CS1951A *Final Project Pre-Proposal*

Rebecca Zuo [rzuo]
Yeon Jun (YJ) Kim [ykim106]
Jessica Dai [jdai6]
Amy Pu [apu1]

Vision

What is your "big idea"?

What are the potential correlations between the demographics of public schools (income, race, dis/ability, ESL/ELL), and the resources/ opportunities available at those schools (edtech products, vocational/tech training programs, arts)?

What might you find at the end of your project?

Our initial intuition suggests that whiter, wealthier schools generally have access to more resources; however, we've noted anecdotally that in some cases, title 1 schools have access to special grants/ funding for additional programs or technology, so it will be interesting to see the extent to which that additional aid is effective. Conversely, the extent to which disparities exist is also unclear.

Data

What dataset do you plan on using?

While we're still working on finding more recent data, our current datasets are from data.gov:

- Data on art education of secondary schools
- Data on education and civil rights issues in public schools
- Data on a range of educational technology resources in public elementary/secondary schools
- Data on career and technical education programs

How big is it?

Really big - includes information from every public school in the nation: 16,500 school districts, 97,000 schools, and 49 million students.

How do you plan to collect it?

Generally, the data has already been collected (by the federal government, via a series of surveys) -- while we will be discussing and accounting for their methodologies, the biggest challenge will be aggregating and correlating the separate datasets.

How do you plan to clean it?

After investigating the specific formatting of the separate data sets, we'll first decide the best way to integrate them and which attributes we want to focus on; then, depending on that decision, we'll check for empty values, misspellings/ typos/ entry errors, and outliers.

Methodology

What do you plan on doing with your data?

The different datasets can be overlayed, such that the public school can be used as a key to corresponding levels of resources, race, income, dis/ability, and education resources. Then different models can be tested on the data in order to find a cleaner visualization to demonstrate the correlation of demographic versus resources.

What techniques do you think you will use to analyze the data?

Sorting can be used to spot initial trends between different classes. Different public schools can be given classifiers based on racial diversity, income and dis/ability so that the dataset can be reduced.

Depending on how quickly we get through other parts of the project, we could also try writing a basic algorithm that would predict resources available given demographics, and vice versa.

How might you visualize your results?

We'll build an interactive visualization of the effects of varying resources in public schools.

The first part of the webapp will have two graphs side by side: the left being some sort of tool where users can choose and change different resources, funding, etc., and see the effects of population, gender, race, etc. in a graph on a graph on the right side. If there are certain trends that are noted in the results of our data, we'll also display that using popup boxes next to the graphs as the user changes the amount of available resources.

The second part of the webapp will generate random classifications of populations, including race, gender, sexuality, background descriptors, and the user will be able to choose from a group of options to predict what resources were available. After guessing, the correct option corresponding to the population will be displayed.

Use of a database

Because we're most familiar with SQL, we're currently planning to use a SQL database; that being said, if there are other database models that would be more effective, we are very open to switching.

Workflow

By the first TA checkin, we hope to have cleaned and sorted through our data. We are using more than 2 datasets, so combining all the datasets will take some time. We hope to have decided on what trends and notable effects to look out for. We also hope to have the structure of our database created.

By the midterm report, we want to be finished with analyzing the datasets, having come up with a decisive result/trend that has either confirmed our denied our hypothesis. We want the most fundamental and main elements of the webapp to be created, including the user interaction features and the data graphs.

Blog

Link to our blog here!