Nadeem Nasimi
Michael Hammer

No8am Datamining Final Project - Phase I

Accomplishments

Exported Data from Database

The data consists of course schedules saved by students using No8am from 2014 to 2016. This data was stored in a NoSQL database. This data was exported to a format similar to JSON. In order to load this data into RStudio we first had to load in the data as plain text, format it as JSON, and then parse the data as JSON using the jsonlite library.

Cleaning Data

In order to clean the data we first had to import it into Rstudio. The data was in JSON format so we had to use a package which could flatten the data so that it would be read in as a table instead of a list of lists. Next, we removed some unnecessary columns and modified their names to nicer formats. At this point the data frame was in an easy to use format, but we still had to clean the data itself. To do this, we first changed NA (for not having a course) to zero-ed out data (['000000','00']). This format is the same as the other data values but it uniquely identifies it as not having a course. Subsequently, we discovered that there was invalid data from other schools that we had to remove and there were empty schedules (rows) that we deleted. Lastly, we parsed the schedule creation date, which was stored as Unix time, into as POSIX timestamp.

Creating Sub-Datasets

After cleaning the data we wanted to create some useful datasets from the cleaned, main one. The first data set we made had the data but courses were represented as True or False if the transaction had the course or did not have it respectively. From that data set, we created another where the True values were replaced with the course names. This allowed us to easily find the course names that each transaction contained. Finally, we created a data set that had only the main sections of the course listed (i.e. there were no labs, recitations, or problem sets). We did this because people in the main section had to be in the other sections, thus the data could be redundant.

Transaction Rules

We were able to create transaction rules for the courses. It told us which courses were strongly related. For example, we saw that if a student is taking CSCI 206 and ENGR 229 then they are likely to be taking MATH 222. This makes sense because CSCI 206, ENGR 229, and MATH 222 were all scheduled in Sophomore year for Computer Science and Engineering major's old catalog.

Challenges

Cleaning Data

Cleaning the data proved to be a harder task than we anticipated. In addition to doing the standard cleaning we did in class such as renaming columns, we had to make design decisions when modifying the structure of our data frame so that it would work on different algorithms. Additionally, when we ran the data through those different algorithms for testing, we would frequently encounter errors that would require us to go back and clean the data again.

Creating Useful Data Frames

While we were cleaning the data, which is all stored one data frame, we decided to make a copy of this data frame at different stages in the process. This will allow us to the data available in multiple formats, depending on the algorithm being used.

Future Work

We plan to use the Apriori algorithm to generate association rules on the course data to discover which courses are most often chosen together. To do this, we will need to group each course by department. This will allow us to filter the generated association rules to eliminate obvious patterns, such as required courses in a given semester for a certain major. The filtered data will show more interested patterns, such as commonly chosen electives for students of different majors.