Dr. Kandru Arun Kumar /International Journal of Engineering & Science Research

TEXT AND IMAGE PLAGIARISM DETECTION

Dr. Kandru Arun Kumar,
Professor, Dept. of CSE,
Malla Reddy Engineering College (Autonomous), Secunderabad,
Telangana State

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: Text, plagiarism, image.

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

Dr. Kandru Arun Kumar / **International Journal of Engineering & Science Research**

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.

PROBLEM DEFINITION

The corpus and the measures form the first controlled evaluation environment dedicated to plagiarism detection. Unlike other tasks in natural language processing and information retrieval, it is not possible to publish a collection of real plagiarism cases for evaluation purposes since they cannot be properly anonymized. Therefore, current evaluations found in the literature are incomparable and often not even reproducible. Our contribution in this respect is a newly developed large-scale corpus of artificial plagiarism and new detection performance measures tailored to the evaluation of plagiarism detection algorithms

OBJECTIVE OF PROJECT

We aimed to create a corpus that could be used for the development and evaluation of plagiarism detection systems that reflects the types of plagiarism practiced by students in an academic setting as far as realistically possible.

EXISTING SYSTEM

The existing methodology maybe sufficient for detecting plagiarism of images when the source and suspected image have not been rotated by a large margin, but in case of rotational changes the existing methodology will fail. The proposed methodology will ensure that even if the image is rotated plagiarism is detected if it has occurred or if an attack of rotational change has been made. Also the existing system is not efficient to detect plagiarism properly for different types of images. The proposed system will ensure that by using adaptive threshold values. The algorithm makes sure that the matching time of the images is less by reducing the search field by a significant factor each time the refinement is done.

PROPOSED SYSTEM

Dr. Kandru Arun Kumar / **International Journal of Engineering & Science Research**

The Proposed Text and Image of images plagiarism detection will take input from the used which will be suspected plagiarized image according to the user. Than the Phash value of that image would be generated using the corpus algorithm. Now the input image would be checked for plagiarism against the images in local database. In Database, image are stored with their respective Phash values. The plagiarism detection engine will follow a series of steps to find out plagiarism]. This would include calculating hamming distance between Phash values of input image and images in database. At the end based on results achieved in detection engine, results will be displayed. In the Same way text file also detected using corpus algorithm.

MODULES

1.New user Signup

Firstly user will register in to Application.It helpful to login into Application with username and password.

2.Login

User will login into Application through username and password.

3.Upload Source File

Folder is created into Upload Source Files' link to load all files from corpus folder.

4.Upload Suspicious files

To load suspicious file and get result.user will upload file to Upload Suspicious files the result is execute. 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.

5. Upload Source Image

In this module 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.

6. Upload Suspicious Image

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. 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. we can both original and uploaded image histogram is matching 100% so plagiarism is detected and now get below result. histogram matching score is 40000 which means all pixels matched so plagiarism is detected in above result
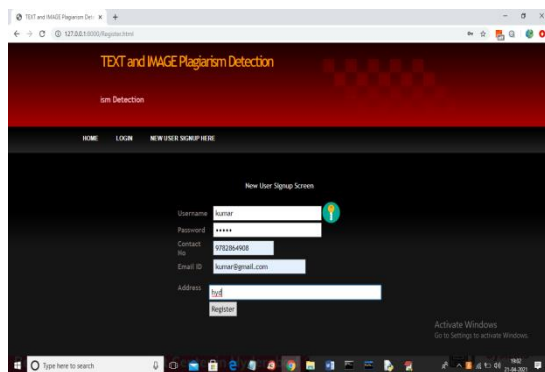
OPERATION

Dr. Kandru Arun Kumar /International Journal of Engineering & Science Research
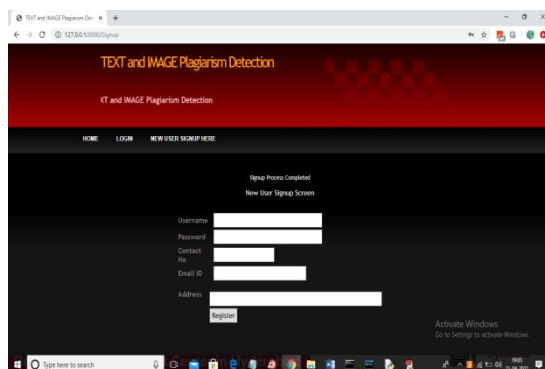
To run project install python 3.7 and then install DJANGO server and deploy

code on that server and run from browser to get below screen
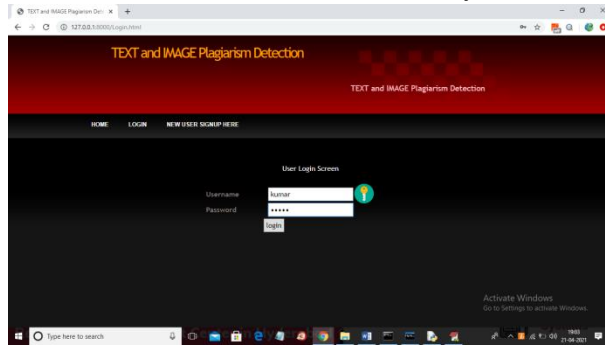

In above screen click on 'New User Signup Here' link to get below screen

In above screen user signup details entered and then click on 'Register' button to get

below screen

In above screen user signup process completed and now click on 'Login' link to get

below screen

4

Dr. Kandru Arun Kumar / **International Journal of Engineering & Science Research**
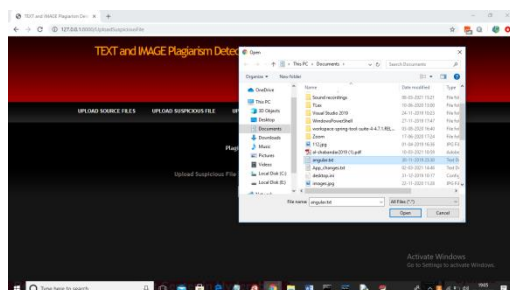

In above screen user is login and then click on button to get below screen

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


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


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

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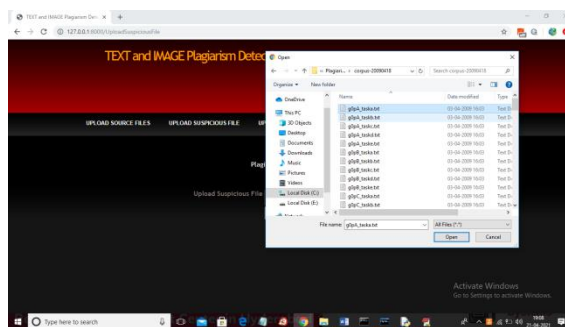
In above screen I am selecting and uploading first file and then click on button to get below result
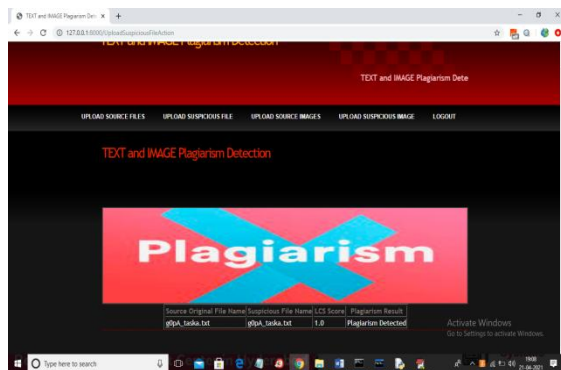
In above screen LCS score is 1.0 which means 100% matched with corpus file so plagiarism detected and similarly not only this u may enter any text file and get result. Now click on 'Upload Source Images' link to upload all images from 'images' folder.

CONCLUSION

corpus is the first standardized corpus dedicated to the evaluation of automatic plagiarism detection and was successfully employed in the First International Competition on Plagiarism Detection. We believe that our corpus and the performance measures will become an effective means to evaluate future plagiarism detection research. Currently, an improved version of the corpus is being constructed.

6

REFERANCES

[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.674- 678.

[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. 315- 343.

[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.