Automated Plagiarism Detection System

Prince Singh Tomar
IIIT-H
Gachibowli, Hyderabad
prince.tomar@students.iiit.ac.in

***Abstract*—Plagiarism is a serious issue in both academic and professional fields. We propose a solution for building a plagiarism detection software that will take scanned documents, typed documents or coding files and will return the percentage of matching.**

I. Introduction

Plagiarism is a major challenge to instructors because Today many students find it easy to copy toppers assignments rather than putting real effort.The spread of the Internet and online academic journals have greatly contributed to the increase of cases of plagiarism among students [1]. This practice has increased more, since classes are held online and children don't want to put effort while they are at home and the access to smartphones and high speed net is also a major factor.Students use social media to access ans from other students [2]. This is affecting students learning capacity and their morals in a major way.Donald McCabe (Rutgers University) found that 64 percent of students admitted to cheating on a test, 58 percent admitted to plagiarism, and 95 percent admitted to cheating in some way, whether it was on a test, plagiarism, or copying homework, in a survey of 24,000 students from 70 high schools [3]. According to an anonymous survey 72.5% students admitted they plagiarised at least once during their graduate study[1]. There are many reasons why people copy codes like a weak student trying to get higher grades, a strong student lacking motivation, less time given to complete work.These students who didn't worked hard in their college days loose jobs prospects in the future this causes loss of money wasted on college studies. It is not limited to students and office workers but organisation whose job is to provide knowledge are also doing this to keep their cost low and high content. For example on Medium about 30,000 stories are plagiarised daily [4]. In Technology and research sector a millions of dollars are lost due to plagiarisation of ideas ,codes and algorithms.

II. Literature Review

A. Typed Text

Some basic methods that can be thought of are Heuristic alignment algorithm like Smith-Waterman Algorithm for finding local alignments but its space complexity is not good [5]. Today they are many tools to checks plagiarism in plain text and the algorithm they use are some of the best algorithms to find matching in documents.Plagiarism.org a free plagiarism detection site uses Fingerprinting to find plagiarism with online resources.Some paid alternatives of plagiarism.com are IntegriGuard and EVE2. EVE2 is also called as a Essay Verification Engine, It claims to conduct a large number of complex searches to identify plagiarised content in the work submitted[5].SCAM which uses Vector-space model claims to have discovered significant cases of science plagiarism in the DBWORLD area of database analysis literature[6]. Some other plagiarism detectors are CopyCatch moslty used for student reports and uses Lexical matching. MatchDetectReveal is used to match plain text using Suffix Tree matching algorithm in databases [7]. Most popular, user friendly and effective plagiarism detection software in the market is turnitin and 81 of the top 100 universities use this tool for plagiarism checking [8].

B. Softwares

MOSS is most popular web-service with user interface to show areas of code flagged as plagiarised. This helps user to see and check whether the code is plagiarised or not. It is based on winnowing algorithm [9]. Moss provides a wide range of languages on which plagiarism detection can be done and it requires just a single command to run it moreover it can also be used to compare two different languages. Another softwares which is said to be better than moss in performance is JPlag [10] which uses Greedy-String-Tiling to check moss [6]. Bplag,Codequery,sim,sid,CodeMatch are other plagiarism detection softwares. CodeMatch uses Fogerprints to find plagiarism [6].

C. Hand Written

Most of the current softwares like turnitin uses ocr to check plagiarism in hand written documents. Research had been done to use convolutional neural network to find plagiarism by checking words images. Later this can be merged with plagiarism detection algorithms to check for plagiarism in written documents. Best performance for word spotting recorded is 0.80 on IAM dataset [11]. "Many approaches such as [2, 29, 35] demonstrated word spotting using fixed length representation based on local features such as SIFT and HOG in a bag of words (bow)framework" [11].

III. System Architecture

A. Overview

The system architecture uses 3-tier client-server architecture which consists of 3 servers one at client side, one middle-ware which hosts the plagiarising algorithms and third server which

hosts database where results will be stored for later access to them. On the client side an application must be stored which will facilitate the functions needed to perform on sending files to server for plagiarism checking. The middle-ware will hold both business logic and data logic. and the database server will use mongo db to store and retrieve data. Mongo is used as its faster compared to MySQL [12] and easy to use.

B. Client Server

The User interface will be written solely in python since its easy and efficient to convert different formats in python. At the Client Side The user will have to enter login details once, after it is verified the middle-ware from , Users login details will encoded and stored in their personal computer and User login details will be filled automatically before sending files. This is proposed as after their access to the plagiarism checker is over they can no longer access it . These files must be named on the basis of students roll-number along with assignment number and Course code. Client need to Upload all the files they want to be checked using user interface. Client will also have to choose whether the assignment is programming or non-programming one. This will help in handling software plagiarism and other type of plagiarism differently.
Software will handle the difference in submission format to some extent. All the image submission of a student will be merged and converted to pdfs, Other typed submissions like .docx,.md will be converted to .txt format. If file is compressed then it will be extracted and file inside it will be submitted to the server. The name used will be the roll-numbers of the students.
If client has manager level access then he can send requests to the middle-ware for requesting the database server to perform specified actions.

C. Middle-Ware

This server will contain data logic and business logic. The middle-ware will listen till some request from client is made. After receiving request it will match clients email and password to the list stored in database server, For that it will send request to database server and receive list as a response. It will also host the plagiarism algorithm and will return the percentage of matching along with the list of starting and ending position of matching lines in code or submitted text. For Scanned pdfs it will return percentage matching and an empty list. Two requests will be sent one to the client side and other to the database to store the result for further reference. This stored data can be requested from the database server whenever user requests for them. Different algorithms will be used for scanned pdfs and other submissions.

D. Database Server

The database servers holds the database. It receives the requests from middle-ware to perform actions on the database. Database server will be responsible to sending the data contained response to the middle-ware.

E. User Stories

Client entered username and password and logged in plagiarism detection software , entered course code, selected non programming assignment and assignment number and Submitted some files which consists of pdfs, zip file containing images, images files these files are converted into pdf and submission of students with more than one images, these images are merged and converted into pdf and send to server. Client get moss Result in Tabulated Form.

Client submits submission of an assignment coded in different different languages after entering course code and assignment number, client selected programming assignment . After getting result from the server client can see the percentage matching in solutions of different students.

Client submits types submissions but some files were in docs format and some in odt form after submitting. These format were changed in text format and the plagiarism results were shoen to the client.

IV. Conclusion and Future Work

A. Conclusion

This solution decreases the load of TA's and instructors on working with people submitting in wrong formats, Hence corrects some of the majors faults. By storing results in database helps from running plagiarism checks again.

B. Future Work

This solution had some flaws and hence we will try to decrease them in future. We also aim at using the stored database to check whether students are copying from their senior's assignment as Some of the assignments are not changed. Moreover we will try to increase out range from submitted assignments to cover most of the resources found on the internet.

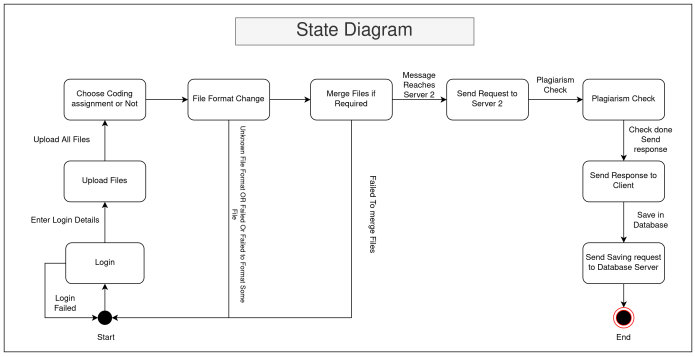

Fig. 1. State Diagram

References

[1] Source code plagiarism by Dejan Sraka, Branko Kaučič
[2] the attitude of students towards plagiarism in online learning: a narrative literature review , researchgate.com, Nwosu Lilian and Joshua Chukwuere
[3] Plagiarism:Facts and stats
[4] Medium Has a Major Plagiarism Problem

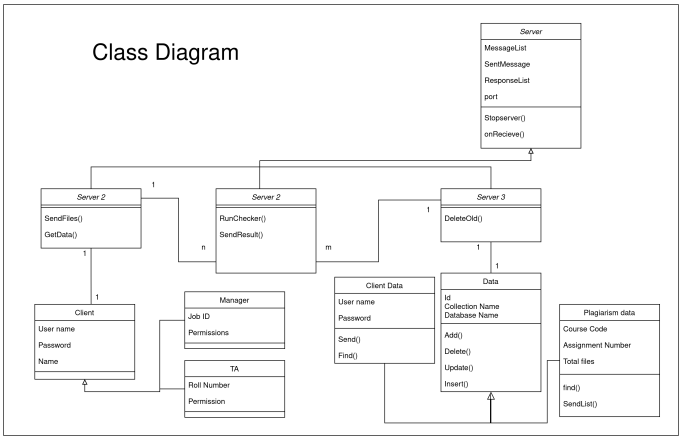

Fig. 2. Class Diagram

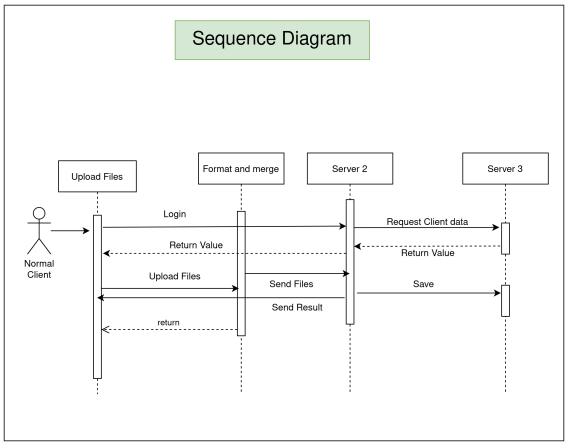

Fig. 3. Sequence Diagram of a Client (TA)

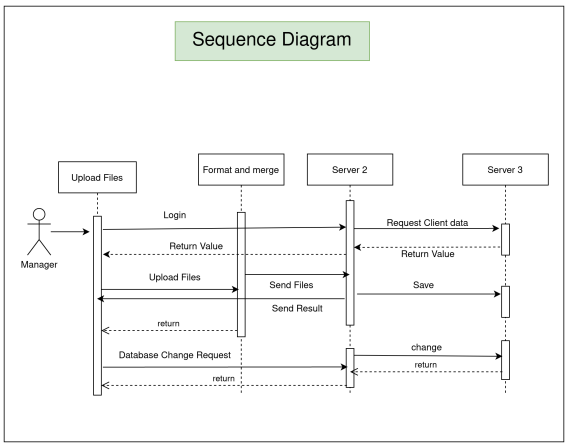

Fig. 4. Sequence diagram of a manager

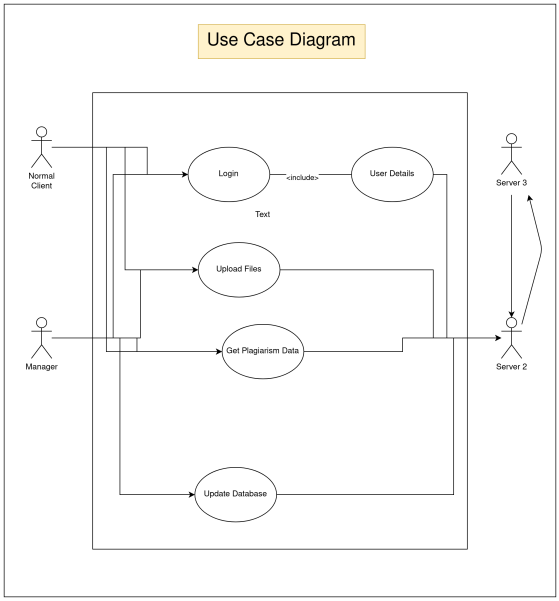

Fig. 5. Use case diagram

[5] Plagiarism Detection with Genetic-Based Parameter Tuning
[6] An Intelligent System for Detecting Source Code Plagiarism Using a Probabilistic Graph Model.
[7] MatchDetectReveal: finding overlapping and similar digital documents.
[8] https://www.turnitin.com/products/similarity
[9] MOSS papers
[10] Academic Source Code Plagiarism Detection byMeasuring Program Behavioural Similarity
[11] Matching Handwritten Document Images
[12] MySQL vs MongoDB