MOOCenimages:

An analytics platform for MOOCDB

*6.UAP Project Proposal*

*Sherwin Wu*

With the current rise of massive open online courses (MOOCs) such as edX, Coursera, and Khan Academy, each of these MOOCs have began to collect large amounts of data from their online students. Hidden within the data are valuable, non-obvious insights about these MOOCs. While each individual MOOC could write their own data analysis scripts on their own data sets to reach their own conclusions, we recognize the benefit for everyone if MOOCs could collaborate on the scripts and analyses they perform. However, the prerequisites for this type of collaboration would be a unified database schema for all data stored by each of the MOOCs. MOOCDB seeks to provide that unified schema for MOOCs. If all the different MOOCs store their data in the MOOCdb format, one script written by anybody could run on all the different MOOCdb instances to generate custom analytics across all data sets. My project, which I will work on closely with my mentor Colin Taylor, under the supervision of Kalyan Veeramachaneni, will involve writing piping scripts to make data in the MOOCdb format, writing analytics scripts to run on them, and creating an analytics platform for MOOCdb.

Before the beginning of my UAP project, the ALFA lab had already narrowed down on a unified schema for MOOCDB in MySQL. For example, the schema requires a “submissions” table that maps a foreign key to a “users” table. The lab had also received a myriad of data dumps from edX and other MOOCs that were not in the MOOCdb format. The first part of my project will be writing piping scripts that will take the data dumps and put them into the MOOCdb format. This is work that is necessary to be completed in order to move onto the second (and more interesting) part of the project.

The second part of the project, which will take up the vast majority of my time for this project, involves writing an analytics platform for anyone using the MOOCdb schema. It will also involve running analytics on the data sets provided to us to identify analytics scripts that are the most interesting and valuable.

The analytics platform at its very core will be a web application that allows for sharing of scripts and visualizations run on various MOOCdb instances. The web application will be built using an existing popular framework, such as Ruby on Rails or Meteor. It will allow users to upload scripts to run on a small mock database that we have stored on our web server and see immediate results of what the script does. Further, the users will also be able to upload the results of their analytics scripts on their own data sets. This will support many visualization formats, such as an HTML/Javascripts documents, a PNG image, or others. Users will be able to upload scripts, and also upload associated visualizations with the scripts. These scripts and visualizations can be upvoted so the most interesting scripts will be ranked higher than others. The web app will also allow for users to download scripts to run on their own MOOCdb instances if they wish.

A rough timeline of my project is as follows. For the first month of the project I plan on writing the piping scripts to assemble at MOOCdb instances for the data dumps that the ALFA lab has created. This will last all of September. Starting in October, I will start running analytics scripts on the MOOCdb instances I helped create. While writing the analytics scripts, I will also start building the script sharing analytics platform with my mentor, Colin Taylor. The first couple weeks of October will be spent planning the application infrastructure and design, and we plan on having a minimal viable product finished in mid-November (before Thanksgiving break). The remaining 3 weeks of the term will be spent polishing up the web application, as well as slowly rolling it out to our partners at edX and Coursera to use. We will iterate based on the feedback we get from them to refine the analytics framework until the end of the project at the end of the semester.

The project will hopefully enable a new method of collaboration for MOOCs that was not available before. For example, if edX finds out that an analytics script they ran showed that students weren’t gaining anything from their online courseware, they can upload the script along with the visualization they generated. Then, Coursera, Khan Academy, or anyone else can download the script and just run it out of the box on their MOOCdb instance (since the schemas should be unified and the same). Then, they could look at the output of the script on their data and see if their data supports the same conclusion or not. By jumping on the MOOC data surplus early like right now, we could avoid the data divergence mess that healthcare is dealing with right now.