CMSC 122 Project Proposal

Natalie Gray, Steven Cooklev, Will Simkins

Our project will focus on the analysis of tournament chess games. Our dataset of chess games will include the following information: 1) every move from each game 2) the result of the game 3) the names of the players 4) the ratings of the players, if it is listed 5) the ECO opening code of the game, 6) where the game was played and 7) what year the game was played. We have two potential sources for this data. One is www.chessgames.com, a free online chess database containing millions of chess games. We could crawl the website, and scrape the data from the site. The second source of data is from ChessBase, a chess software. The software already has millions of games in html format.

We have several potential goals in analyzing these chess games. Our primary goal is to produce a heatmap for each type of piece, from a queried portion of the dataset. The heatmap would show how long each type of piece is on a certain square. For example, we could produce a heat map that describes the location of white knights, a heat map that describes the locations of black pawns, and a heat map that describes the location of all captures. We can query the data in many ways, including by playing strength, by player, by country, by opening, by the length of the game, by the outcome, etc. We could use the heatmap to answer many interesting questions. For example, how do heatmaps differ for grandmasters vs. amateur players? How do heat maps differ by decade? How do heatmaps of different world champions differ? How does the expected outcome of a game change depending on if white has a knight is on the edge of the board? Do German players tend to move their knights differently than English players?

There are many possible side-goals we could accomplish with the data as well. We could look at expected win rates of each opening, how common different openings are in each country, and how opening choices have changed through each decade. We're less interested in these goals because analysis on these questions has already been done.

The first step of our project will be getting our data in a usable format. If we use the chessbase dataset, then we are able to convert data to HTML easily. We will then need to parse this HTML file in

order to extract the data we are interested in analyzing (particularly moves so we can determine the piece locations for the heatmap). We will likely use BeautifulSoup to do this. If we decide to also incorporate data from Chessgames.com, we will need to scrape the website to get an HTML file, and then we will use BeautifulSoup to parse that HTML file. We would also need to merge the two sources of data if we decide we need to use both sources. We hope to complete these tasks during weeks 4 and 5, as well as deciding on a data format to use to store the games. Ideally we would have a complete data set at the end of week 5.

The second step will be deciding how to analyze the data and visualize it in a heatmap. We would then decide on exactly which variables we want to focus on in the heatmaps, and make sure we can calculate everything we need to display the heatmaps properly. For visualization of the data, we would likely use a Python visualization library (ex: http://seaborn.pydata.org/index.html) to help us get a clean visualization of the chessboard. We likely will make alterations to the visualization we get from the library we use to make it fit our project, but a library will give us a good foundation to start working on the visualization. In weeks 6-7, we hope to have our data analyzed, know how we are going to visualize it, and be able to do trial runs of our visualization.

The third and final step will be for us to build the actual user interface for our project. We plan to will make a website which will display the heatmap, and also have a menu where the user can input how they want to restrict the data represented by the heatmap. The user would pick a piece that they are interested in, and then could put in further specifications such as player, decade, region, opening, ranking, etc. We hope to design a page where the user could input their specifications and quickly see a large variety of heatmaps of interest. Weeks 7-9 will be spent working on this interface and ensuring that everything functions well before we do our final presentation for the class.