Tichafa Andrew Rinomhota,

This is my report on the lexical analyzer for COVID-19 case surveillance data which uses Flex. The analyzer recognizes valid tokens from an input dataset and reports lexical errors when and if they are there.

The two regular expressions from my exp-rules.l file I chose to explain are Date and Gender



This rule recognizes dates that are in the format YYYY/MM/DD. It matches four digits for the year and then a forward slash and then two digits for the month then two digits for the day. This will return the date token.

A black rectangle with white text

AI-generated content may be incorrect.

These rules take the words "Male," "Female," and "Other" and it takes into consideration if the first letter is uppercase or lowercase. Once a match is found, the correct gender token is returned.

How Flex Works

Flex is a tool in software engineering that generates a lexical analyzer using tokens which are focused on user-defined rules. It takes in an input file containing a list of regular expressions in C/C++. When executed, flex creates a C source file named lex.yy.c, which contains the yylex() function. This function scans an input stream, and matches lexemes to their specified regular expressions, and performs actions. Looking at the lex.yy.c which is really long, flex uses finite state machines. Flex, fits into the software development workflow as a preprocessing tool for compilers and interpreters. It converts a sequence of characters into tokens and allows subsequent phases. I feel like for me Flex was easy and simple to learn after looking at its documentation. Although simple I believe Flex can be very powerful tool to analyze data and makes the task easier.

References

https://www.geeksforgeeks.org/flex-fast-lexical-analyzer-generator/

https://web.stanford.edu/class/archive/cs/cs143/cs143.1128/handouts/050%20Flex%20In%20A%20Nutshell.pdf