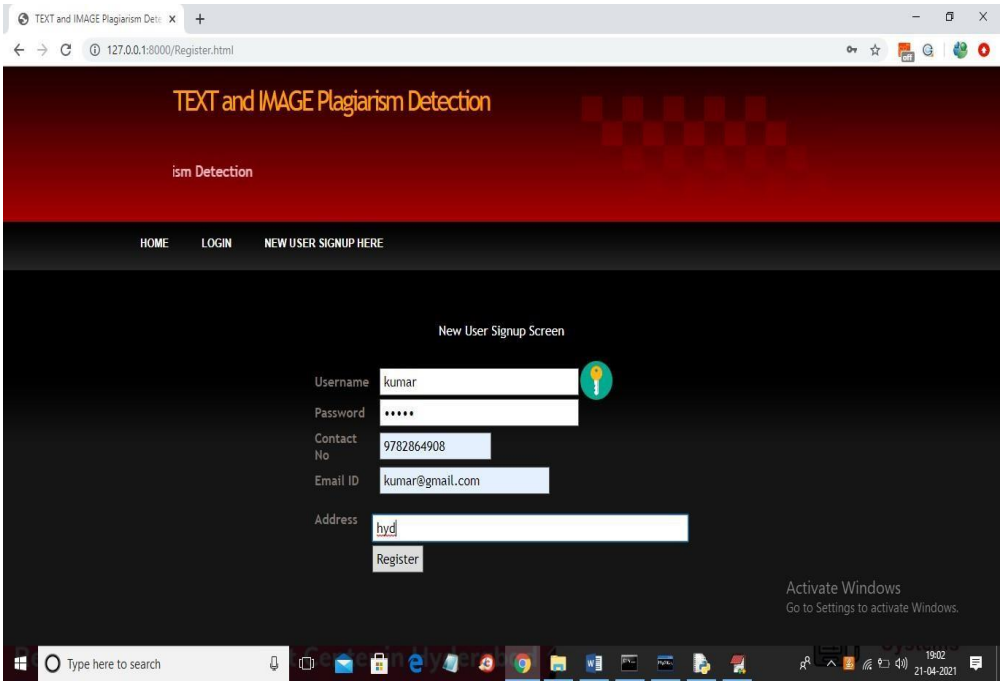

Fig.5 New User Sign Up

In above screen user signup details entered and then click on ‘ Register’ button to get below screen

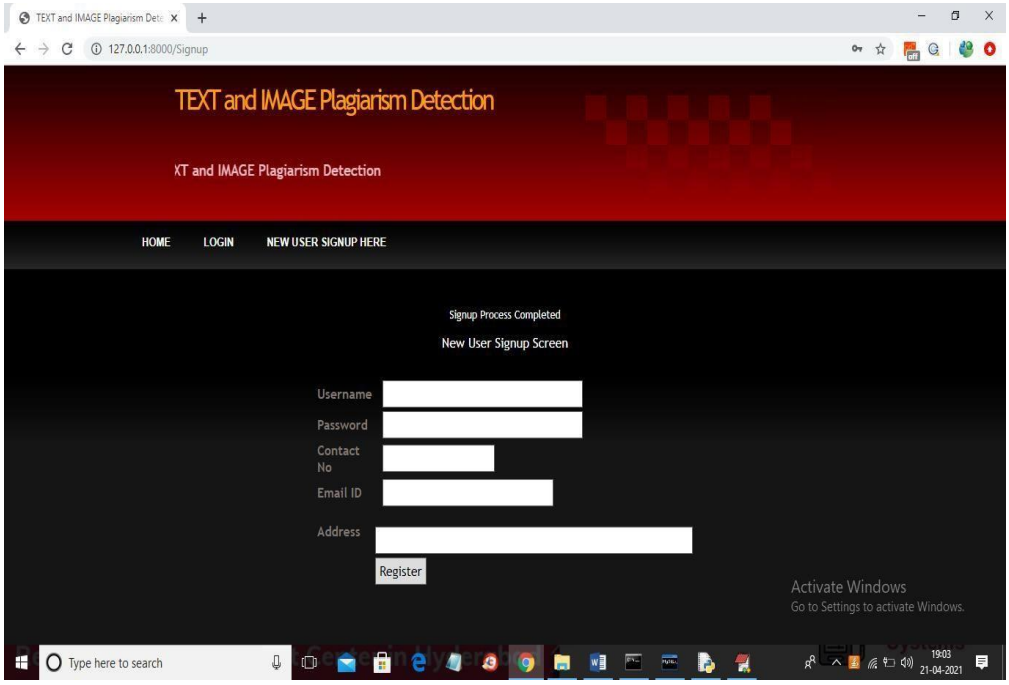

Fig. 6 signup process completed

In above screen user signup process completed and now click on ‘ Login’ link to get below screen

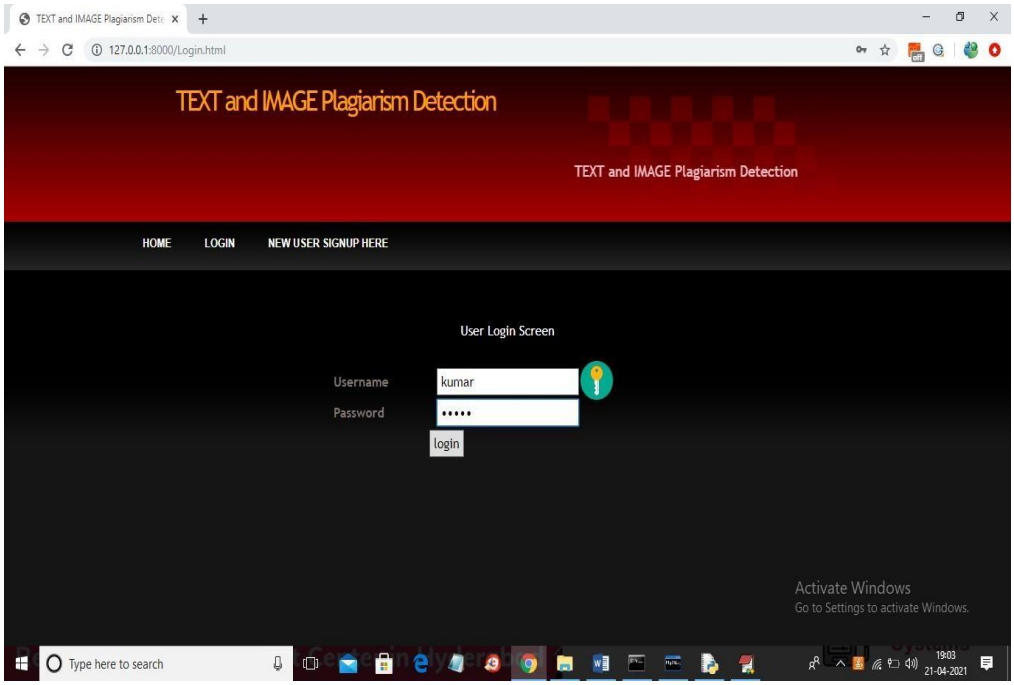

Fig. 7 User Login

In above screen user is login and then click on button to get below screen

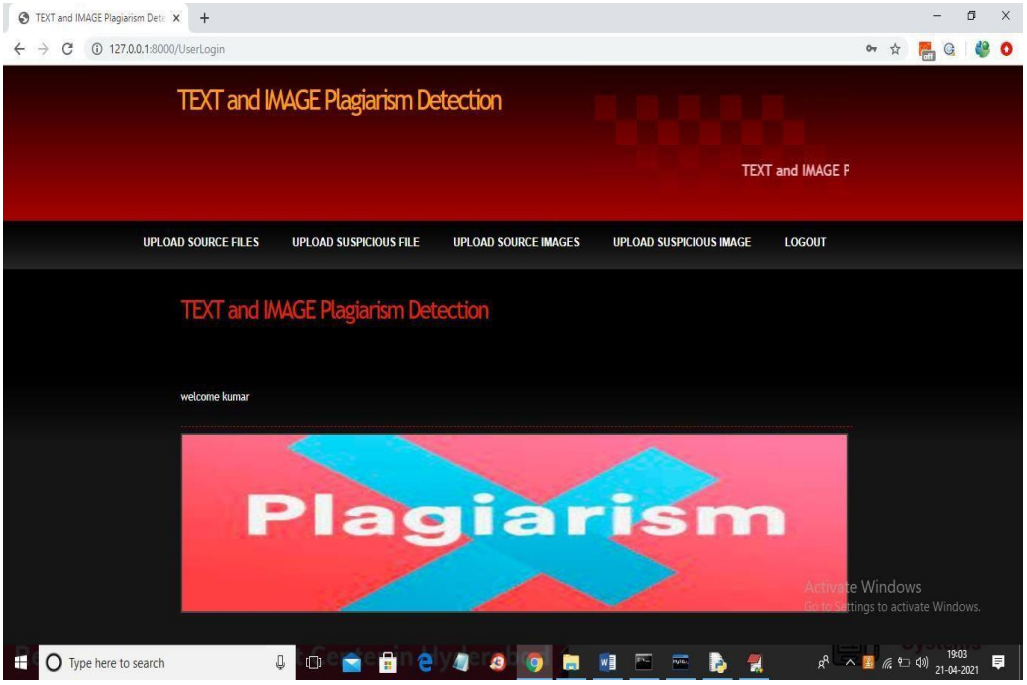

Fig. 8 Upload Source Files'

In above screen click on ' Upload Source Files' link to load all files from corpus folder

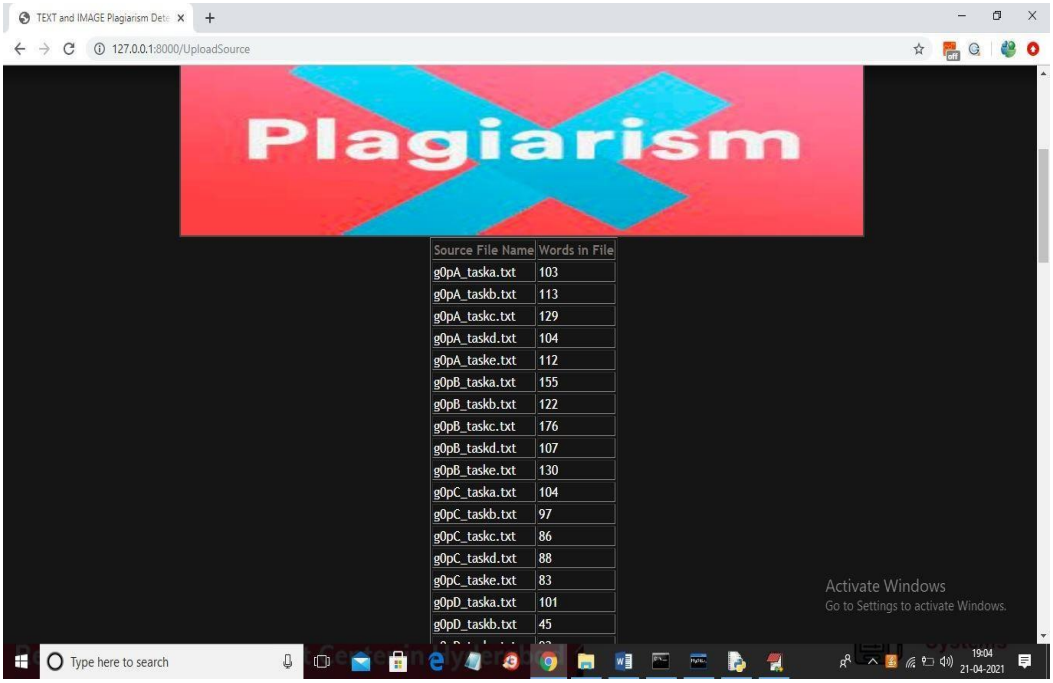

Fig. 9 Upload Suspicious File'

In above screen all files are loaded now click on ' Upload Suspicious File' button to load suspicious file and get result

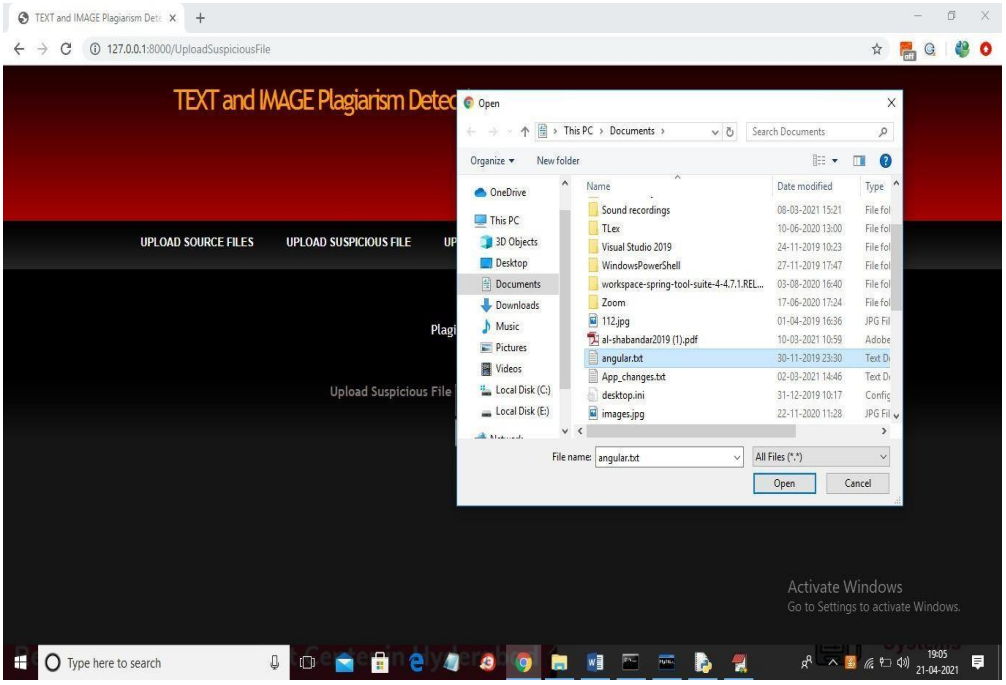

Fig. 10 selecting and uploading ‘ angular.txt’ file

In above screen I am selecting and uploading ‘ angular.txt’ file and then click on ‘ Open’ button to get below result and then click on ‘ Check Plagiarism’ button to get result

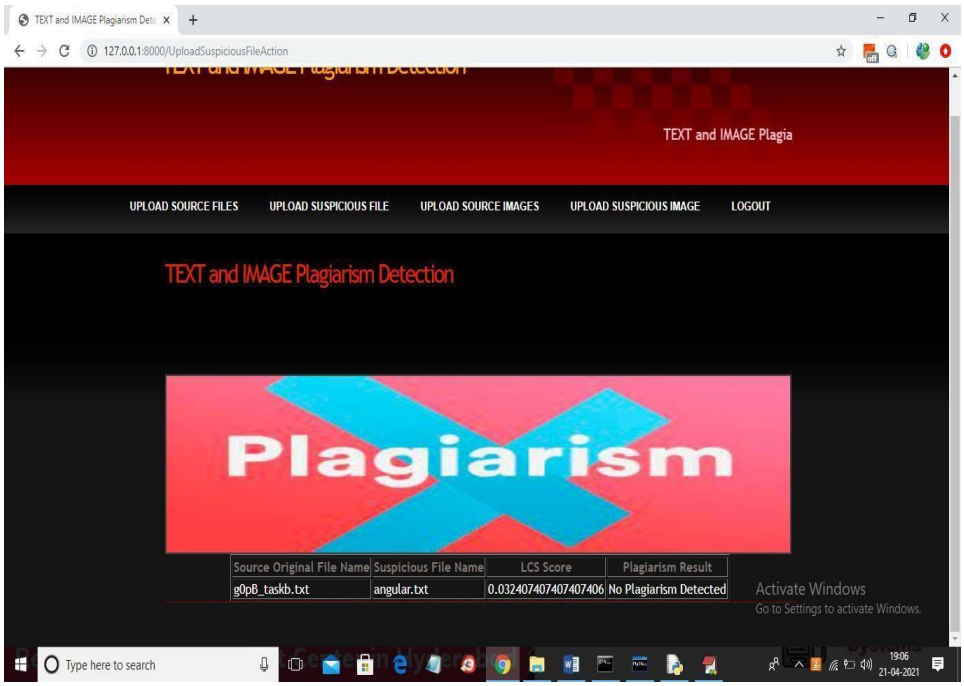

Fig. 11 angular.txt file matched

In above screen angular.txt file matched very little with g)pB_taskb.txt corpus file and we got similarity score as
0.03 so no plagiarism detected and now upload any file from corpus and see result

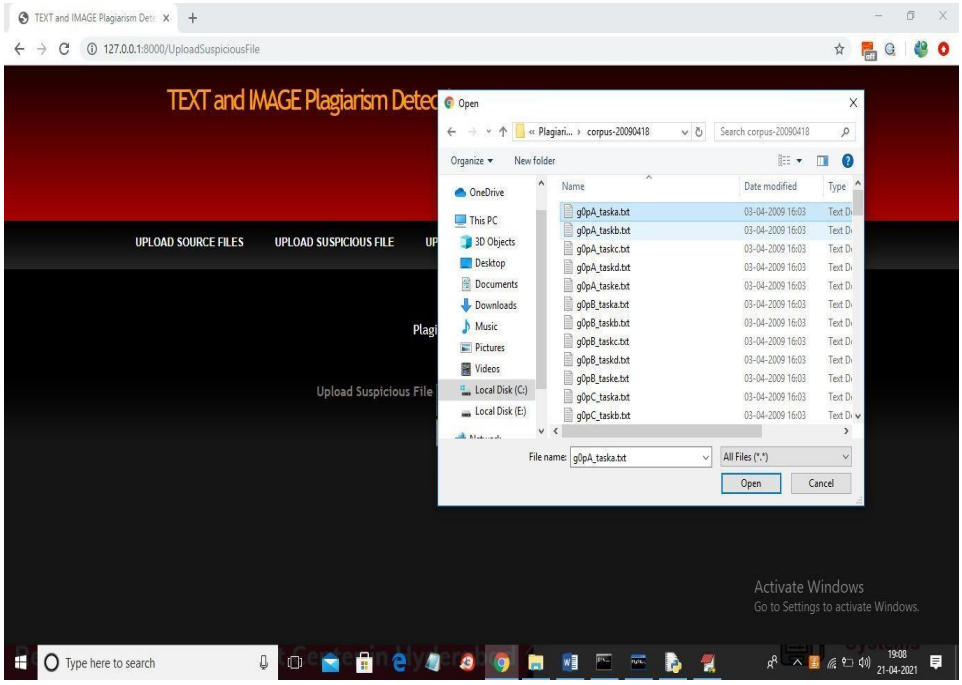

Fig. 12 selecting and uploading first file

In above screen I am selecting and uploading first file and then click on button to get below result

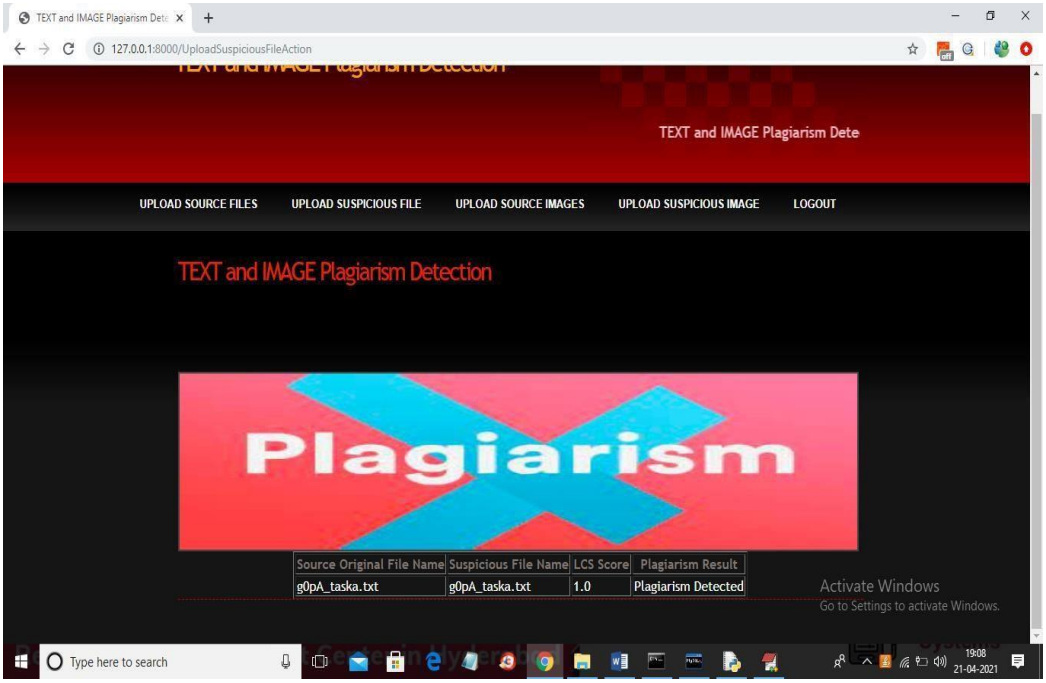

Fig. 13 LCS score

In above screen LCS score is 1.0 which means 100% matched with corpus file so plagiarism detected and similarly not only this u may enter any text file and get result. Now click on ‘ Upload Source Images’ link to upload all images from ‘ images’ folder

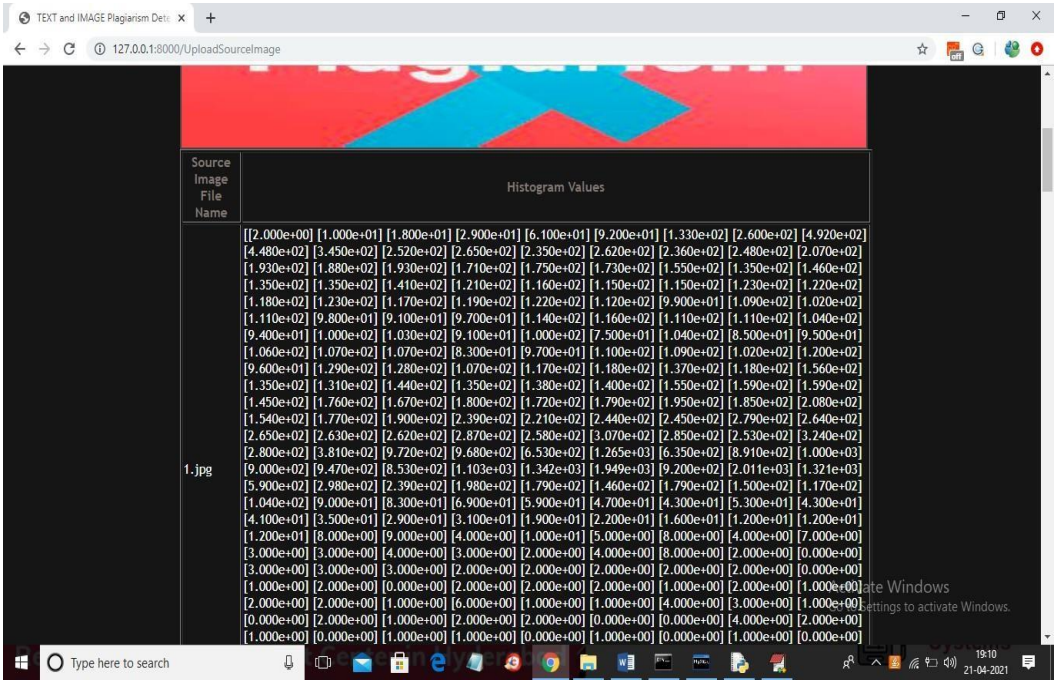

Fig. 14 Upload Suspicious Image

In above screen from all database images histogram will be calculated and store in array and whenever we upload new test image then both histogram will get matched and now click on ' Upload Suspicious Image' link to upload some image

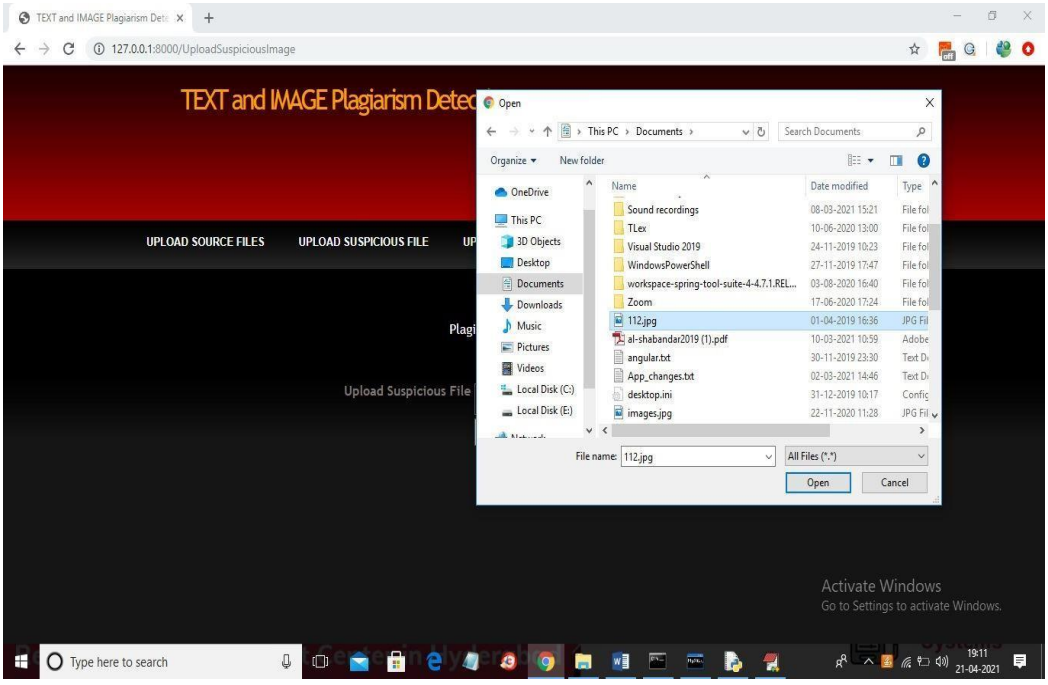

Fig. 15 selecting and uploading ' 112.jpg' file

In above screen I am selecting and uploading ' 112.jpg' file and then click on ' Open' button to get below result

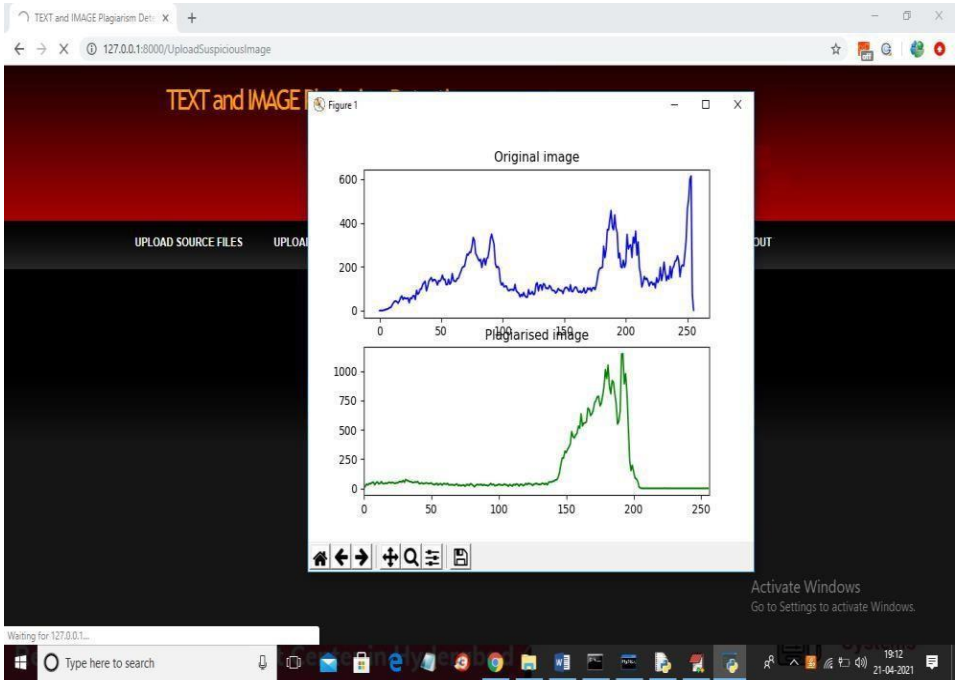

Fig. 16 Generating histogram

In above screen we can see for database image and uploaded image we generated histogram and we can see there is no match in histogram so no plagiarism will be detected and now close above graph to get below result

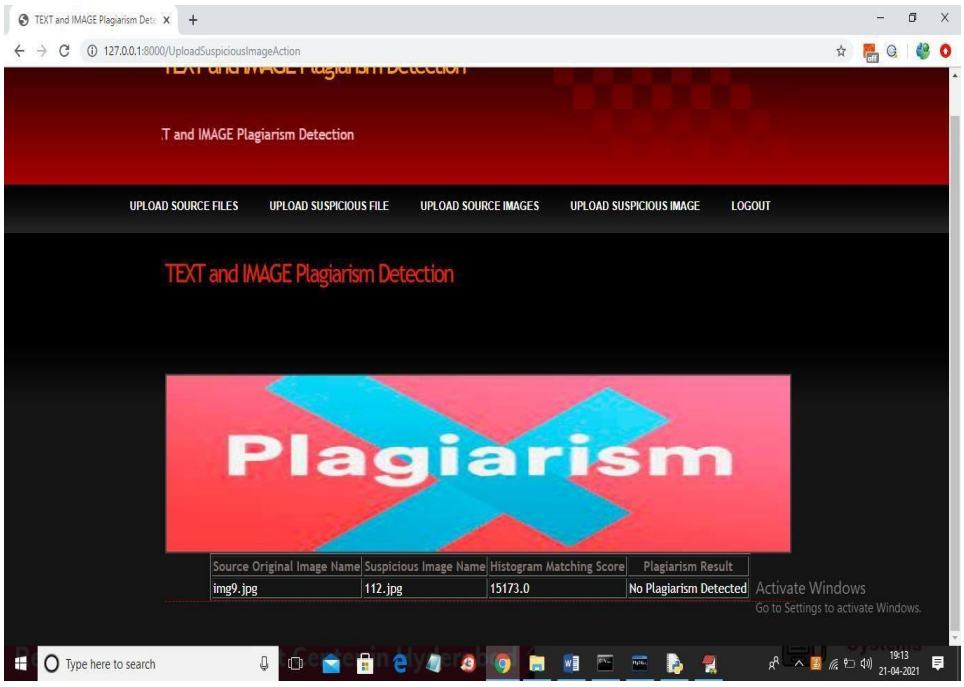

Fig. 17 histogram pixel matching score

In above screen histogram pixel matching score is 15173 out of 40000 pixels so image is not plagiarised and now upload image from " images" folder and see result

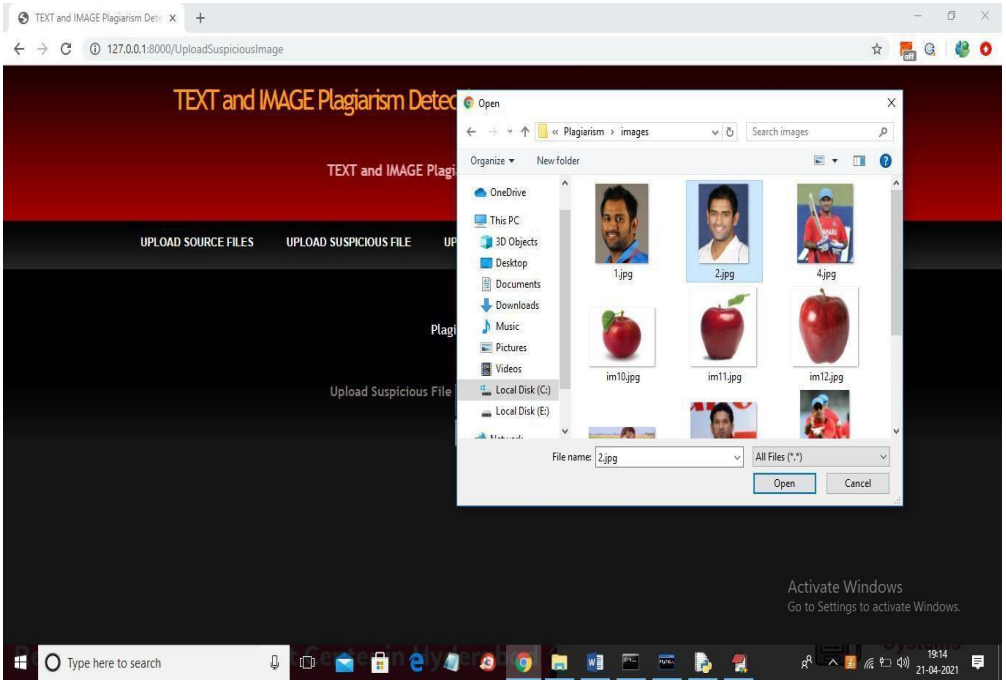

Fig. 18 selecting and uploading ‘ 2.jpg’ file

In above screen I am selecting and uploading ‘ 2.jpg’ file from “ images” database folder and below is the result

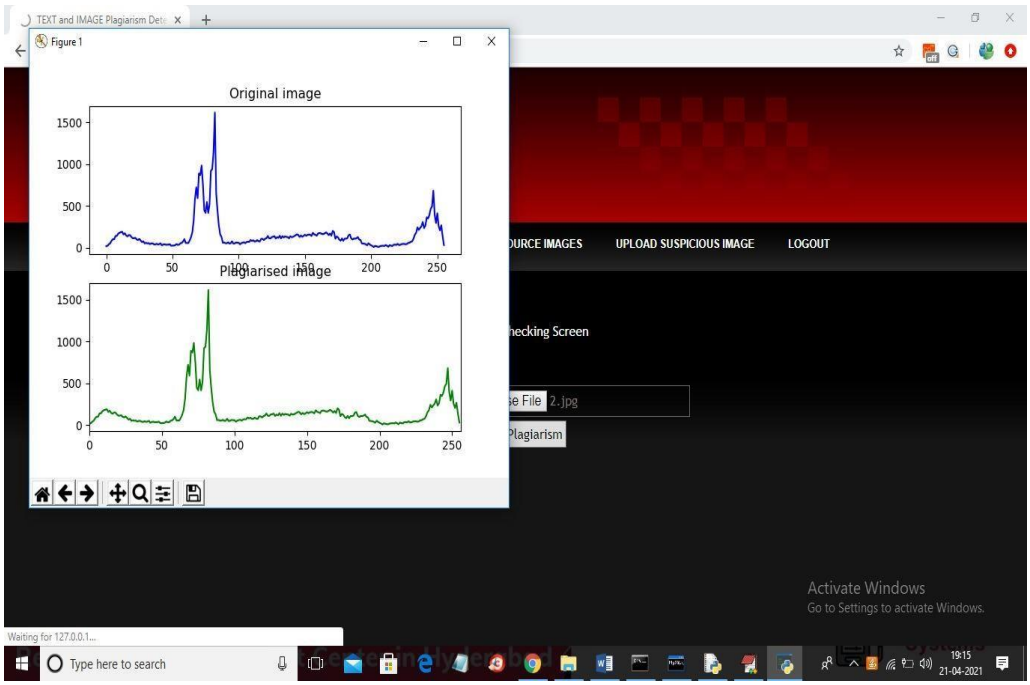

Fig. 19 original and uploaded image histogram

In above screen we can both original and uploaded image histogram is matching 100% so plagiarism is detected and now close above graph to get below result

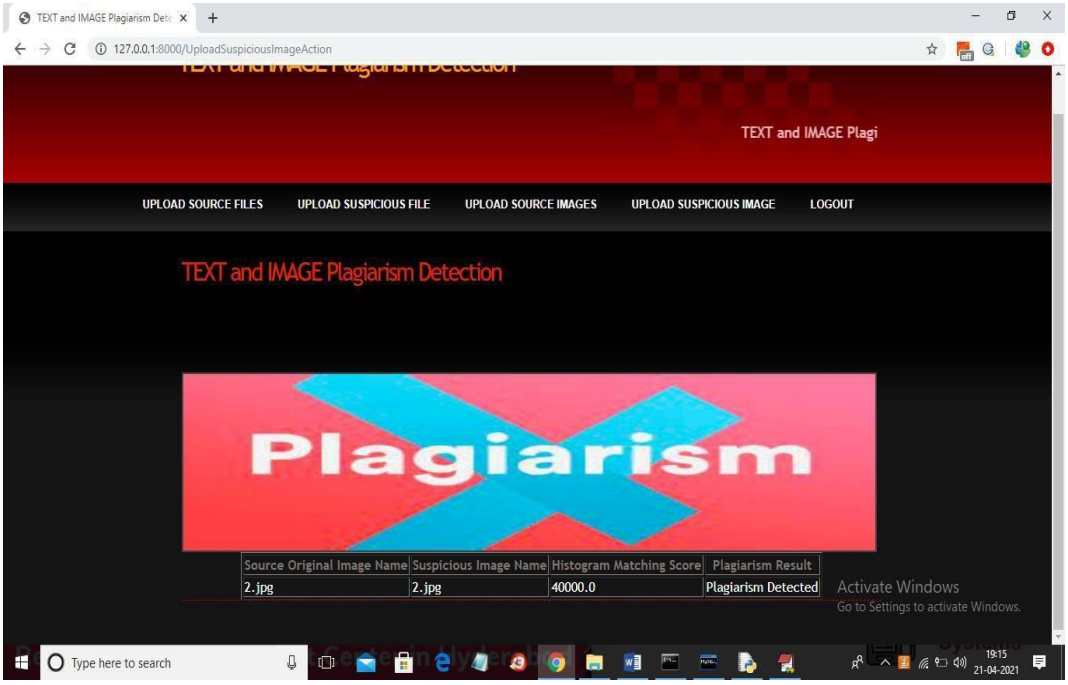

Fig. 20 histogram matching score

In above screen histogram matching score is 40000 which means all pixels matched so plagiarism is detected in above result. Similarly u can upload any text file and image and test the application

V. FUTURE SCOPE AND CONCLUSION

The issue of plagiarism in academic research is receiving more attention than ever. Web conditions and the capacity to do complex and sophisticated searches in a short amount of time have had a significant impact on research. Visuals are ignored by text-focused plagiarism detection programmes. When it comes to conveying the huge quantities of information included in a research paper or other academic writing, images are an important part of the process. It's probable that computer-generated texts include plagiarism due to the large quantity and diversity of images available, as well as the fact that flowcharts contain a great deal of information. Our goal is to detect how many images in a paper have been plagiarised using the Histogram Model.

REFERENCES

[1] Imam Much IbnuSubroto and Ali Selamat, "Plagiarism Detection through Internet using Hybrid Artificial Neural Network and Support Vectors Machine," TELKOMNIKA, Vol.12, No.1, March 2014, pp. 209-218.

[2] UpulBandara and GaminiWijayrathna ,"Detection of Source Code Plagiarism Using Machine Learning Approach," International Journal of Computer Theory and Engineering, Vol. 4, No. 5, October 2012, pp.674678.

[3] SalhaAlzahrani, Naomie Salim, Ajith Abraham, and Vasile Palade," iPlag: Intelligent Plagiarism Reasoner in Scientific Publications," IEEE World Congress on Information and Communication Technologies, 2011.

[4] BarrónCedeño, A., & Rosso, "On automatic plagiarism detection based on n-grams comparison," In Advances in Information Retrieval, Vol. 5478. Lecture Notes in Computer Science, pp. 696– 700, Springer.

[5] Ahmad Gull Liaqat and Aijaz Ahmad, "Plagiarism Detection in Java Code," Degree Project, Linnaeus University, June 2011, pp. 1-7.

[6] A Selamat, IMI Subroto and Choon-Ching Ng, "Arabic Script Web Page Language Identification Using HybridKNN Method," International Journal of Computational Intelligence and Applications, 2009, pp. 315343.

[7] Michael Tschuggnall and Gunther Specht , "Detecting Plagiarism in Text Documents through GrammarAnalysis of Authors," pp. 241-255.

[8] Bill B. Wang, R I. (Bob) McKay, Hussein A. Abbass and Michael Barlow, "Learning Text Classifier using the Domain Concept Hierarchy," ACT 2600, pp. 1-5.

[9] Francisco R., Antonio G., Santiago R., Jose L., Pedraza M., and Manuel N., ―Detection of Plagiarism in Programming Assignments,‖ IEEE Transactions on Education, vol. 51, No.2, pp.174-183, 2008.