Runestone Anonymous Data Proposal

Bradley Miller, PhD President and Founder of Runestone Interactive LLC

April 20, 2020

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 April 2020 we have collected 315 million rows of clickstream information and 19 million code samples. I think this is a very valuable data set that holds a huge potential for better understanding how students learn. I have always wanted to get this data into the hands of researchers who can analyze the data and suggest improvements to the platform and books. In particular the data from the how to think like a computer scientist book spans 8 years and many documented improvements to the books, so that longitudinal studies could be done across several years. The terms of service on Runestone Academy allow for me to share the data for educational research purposes, and in order to improvement the

effectiveness of the site. I believe your uses of the data would fall under those allowances.

Organization

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

1. How to think like a computer scientist
2. Foundations of Python Programming
3. Java Review Course
4. CS Awesome Course
5. Python Data structures course
6. Data Structures in C++ course
7. How to think like a data scientist

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 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 students 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.

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. I will make all of the data available to your project in the form of CSV files or some other agreeable data format. I will publish the data on the Runestone Interactive website to make it available for other researchers.