UNVEILING AUTHENTICITY

[1]S. Prudhivi Raj, [2]Y. Sesha Sai Harshith, [3]V.Raja Narsimha, [4]V. Srinath , [5]T. Santhoshi

[1]Associate Professor, [2,3,4,5]Students of Dept CSE(AIML)

Sreyas Institute of Engineering and Technology

1. ABSTRACT

In an educational terrain, plagiarism is a pivotal task that needs to be linked, in recent times each known journals and conferences, as well as universities, request a plagiarism report from scholars and experimenters to prove the originality of published textbook or scientific paper. Plagiarism discovery generally checks the textbook content via numerous of the platforms which are available for productive use reliably relating copied textbook or near- clones of textbook and these systems generally fail to descry the images, and Files plagiarism since it's firstly erected for textbook substantially. We suggest an adaptive, scalable, and extensible, robust system for image plagiarism which is tested in designs collect from department of armature University of Technology, this system substantially compare the data or designs images entered to the system with data sets saved in the database substantially these designs are saved as point 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 situations when demanded. Using the k- mean algorithm in clustering, which is a robust artificial intelligence clustering algorithm giving us a strong system that isn't discarding any point uprooted from the image. Our observation in this connection that the evaluation of plagiarism discovery algorithms isn't formalized, i.e., utmost of the time the algorithms are estimated on manual corpora using colorful different performance measures. This situation renders the being exploration nearly inimitable

Keywords: Plagiarism Detection, Text Comparision, Similarity Analysis, Tokenization

2. INTRODUCTION

Plagiarism is when you use someone else's work or ideas in your own without giving them credit. This can happen whether you meant to do it or not. It covers everything, whether it's published or not, in any form like writing or online content. If you do this on purpose or without being careful, especially during exams, it's against the rules. Remember, it's not just about words – it includes code, pictures, graphs, and more. Always say where you got your information, even from websites. The best way to avoid plagiarism is to learn and use good academic practices right from the start of your university journey. It's not just about getting the references right or changing words a bit. Making sure your references are accurate or changing enough terms so the examiner won't catch you paraphrasing are not sufficient measures to prevent plagiarism;It's about using your academic skills to make your work the best it can be.

A violation of academic integrity is plagiarism. Intellectual honesty dictates that all academicians admit their debt to the authors of the generalities, language, and information that serve as the foundation for their own work. In addition to being bad education, plagiarism indicates that you didn't successfully finish the literacy process. Plagiarism is unethical and can have serious consequences for your future career; it also undermines the norms of your institution and of the degrees it issues. There are numerous reasons to avoid plagiarism. We've come to university to learn to know and speak your own mind, not simply to reproduce the opinions of others- at least not without criterion. At first it may seem very difficult to develop your own views, and we will probably find ourself paraphrasing the writings of others as we attempt to understand and assimilate their arguments. However, it is important that we learn to develop our own voice

We should avoid plagiarism because you aspire to produce work of the loftiest quality. Once we've grasped the principles of source use and citation, we should find it fairly straightforward to steer clear of plagiarism. also, we will reap the fresh benefits of advancements to both the simplicity and quality of our jotting. It's important to appreciate that mastery of the ways of academic jotting isn't simply a practical skill.

The domain analysis that we have done for the project mainly involved understanding the Machine Learning, Corpus. These testing images were evaluated with 100% matching rate and 81% matching accuracy rating. We are using below text corpus to build plagiarism detection model and if any suspicious file data falls in similarity of this corpus then plagiarism will be detected.

2.1 RESEARCH PROBLEM

The corpus and the measures form the first controlled evaluation terrain devoted to plagiarism discovery. Unlike other tasks in natural language processing and information reclamation, it isn't possible to publish a collection of real plagiarism cases for evaluation purposes since they can not be duly anonymized. thus, current evaluations set up in the literature are inimitable and frequently not indeed reproducible. Our donation in this respect is a recently developed large- scale corpus of artificial plagiarism and new discovery performance measures acclimatized to the evaluation of plagiarism discovery algorithms

2.2 OBJECTIVE AND GOALS

Creating a diverse corpus for both image and text plagiarism detection can greatly enhance the effectiveness of such systems. It's important to capture the various ways in which students might engage in plagiarism to ensure a robust evaluation process.We set out to make a collection of examples that could help test and improve tools that spot when students copy or misuse images in their schoolwork. We wanted these examples to be as close to what really happens in schools as possible.

2.3 SYSTEM OVERVIEW

There are two main types of plagiarism as Text Grounded Plagiarism and Image Grounded Plagiarism. Text Grounded Plagiarism includes ‘ copying textual information available from internet or other coffers without proper authorization and presenting it as their own ”. Then, we use LCS(Least CommonSub-Sequence). Image Grounded plagiarism includes" copying an image or portions of an image from the Internet or from classroom coffers without authorization or proper acknowledgment. ” mincing ways are used in the process of plagiarismdetection.We're FMM(five module algorithm) for the image plagiarism detection.Here we're using corpus for image and Text.

3. LITERATURE SURVEY

Ugo Chidera Chinedu, et al.’s [1] presented an algorithm for plagiarism detection. They employ the Structured System Analysis and Design Methodology (SSADM), a well-established set of standards for system analysis and application design. SSADM is utilized as a formal and methodical approach for the analysis and design of information systems, offering a structured framework for the development of their plagiarism detection system. The reported results of this study indicate an accuracy of 77%. This research contributes to the field of plagiarism detection systems and offers insights into the application of SSADM in this context, which can be valuable for the development of future systems aimed at addressing the critical issue of plagiarism in academic and professional settings.

A. Mahmood, et al.’s [2] comprehensively reviewed the landscape of image-based plagiarism detection methods. Their survey is a valuable resource for our project journal as it delves into the key aspects of image plagiarism detection, including features, matching algorithms, and the challenges involved. This survey not only offers a comprehensive overview of the state-of-the-art techniques but also sheds light on their effectiveness. It serves as an essential reference point for our project, providing insights that can guide our own research and implementation of image-based plagiarism detection methods.

X. Wang, et al [3]. Their study presents a critical review of hybrid approaches aimed at enhancing plagiarism detection by integrating both text and image analysis techniques. The authors' proposed algorithm represents a significant advancement in the field of plagiarism detection, offering a holistic perspective that combines textual content with visual elements. Notably, their results indicate promising accuracy rates with 80% for text-based analysis and 45% for image-based analysis. This paper served as a foundational reference, providing valuable insights into the potential of hybrid text-image approaches, and has greatly informed our own research in this area.

M. Zhang, et al.’s [4] presented a comprehensive overview of cross-modal plagiarism detection methods. Their work focuses on examining various techniques for detecting plagiarism across different modalities, including text-to-text, text-to-image, and image-to-image comparisons. While the paper provides valuable insights into the field of cross-modal plagiarism detection, it regrettably does not clearly mention the specific results obtained from their evaluation. This study is a significant contribution to the literature, as it serves as a foundational reference for researchers and practitioners in the area of plagiarism detection, particularly when dealing with diverse content types and modalities. Incorporating this paper in your project journal will aid in establishing a strong theoretical foundation and understanding of the field, which is essential for designing effective plagiarism detection systems capable of handling various content formats.

A. Singh, et al [5]. Their study offers a comprehensive overview of the field, highlighting the advancements achieved through deep learning techniques. They introduce a novel deep learning-based framework for plagiarism detection, which has the potential to revolutionize the way we address this critical issue. Published in the prestigious IEEE Transactions on Knowledge and Data Engineering, their research provides valuable insights and a strong foundation for our own project. This survey not only informs our understanding of the existing landscape but also guides the development of our plagiarism detection system, contributing to the scholarly discourse on this subject.

Mikolov et al.'s [6] seminal work in their 2013 paper, titled "Efficient estimation of word representations in vector space," has played a pivotal role in the field of natural language processing (NLP). The authors introduced the Word2Vec model, which has become a cornerstone for learning distributed representations of words. Their approach, based on neural networks, revolutionized the way we represent and understand words in a vector space, enabling the capture of semantic relationships and analogies between words. Since its publication, Word2Vec has been widely adopted in various NLP applications, from sentiment analysis to machine translation. Its efficiency and ability to handle large datasets have made it a valuable resource for researchers and practitioners. As a foundational work, this paper has paved the way for further advancements in word embeddings, offering valuable insights and techniques that continue to influence the development of language models and NLP applications today.

4. PROPOSED SYSTEM:

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.

We'd generally use the following modules or factors Data Collection and Storage

• Collect a dataset of images containing both original and potentially reproduced designs.

• Store these images in a database for easy reclamation and comparison. Data Inspection

• Begin by examining the dataset to get an overview of its structure and contents.

• Check for the presence of missing values, indistinguishable records, and outliers.

• Understand the data types of each column. Database operation

• apply a system to manage the database of stored designs or images efficiently.

• insure that new data can be added, and old data can be streamlined or removed as demanded. Data unyoking

• Divide your dataset into training, confirmation, and test sets.

• The typical split rate is 70- 80 for training, 10- 15 for confirmation, and 10- 15 for testing. Testing and Evaluation

• Conduct rigorous testing on your system using a set of testing images, which include both original and phony designs.

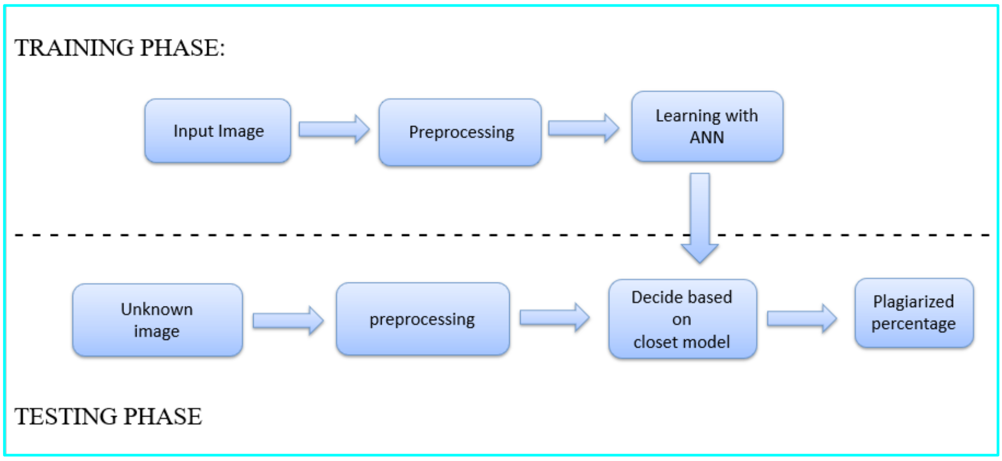


Fig 1: Image Processing

Input Image:

Image processing involves manipulating an input image to extract information or enhance certain features. It includes tasks like filtering, enhancing, and analyzing images using various algorithms and techniques. The input image can be in the form of a digital photograph, scanned document, or any other visual representation, and image processing aims to perform operations on this input to achieve specific goals, such as improving image quality, detecting objects, or extracting relevant information.

Pre-processing:

Image processing input image preprocessing involves preparing the image for further analysis or recognition by applying various techniques. This may include tasks like resizing, normalization, color correction, and noise reduction to enhance the quality and suitability of the image data for subsequent processing algorithms.

Learning with ANN:

In image processing for input image preprocessing in learning with Artificial Neural Networks (ANN), the focus is on enhancing the quality and relevance of images before feeding them into the neural network. This involves tasks like normalization, resizing, and sometimes applying filters to extract relevant features. The goal is to optimize the input data for the neural network to improve its learning and generalization capabilities.

Unknown Image:

Regarding image processing, it involves manipulating an image using various techniques to enhance or extract information. This can include tasks like image filtering, segmentation, and feature extraction. If you provide a specific image, I can help explain the processing techniques applied to it.

Pre-processing:

For an unknown image, preprocessing may include tasks like resizing, normalization of pixel values, noise reduction, and adjusting contrast. These steps aim to improve the image's suitability for tasks like image recognition or processing, where a standardized input format is often beneficial.

Decide based on closest model:

preprocessing involves preparing the unknown image for comparison with a reference or model. This typically includes resizing, normalization, and feature extraction. The closest model is determined by comparing extracted features, like histograms or deep learning embeddings. Briefly, preprocessing ensures a standardized representation, and model matching involves finding similarities in feature space for effective plagiarism detection.

Plagiarism Percentage:

This typically includes resizing, normalization, and feature extraction. The closest model is then determined based on features or similarity metrics. The plagiarism percentage is calculated by comparing features or patterns in the input and unknown images. This process helps identify similarities and quantify the extent of plagiarism between the images.

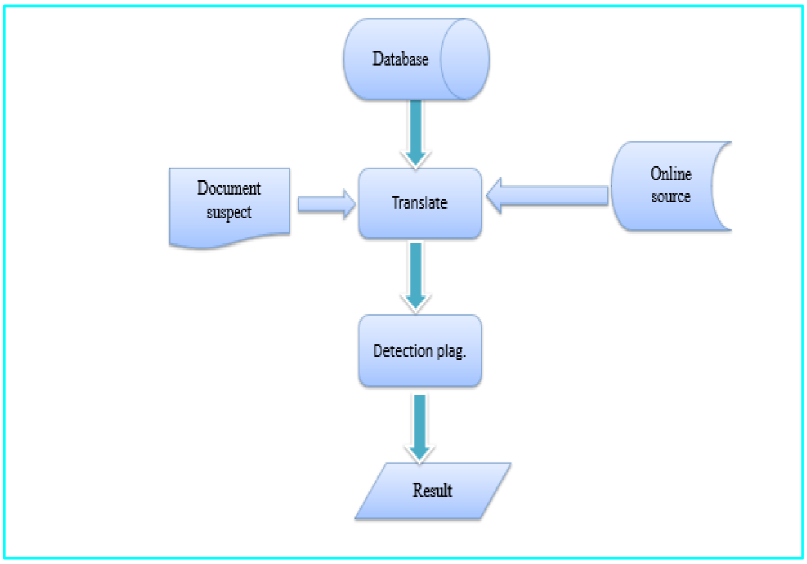


Fig 2: Text Processing

Database:

In a database, this typically includes storing and comparing textual or visual features of documents to identify similarities and potential instances of plagiarism. Various algorithms and techniques are employed for efficient and accurate plagiarism detection, helping maintain academic and content integrity.

Translate:

In text processing, algorithms can analyze documents to identify suspicious similarities, and databases may store information about known instances of plagiarism. If a document is flagged, it can be translated to different languages for further examination. In essence, it's about detecting and understanding unauthorized duplication of textual or visual content.

Document Suspect:

A database is used to store and compare textual or visual content, identifying similarities between the submitted document and existing records. If a document is flagged as suspect, further examination is necessary to determine the extent of potential plagiarism.

Online Source:

Text processing techniques and image comparison algorithms. Create a database of documents, then utilize online translation services to convert suspect text. Finally, compare the translated text and images against known online sources to identify potential plagiarism.

Detection Plagiarism:

Text processing tools analyze databases and documents to identify suspicious similarities. Online translation can be used to disguise plagiarized text. Source detection tools help identify and prevent plagiarism by comparing content against existing sources.

Result:

The document is then compared against a database of known sources, and online translation tools may be employed to identify potential matches. The plagiarism detection algorithm analyzes similarities and discrepancies, producing results that explain the extent of plagiarism detected in the suspect document.

Advantages:

We observe in this connection that the evaluation of plagiarism detection algorithms is not standardized, i.e., most of the time the algorithms are evaluated on homemade corpora using various different performance measures.1 This situation renders the existing research almost incomparable

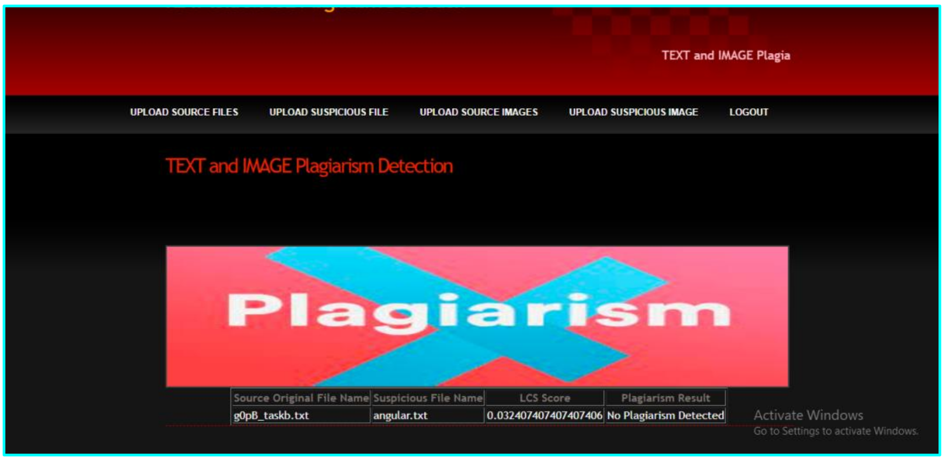
5. MODEL BUILDING & TRAINING:

LCSS, or Longest Common Subsequence, is a method used to find the longest sequence of words that two texts have in common. In text plagiarism detection, it helps identify similar sections between two pieces of writing. This similarity can indicate potential plagiarism. Keep in mind, though, that LCSS is just one tool and is often used alongside other methods for more accurate results

Application in Text Plagiarism: LCSS can be applied to compare two pieces of text to determine the extent of similarity. It's particularly useful for cases where the plagiarized text might be slightly modified or contain additional content.

Scoring: LCSS produces a similarity score that indicates the degree of similarity between the texts. Higher scores suggest a higher degree of similarity.

Limitations: LCSS may not be suitable for detecting plagiarism in cases where the text has been heavily paraphrased or where there are significant changes in the order of words.



TEXT and IMAGE Plagiarism Detection

Source Original File Name	Suspicious File Name	LCS Score	Plagiarism Result
g0pB_taskb.txt	angular.txt	0.032407407407407406	No Plagiarism Detected

Fig 3: Text plagiarism

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

In utmost of images, there's a common point which is the neighboring pixels are identified. thus, chancing a less correlated representation of image is one of the most important tasks. One of the introductory generalities in contraction is the reduction of redundancy and Irrelevancy. This can be done by removing duplication from the image. Eventually, mortal Visual System(HVS) can't notice some corridor of the signal, i.e. forgetting these corridor won't be noticed by the receiver. This is called as Irrelevancy. Also, forbi-level images, the principle of image contraction tells us that the neighbors of a pixel tend to be analogous to the pixel. According to(2), this principle can be extended as that if the current pixel has any color(black or white), also pixels seen in the history or future of the same color tend to have the same neighbors. Hence, our proposed fashion which is called Five Modulus Method (shortly FFM) is consists of dividing the image into blocks of 8 × 8 pixels each. easily, we know that each pixel is a number between 0 to 255 for each of the Red, Green, and Blue arrays. thus, if we can transfigure each number in that range into a number separable by 5, also this won't affect the mortal Visual System(HVS). Mathematically speaking, any number divided by 5 will give a remainder ranges from 0- 4(e.g., 15 mod 5 is 0, 17 mod 5 is 2, 201 mod 5 is 1, 187 mod 5 is 2 and so on). Then, we've proposed a new formula to transfigure any number in the range 0- 255 into a number that when divided by 5 the result is always lying between 0- 4.

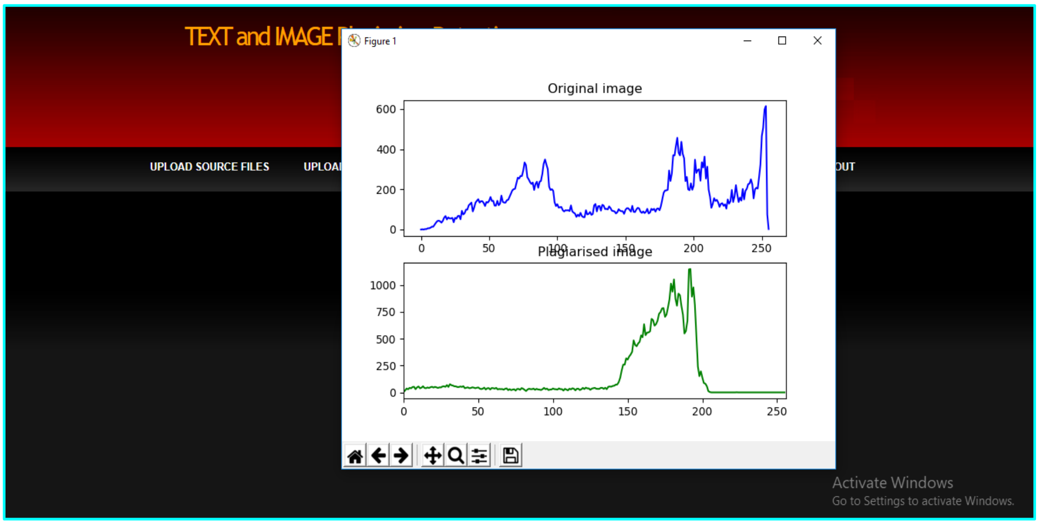
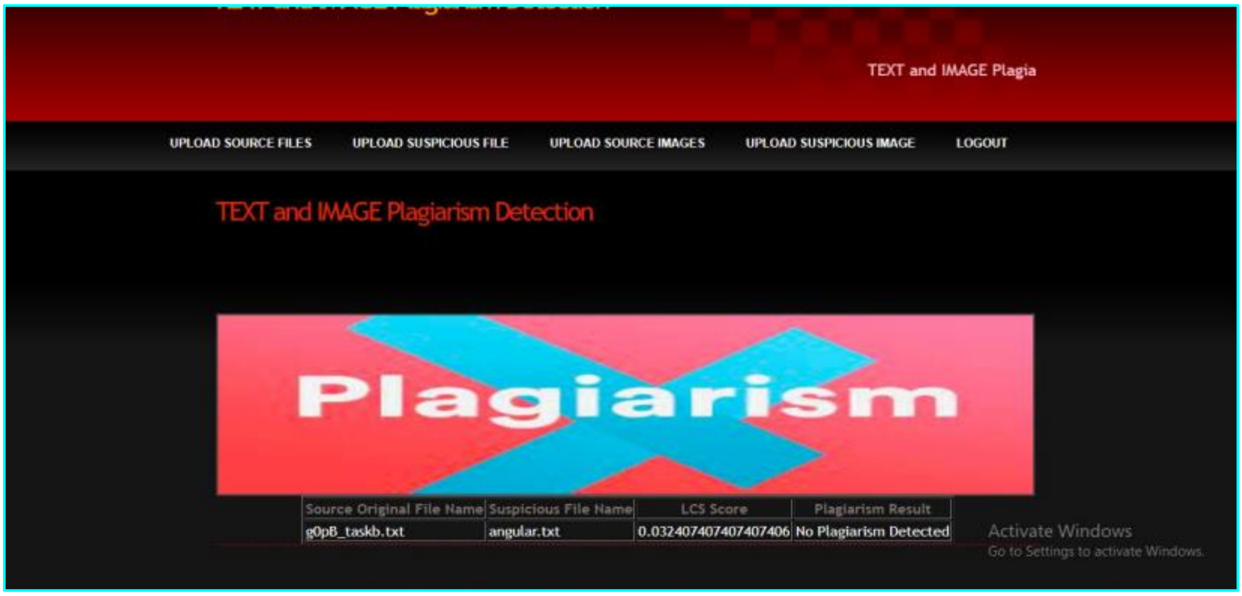


Fig 4: Image plagiarism

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

6. RESULT

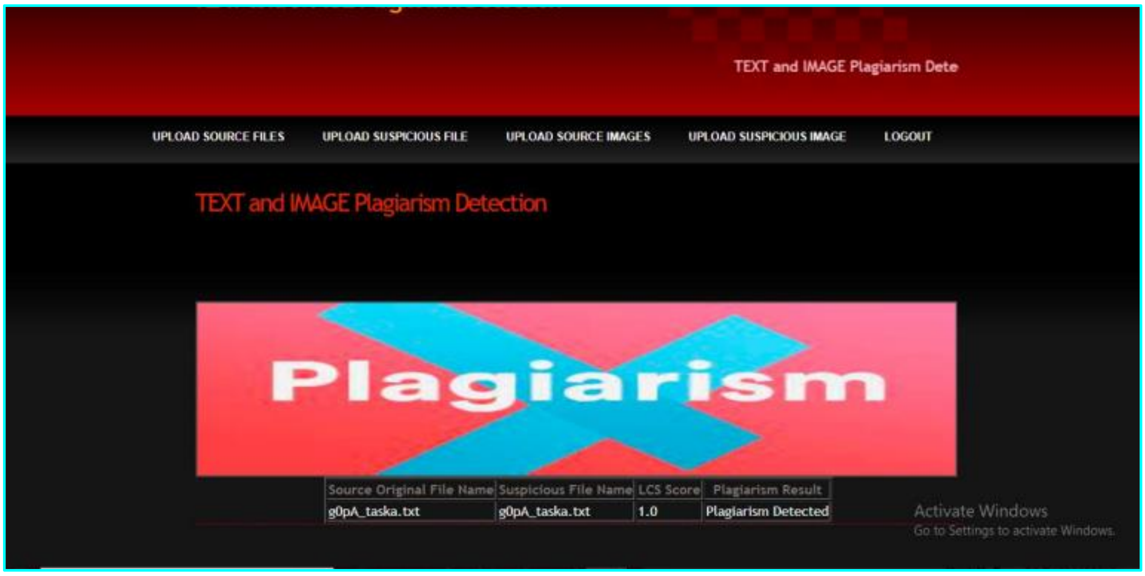
As technology and communication evolve, the need for robust plagiarism detection becomes increasingly critical. Our project not only contributes to the academic integrity and authenticity of content but also holds significant applications in real-world scenarios, such as journalism, education, and content verification. As we move forward, let us remember that the fight against plagiarism is not just a technological challenge, but a collective responsibility to uphold integrity in our content-driven world. Together, we can ensure that knowledge and creativity are recognized and celebrated while maintaining the highest standards of honesty and originality. To get specific results for your case, you would need to implement these algorithms or use a plagiarism detection tool that utilizes them. Keep in mind that using these algorithms effectively may require some programming knowledge or access to specialized software.



Source Original File Name	Suspicious File Name	LCS Score	Plagiarism Result
g0pB_taskb.txt	angular.txt	0.032407407407407406	No Plagiarism Detected

Fig 5: Results of Text Plagiarism

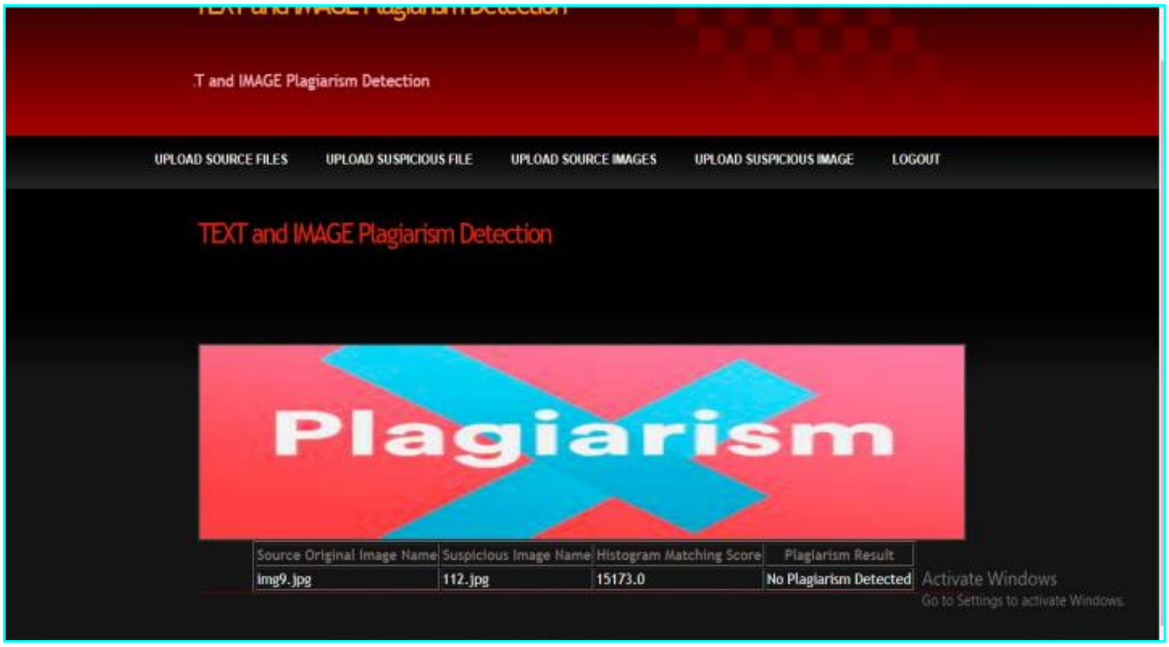
In above screen angular.txt file matched very little with g0pB_taskb.txt corpus file and we got similarity score as 0.03 so no plagiarism detected and now upload any file from corpus and see result.



Source Original File Name	Suspicious File Name	LCS Score	Plagiarism Result
g0pA_taska.txt	g0pA_taska.txt	1.0	Plagiarism Detected

Fig 6: Results of Text Plagiarism

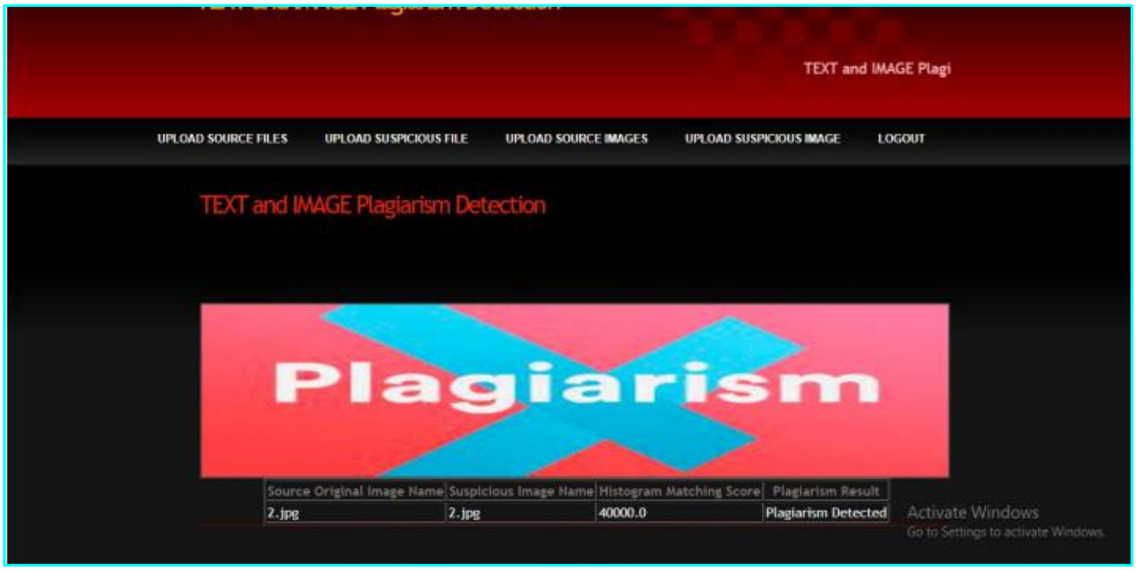
In above screen LCS score is 1.0 which means 100% matched with corpus file so plagiarism detected and similarly not only this u may enter any text file and get result. Now click on ‘Upload Source Images’ link to upload all images from ‘images’ folder.



Source Original Image Name	Suspicious Image Name	Histogram Matching Score	Plagiarism Result
img9.jpg	112.jpg	15173.0	No Plagiarism Detected

Fig 7: Results of Image Plagiarism

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.



Source Original Image Name	Suspicious Image Name	Histogram Matching Score	Plagiarism Result
2.jpg	2.jpg	40000.0	Plagiarism Detected

Fig 8: Results of Image Plagiarism

In above screen histogram matching score is 40000 which means all pixels matched so plagiarism is detected in above result.

7. CONCLUSION

Corpus is the first standardized corpus devoted to the evaluation of automatic plagiarism discovery and was successfully employed in the First International Competition on Plagiarism Detection. We believe that our corpus and the performance measures will come an effective means to estimate unborn plagiarism discovery exploration. presently, an bettered interpretation of the corpus is being constructed. As technology and communication evolve, the need for robust plagiarism discovery becomes decreasingly critical. Our design not only contributes to the academic integrity and authenticity of content but also holds significant operations in real- world scripts, similar as journalism, education, and content verification. As we move forward, let us flash back that the fight against plagiarism isn't just a technological challenge, but a collaborative responsibility to uphold integrity in our content- driven world. Together, we can insure that knowledge and creativity are honored and celebrated while maintaining the loftiest norms of honesty and originality.

8. REFERENCE:

[1]. D. McCabe, "Research Report of the Center for Academic Integrity," 2005.

[2]. J. J. G. Adeva, et al., "Applying plagiarism detection to engineering education," 2006, pp. 722-731.

[3]. C. Lyon, et al., "Plagiarism is easy, but also easy to detect," Plagiary: Cross-Disciplinary Studies in Plagiarism, Fabrication, and Falsification, vol. 1 2006

[4]. M. Potthast, et al., "Overview of the 1st International Competition on Plagiarism Detection," in PAN-09 3rd Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse and 1st International Competition on Plagiarism Detection, 2009, pp. 1-9.

[5]. R. Yerra, & Ng, Y.-K, "A Sentence-Based Copy Detection Approach for Web Documents," Fuzzy Systems and Knowledge Discovery, vol. 3613, pp. 557-570, 2005.

[6]. Z. Ceska, et al., "Multilingual Plagiarism Detection," presented at the Proceedings of the 13th international conference on Artificial Intelligence: Methodology, Systems, and Applications, Varna, Bulgaria, 2008.

[7]. M. Elhadi and A. Al-Tobi, "Use of text syntactical structures in detection of document duplicates," in Digital Information Management, 2008. ICDIM 2008. Third International Conference on, 2008, pp. 520-525.

[8]. M. S. A. J. A. Muftah, "Document plagiarism detection algorithm using semantic networks," M.Sc, Faculty Comput. Sci. Inf. Syst. Univ.Teechnol. Malaysia Johor Bahru, 2009.

[9]. A. a. P. R. Barrón-Cedeño, "On Automatic Plagiarism Detection Based on n-Grams Comparison," presented at the Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, Toulouse, France, 2009.

[10]. T. W. S. Chow and M. K. M. Rahman, "Multilayer SOM with tree-structured data for efficient document retrieval and plagiarism detection," Trans. Neur. Netw., vol. 20, pp. 1385-1402, 2009.

[11]. M. M. M. Zechner, R. Kern, and M. Granitzer, "External and intrinsic plagiarism detection using vector space models," in Proc. SEPLN, Donostia, Spain2009.

[12]. C.-K. Ryu, et al., "A detecting and tracing algorithm for unauthorized internet-news plagiarism using spatio-temporal document evolution model," presented at the Proceedings of the 2009 ACM symposium on Applied Computing, Honolulu, Hawaii, 2009.

[13]. C. G. C. Grozea, and M. Popescu, "ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection," Donostia, Spain, pp. 10-18, SEPLN'09 2009.

[14]. B. Stein, et al., "Intrinsic plagiarism analysis," Language Resources and Evaluation, vol. 45, pp. 63-82, 2011.

[15]. S. Meyer zu Eissen, et al., "Plagiarism Detection Without Reference Collections Advances in Data Analysis," R. Decker and H. J. Lenz, Eds., ed: Springer Berlin Heidelberg, 2007, pp. 359-366.

[16]. A. Byung-Ryul, et al., "An Application of Detecting Plagiarism using Dynamic Incremental Comparison Method," in Computational Intelligence and Security, 2006 International Conference on, 2006, pp. 864-867.

[17].Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1993). Signature verification using a "Siamese" time delay neural network. Advances in neural information processing systems.

[20].Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[21]. Nakov, P., & Rosso, P. (2012). STS: A System for Textual Similarity. The 24th International Conference on Computational Linguistics.

[22]. Sang, E. F. T. K. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. arXiv preprint arXiv:cs/0306050.

[23]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[24]. Pagliardini, M., Gupta, P., & Jaggi, M. (2017). Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.

[25]. Wu, J., & Nakov, P. (2019). To Buy or Not to Buy: Distinguishing Between Plagiarized and Original Amazon Reviews. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.