CS 254
Design and analysys of algorithm

Plagiarism Detection

This project is intended to implement the existing algorithm for plagiarism detection in source code and also improving the existing algorithm.

Rahul kumar-160001047
Harsh vardhan-160001025

Plagiarism Detection

code duplication is a commmon probem in software system especially for the large ones.
Every code should be checked for plagiarism in it .But simple string matching would not
work. Sometimes simple string matching can be fooled by simply changing the variable names and other parameters i.e. string matching would find only exact copy of the code but we need to include those codes also which are changed purposely by the programmer.
We need to include the changed version of the code in plagiarism detection.
For this, we would be implementing existing algorithm for "Abstract syntax tree" and try to improve the existing one .

Abstract Syntax Tree:-

Abstract syntax tree is a tree representation of of abstract syntactic structure of source code written in a programming language.Each node of the tree denotes a construct occuring in the source code.

```
x := a + b;
y := a * b;
while (y > a) {
    a := a + 1;
    x := a + b
}
```
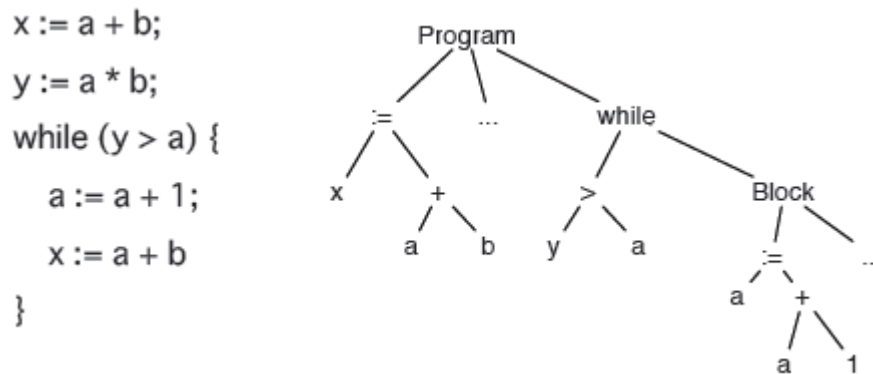

Fig. Example of a abstract syntax tree

we can also use the google api to check if the program is copied from internet.


```
int function (int parameter) {
    int k = 10;

    printf("Hello, this is the function");

    int i = 0;
    while (i < 7) {
        i++;
        // do something cool
    }
}
```

AST embeds both **Syntactic** and **Lexical** Information

• Program Instructions

• Name of Variables, Literals...