Nadeem Nasimi
Michael Hammer

No8am Datamining Final Project Proposal

Problem Statement

No8am is a course scheduling web site available to students at Bucknell University. It has been running since our sophomore year in 2014. Over the past few semesters, it has collected course information from hundreds of students. We believe that correlations exist within this data. For example, people in certain majors most likely share similar interests. Thus, they could be in the same classes either because their major requires it or because they are interested in it. We aim to use the data to suggest relevant courses to students as they create their schedules in No8am.

Data Description

A user's schedule is sent to a database each time the user clicks the save button for easy retrieval of the schedule in the future.

The data has been exported from the database as a list (similar to JSON) of saved course schedules over multiple semesters. The image on the right shows an example of what the data looks like. We can see each course that the user chose as well as specific sections of those courses. Additionally, we can use the time the schedule was saved to determine the semester it was scheduled in.


```
{
    "path": {
        "key": "0DKC",
        "ref": "6b01567b98010ea2",
        "reftime": 1478719368410
    },
    "value": {
        "CSCI 204": {
            "main": ["55405", "01"],
            "L": ["57689", "62"]
        },
        "ECON 103": {
            "main": ["57567", "02"]
        },
        "MATH 201": {
            "main": ["50094", "01"]
        },
        "SPAN 207": {
            "main": ["58755", "03"],
            "R": ["58773", "44"]
        }
    },
    "reftime": 1478719368410
}
```


Approach

In order to be able to suggest courses to students, there are several tasks we need to accomplish: preprocessing the data, exploring the data, and developing a prediction algorithm. To successfully accomplish these tasks we will use RStudio to apply various algorithms from class and from other sources we find.

Preprocessing

Preprocessing the data includes uploading the data into RStudio and cleaning the data. We will need to use the summary() and str() functions in R in order to understand the data more. We will then use what we learn to make the data meaningful and allow us to use it in our analysis.

Exploration
Exploring the data includes finding correlations and patterns. Some techniques we will use include creating ggplots and developing association rules.

Prediction Algorithm
Developing a good prediction algorithm will involve applying the various algorithms we learned from class such as Neural Nets, Naive Bayes, and Decision Trees. We will then use prediction evaluation techniques, like looking at ROC curves, to determine which algorithm predicts our data the best.

Success Criteria

Our project will be considered successful if we have developed a strong understanding of any correlations in the data and if we have developed the ability to successfully suggest relevant courses for students. Relevant courses will be determined by several criteria such as upper level courses should not be mixed with lower level ones and seeing if courses in similar majors are grouped together. We will also use human testing to see if they agree with the suggested courses.

Milestones

Data Pre-Processing - 11/16/2016
Have the data uploaded and cleaned.

Explore and Organize Data - 11/18/2016
Have a good understanding of the data. This includes creating plots and statistics.

Prediction Algorithm - 11/28/2016
Have a deep understanding of the data. Also have implemented and tested several prediction algorithms.

Analyze Results - 12/1/2016
Have a strong understanding of what we found from the data and its ability to predict courses.