02/25/2020 - 03/09/2020

E E/Cpr E/S E 492 Bi-Weekly Report 4

Intelligent Code Editor

Client & Advisor: Ali Jannesari

sdmay20-46

John Jago – Software Lead

Keaton Johnson – Systems Lead

Jon Novak – Machine Learning Lead

Matthew Orth – Meeting Facilitator

Garet Phelps – Report Manager

Isaac Spanier – Test Lead

Weekly Summary

This sprint has been focused on finishing up the processing scripts, and getting the mined dataset ready for labelling. We were able to complete the Java method invocation dataset GitHub mining, which mined about 5,000 unique Java method invocations. We also completed the Java type translation automation script that ran the mined Java method invocations through and their files to return a list of Java method invocations with their parameters and variables converted to their type. Finally, we also finished the AWS hosting for the NLTK scripts on a Lambda server. This work will allow us to begin labeling our 1,000 sample Java method invocation dataset next sprint.

Past week accomplishments

John Jago

- Peer evaluation presentation
  - Updated our prototype implementation overview based on the progress made since the beginning of the semester
- Dataset processing for Java statements
  - Adjusted the preprocessing script so that it strips the package name
  - Added more tests cases
  - Investigated edge cases (see Sprint 11 folder, Preprocessing Edge Cases)
- Natural language statement preprocessing adjustments
  - Adjusted our script so that it does not remove information that may be helpful for the model, such as arithmetic operators and Java types
- Mined Java statement preprocessing

    - Created a Bash script to automate repo downloading
    - Ran Keaton’s mined Java statements through the invocation formatter
        - The real data revealed many edge cases, most of which have been resolved

Keaton Johnson

- Continued work on dataset mining
    - Created 2 additional iterations to correct the formatting of resultant data.
    - Added another output file that lists download links for all the scanned repos.
    - Removed duplicant methods from output file.

Jon Novak

- Continuing to work on dataset mining.

Matthew Orth

- Evaluation Dataset Update:
    - Updated the 100 sample evaluation dataset's natural language statements to include the new chained and nested format of requiring the user to specify this information in the input natural language statement
        - Created 1 file that gives samples of what the user would input (using original variable and method names)
        - Create 1 file that contains the dataset labeling (containing only Java types)
- Evaluation Dataset System Testing:
    - Run the 100 sample evaluation dataset through our whole system to ensure our plan for the 5,000 sample dataset will work as we expect
        - Based on labeling and testing process, a new dataset labeling process and pointers were created to ensure uniform dataset labeling for the 5,000 sample dataset
        - Also confirmed that training and translating a model using the 100 sample evaluation dataset works well, so our current plan has been confirmed
        - Resolved the <unk> token error
- Updated Peer Evaluation Presentation Slides:
    - Updated our presentation from last semester with updated progress information

Garet Phelps

- Peer evaluation presentation
    - Worked on part of the peer evaluation presentation and recorded my part.
- Plugin Preprocessing
    - Worked on the preprocessing that happens within the intellij plugin.
    - Not bulletproof yet but the main idea is there

Isaac Spanier

- Peer Evaluation Presentation
    - Updated our presentation slides and then recorded my part

- AWS Lambda Server Setup
    - Moved preprocessing script from a lambda function to an application that can be run via the AWS server.
    - Setup access for all members of the team, and gave them permissions to both the EC2 and Lambda servers.

Individual contributions

Name	**Contributions**	**Bi-Weekly Hours**	**Cumulative Semester Hours**
John Jago	● Peer evaluation presentation ● Dataset processing for Java statements ● Natural language statement preprocessing adjustments ● Mined Java statement preprocessing	18	38
Keaton Johnson	● Continued work on dataset mining ● Mainly worked on reformatting data	11	30
Jon Novak	● continue working on dataset mining ● Peer evaluation	10	28
Matthew Orth	● Evaluation Dataset Update ● Evaluation Dataset System Testing ● Updated Peer Evaluation Presentation Slides	14	53
Garet Phelps	● Peer-eval presentation ● Preprocessing for intellij plugin	12	31
Isaac Spanier	● Peer Evaluation Presentation ● Lambda application work ● AWS Server setup	13	33

Plans for the upcoming sprint

John Jago

- Label 170 samples from 1,000 sample dataset

Keaton Johnson

- Label 170 samples from 1,000 sample dataset

Jon Novak

- Label 170 samples from 1,000 sample dataset

Matthew Orth

- Label 170 samples from 1,000 sample dataset

Garet Phelps:

- Label 170 samples from 1,000 sample dataset

Isaac Spanier

- Label 170 samples from 1,000 sample dataset

Summary of weekly client/advisor meeting

Meeting with Hung Phan on 2020-03-02 at 4:30 pm

We reviewed the finalized steps for labeling the dataset to ensure consistency when we begin labeling our large method invocation dataset next week. We reviewed the results from the evaluation dataset, and it appears that the way we decided to format the data will allow for a relatively high accuracy in translation. Keaton and John agreed on a file format that will allow John's Java method invocation preprocessing script to work with Keaton's mined data. We also cut down our method invocation dataset size from 5,000 down to 1,000 since it would have taken 50 hours per person to do the labeling, but we plan to permute the English through word choice and sentence structure to grow the dataset to 5,000 or more.

Meeting with Hung Phan on 2020-03-09 at 4:30 pm

We met with Hung to discuss our progress since the last week. We resolved a few outstanding issues with labeling the mined Java statements and prepared a plan for labeling the dataset during the next sprint. Professor Jannesari did not attend this meeting due to another appointment.