IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

TEXT SIMILARITY IDENTIFICATION FROM DRAWINGS AND GRAPHS

[1]Varsha Bhosale, [2]Shriya Kadam, [3]Abhishek Joshi, [4]Sagar Kulkarni

[123]Student,[4]Professor
[1]Information Technology,
[1]Pillai College of Engineering, University of Mumbai, New Panvel, India

Abstract: Copying someone else's text, or data without their permission may be a serious misdeed. Using image-to-text conversion techniques such as the Heroku app which is an OCR technique we extract the text from the image contents by checking the Plagiarism and discover the Similarity of the image with the present images replaced by synonyms or antonyms. For Plagiarism detection, we have used Cosine Similarity which can check the similarity between documents.

Index Terms **-Heroku App, OCR(Optical character recognition), Plagiarism Detection, Plagiarism.**

I. INTRODUCTION

Text recognition has gained a lot of prominence in recent years as it has entered a large arena of applications. Most of the information transfers these days take place via images or scanned documents. Digital images are getting very popular Text in the Image can be found in magazines, posters, newspapers, scanned documents, etc. Most of the time the text in the image is highly plagiarized. Many free downloadable software, paid tools, and online tools are present in the market with the aim of detecting and providing similarity reports Turnitin, iThenticate, etc. but the main drawback of this software is they fail to check the plagiarism in the image. So, we propose a new similarity measure by using the idea we will extract the text from the image contents by checking the Plagiarism and discover the Similarity of the image.

II. LITERATURE REVIEW

2.1. Extracting Text From Image Document and Display Related Information:

This paper uses Edge Based Algorithm for extraction of text from images and Connected Component Algorithm which uses geometrical analysis to merge these text components which alters non-text components.

2.2. Text Recognition from Images: A Study:

This paper presents a brief summary of various steps used in text recognition from images. A review of the basic model of the text recognition system is also given which describes the flow of text recognition from images. Finally, the various fields where text recognition could be used are discussed. This paper discusses the text recognition module and also presents the various applications of text recognition from images.

2.3. Text Information Extraction in Images and Video: A Survey:

In this paper, we are provided with a comprehensive literature review of text extraction in images and video as well as text-based

image and video retrieval. Text data present in images and videos contain useful information for automatic annotation, indexing, and structuring of images. Extraction of this information involves detection, localization, tracking, extraction, enhancement, and recognition of the text from a given image. However, variations of text due to differences in size, style, orientation, and alignment, as well as low image contrast and complex background problems of automatic text extraction, were discussed in this paper.

2.4. Texts Semantic Similarity Detection Based Graph Approach.:

They offered a new approach using graph theory for computing text semantic similarity and using WordNet as a knowledge base. By

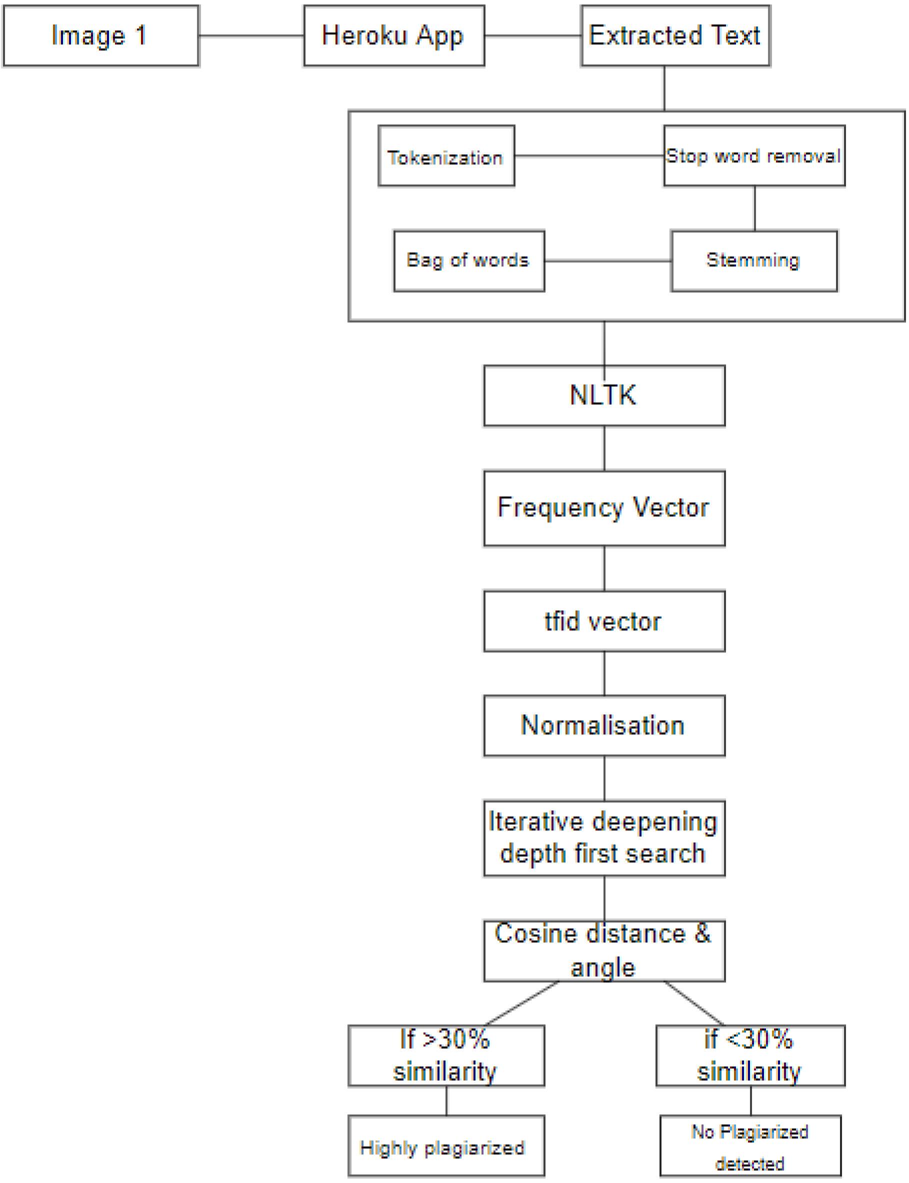
selecting specific edges, only the specific weight of similarity is selected for pairs of words.

III. PROPOSED WORK

The proposed Text Similarity Identification extracts the text from Images and check the plagiarism and gives the output that whether how many percent of the text is plagiarized. If the plagiarism is more than 30% than it is considered as highly plagiarized and if the plagiarism is less than 30% then no plagiarism is detected.

IV. SYSTEM ARCHITECTURE

The system architecture is given in Figure 2. Each block is described in this Section.



a. Extracting Text: The Image File will be sent over the Heroku app, which will then give a response object which contains the text from the image. That text will be stored in a .txt file. then that .txt file is taken as input and given for preprocessing.

b. Pre-processing: The first step in the pre-processing is to present the English documents in a clean word format and the output data will only consist of useful phrases.

c. Tokenization: Tokenization could be a method of changing sentences into a sequence of words so that processing word by word can be easily performed.

d. Stop Word Removal: The most frequently occurring words which unnecessarily slow down the processing of documents are called stop words. These words are irrelevant. Such words include articles, conjunctions, 22 prepositions, and other function words.

e. Stemming: Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. Stemming is important in natural language understanding (NLU) and natural language processing (NLP). When a new word is found, it can present new research opportunities.

f. **A Bag of Words:** A bag of words is a representation of text that describes the occurrence of words within a document. We just keep track of word counts and disregard the grammatical details and the word order. It is called a "bag" of words because any information about the order or structure of words in the document is discarded.

g. **Frequency Vector:** It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. It is often used to measure document similarity in text analysis.

h. **TF-IDF:** TF-IDF is an abbreviation for Term Frequency-Inverse Document Frequency and is a very common algorithm to transform text into a meaningful representation of numbers. The technique is widely used to extract features across various NLP applications

i. **Iterative deepening search:** iterative deepening search or more specifically iterative deepening depth-first search (IDS or IDDFS) is a state space/graph search strategy in which a depth-limited version of depth-first search is run repeatedly with increasing depth limits until the goal is found.

V. CONCLUSION

In this, the study of different domain techniques is presented. Plagiarism detection systems available in the market and online, at the present time, do not have very good accuracy due to a lot of factors one such factor is that they do not compare images. Named Entity Recognition (NER) is one such factor that most of the systems are not able to avoid while considering the plagiarized content. A solution to it is using Natural Language Processing and Machine Learning methods. It will be able to extract the text from the image contents and discover the similarity of the image with the existing images replaced by synonyms or antonyms. The final output decides whether the documents have similar content or not.

VI. ACKNOWLEDGMENT

We would like to thank our Guide Prof. Sagar Kulkarni, who gave us this golden opportunity to do this wonderful project on the topic "Text Similarity Identification using Drawings and Graphs", which also helped us in doing a lot of research and thereby increasing our knowledge.

We would like to express our gratitude to our H.O.D of Information Technology Dr. Satish Kumar Verma for giving us this opportunity and for motivating us to do innovative things that will be beneficial for our future.

We would also like to thank our principal Dr. Sandeep Joshi who always encouraged us and motivated us to study in this great college and also helped us in various things. We take this opportunity to thank all our faculties who have directly or indirectly helped in our project.

REFERENCES

[1] Extracting Text from Image Document and Displaying Its Related Information, K.N. Natei, J. Viradiya, S. Sasikumar, K.N. Natei Journal of Engineering Research and Application, Vol. 8, May 2018

[2] Graph-based Representation for Sentence Similarity Measure: A Comparative Analysis, Siti Sakira Kamaruddin, Yuhanis Yusof, Nur Azzah Abu Bakar, Mohamed Ahmed Tayie, Ghaith Abdulsattar A.Jabbar Alkubaisi, International Journal of Engineering & Technology, Vol 7, No 2.14 (2018)

[3] Texts Semantic Similarity Detection Based Graph Approach Majid Mohebbi and Alireza Talebpour Department of Computer Engineering, Shahid Beheshti University, Iran

[4] Text Recognition from Images: A Study Sahana K Adyanthaya Assistant Professor: Department of ECE A. J. Institute of Engineering and Technology Mangaluru, India