Daniel Schwartz
CS 123 – Professor Rogers
4-14-14

BIG DATA PROJECT PROPOSAL:

Predicting User Retention on Khan Academy

Data. Khan Academy is a very popular online learning platform that focuses on STEM topics. The site gets upwards of 10 million unique users per month, who spend the majority of their site time watching short video lectures and working on practice exercises (about 4 million are completed every day). I've contacted Khan Academy's research team, and if I get what I want, I will be working with the complete video, exercise and assessment logs of a random sample of users (I've requested a minimum of 20GB of data, which would require MapReduce at the minimum for any analysis I try to do through R, e.g. using rpy2 — if they won't give me enough data to satisfy the purposes of this project, I'll just switch to my other idea, described below).

Exploration and Prototype. In terms of language, I plan on using Python (and also the rpy2 package to implement R-specific methods and data structures as needed). For the prototype analysis I will run some naïve models (generalized linear models for simple predictions of retention and user activity, and mixed effects models to look at "growth" of users' activity over time) and create some exploratory visualizations (mainly focused on developing intuition for what different use cases in the data look like). One of the main goals of the prototyping will be to figure out how I want to operationalize the important concepts (i.e. what are the most interesting and useful ways to define "retention" and "activity").

Final Product. My main goal is to build a small set of models and visualizations that explain what makes first-time Khan Academy more likely to never return or be inactive on the site, and I will communicate these results in the report. Ideally my analysis could help Khan Academy identify potential strategies to increase retention of first-time users, particularly those who may be "low-hanging fruit".

Note: If I am unable to get any Khan Academy data, or I am unable to get a sufficient amount to satisfy the requirements for this project, I will analyze the 166GB Yahoo! Answers Browsing Behavior dataset on webscope.sandbox.yahoo.com. I would use this dataset, which is quite close in nature to the Khan Academy data, for a similar analysis (modeling and visualization) of user retention. The Yahoo data have some additional interesting features, including a point system, to explore.