NYU TANDON SCHOOL OF ENGINEERING

Patterns in programming: identifying recurring mistakes made by first year computer science students

Darryl Reeves, Gaurang Ruparelia

Abstract

Starting from a hypothesis that there are a set of common mistakes that students make when first learning to program, this project is focused on using coding submissions as a dataset for identifying, classifying, and understanding certain recurring patterns of mistakes or misunderstandings first time programmers make. The dataset of code student submissions was converted Pandas dataframes after removing all code comments, so that further statistical analysis can be conducted on the dataset.

Background on the Project

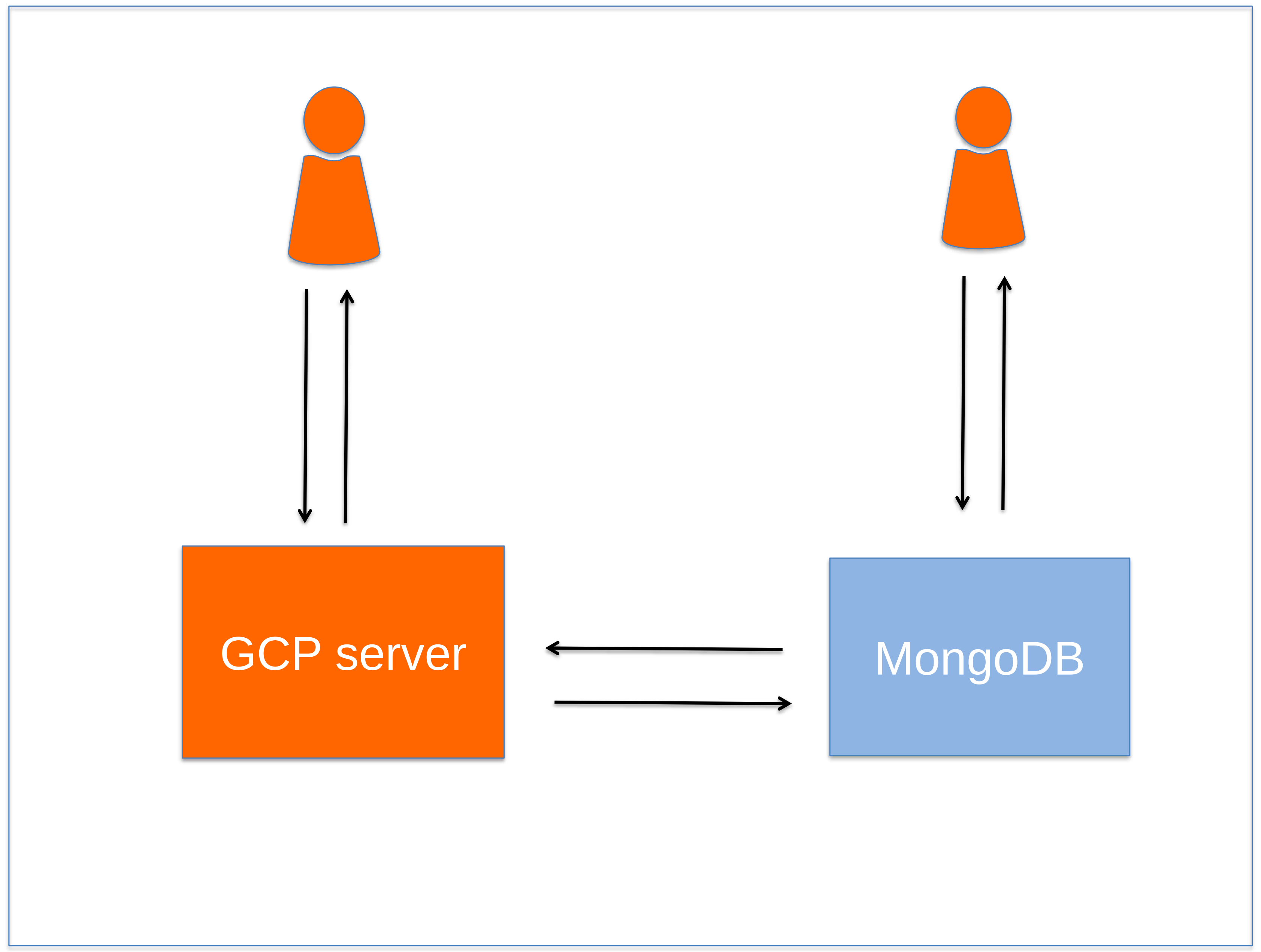
For students learning programming for the first time, the lack of familiarity with the rules of programming languages and the syntax required can make initial efforts to solve even simple problems frustrating and demotivating. Instructors do not always have insight into the types of problems students face and, often, students are intimidated by the idea of asking for help. The series of steps that leads to the students' final submission for a programming assignment represents an evolution of understanding that can prove insightful.

For the dataset, homework submissions from students that both give consent and took the CS1114: Introduction to programming class in academic year 2020, is used.

Methodology

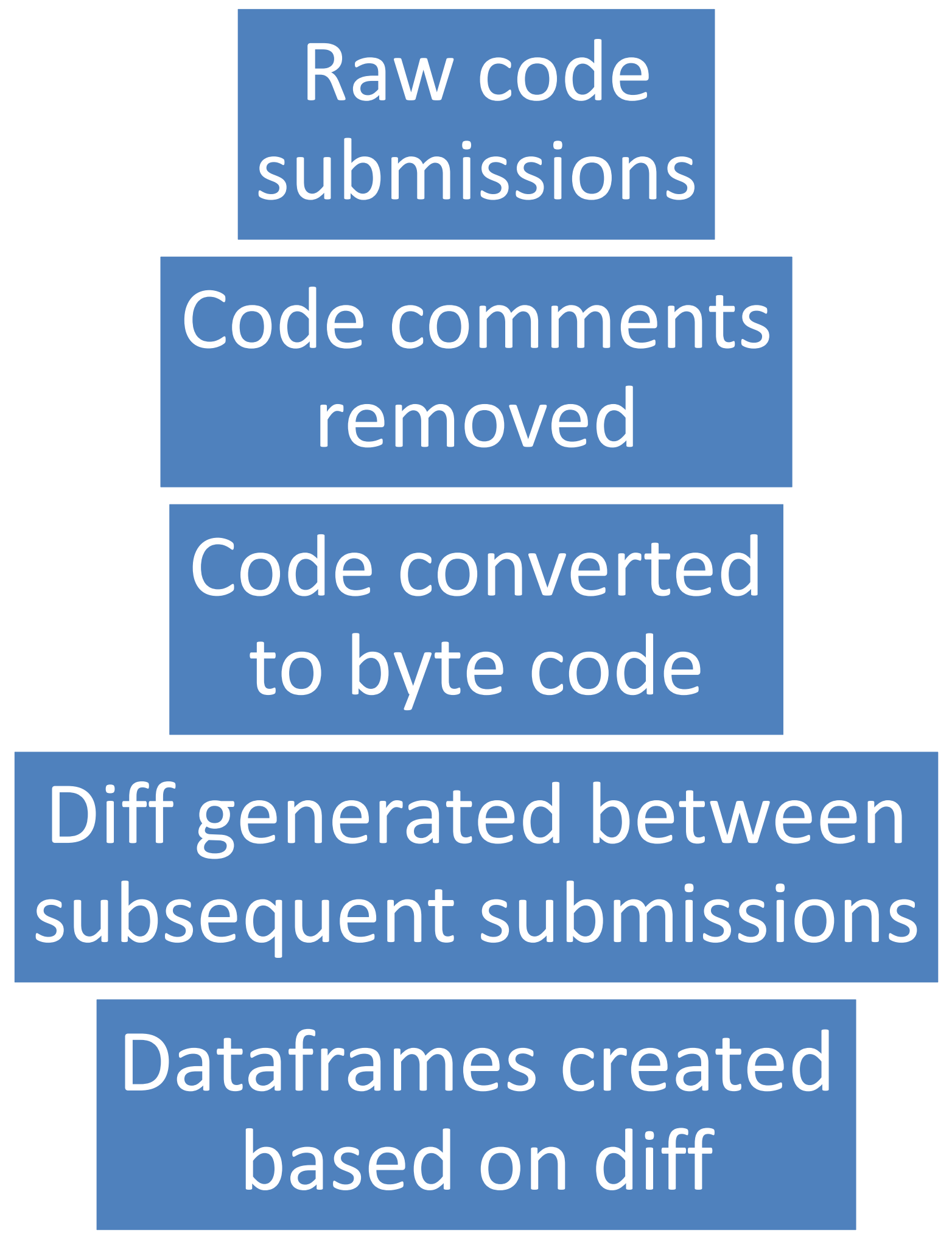
After removing all the comments and any traces of identification from the assignment, the code from the assignment is converted into byte code using the disassembly module in Python. The diff module is used to reveal the progress a student makes over subsequent submissions in the form of byte code.

Figure 1: Visualizing the interaction between different services and the user



This data is stored in MongoDB and retrieved to create two dataframes: instruction_count_dataframe, counting the number of times a byte code instruction appeared in a submission, and test_results dataframe, to identify how many test cases a student manages to pass over the course of code submissions. Majority of the computation for this project is done over cloud using the Google Cloud Platform services.

Figure 2: Flowchart showing data pipeline


Raw code submissions

Code comments removed

Code converted to byte code

Diff generated between subsequent submissions

Dataframes created based on diff

Future

With the submission code successfully transformed into dataframes, it is possible to perform statistical analysis on this code to identify patterns of mistakes for future intelligent tutoring programs.

Acknowledgement

The authors thank NYU Tandon School of Engineering's Office of Undergraduate Academics for generous funding of the project.