

Runestone Anonymous Data Proposal II

Bradley Miller, PhD

President and Founder of Runestone Academy LTD

November 1, 2024

Runestone Interactive has been collecting data on students use of interactive textbooks since Fall of 2012. The data we have been collecting includes clickstream data on nearly every interaction with the book including:

- page views
- Multiple choice question answers
- fill in the blank answers
- Stepping through code in CodeLens
- Rearranging blocks in Parsons problems
- The results of running unit tests on code
- answers to drag-n-drop questions, clickable area questions
- short answer questions
- The state of their code whenever they run an exercise or modify an example from the textbook

As of November 2024 we have collected over 1 billion rows of clickstream and 80 million code samples. This is about 4 times the amount of data I reported in 2020 in a similar proposal. One thing that we have learned from the analysis of the previous data sets is that more density, especially more density of data in the latter chapters of books would

be better. Also, as we now have many years worth of data I would propose that we create a longitudinal data set spanning several years for the CS Awesome and one python based book, that I will choose after analyzing the available data. This longitudinal data could even track courses from several institutions that have consistently used a runestone book for several years. A few institutions have used runestone for nearly 10 years.

Organization

As Runestone offers several different courses I propose to organize the data sets into groups according to the base course:

1. CS Awesome in Java
2. How to think like a computer scientist
3. Python for Everyone
4. Python Data structures course
5. Data Structures in C

Click Steam Data

The Clickstream data will consist of the following

1. Anonymized username
2. Anonymized course name
3. Base Course
4. Anonymized institution
5. Type of institution (high school, college, grad school, etc.)
6. Start date of the course
7. Timestamp of the event

8. Event type - e.g. page view, activecode, multiple choice, etc.
9. Event Details - specific answer and correctness
10. Unique identifier of the component acted upon
11. The chapter identifier where the event is located
12. The identifier of the page within the chapter

The code data consists of the institutional and course data along with the following:

1. Timestamp
2. Anonymized user id
3. Source code that was run — sanitized as much as possible for the inadvertent inclusion of a student's name in the code.
4. The unique identifier for this coding exercise
5. A problem statement for this exercise
6. Any error messages that were generated by the run
7. The programming language of the code
8. Any unit test data that is available for this run.

All data will be converted so that it is compatible with the datashop schema.

Time Line and Budget

I estimate that organizing all of this will take 2-3 weeks of work. I propose \$7500 to pay for my time in doing this work. A large amount of the work for this proposal will be analyzing the data we have available to find courses that contain sufficient density. Secondly, data that is more than two years old has been archived to csv files and will have to be reconstituted into a useable database to produce the longitudinal data sets.

I will make all of the data available to your project in the form of CSV files or some other agreeable data format. The data will be made available on the data shop website alongside the previous data.

Making sure that the data include questionnaires