**Senior Project Proposal**

Cody Hock

Fall 2014

**Objective**

For my senior project I will be doing an analysis on the outcomes of NFL games that have been played from 2000-2013. I will gather data from the web and place this into my own database for quick access during my analysis stage. One of the two larger segments of my project will be in the R programming language. I have never spent any time with R, so this will be a great opportunity to learn a language that is not taught in the traditional Computer Science curriculum at NMU. This project will have a significant research component involving many trials that result in failures (predicting sporting events, as it turns out, is very difficult). It is my hope that the Computer Science Committee takes this into account when reviewing my proposal.

**Project Details**

My project is broken up into three main components: PHP regular expressions; C# file manipulation with MySQL data storage; and the open-source, R programming language.

The first component, written in PHP, is to collect the data from Regular Expressions. This will feed into file manipulation with C# and loading the data into my own MySQL database. The final area of focus will be with R, which I have just recently been exposed to. As I begin to learn about R, I have noticed one of the features it supports is running regular expressions on the web for data mining and even storage. I was not aware of this until I wrote my program in PHP to do the data collection already. However, I feel that this allowed me to cover a wider range of topics/languages for this project and I learned more because of it

I know that one of requirements that our website says is *“we hope for about maybe 1,000 lines of code”*. However, I do not think I will reach this number for my final implementation of this project (though I plan for a good deal of code writing for practice in R that won't be directly used). I will be focusing more on the *“credit for projects with a significant research component, difficult algorithms or data structures, or some other technical challenge”.* R programming will likely not result in the 1000 lines of code requirement, but it is a very dense language that will cover the research and algorithm component of requirements.

I will be using the data collected from years 2000-2011 as a Training set to test and refine my algorithms and then use the final 2 years of data as a Testing set. I will run appropriate statistical validations, including producing r-squared values, to determine that the correlation provides some form of legitimacy. Some of the popular methods for game prediction, that I will also be trying, include: home field advantage, turnover ratio along with time of possession, and defense wins championships. I will look for correlations in the results to go along with these popular conjectures. I would also like to try and create an algorithm that compares the results of the entire team with how elite the quarterback is. Many “experts” claim the best quarterbacks find ways to help their teams win. Well, I would like to see for myself if this is true.

As for a deliverable, I cannot promise a working predictor (if I could I would be spending a lot more time in Vegas). However, I plan to have all of my trials on display to compare their effectiveness when predicting the winning team in an NFL game.

**Grading Rubric**

**PHP**

|  |  |
| --- | --- |
| **Feature** | **Points** |
| Program can index multiple pages for data collection | 2 |
| Regular Expressions gather the data required for the project (without this there is no project!) | 15 |
| Program can parse the results from each Regex into a .csv file for use later on | 5 |
| Refactoring code (usefully; I have a lot of regex to run)! | 3 |

**MySQL (and C#)**

|  |  |
| --- | --- |
| **Feature** | **Points** |
| C# program can parse all of the separate .csv files into the two that are needed for each year | 5 |
| Create and manage my own MySQL database (1 database, 2 tables) | 3 |
| Can load the .csv files into the proper tables in my NFL database (1 per table) | 2 |

**R**

This is the section of my project where I do not know where a greater understanding of R will take me. A large part of this project will be in the area of research and developing algorithms for game predictions. I know R has some very cool features such as allowing for linear regression and machine learning. I have reserved an equal amount of points as the PHP component of my project (25 points) for R. I do not know what will be used to get there however because much of what I will write will not be included for the functioning final project. These points can be discussed upon completion based upon the work that is completed, but a lot of what I should accomplish includes what was discussed in my project details.

**Total:**

**PHP** 25

**MySQL** 10

**R** 25

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **A** | **B** | **C** | **D** | **F** |
| > 52 | 45 | 38 | 30 | 30 > |