ARTIFICIAL INTELLIGENCE MODEL BASED TEXT AND IMAGE PLAGIARISM DETECTION BASED USING K-MEAN CLUSTERING ALGORITHM

Dr. Santosh Kumar Byraboina Associate professor, CSI Wesley Institute of Technology and Sciences, Hyderabad

Dr. P. M. Yohan Professor and Principal, CSI Wesley Institute of Technology and Sciences, Hyderabad

Abstract: In an educational environment, plagiarism is a crucial task that needs to be identified, in recent years all known journals and conferences, as well as universities, request a plagiarism report from students and researchers to prove the originality of published text or scientific paper. Plagiarism detection usually checks the text content via many of the platforms which are available for productive use reliably identifying copied text or near-copies of text and these systems usually fail to detect the images, and Files plagiarism since it is originally built for text mainly. In this paper, we suggest an adaptive, scalable, extensible, robust method for image plagiarism which is tested in designs collecollected the department of architecture University of Technology, this method mainly compare the data (designs images) entered to the system TEXT and IMAGE Plagiarism Detection. with data sets saved in the database mainly these designs are saved as feature which is one of the artificial intelligence algorithms and match by using k-mean clustering and the similarity check is done with threshold used 40% which can be changed to an accepted levels when needed. Using the k-mean algorithm in clustering, which is a robust artificial intelligence clustering algorithm giving us a strong system that is not discarding any feature extracted from the image. In this paper, data sets consist of 45 samples as training images saved and used in the system as the system database and using 48 samples as testing images which consist of original and forgery designs. These testing images were evaluated with 100% matching rate and 81% matching accuracy rating. We are using below text corpus to build plagiaism detection model and if any suspicious file data falls in similarity of this corpus then plagiarism will be detected. This corpus you can see inside „corpus-20090418‟ folder. 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.

Index Term: - IMAGE Plagiarism,k-means, published text, scientific paper

I INTRODUCTION

There are two main types of plagiarism as Text Based Plagiarism and Image Based Plagiarism. Text Based Plagiarism includes „copying textual information available from internet or other resources without proper permission and presenting it as their own" Image Based plagiarism includes "copying an image or portions of an image from the Internet or from classroom resources without permission or proper acknowledgment." Hashing techniques are used in the process of plagiarism detection. There are different algorithms for plagiarism.here we are using corpus for image and Text. 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 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.

2 LITERATURE SURVEY

This paper gives a brief idea about classification, the classifications done based on language in the documents. Languages are classified as Mono-lingual and cross-lingual or multi lingual[1]. Mono-lingual plagiarism detection identifies and extracts texts from the document and detects language of same kind i.e English-English plagiarism. Cross-lingual or multi-lingual plagiarism detection also deals with identification and extraction of text from document and detects language of different kind's i.e English-Arabic plagiarism. Shape-Based Plagiarism Detection for Flowchart Figures in Texts does pre-processing by determining the boundaries, edges, distance and the figures are stored in database by eliminating the text from the figures[2]. The system takes the sample figure and pre- processes it to build the query vector that will be compared with the figure-document stored in the database, this will be the training phase[3]. Then the test figures are given as input to the system and compares with the figures stored in the database. The result is the number of figures copied from Revised Manuscript Received on June 7, 2019 Akshay S, Department of Computer Science, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Mysuru, India. Chaitanya B N, Department of Computer Science, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Mysuru, India[4]. Rishabh Kumar, Department of Computer Science, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Mysuru, India. the original paper. In this paper, we compare Set A, B as two RGB images with same size, comparing A and B is to detect the same color of pixels with same location, the steps and algorithms[5]. C is an image matrix from color matrix A subtracting color matrix B, then C=A-B, if the corresponding pixels of A and B have same color, then the RGB significance of corresponding pixels in image C should be 0, which means black, so the copied pixels between image A and B should be black. H is a set with black pixels extracted from image C, so all Copied pixels between image A and B should be contained in H. As the images A, B may have the same background color, when comparing A and B, the part with same background color will be extracted to set H, therefore the therefore part must be eliminated recurring to the character of background color that it‟s usually monochrome[6]. The study of Histogram describes about the applications of how histograms can be used to know the properties of the image, enhancement of the image, to detect exposure saturation, brightness, gaps etc, and it also helps in thresholding. This paper also deals with Histogram stretching which determines the contrast of the images. Histogram sliding shows the intensity and brightness of the images[7]. Histogram Equalization equalizes all the pixels of the image to one form which gives us the flat graph. Flowchart Plagiarism Detection method uses area detection technique to detect plagiarism, the flowchart images are given as input to the system which are pre- processed by detecting the edge using"cannyedge detection". For each shape in the image, the centroid and boundary is detected. Euclidean distance is calculated from centroid to boundary and a graph is generated[8]. Then the generated graph is compared with the original image graph. The result is an alert displaying whether the image is plagiarized or not. This drawback of this approach is that it only works on flowchart images[9]. The paper "Edge Detection Methods" describes about the edge detection which an important feature extraction method, this method can be used to determine the lines in the images. The Author classifies different edge detection techniques like Sabel, Prewitt"s, Robert"s Cross, Laplacian of Guassian and Canny. Sample images are converted to grayscale on which the experiments are performed on all the techniques. The comparison between these techniques is explained and concluded that Canny is best compared to all the techniques. Content-Based Image Retrieval (CBIR) is a kind of feature extraction method which uses may contents of an image like shape, color, texture for representation and indexing of image[10].

3.IMPLEMENTATION STUDY

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 methodology

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. 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.

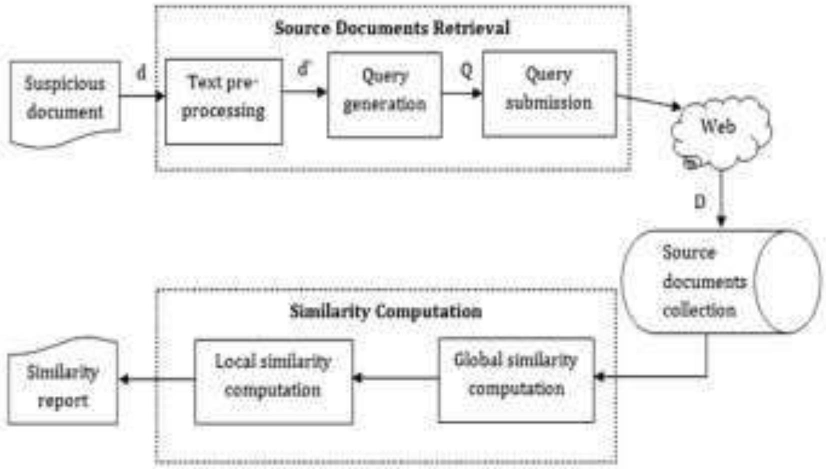

Fig 3.1:- system architure

Methodology and Algorithms

MODULES:

1. **New user Sign up**

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.

4.RESULTS AND EVOLUTION METRICS

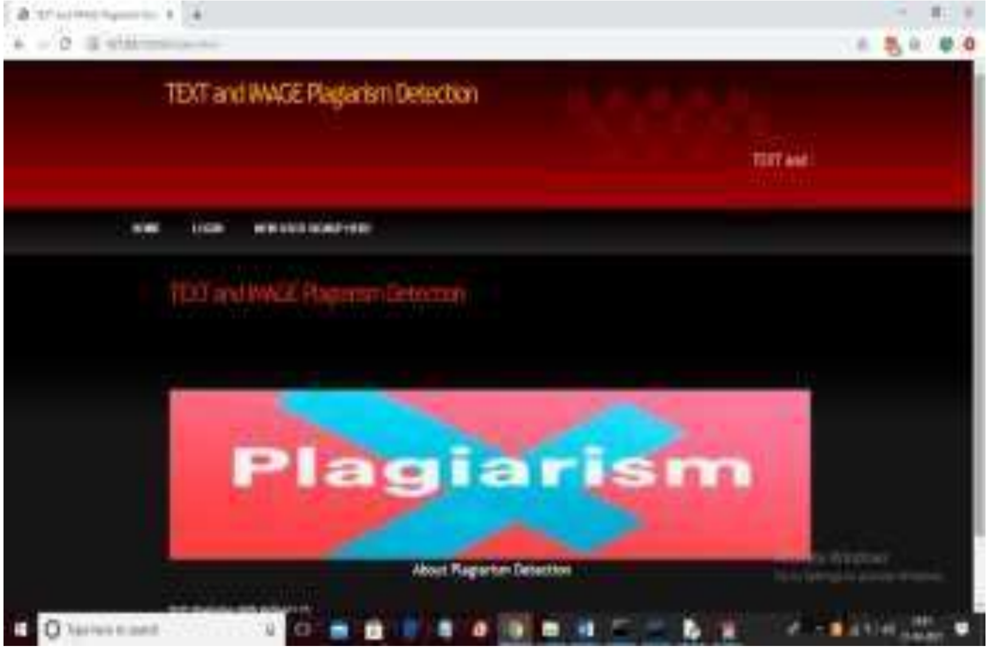
TEXT and IMAGE Plagiarism Detection
TEXT and IMAGE Plagiarism Detection
Plagiarism
About Plagiarism Detection

Fig 4.1:- home page

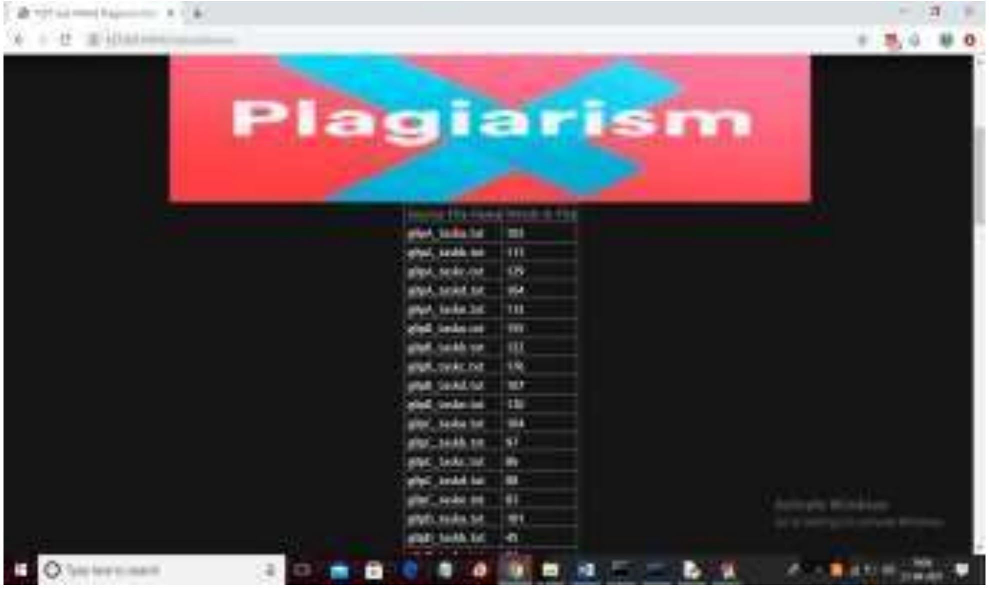

Fig4.2:- In above screen all files are loaded now click on „UploadSuspicious File‟ button to load suspicious file and get result.

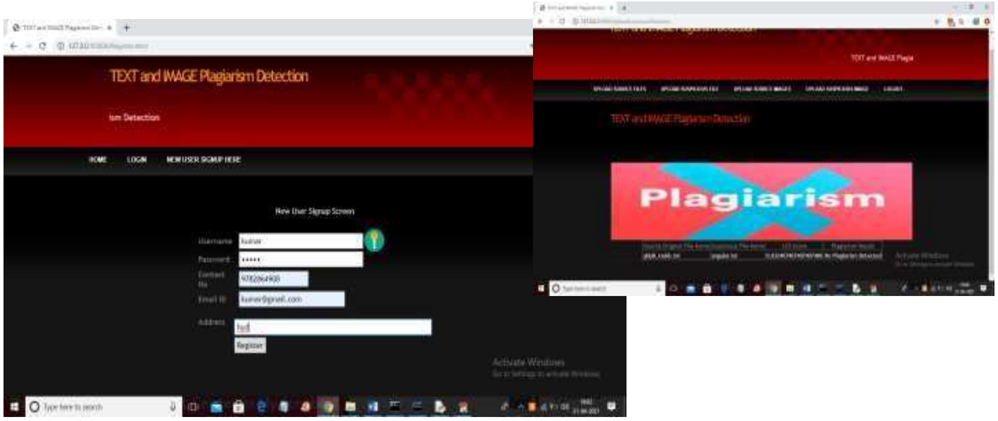

Fig 4.3:- In above screen angular.txt file matched very little with g)pB_taskb.txt corpus file and we got similarity score as 0.03 so no plagiarism detected and now upload any file from corpus and see result

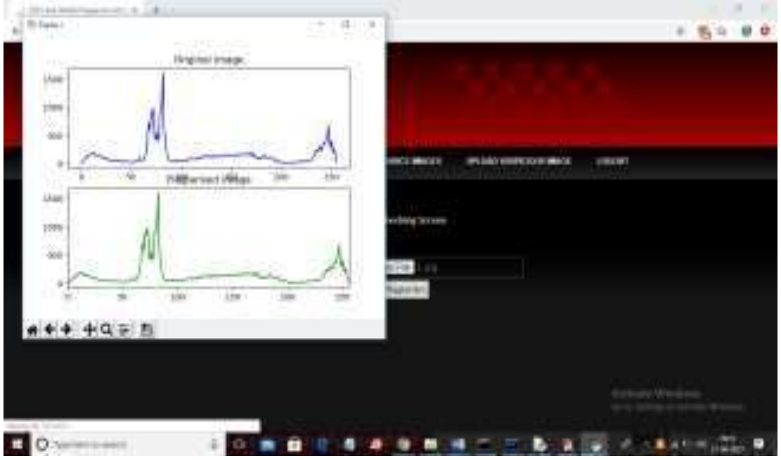

Fig 4.4:- In above screen we can both original and uploaded image histogram is matching 100% so plagiarism is detected and now close above graph to get below result.

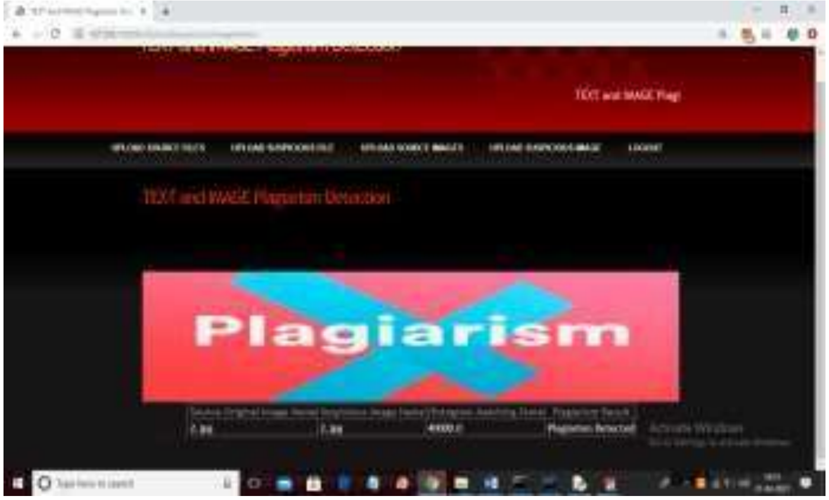

Fig 4.5:-In above screen histogram matching score is 40000 which mean all pixels matched so plagiarism is detected in above result. Similarly upl.

5.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.

References

[1]. Green, Stuart P, Plagiarism, Norms, and the Limits of Theft Law: Some Observations on the Use of Criminal Sanctions in Enforcing Intellectual Property Rights, Hastings Law Journal, 2002, 54 (1)

[2]. How Plagiarism Detection Works. (2016, May 18). Retrieved April 17, 2018, from https://www.plagiarismtoday.com/2016/05/03/plagiarismdetection- works

[3]. Vinod K.R.*, Sandhya.S, Sathish Kumar D, Harani A, David Banji and Otilia JF Banji, Plagiarism – History, Prevention and Detection, journal for drugs and medicines, 2011, 3(1), 1-4

[4]. AC Popescu and H Farid, Exposing Digital Forgeries by Detecting Duplicated Image Regions, Dept Computer Science, Dartmouth College, Hanover, 2004, 515.

[5]. B Mahdian and S Saic, Detection of Copy–Move Plagiarism using a Method based on Blur Moment Invariants, Forensic

[6]. JW Wang, GJ Liu, Z Zhang, Y Dai and Z Wang, Fast and robust forensics for image region-duplication Plagiarism, Acta Automatica Sinica, 2009, 35(12), 1488-1495.

[7]. M Zimba, and S Xingming, DWT-PCA (EVD) Based Copy-move Image Plagiarism Detection, International Journal of Digital Content Technology and its Applications, 2011, 5(1), 251-258.

[8]. S Bravo Solorio and AK Nandi, Automated Detection and Localisation of Duplicated Regions Affected by Reflection, Rotation and Scaling, Image Forensics Signal Processing, 2011, 91(8), 1759-1770.

[9]. M Sridevi, C Mala and S Sandeep, Copy–Move Image Plagiarism Detection, Journal of Computer Science and Information Technology, 2012, 52, 19-29.

[10]. Senosy Arrish, Fadhil Noer Afif, Ahmadu Maidorawa and Naomie Salim, Shape-Based Plagiarism Detection for Flowchart Figures in Texts, International Journal of Computer Science & Information Technology (IJCSIT), 2014, 6(1).