Lab One

Alex Smith
alex.smith1@Marist.edu

January 28, 2019

1 Crafting a Compiler

1.1 Exercise 1.11

The Measure Of Software Similarity (MOSS) [SWA03] tool can detect similarity of programs written in a variety of modern programming languages. Its main application has been in detecting similarity of programs submitted in computer science classes, where such similarity may indicate plagiarism (students, beware!). In theory, detecting equivalence of two programs is undecidable, but MOSS does a very good job of finding similarity in spite of that limitation. Investigate the techniques MOSS uses to find similarity. How does MOSS differ from other approaches for detecting possible plagiarism?

In order to detect similarity between documents, MOSS divides the document into k-grams, which are contiguous substrings of length k, where k is chosen by the user. Then each k-gram is hashed and a subset of these hashes are retained to be the fingerprints of the document. The way MOSS differs from other approaches is the implementation of Winnowing. Winnowing is an algorithm that defines a window of size w consecutive hashes of k-grams and ensures that at least one fingerprint is selected from each window. MOSS also compares a list of matching fingerprints to other documents and reports which documents have the greatest number of matches [1]. By ensuring that fingerprints are selected from differing groups, there is a greater chance MOSS will detect similarity between documents.

[1] Schleimer

1.2 Exercise 3.1

Assume the following text is presented to a C scanner:

```
main(){
    const float payment = 384.00;
    float bal;
    int month = 0;
    bal=15000;
    while (bal>0){
        printf("Month: %2d Balance: %10.2f\n", month, bal);
        bal=bal-payment+0.015*bal;
        month=month+1;
    }
}
```

What token sequence is produced? For which tokens must extra information be returned in addition to the token code?

The token sequence that is produced is (separated by line for ease of reading):

Identifier Token, Left Parenthesis Token, Right Parenthesis Token, Left Bracket Token

Const Token, Float Token, Identifier Token, Assignment Token, Decimal Token, Semicolon Token

Float Token, Identifier Token, Semicolon Token

Integer Token, Identifier Token, Assignment Token, Number Token, Semicolon Token

Identifier Token, Assignment Token, Number Token, Semicolon Token

While Token, Left Parenthesis Token, Identifier Token, Greater Than Token, Number Token, Right Parenthesis Token, Left Bracket Token

Print Token, Left Parenthesis Token, Quotation Token, Identifier Token, Colon Token, Percent Token, Number Token, Char Token, Identifier Token, Colon Token, Percent Token, Decimal Token, Char Token, Quotation Token, Comma Token, Identifier Token, Comma Token, Identifier Token, Right Parenthesis Token, Semicolon Token

Identifier Token, Assignment Token, Identifier Token, Subtraction Token, Identifier Token, Addition Token, Decimal Token, Multiplication Token, Identifier Token, Semicolon Token

Identifier Token, Assignment Token, Identifier Token, Addition Token, Number Token, Semicolon Token

Right Bracket Token

Right Bracket Token

The Identifier Tokens, Number Tokens, and Decimal Tokens need additional information such as the value (for the Number and Decimal Tokens) and the name (for the Identifier Tokens). Other useful information that can be returned for all token types include line number and index in order to provide more in depth error messages.

2 The Dragon Book

2.1 Exercise 1.1.4

A compiler that translates a high-level language into another high-level language is called a source-to-source translator. What advantages are there to using C as a target language for a compiler?

The advantage of using C as a target language for a compiler is that certain programs or applications may not be compatible with other languages but is compatible with C. Translating directly to C can allow the user to easily link the newly created C file with other existing C files as well. Also, if the language is new or has a small user base and does not have a compiler, translating to C will allow the use of a C compiler.

2.2 Exercise 1.6.1

For the block-structured C code of Fig. 1.13(a), indicate the values assigned to w, x, y, and z.

```
int w, x, y, z;
int i = 4; int j = 5;

{ int j = 7;
i = 6;
w = i + j;
}

x = i + j;

{ int i = 8;
y = i + j;
}

z = i + j;
```

w = 13 (since it is using i = 6 and j = 7)
x = 11 (since it is using i = 6 and j = 5)
y = 13 (since it is using i = 8 and j = 5)
z = 11 (since it is using i = 6 and j = 5)

References

[1] S. Schleimer, D. S. Wilkerson, and A. Aiken. *Winnowing.* Proceedings of the 2003 ACM SIGMOD international conference on on Management of data - SIGMOD 03, 2003.