Text Based Plagiarism detection On Electronic Submissions

Bachelor of Engineering
in
Computer Engineering

Submitted by

Ms. Ankita Salavi 17CE5001

Ms. Komal Sonawane 17CE5012

Ms. Shweta Tarmale 17CE5022

Guided by

(Mrs. Smita Bhoir)

Department of Computer Engineering

Ramrao Adik Institute Of Technology

Dr. D. Y. Patil Vidyanagar, Sector-7, Nerul, Navi Mumbai-400706.
(Affiliated to University of Mumbai)

April 2020

Text Based Plagiarism detection on Electronic Submissions

B.E. Project Report

Submitted in partial fulfillment of the requirements

For the degree of

Bachelor of Engineering
in
Computer Engineering

Submitted by

Ms. Ankita Salavi 17CE5001

Ms. Komal Sonawane 17CE5012

Ms. Shweta Tarmale 17CE5022

Guided by

(Mrs. Smita Bhoir)

D Y PATIL
RAMRAO ADIK
INSTITUTE OF
TECHNOLOGY
NAVI MUMBAI

Department of Computer Engineering

Ramrao Adik Institute Of Technology

Dr. D. Y. Patil Vidyanagar, Sector-7, Nerul, Navi Mumbai-400706.
(Affiliated to University of Mumbai)

April 2020

Ramrao Adik Institute of Technology

(Affiliated to the University of Mumbai)

Dr. D. Y. Patil Vidyanagar, Sector-7, Nerul, Navi Mumbai-400706.

CERTIFICATE

This is to certify that, the project 'B' titled

Text Based Plagiarism detection on Electronic Submissions

is a bonafide work done by

Ms. Ankita Salavi 17CE5001

Ms. Komal Sonawane 17CE5012

Ms. Shweta Tarmale 17CE5022

and is submitted in the partial fulfillment of the requirement for the degree of

Bachelor of Engineering
in
Computer Engineering
to the
University of Mumbai

Supervisor

(Mrs. Smita Bhoir)

Project Co-ordinator
(Mrs. Smita Bharne)

Head of Department
(Dr. Leena Ragha)

Principal
(Dr. Mukesh D. Patil)

Project Report Approval for B.E

This is to certify that the project ‘B’ entitled ***“Text Based Plagiarism detection on Electronic submissions”*** is a bonafide work done by ***Ms. Ankita Salavi***, ***Ms. Komal Sonawane***, and ***Ms. Shweta Tarmale*** under the supervision of ***Mrs. Smita Bhoir***. This dissertation has been approved for the award of ***Bachelor’s Degree in Computer Engineering, University of Mumbai***.

Examiners:

1. . . . .

2. . . . .

Supervisors:

1. . . . .

2. . . . .

Principal:

. . . .

Date: . . . / . . . /. . . . . .

Place: . . . .

Declaration

We declare that this written submission represents my ideas in my own words and where other’s ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Ms. Ankita Salavi	**17CE5001**	_____________________
Ms. Komal Sonawane	**17CE5012**	_____________________
Ms. Shweta Tarmale	**17CE5022**	_____________________

Date : . . . /. . . /. . . . . .

Abstract

Plagiarism is characterized as the practice of claiming or suggesting original authorship of the written or artistic work of someone else, in whole or in part, into one's own without proper recognition. Plagiarism is one of the rising problems in universities and other academic institutions nowadays, and is still a concern. False copying of another's work is known as plagiarism. Plagiarism is difficult to identify manually and this method should be automated. Field data mining that can help identify the plagiarism. Plagiarism can be observed using a number of data mining techniques. The techniques that can aid in this process include text mining, clustering, bi-gram, tri-gram, n-grams. These days, the plagiarism detection techniques focus not only on exact copying but also on catching intelligent plagiarism like paraphrasing. The proposed work provides an efficient system for plagiarism detection.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Overview

Presenting someone else's work as your own is known as plagiarism, or copying the words, thoughts, or writing of another person, and claiming that they are his own work. Stealing someone else also works without complete acknowledgment Known as Plagiarism. Already this topic is raised a day quite vast as students need to develop skills and gain their own expertise while they are studying. That means writing their own essays without copying the work of someone else who would not encourage them to know. Plagiarism is needed because Internet use is increasing today and knowledge is accessible through a single click, but the original author is working hard to produce his work. Therefore it is disrespectful to copy it without giving credits. Other students also work hard for their findings. The originality of the authors work can be easily identified by using the plagiarism checker.

Plagiarism is critical issue now days. By solving this issue we know the original work, thought and think which are created by the author. Plagiarism checker enables students a way to demonstrate their understanding of the material and develop the ability to communicate with others with your own writing and thought. Commonly there are different methods are used for plagiarism. Many of them are plagiarism for copy-paste, paraphrasing plagiarism means restating certain material in different terms, translated plagiarism implies reproduction of material and use without relation to original job.

Plagiarism can be identified through the implementation of any method and algorithm to locate the document's originality. Different methods are available to identify the plagiarized content

Chapter 1

from the document such as natural language processing, symentatic analysis. Using 'Text Based Plagiarism Detection by Electronic Submission' we provided the solution to the issue of plagiarism detection. Data Mining is primarily the area that is used to detect plagiarism. There are various data mining algorithms are used to detect plagiarized content like wise tri-gram, bi-gram but we are using the most effective algorithm known as n-gram. This algorithm is works in three different phases to solve the problem of plagiarism and to generate the originality report with the percentage of plagiarism in the document with referred source links.

Plagiarism detection software is mainly used in all are where the author, student or any person want to represent there original article, reports, assignments, presentation or any other research papers. It will be beneficial for the institutes, company, schools where originality of work is most important.

1.2 Objective

The major objectives of this project are:

- To recognize any act of dishonesty in academic research is academic misconduct To check the originality of given document, text.
- To highlight content which is an exact match for the words of the original author.To show the result immediately and the percentage of the compared text.
- To provide some advanced level plagiarism checker for accuracy of grammar and paragraphing of the written text to improve the document.

1.3 Motivation

The rapid development of information technology, especially the Internet is pointed out to be factor driving the student to practice plagiarism. Plagiarism severely affects the educational process in a variety of areas, where students are discouraged from cultivating the innovative thinking and analytical skills that adversely affect a college / university student's overall educational experience.

In different fields such as literature, music, software, scientific articles, research papers, magazines, advertising, websites etc., plagiarism can be found. A study carried out in the United States shows that nearly 40 per cent of the 18,000 university students plagiarized at least once.

The proposed system is designed to create plagiarism identification that is based on matching characters, n-gram, chunks or words to resolve plagiarism.

1.4 Organization of report

The report is organized into 8 chapters after the Abstract.

i. The chapter 1 is Introduction, which contains sections like Overview of Project, Objectives and Motivation of the project.

ii. The chapter 2 is Literature Survey, which contains sections like Survey of Existing System, its Limitations and the Problem Definition.

iii. The chapter 3 named Proposal contains Proposed Work, Proposed Methodology and the Details of Hardware and Software Requirements.

iv. The chapter 4, Planning and Formulation describes the scheduling and flow of the project with help of Gantt chart.

Chapter 1

v. The chapter 5, Design of System, describes the overall design of the proposed system.

vi. The chapter 6 named Expected Results, gives the description of the expected Project Outcomes.

vii. The chapter 7 is Conclusion of Proposed System.

viii. The chapter 8 is Future Work of the project. The References of the Literature Survey are given at the end of Report.

Chapter 2

Literature Survey

2.1 Survey of Existing System

Research Paper survey

The extensive literature survey of various research papers carried out and findings are as follows-

1. MAC Jiffriya, MAC Akmal Jahan, Roshan G Ragel and Sampath Deegalla “ANTIPLAG: PLAGIARISM DETECTION ON ELECTRONIC SUBMISSIONS OF TEXT BASED ASSIGNMENTS”[1] .This paper provides an efficient and fast tool for plagiarism detection of electronic documents based on for text. The key emphasis is the identification in the document of the related word.

2. Muhammad Usman, Muhammad Waleed Ashraf ' PLAGIARISM DETECTION Method USING DATA MINING TECHNIQUES.[2] 'This paper explains the Tri-gram and clustering method used in data mining to detect plagiarism.

3. Vani K and Deepa Gupta "Research ON EXTRINSIC TEXT PLAGIARISM DETECTION TECHNIQUES AND Methods."[3] This paper defines tools and techniques for the detection of extrinsic text plagiarism, a fuzzy semantic-based approach to string similarities.

4. Kamalpreet Sharma and Balkrishan Jindal "AN Enhanced PLAGIARISM Identification Method FOR SEMANTIC ANALYSIS USING CUSTOM SEARCH ENGINE"[4] This paper explains extra-corpal semantic analysis using custom search engine and crawling, Custom search engine using semantic analysis.

5. Mansi Sahi and Vishal Gupta "Efficiency comparison of different plagiarism detection techniques," [5] This paper discusses different plagiarism techniques, including semantic-based, enhanced ranking based on semantic, semantic and syntactic, metrics-based, semantic role labeling with sentence ranking

Chapter 2

2.2 Summary of Literature Survey

Paper Title	Author	Year	Focus	Methods	Purpose
AntiPlag Plagiarism Identification of Text Based Assignments on Online Submissions.	MAC Jiffriya, MAC Akmal Jahan, Sampath Deegalla, Roshan G Ragel	2018	Efficient and fast tool for plagiarism detection when sending text based electronic documents	Tri-gram sequence matching algorithm.	Detecting the similar term into the document.
Detection of plagiarism using Data Mining Techniques	Muhammad Usman, Waleed Ashraf	2017	This paper shows how software tools detect the plagiarism.	Data mining techniques	Study the different data mining techniques that use to detect plagiarism.
Study on Identification Methods and Tools for Extrinsic Text Plagiarism	Deepa Gupta and Vani K.	2016	Plagiarism of extrinsic text tools and techniques Detection	Fuzzy semantine comparison approach to strings	Study and analysis of some of the Plagiarism detection algorithms.
Effectiveness Comparison of different plagiarism detection techniques	Vishal Gupta and Mansi Sahi	2016	This paper addresses various plagiarism methods based on metrics, including semantic based, enhanced ranking based semanthetic, semanthetic and syntactic.	Semantic Role Labelling with Sentence Ranking	In this paper the emphasis is on techniques to detect extrinsic plagiarism
An enhanced Way to Detect Plagiarism for Semantic Research using a Custom Search Engine.	Kamalpreet Sharma, and Jindal Balkrishan	2016	This paper uses custom search engine and crawling method to address extra-corpal semance analysis.	Custom search engine using semantic analysis.	Semantic plagiarism detection system using crawling service provided by custom search API

Table 2.2: Summary of Literature Survey

Chapter 2

Table 2.2 shows the analysis of various research papers and their methodologies for plagiarism detection. Each method has its advantages and disadvantages which are analyzed based on their accuracy value. In existing system Pre-processing and clustering techniques can be used to decrease the overhead of the process. Moreover, similarity score can be calculated through the clusters of plagiarized data so that efficiency can be improved.

2.3 Problem Statement

This project is a content-based approach for analyzing and visualizing similarities between text documents. Field data mining that can help identify the plagiarism. The processing of this technique will be stored in electronic form as all the text documents, the text document is converted into a suitable format Furthermore, the data is passed through the text analysis step, then the word count and no characters are generated from the uploaded document, then the plagiarized text highlighted with the original referenced source after that graph is generated showing the percentage of unique and plagiarized text.

2.4 Scope

Plagiarism can occur with regard to all forms of sources and media: not only text, but also images, musical quotes, computer code, etc.; not only text published in books and journals, but also text downloaded from websites or from other media, including in published content, unpublished plays, including lectures, handouts and other staff members' work. Submitting a study or draft of someone else's work as part of your own report without explicitly specifying who did the work (for example, where research was applied by others to a collaborative project).

Chapter 3

Project Proposal

3.1 Proposed Work

Detection of plagiarism is the mechanism for ensuring the authenticity and originality of the documents. This plagiarism detection program offers statistics and/or thorough analysis based on their contents and, optionally, their metadata, on the similarity of input documents. In general, the input documents are pre-processed before making any comparisons to eliminate irrelevant information from them, such as headers and footers, blank pages, small or all images, transform mathematical equations into textual form, etc.

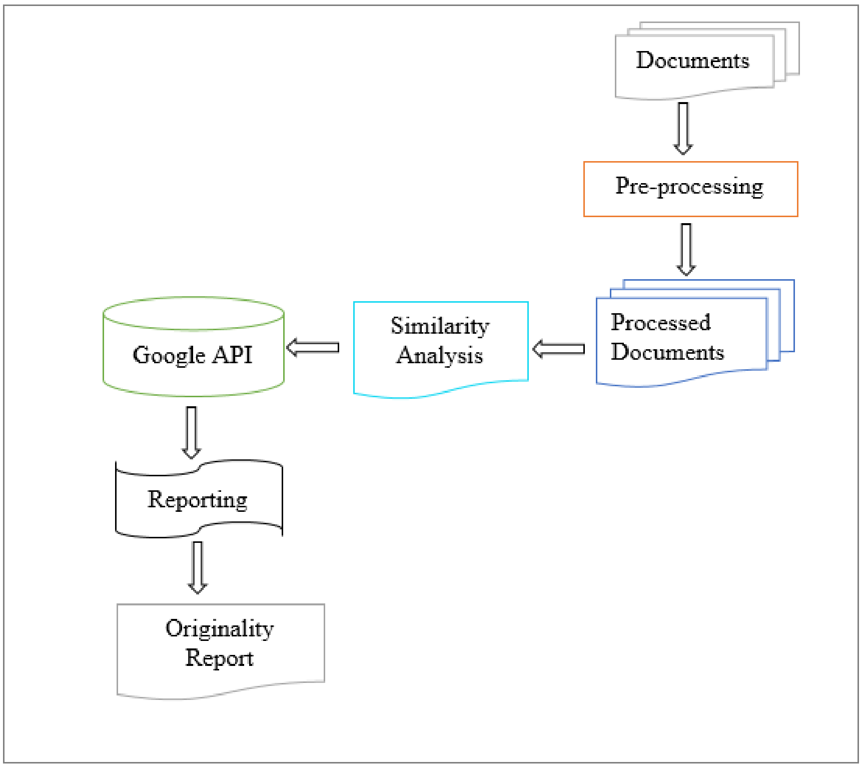


Fig 3.1. Proposed System

Chapter 3

Finally, convert the documents to a standard format appropriate for the Similarity analysis algorithm as input.

The similarity analysis module parses the standard format, breaking the text content into text fragments, and scans Google's APIs for copies of each fragment. The similarity analysis module can use the document's plain text format or alternative encoding, such as n-grams, throughout the quest.

Once the quest for plagiarized text has been completed, a reporting module receives the similitude analysis module output and summarizes the results, including information on the originality level, top plagiarized sources and more.

3.2 Proposed Methodology

It’s a system used to test a material if it has plagiarism or not, this material could be scientific article or technical report or essay or others, also the system can emphasize the parts of plagiarism in the material and estate from where it’s copied with the source link. Following are the steps involved in plagiarism detection:

1. The first phase of document preprocessing happens only once per document.

2. 2. The irrelevant information is extracted from the document and transforms the input documents into a common format.

3. Then the processed text is converted into the clusters depending upon the document size and number of sentences.

4. In the next step the clusters are used to check the similarity of content with in the dataset by using the Google Api keys.

5. The Crawler is used to extract the links which are used by the user and to compare the

Chapter 3

clusters with the data to find the plagiarized text and the original text.

6. In the final step the originality report is generated in which the plagiarized text is highlight with the source link.

Current Methodology

- Clustering is an unsupervised machine learning method, and a powerful data mining technique, it is the process of grouping similar objects together.
- This technique can theoretically speed up the information retrieval process by a factor of K where K is the number of clusters.
- This may be achieved by clustering similar paragraphs together and measure the similarity between each new query and the centroids of the clusters.
- Then measure the similarity between the query and paragraphs in one cluster, this is much faster than measuring similarity of query against each paragraph in the data set.
- The n-gram processing algorithm includes the pre-processing module which is key word extraction, search and streaming process.

3.3 Details of Hardware/Software Requirement

- **HARDWARE REQUIRMENT:**
  Processor: Intel Core i5,
  Memory: total capacity of 2 GB RAM,
  Storage: 400 MB hard disk space

- **SOFTWARE REQUIRMENT:**
  Operating System: Windows 7/8/10,
  Web Browser: Google Chrome,
  Internet : Minimum 2 Mbps

Chapter 3

- **TECHNOLOGIES USED:**

  Java version "1.8.0_221"

  Eclipse IDE

  Server: Tomcat v8.5

  Front-End: HTML, CSS, Javascript

Chapter 4

Planning and Formulation

4.1 Schedule for Project and Gantt Chart

In this we show a tentative timeline required for the development of this project. The start of this project is from July 2019 when the guides were assigned to the team. The project is scheduled to be complete in March 2020.The Gantt Chart below show the different phases and the tentative time allotted to each phase.

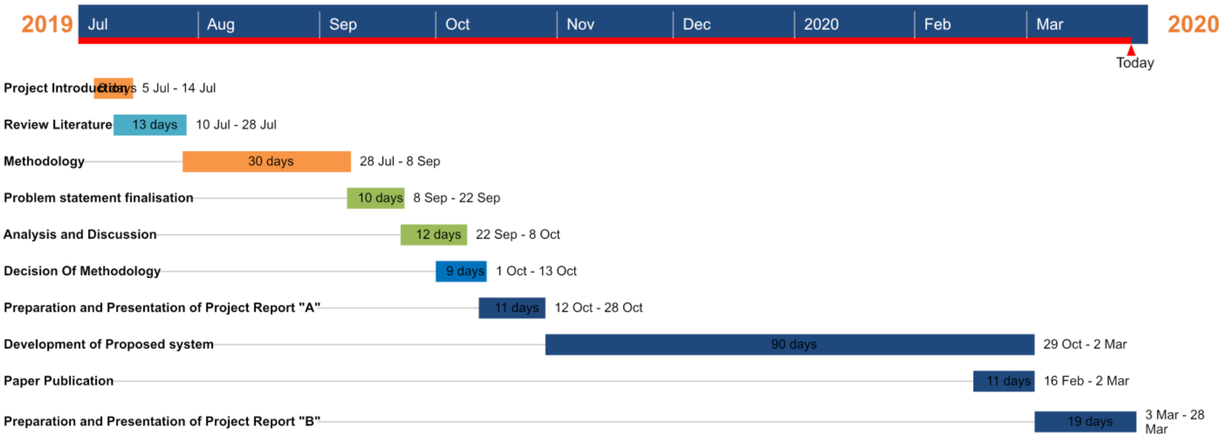


Fig 4.1: TimeLine of The proposed system

Chapter 5

Design of System

5.1 Architecture of Proposed Plagiarism detection

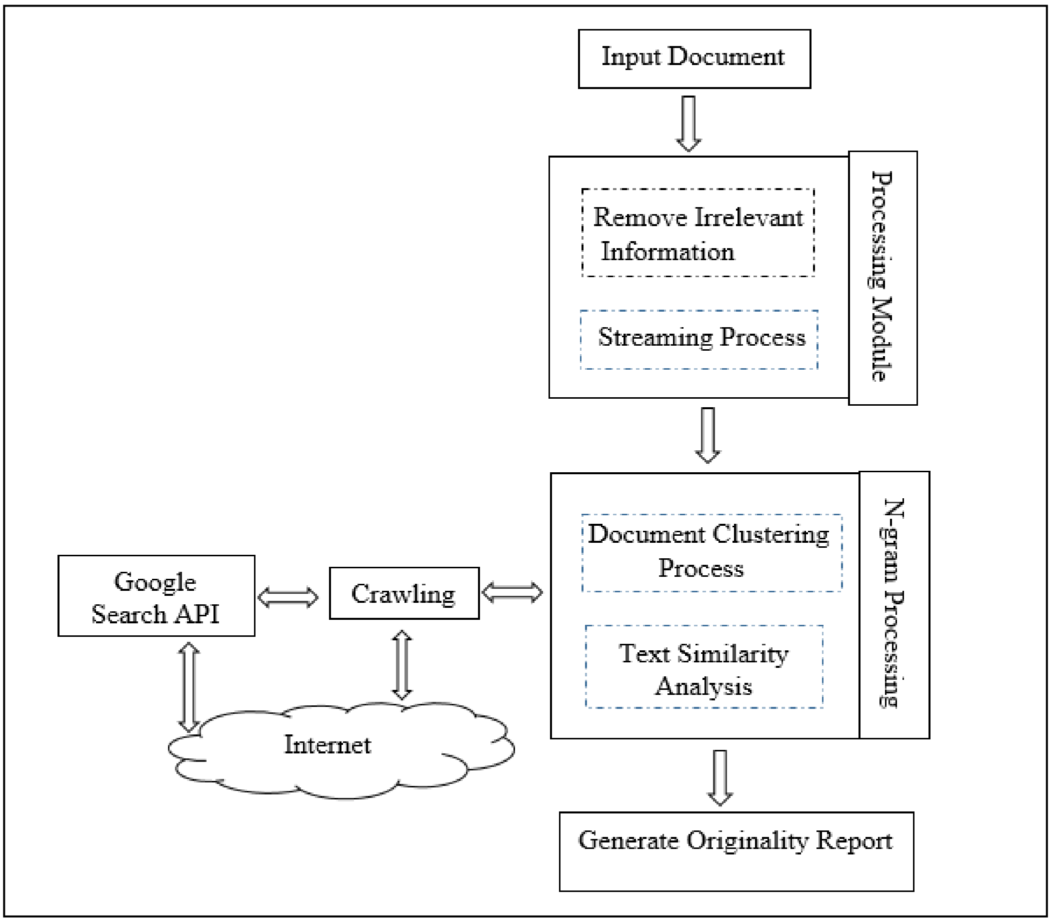


Fig 5.1 Architecture of Plagiarism Detection

Chapter 5

The input documents are pre-processed to delete unnecessary information from them, such as headers and footers, blank pages, small or all images, transform mathematical equations into textual form, etc., and finally convert the documents into a standard format that is

suitable as input for the algorithm of similarity analytics.

The similarity analysis module parses the common format, splitting text content into text fragments, and searches Google's APIs for copies of each fragment.

The similarity analysis module may use the document's plain text format or alternative encoding, such as n-grams, during search.

Once the search for plagiarized text is completed, a monitoring module collects the output of the similarity analysis module (i.e. the list of non-original fragments and their source information) and analyses the findings, including data on the degree of originality, top plagiarized sources and more.

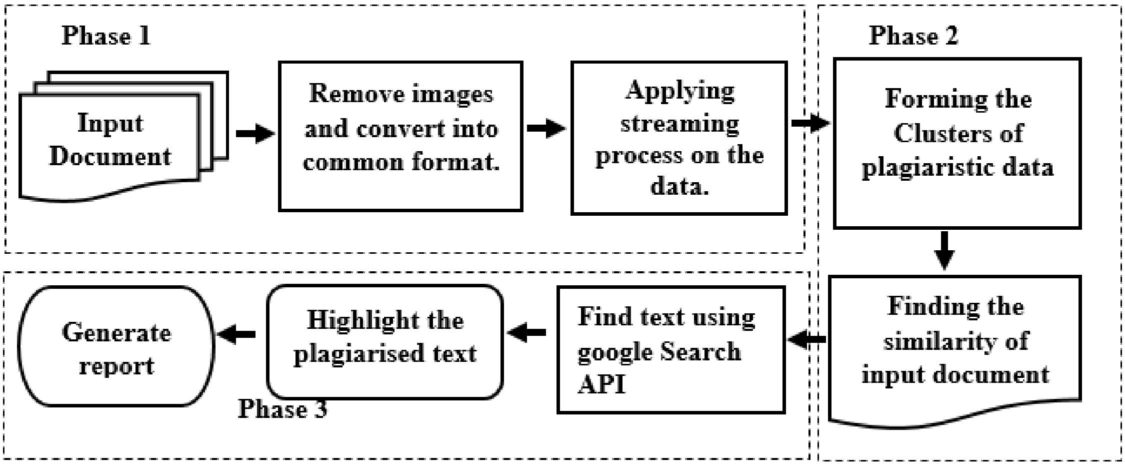

Phase 1
Input Document
Remove images and convert into common format.
Applying streaming process on the data.
Phase 2
Forming the Clusters of plagiaristic data
Finding the similarity of input document
Find text using google Search API
Highlight the plagiarised text
Generate report
Phase 3

Fig 5.2 Flow Diagram for Plagiarism Detection

In Phase 1: We first upload the document. The document is further processed and the irrelevant data from the document like images, screenshots, header, footer are removed from the document. After this process the streaming is applied on the document. Streaming is the

Chapter 5

process to detect the continuous data streams and find the similar data and keywords in the given text.

In Phase 2: Clusters of the document is created based on the length of the document and this cluster is used for checking the similarity of the text with the dataset.

In Phase 3: The similarity of clusters is checked with the help of Google API and the referred links are detected using crawler and finally the originality report is generated.

Chapter 6

Results and Discussion

6.1 Implementation Details

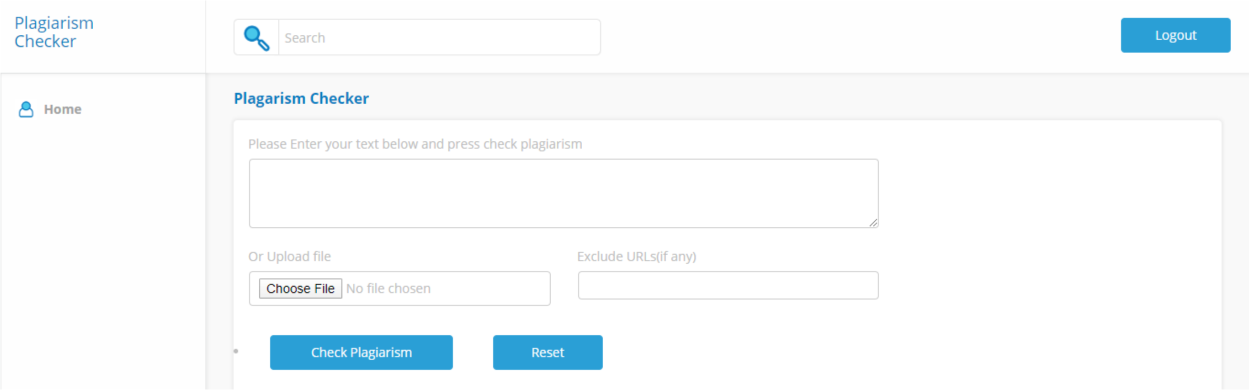
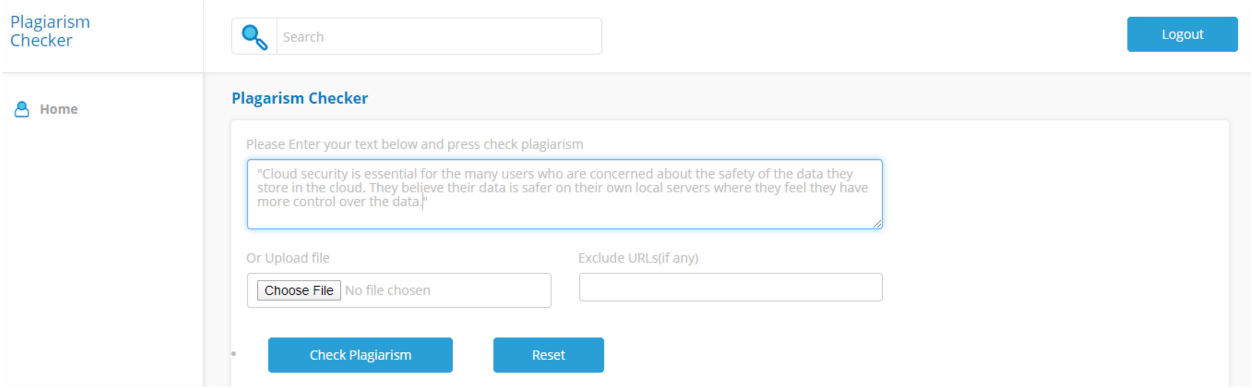


Figure 6.1: GUI: Home Screen

Fig 6.1 figure shows the home screen of proposed system, the interface include the input box to enter the text input. System also provides the other option that is give input through text file for that we have to click on choose file option.

Chapter 6



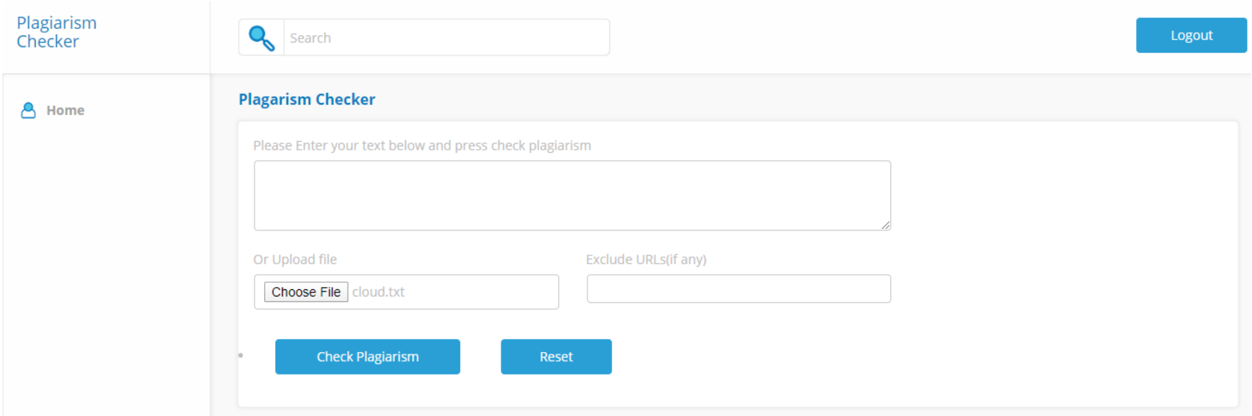
Please Enter your text below and press check plagiarism

"Cloud security is essential for the many users who are concerned about the safety of the data they store in the cloud. They believe their data is safer on their own local servers where they feel they have more control over the data."

Choose File No file chosen

Figure 6.2: GUI: Submitting an input in box

Fig 6.2 shows the interface of the system where the user will input the text to be tested and after choosing an check plagiarism option, results is generated.



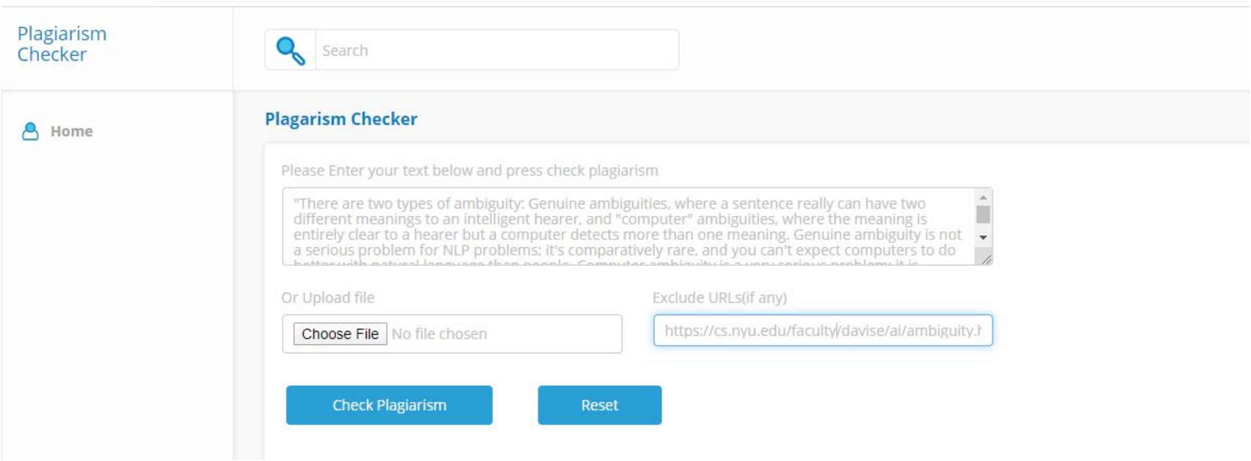
Please Enter your text below and press check plagiarism

Choose File cloud.txt

Figure 6.3: GUI: Submitting an input through text file

Chapter 6

Fig 6.3 shows the interface where user provides the input through the file by uploading the text file.

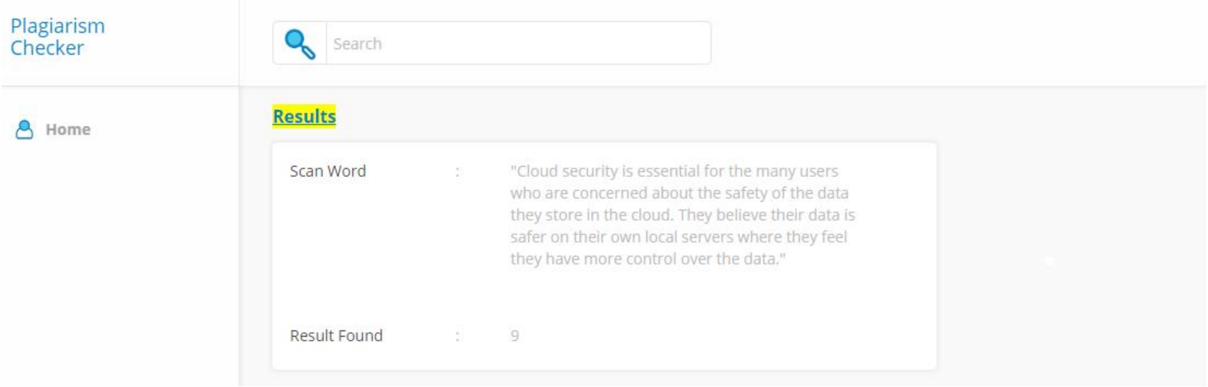


"There are two types of ambiguity: Genuine ambiguities, where a sentence really can have two different meanings to an intelligent hearer, and "computer" ambiguities, where the meaning is entirely clear to a hearer but a computer detects more than one meaning. Genuine ambiguity is not a serious problem for NLP problems; it's comparatively rare, and you can't expect computers to do better with natural language than people. Computer ambiguity is a very serious problem; it is

https://cs.nyu.edu/faculty/davise/ai/ambiguity.h

Figure 6.4: GUI: Submitting an input and exclude URL

Fig 6.4 shows the interface where user submit the input for plagiarism checking in text box and also applies the exclude URL function by providing URL which user want to not analyze while plagiarism checking.



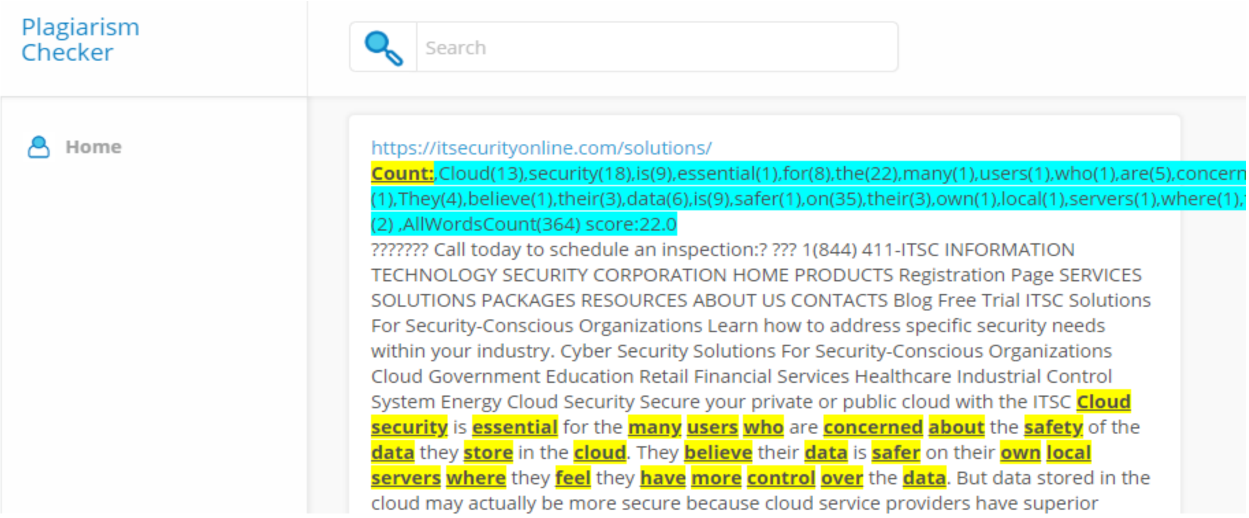
Scan Word : "Cloud security is essential for the many users who are concerned about the safety of the data they store in the cloud. They believe their data is safer on their own local servers where they feel they have more control over the data."

Result Found : 9

Figure 6.5: GUI: The Result of the Process Part 1

Chapter 6

The result appears as shown in fig 6.5 gives that 9 source of results are found for submitted input.



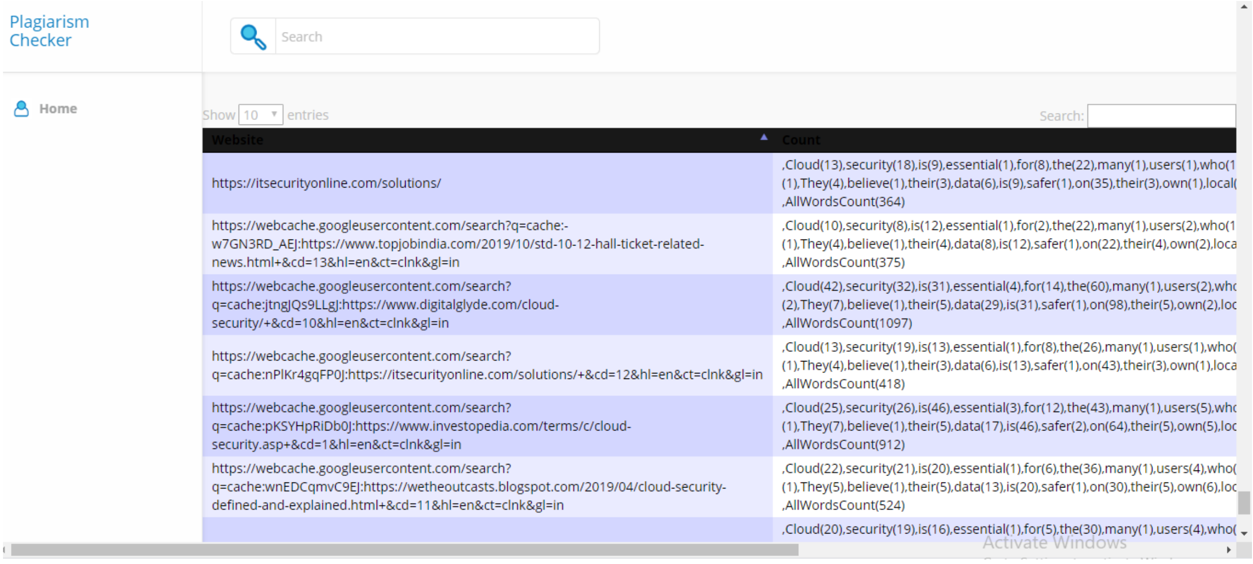
https://itsecurityonline.com/solutions/
Count:,Cloud(13),security(18),is(9),essential(1),for(8),the(22),many(1),users(1),who(1),are(5),concern
(1),They(4),believe(1),their(3),data(6),is(9),safer(1),on(35),their(3),own(1),local(1),servers(1),where(1),
(2) ,AllWordsCount(364) score:22.0
??????? Call today to schedule an inspection:? ??? 1(844) 411-ITSC INFORMATION TECHNOLOGY SECURITY CORPORATION HOME PRODUCTS Registration Page SERVICES SOLUTIONS PACKAGES RESOURCES ABOUT US CONTACTS Blog Free Trial ITSC Solutions For Security-Conscious Organizations Learn how to address specific security needs within your industry. Cyber Security Solutions For Security-Conscious Organizations Cloud Government Education Retail Financial Services Healthcare Industrial Control System Energy Cloud Security Secure your private or public cloud with the ITSC **Cloud** **security** is **essential** for the **many** **users** **who** are **concerned** **about** the **safety** of the **data** they **store** in the **cloud**. They **believe** their **data** is **safer** on their **own** **local** **servers** **where** they **feel** they **have** **more** **control** **over** the **data**. But data stored in the cloud may actually be more secure because cloud service providers have superior

Figure 6.6: GUI: The Result of the Process Part 2

The result appears as shown in fig 6.6:

1. The source from where input is plagiarized.
2. The Yellow highlighted text is the plagiarized text in the input the text.
3. The Aqua highlighted text shows the count of words of input text present in the source.

Chapter 6



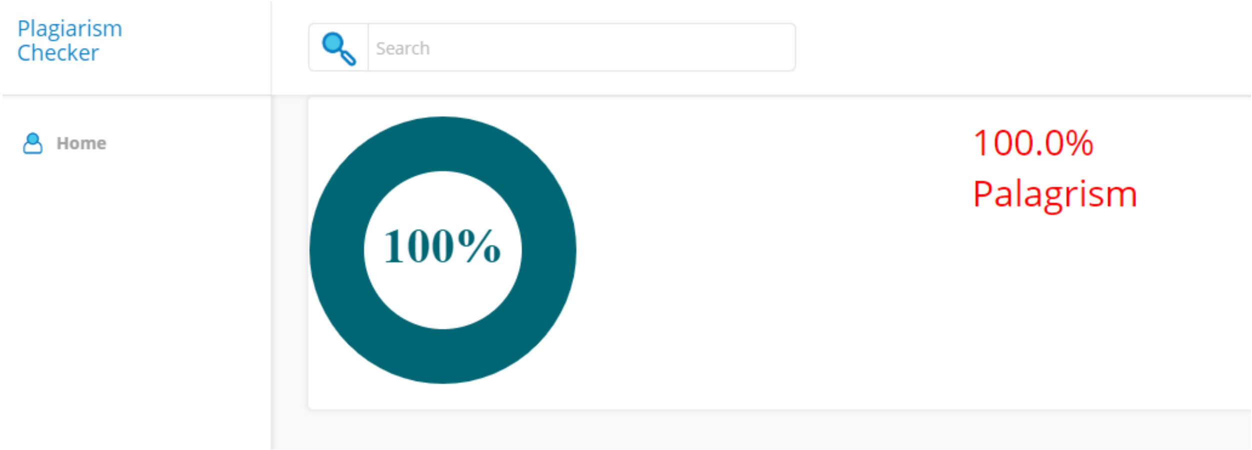
Website	Count
https://itsecurityonline.com/solutions/	,Cloud(13),security(18),is(9),essential(1),for(8),the(22),many(1),users(1),who(1 (1),They(4),believe(1),their(3),data(6),is(9),safer(1),on(35),their(3),own(1),local(,AllWordsCount(364)
https://webcache.googleusercontent.com/search?q=cache:-w7GN3RD_AEJ:https://www.topjobindia.com/2019/10/std-10-12-hall-ticket-related-news.html+&cd=13&hl=en&ct=clnk&gl=in	,Cloud(10),security(8),is(12),essential(1),for(2),the(22),many(1),users(2),who(1 (1),They(4),believe(1),their(4),data(8),is(12),safer(1),on(22),their(4),own(2),loca ,AllWordsCount(375)
https://webcache.googleusercontent.com/search?q=cache:jtngJQs9LLgJ:https://www.digitalglyde.com/cloud-security/+&cd=10&hl=en&ct=clnk&gl=in	,Cloud(42),security(32),is(31),essential(4),for(14),the(60),many(1),users(2),whc (2),They(7),believe(1),their(5),data(29),is(31),safer(1),on(98),their(5),own(2),loc ,AllWordsCount(1097)
https://webcache.googleusercontent.com/search?q=cache:nPlKr4gqFP0J:https://itsecurityonline.com/solutions/+&cd=12&hl=en&ct=clnk&gl=in	,Cloud(13),security(19),is(13),essential(1),for(8),the(26),many(1),users(1),who((1),They(4),believe(1),their(3),data(6),is(13),safer(1),on(43),their(3),own(1),loca ,AllWordsCount(418)
https://webcache.googleusercontent.com/search?q=cache:pKSYHpRiDb0J:https://www.investopedia.com/terms/c/cloud-security.asp+&cd=1&hl=en&ct=clnk&gl=in	,Cloud(25),security(26),is(46),essential(3),for(12),the(43),many(1),users(5),whc (1),They(7),believe(1),their(5),data(17),is(46),safer(2),on(64),their(5),own(5),loc ,AllWordsCount(912)
https://webcache.googleusercontent.com/search?q=cache:wnEDCqmvC9EJ:https://wetheoutcasts.blogspot.com/2019/04/cloud-security-defined-and-explained.html+&cd=11&hl=en&ct=clnk&gl=in	,Cloud(22),security(21),is(20),essential(1),for(6),the(36),many(1),users(4),who((1),They(5),believe(1),their(5),data(13),is(20),safer(1),on(30),their(5),own(6),loc ,AllWordsCount(524)
	,Cloud(20),security(19),is(16),essential(1),for(5),the(30),many(1),users(4),who(

Figure 6.7: GUI: The Result of the Process Part 3

The result appears as shown in fig 6.7:

1. Column one shows, the source from where input is plagiarized.
2. Second column gives, the count of words of input text present in the source.

Chapter 6



100.0% Palagrism

Figure 6.8: GUI: The Result of the Process Part 4

Fig 6.8 shows the result in graph format that is input text is 100% plagiarized

6.2 Result Analysis

Proposed Program checked five samples of input and compared the findings with current commercial plagiarism detection tool such as "duplichecker", " smallseotools " and quetext. Figures 6.9, 6.10 and 6.11 represent links found in Proposed Framework for each input sample Sample 1, Sample 2, Sample 3, Sample 4, Sample 5, and smallseotools, duplichecker, quetext, respectively. The proposed program shows more plagiarized connections than the smallseotools, duplichecker, quetext. The proposed system showed better performance than the current system in proper detection.

Chapter 6

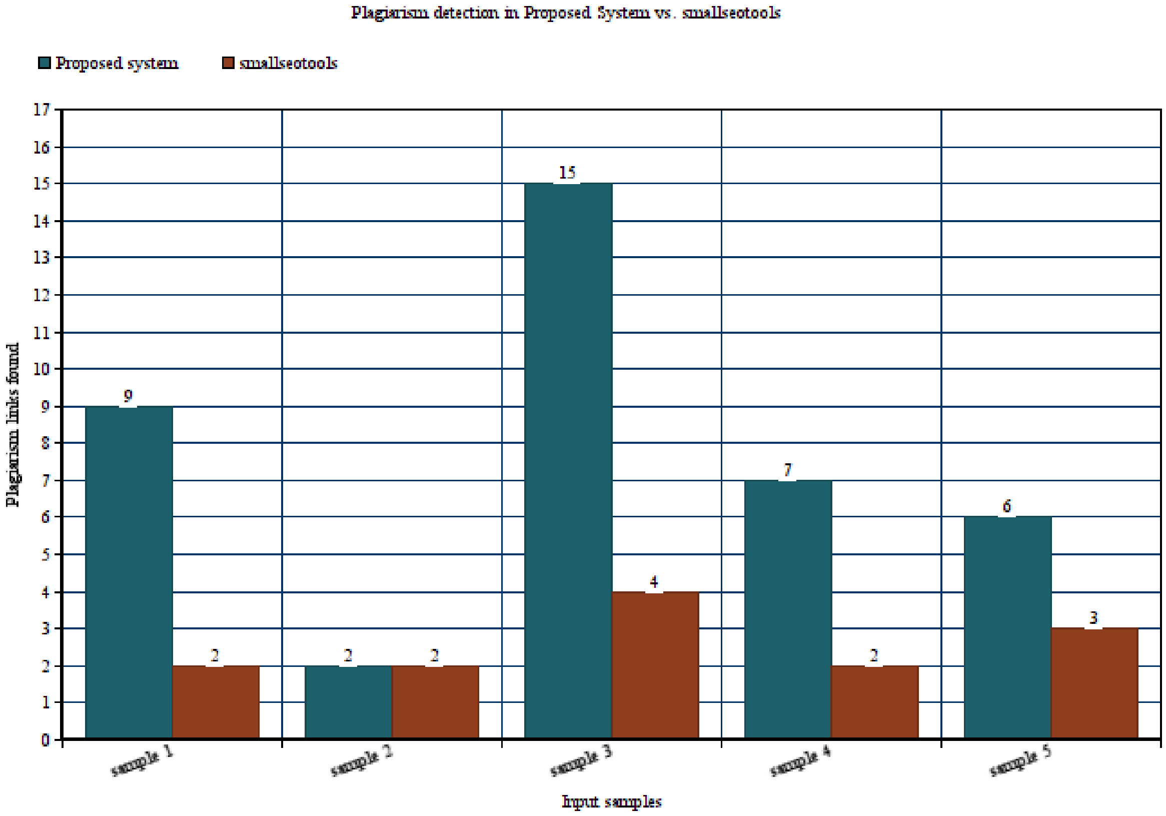


Fig 6.9: Plagiarism detection in Proposed System vs. smallseotools

Fig 6.9 shows the comparisons of proposed system with current plagiarism detection tool “smallseotools”, which shows that for sample 1 proposed system gives 9 links for plagiarized text whereas in smallseotools gives 2 links. For sample 2 both existing system and proposed system gives 2 links. Sample 3 is extremely plagiarized in the proposed system as 15 plagiarized links found compared to smallseotools that show just 4 plagiarized links.

Chapter 6

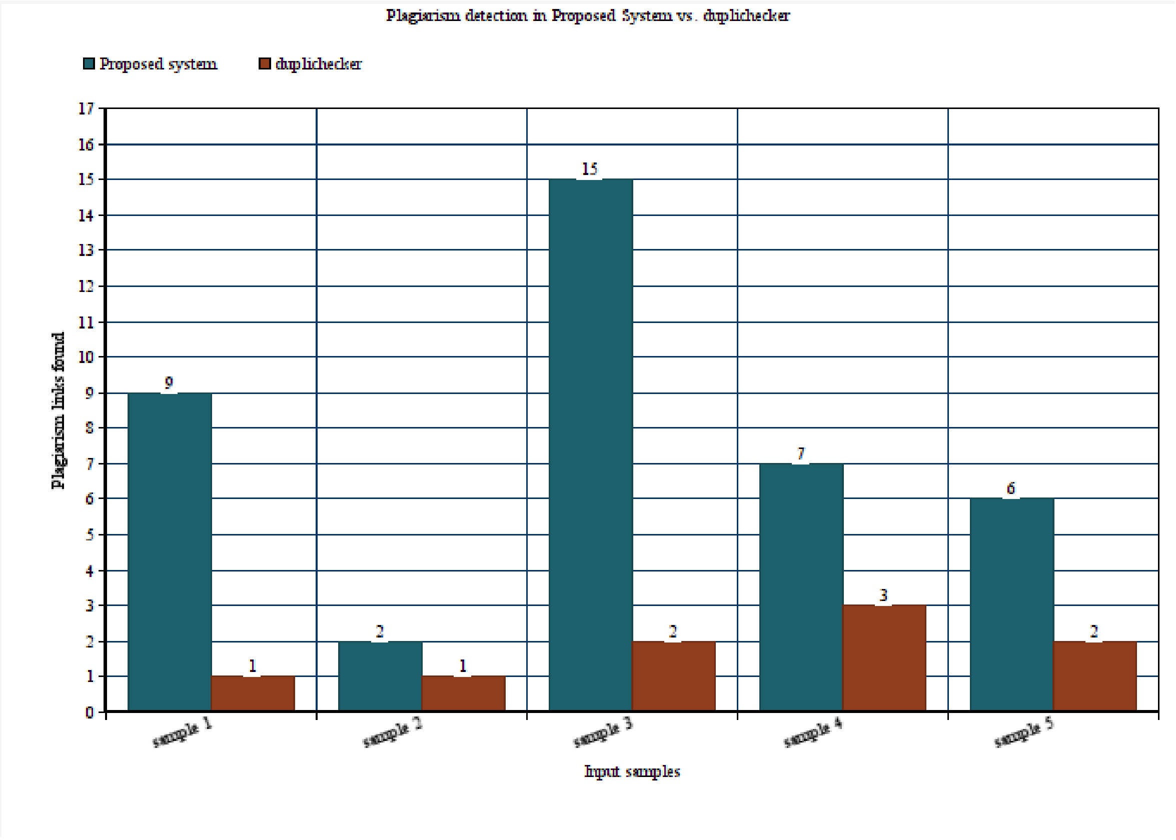


Fig 6.10: Plagiarism detection in Proposed System vs. duplichecker

Fig 6.10 shows the comparisons of proposed system with current plagiarism detection tool “duplichecker”, which shows that for sample 2 proposed system gives 2 links for plagiarized text whereas in duplichecker gives 1 links. For sample 4 existing system gives 3 links proposed system gives 7 links. Sample 1 and sample 3 is extremely plagiarized in the proposed system 9 and 15 plagiarized links found compared to duplichecker that shows just 1 link for sample 1 and 2 links for sample 3.

Chapter 6

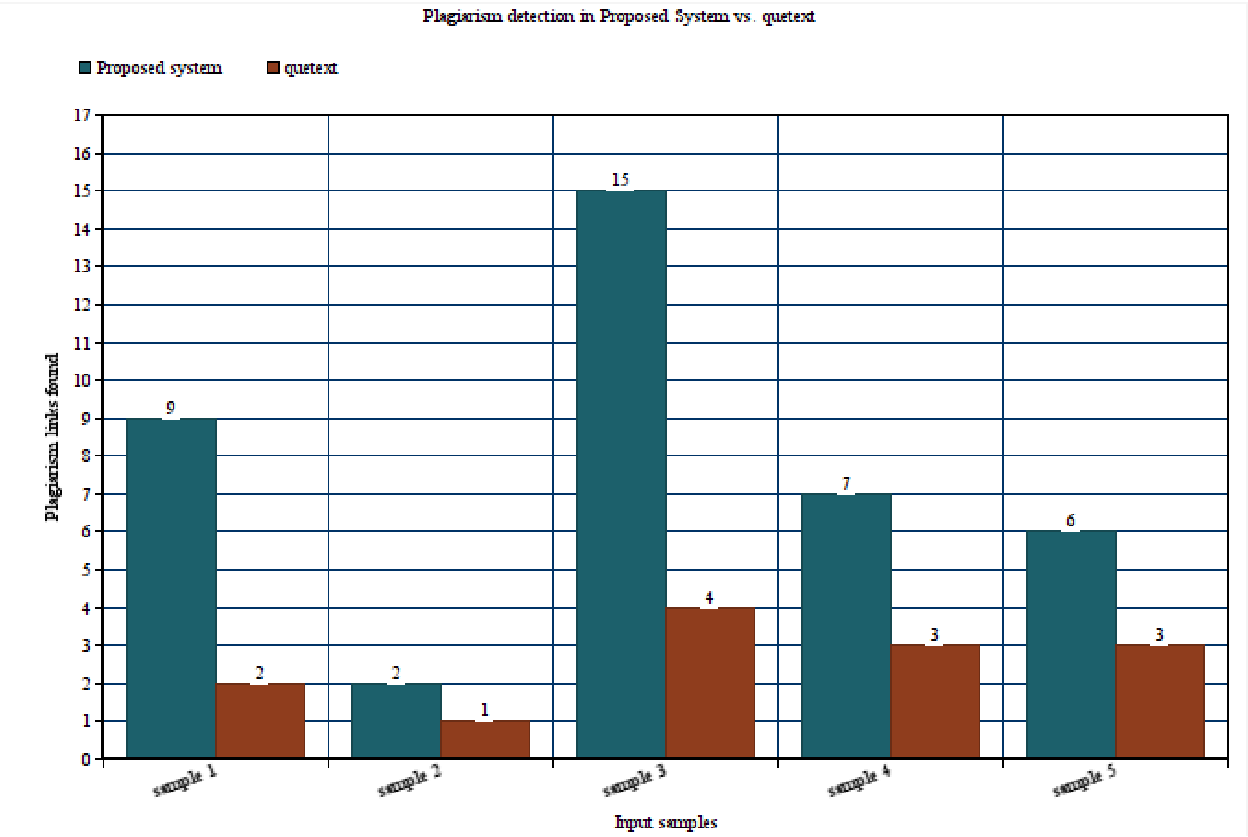


Fig 6.11: Plagiarism detection in Proposed System vs. quetext.

Fig 6.11 shows the comparisons of proposed system with current plagiarism detection tool “quetext”, which shows that for sample 4 and sample 5 quetext gives 2 links for plagiarized text whereas in proposed system gives 7 links for sample 4 and 6 links for sample 5. Sample 1 and sample 3 is extremely plagiarized in the proposed system 9 and 15 plagiarized links found compared to quetext that shows just 2 link for sample 1 and 4 links for sample 3.

Chapter 6

Input Sample	Proposed system	smallseotools	Duplic hecker	quetext
Sample 1	9	2	1	2
Sample 2	2	2	1	1
Sample 3	15	4	2	4
Sample 4	7	2	3	3
Sample 5	6	3	2	3

Table 6.1: Analysis of input samples.

According to Table 6.1, For Sample 1 Proposed system gives 9 plagiarized links whereas smallseotools, duplichecker, quetext gives 2, 1 and 2 links respectively. Sample 3 is extremely plagiarized in the proposed system as 15 plagiarized links found compared to other systems that show just 4,2and 4 plagiarized links. Plagiarized links found for sample 2 is same for both proposed system and smallseotools i.e. 2 and 1 link is found in duplichecker and quetext.

Chapter 7

Conclusion and Future Work

Conclusion

It is an growing attraction among students and an invariable difficulty in dealing with the issue for the professors. The academic community undoubtedly respects the appreciation of the contributions other people bring to knowledge. And so the penalty may be serious for those arrested for plagiarism. The detection mechanism of plagiarism should be automated so that it can be accurate. Data mining techniques can be used to enhance the plagiarism detection process. In proposed work a methodology is given using data mining techniques from which it is assumed the efficiency of the process can be improved .Pre-processing and clustering methods may be used to reduce process overhead. In addition, similarity score can be measured via plagiarized data clusters so as to increase performance.

Chapter 7

Future Work

The identification of plagiarism is at present restricted to text material. Program allows authors, consumers, businesses, and others to search vast databases precisely for duplicate and plagiarized content. In order to overcome limitations some enhancement are to find the amount of time taken to find similar contents, to reduce the computation time and find the relevant content from the list of the contents and improve efficiency. Also detect the self plagiarized data with different techniques.

References

1. MAC Jiffriya, MAC Akmal Jahan, Roshan G Ragel and Sampath Deegalla Antiplag: Plagiarism Detection on Electronic Submissions Of Text Based Assignments, IEEE (2018).

2. Muhammad Usman, Muhammad Waleed Ashraf . Plagiarism Detection Process Using Data Mining Techniques, IEEE (2017).

3. Vani K and Deepa Gupta . Study On Extrinsic Text Plagiarism Detection Techniques And Tools, IEEE January (2016).

4. Kamalpreet Sharma and Balkrishan Jindal. An Improved Plagiarism Detection Approach For Semantic Analysis Using Custom Search Engine, IEEE (2016).

5. Mansi Sahi and Vishal Gupta. Efficiency Comparison of various Plagiarism Detection Techniques, IEEE (2016).

6. Manav Bagai , Vibhanshu , Siddharth Gupta, Rashid Ali. Text Based Plagiarism Detection, International Journal For Technological Research in Volume 3, Issue 8, April-(2016).

7. Liang Zhang, Zhuang Yueting, Yuan Zhen-ming. A model of program detection of plagiarism based on distance information and clustering. (2010).

8. https://www.plagiarismsoftware.net/

9. https://www.duplichecker.com/

10. http://en.writecheck.com/blog/2012/07/26/5- reasons-to -use-a-plagiarism-checker.

11. IEEE 8 th International Conference on Industrial and Information Systems,ICIIS (2018),Sri Lanka.

12. https://www.plagramme.com/useplagiarism- checker-properly

Appendix A

Weekly Progress Report

Appendix B

Paper Publication

Appendix C

Project Competition

Acknowledgement

We take this opportunity to express my profound gratitude and deep regards to my guide **Mrs. Smita Bhoir** for his/her exemplary guidance, monitoring and constant encouragement throughout the completion of this report. We are truly grateful to his/her efforts to improve my understanding towards various concepts and technical skills required in our project. The blessing, help and guidance given by her time to time shall carry us a long way in the journey of life on which we are about to embark.

We take this privilege to express my sincere thanks to **Dr. Mukesh D. Patil, Principal, RAIT** for providing the much necessary facilities. We are also thankful to **Dr. Leena Ragha**, Head of Department of Computer Engineering, Project Co-ordinator **Mrs. Smita Bharne** and Project Co- ordinator **Mrs. Bhavana Alte**, Department of Computer Engineering, RAIT, Nerul Navi Mumbai for their generous support.

Last but not the least we would also like to thank all those who have directly or indirectly helped us in completion of this thesis.

Ms. Ankita Salavi

Ms. Komal Sonawane

Ms. Shweta Tarmale