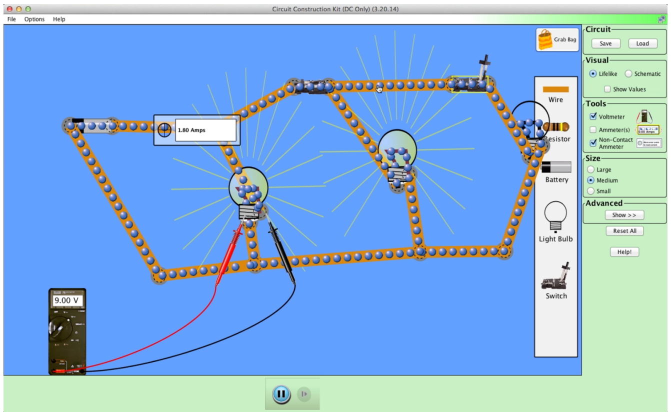
Analyzing n-grams in raw log data

Lauren Fratamico
fratamic@cs.ubc.ca

Task Description



Visual: ● Lifelike ○ Schematic
□ Show Values

Tools: ☑ Voltmeter □ Ammeter(s) ☑ Non-Contact Ammeter

Size: ○ Large ● Medium ○ Small

My visualization will be built on data collected from student interactions with an online circuit building simulation. In this simulation, they can both build their own circuits and test them. The picture on the right shows some of what can be done in the sim. We ran a study with 100 first year physics students, where we had them interact with the simulation for an hour. They were also given a pre and post test so we can evaluate how much was learned during the sim interaction and can group them into "high learners" and "low learners". The sim logged actions taken by the student. A short example of the logs is shown below. We'd like to see if different actions are taken by different groups of students. This could allow us to give adaptive interventions to help students learn more while playing with the circuit simulation.

The logs look like this:

Time	**Actor**	**Component**		**Action**
1020399470968	user	battery.0	sprite	addedComponent
1020399472687	user	wire.1	sprite	addedComponent
1020399473812	model	junction.41	junction	junctionFormed

Personal Experience

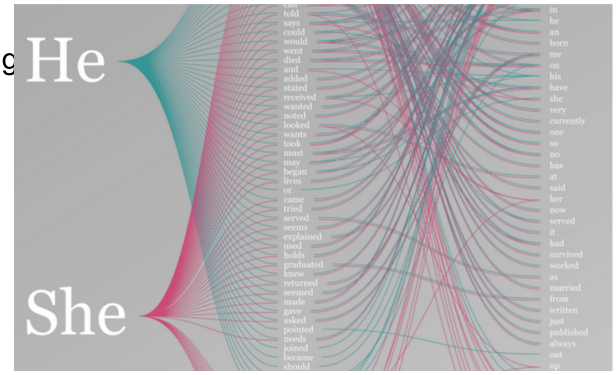
I have been working with this dataset for the past year, however we have not yet analyzed the raw log files. My approach so far with this research has been to come up with a couple different higher level hierarchies that combine actions, but I am interested in seeing if similar results are found with just the raw logs in terms of frequencies of actions. With this visualization, I am also hoping to learn about the sequences of actions taken by students. Through my research, I analyzed sequences only at a higher level and have not found anything so far at that level. The results we have so far indicate that testing more frequently is associated with learning more, and organizing the components on the screen is associated with learning less.

Proposed Solution

I propose a solution that allows users of the visualization to explore the most common actions taken by students using the sim. He will also be able to explore the actions taken before and after that action, along with how the actions differ amongst two populations of users (high and low learners). Interactions with the system will be broken up into two screens. The first shows frequencies of 1-grams (actions). And the second screen shows transition frequencies between the actions. I thought about having both views on one screen, but in the interest of available pixels, I think it would be best to have them on separate screens so the whole width can be used for the n-gram visualization. The user should not need to look at both views at the same time.

As can be seen in the Interface Mockups, the low learners column is red and the high learners column is green. This corresponds to the colors on the second screen, where the colors are used to show the transition frequencies. The colors were also chosen to represent good (green - high learners) and bad (red - low learners') actions.

The second screen will look like the example picture to the right found on Chris Harrison's blog about visualizing web trigrams. However, in my visualization, the thickness of the lines will represent the transition frequencies. In this way the user will be able to visually compare the transition frequencies of high and low learners. If this is insufficient for comparison (eg, if the lines are too close in size), then another method will be used, perhaps varying the size of the text.

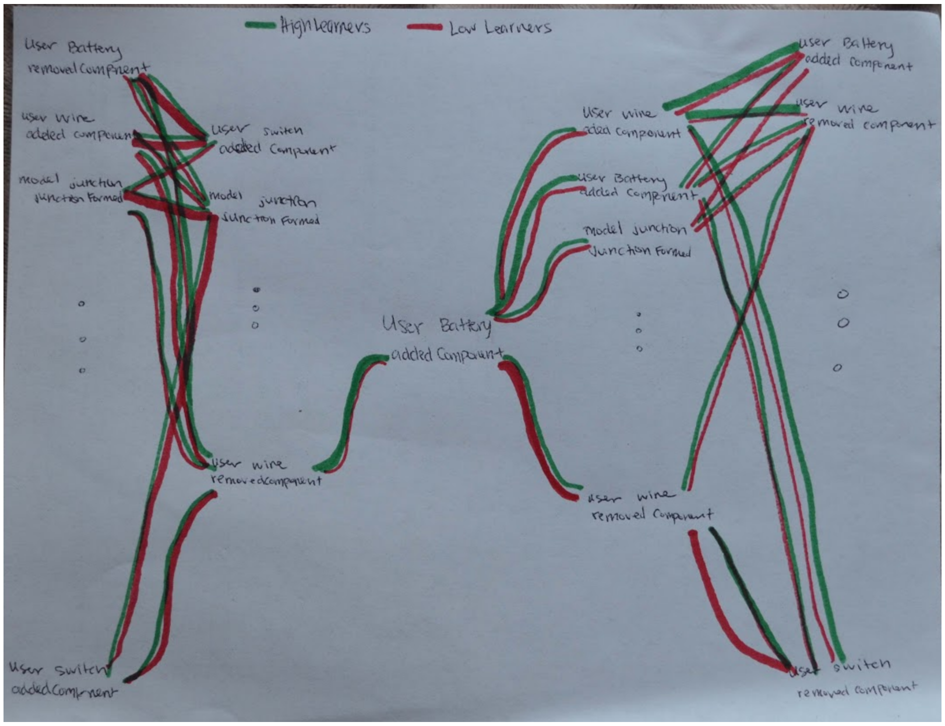


Interface Mockups

Screen 1 - Users can sort by any of the columns and can click on any of the actions. Clicking on a action brings them to the second view:

Action	**Low Learner Frequency ↓**	**High Learner Frequency**	**Statistical Significance**
user battery removedComponent	.0140	.0009	.00005
model junction junctionFormed	.0100	.0087	.34000
user wire removedComponent	.0099	.0150	.04400
user battery addedComponent	.0087	.0023	.09235
user resistor addedComponent	.0077	.0047	.16540
user wire addedComponent	.0043	.0052	.29840
user switch removedComponent	.0040	.0122	.02998
⋮	⋮	⋮	

Screen 2 - In this view the users can explore the actions that take place before and after the selected action. The widths of the lines correspond to the transition probabilities between actions. The colors are chosen to be consistent with the view in screen 1:


High Learners
Low Learners
User Battery removal Component
user wire added component
model junction junction Formed
User switch added Component
model junction Junction Formed
User wire removed component
User switch added Component
User Battery added Component
User wire added component
User Battery added Component
model junction Junction Formed
User wire removed Component
user Battery added Component
user wire removed component
user switch removed Component

Scenario of Use

The user starts the system to explore the differences between high and low learners. He start on the first screen and sort by the Statistically Significant column so that he can see which actions are statistically significantly different between high and low learners. He can also see the frequency of the action by low and high learners by looking in the columns to the left. He scrolls down to an action that he has interest in. He chooses “user battery addedComponent” because it is the most significant. Once he clicks on this, a new view is opened. This shows the chosen action in the center and two columns of action to the right and left of center. These actions are the ones most commonly taken before and after the selected action. This list will be limited to ~10 so that it will be a manageable number for the user to view. From this screen the user can trace before and after the selected action to see the sequences that surround it and he can compare the relative transition frequencies of the actions by groups.

Proposed Implementation Approach

I plan to use d3 to create my visualizations. I have seen a couple examples, including code, of semi related designs, so I will build off of some of those if I am able to. I will be using python to do the data processing.

Milestones and Schedule

Date	Work completed
Nov 3 (W)	Preprocess log data. This involves creating separate files for each student and activity and removing specific information about which item the student was using (eg, I don't care if it was light bulb 1 or 5)
Nov 7 (F)	Calculate numbers needed for my n-gram visualization. This likely includes frequencies of 1-grams as well as transition frequencies.
Nov 14 (F)	Have an initial n-gram visualization - the ability to visualize some of the data. Both my 1-gram frequencies and the n-gram tree for a specific start or end point.
Nov 21 (F)	Build a view that allows for the comparisons between multiple groups - Initially I will be focusing on differences between high and low learners.
Nov 28 (F)	Have the transitions working, i.e., the system should allow users to click on a 1-gram and load the correct n-gram tree.
Dec 5 (F)	2nd iteration of the visualization layout. Think about other views potentially helpful to add.
Dec 10 (W)	Finalize presentation and prepare a demo video
Dec 12 (F)	Final Presentation
Dec 15 (M)	Turn in final paper

Previous work

Academic work and related blogs

http://hint.fm/projects/wordtree/
http://www.chrisharrison.net/index.php/Visualizations/WebTrigrams
http://www.visualizing.org/stories/visualizing-ngrams

Relevant code examples

http://bl.ocks.org/mbostock/4063570
http://bl.ocks.org/mbostock/4339083
https://developers.google.com/chart/interactive/docs/gallery/wordtree
http://www.jasondavies.com/wordtree/wordtree.js?20130312.1