EE / CprE / SE 492 Weekly Report 04: SDMay21-29
Intelligent Code Editor

Mar. 2 - Mar. 15
Client: Hung Phan
Advisor: Dr. Ali Jannesari

Team Members & Roles:

Evan Christensen - Meeting Scribe
Ben Gonner - Report Manager
Jacob Puetz - Chief Engineer Software Systems
Jordan Silvers - Meeting Facilitator
Cory Smith - Test Engineer

Weekly Summary:

In the previous period, we continued work on the development of our plugin and classification data, and began focusing more on the actual data we're going to use. Specifically, the SPoC dataset we're working with is pseudocode to code, but our end goal of the project is to do natural language to code, so we need to manually translate enough of that pseudocode into actual English. Additionally, we're working on training an Open-NMT neural network to do translations rather than classification to act as a baseline for our classification GNN, and some extra preprocessing of the data to use with this (and potentially the classification GNN too).

Past Week Accomplishments:

- Open-NMT & Preprocessing - Cory
  - I have set up an Open-NMT neural network for the baseline to compare our results to
    - Our client ran it without any preprocessing done on the data and got an accuracy of about 42%
    - Preprocessing should help quite a bit
  - The main form of preprocessing I'm doing is replacing variables, functions, and literals with placeholder values
    - Because function names, variable names, and literals can all have an effectively infinite number of variations, holding them all in the tokenizer separately is extremely inefficient and not flexible enough for regular use
    - Instead of doing that, I'm replacing every name and literal with a placeholder (like @FUNCTION_NAME_0 or @STRING_LITERAL_1)
    - The placeholders all have numbers attached to them based on the order in which they are found in the pseudocode. This will hopefully help the network identify what order these are supposed to go in at the end
- GNN Model Exporting & Reuse - Cory

    - Haven’t done much on this in the last two weeks, but I have determined that the encoding process occurs in the build_graph.py file, so that’s where I’ll be focusing for the next week
- Classification Sorting - Jacob/Evan
    - Continuing to sort the dataset into various loop/conditional/variable/etc categories
    - Need to manipulate some of the dataset for natural language processing as opposed to pseudocode
- Visual Studio plugin - Ben/Jordan
    - Added plugin command to context menu
- Dataset Creation - Everyone
    - Modified hundreds of lines of data for client and advisor

Individual Contributions:

Name	Contributions	Hours This Week	Total Hours
Evan Christensen	Classification Sorting and Dataset Creation	6	17
Ben Gonner	Plugin development and dataset creation	7	16
Jacob Puetz	Classification sorting	4	14
Jordan Silvers	Plugin development and dataset creation	7	19
Cory Smith	Open-NMT & Preprocessing, GNN Model Exporting & Reuse	6	32

Upcoming Plans:

- Continue work on creating an easily usable model that can take in a string and put out a string
- Continue to work on sorting the training data.
- Set the plugin command to take a line of text or a highlighted section of text.