## **Using Hashing Algorithms to Detect Academic Plagiarism in Student Submissions**

### **I. Introduction/Background**

Due to the rise of various online resources, students tend to copy and paste data from online sources or any other source available for that matter right into their assignments. Which makes it harder for educators who have to ensure the integrity of their robocopters. Manual checking and keyword searches have been the traditional methods of detection but are time-consuming and usually inefficient. Contemporary plagiarised text detection tools utilise hashing algorithms to calculate fingerprints which are subsequently compared against the vast data repositories of previously-fed texts to seek signs of overlap.

Firstly, plagiarism in academic submissions is an ever-increasing problem as well due to many students having access to the internet in a digital education environment and being able to find most content that answers their questions. There are plagiarism checkers, but most need only to do basic string matching and a simple modification of the text or formatting can easily bypass these checks. However, more sophisticated approaches would utilise hashing and cryptography to provide a much stronger solution by being able to detect micro-changes in text, as well as finding approximate matches more quickly and surely.

### **II. Project Scope**

In this project we will learn about usage of hashing algorithms in academic plagiarism detection. In particular, it will describe the use of SHA-256 and MD5 to create a single hash value (or fingerprint) for each submission. It will also look into the rolling hash functions applications when identifying similar/paraphrased text, and how these techniques can be used for comparing student submissions to a database of known sources.

**(What Will Be Covered)**

* An Overview of Hash Functions ( SHA-256, MD5)
* Using rolling hashes and comparing the hash to determine if very close duplicates or paraphrased content are present.
* Build a basic plagiarism check that compares student submissions against a database of documents.
* A simplistic system to hash student submissions and documents then match both hashes or compare the similarity of documents.
* Advantages and Disadvantages of Hashing Algorithm for Plagiarism Detection

**(What Will Not Be Included)**

* Commercial-grade plagiarism detection systems like Turnitin or Copyscape development.
* It consists of a background study on the machine learning-based plagiarism detection approaches (e.g., Deep learning, natural language processing) with high-level details

### **III. Benefits Forecasted**

In this project we will explore a concept in computer science, hashing algorithms and how they can be used to solve a problem in the real world–specifically by detecting plagiarism at scale. The project will demonstrate how, using efficient hashing techniques, the detection of plagiarism can be automated to make it convenient for educators to catch those who are attempting academic dishonesty rather than having to manually check each submission.

**Who Will Benefit and What They Will Gain**

* Students: To learn more about how plagiarism detection works and why academic integrity matters.
* For Teachers and Educators: Utilising hashing algorithms allows for faster detection with less computational resources in cases of large classes or plagiarised submissions placed on online education platforms.
* Educational Institutions: Develop more trustworthy and automated approaches to ensure academic honesty in crowds of student submissions.
* Becoming a Computer Science Researcher: Know how we can find the use-cases of algorithms and data-structure like hashing to solve real-world digital education problems, etc.

### **IV. Project Plan**

**Tasks to Accomplish the Project:**

1. **Research Phase**

* Study the basic concepts of hashing algorithms (SHA-256, MD5) and how they are being used for data integrity and security
* Investigate plagiarism detection techniques especially hash-based for exact or near-exact matches.
* Learn how rolling hash functions could be used to identify slight changes in text or paraphrasing.

1. **Development Phase**

* Build a simple plagiarism detection system where student submissions as well as a bank of reference documents are hash using algorithms like SHA-256.
* Check similarity of text and find small changes in plagiarism using rolling hashes
* Create a basic comparison application that runs the hashes of student submissions against those of custom content to check for potential plagiarism.
* Create an interface to allow teachers to upload students' work and documents to compare it against to automate plagiarism detectors.

1. **Testing Phase**

* Use a sample of original and plagiarised copies written by students to evaluate the accuracy of plagiarism detection.
* Performance and detection rates of plagiarism using different hashing algorithms (SHA-256 vs MD5)
* Check out what rolling hashes can do or not on subtle plagiarised content

1. **Documentation Phase**

* Prepare a report on the theory of hashing algorithms, designing plagiarism detection systems and Testing.
* Elaborate on the ability of hashing algorithms to integrate and involve inside plagiarism detection tools in a broader context, as well as the limits of simply applying hashes when it comes to less naive types of plagiarising (e.g., deep paraphrasing).

**Final Output**

The project will help in building a faster plagiarism detection tool for educators which will be scalable and easier to use, using fundamental computer science concepts such as hashing and rolling hash functions, this can provide a solution to real-world problems that the education sector has been facing. Automating the process of plagiarism detection saves teachers time and helps enforce academic integrity in the classroom. A working plagiarism detection tool for extracting hash algorithms (SHA-256, MD5) and rolling hashes in detecting academic dishonesty in student submissions. An article that details the mathematics behind hashing, how it was developed, findings from tests conducted on the tool and what this means for academic integrity within digital education.