Jacob Dineen & Mason David
Scripting for Data Analysis
8/12/2018
Final Project Proposal

Assignment:
1. Choose whether to work individually or to work in a team of two or three people. If you wish to work in a team, specify the people that you have talked with to form a team.
Mason David and Jacob Dineen to be working together.

2. Pick a topic of investigation and the data that you will use, ideally from more than one source. The topic could focus on one main data set but also have supporting data. Your topic may focus on a single target topic or person, combinations of them, a comparison of more than one target topic and person, or comparisons over time.
The data may come from any source: those that you have found online, collected from social media, or obtained through other means.
We will be exploring baseball data collected by Sean Lahman and a team of researchers[1]. The dataset is comprehensive and dates from 1871 through 2014. Containing over 20 CSV files, the data ranges from an individual to a team level and is inclusive of player and team personnel statistics at a low grain. We feel that this aggregated collection of data across many sources will be an exercise of some of the key learnings acquired throughout this course.

3. Pick several possible methods of analysis in order to give some initial idea of what analysis you will try. This analysis will be to answer the types of questions that you have worked on for the homework assignments. Since we are not focused on visualization, the results of your analysis can be reported as structured tables with a unit of analysis and collected, summarized, or computed values for those units.

We will explore with a number of different techniques to draw out insights from this dataset, including but not limited to:

1. **Functional programming to read in data in appropriate format. Most of our data appears to be in tabular/structured form at the moment, but we may explore bringing in some additional, unstructured data if it makes sense to do so. This includes use of loops/functions/dictionaries/mapping/arrays/lists. We will be using some/all of the following modules:**
   Pandas/Numpy/CSV/OS/Sys/Sklearn/Matplotlib/Seaborn.
2. **Mapping multiple csv files to each other based on shared keys/indices.**
3. **Time series analysis - Show how statistics/salary change over time.**

4. **Hierarchical grouping of central tendency- Show important statistics by key categories.**
5. **Explore the use of binning categorical/numerical variables to reduce dimensionality/variance.**
6. **Prepare/Clean that data for some algorithmic approach to predict Hall of Fame probability distribution (Logits/Random Forests/Gradient Boosted Trees). Ultimately show which features drive Hall of Fame candidacy.**
7. **Potentially explore regression techniques for predicting salary based on performance. Additional questions will be formulated and reported on throughout the exploratory phase of our project.**

4. If you know of places where you may need help with development, try to list that now. This can range from big things (I want to get information from FourSquare comments) to small things (I'd like a program that helps me to get dates from the documents in my collection and be able to compare dates).
[Potential remaining topics: Social network analysis, geographic locations and maps, getting all Facebook comments. You may request to add to this list.]
At the moment, we look to have all of the data we will need, but that is subject to change. We should be able to read in all statistical data directly from Git. Most of our initial exploratory work will center around preparing our data from a predictive model - We will need to work to understand which features are relevant, which are subject to multicollinearity and which are not relevant. It will be important for us to reduce dimensionality to a point that doesn't tax our ability to compute. Essentially we are trying to maximize/optimize insights based on statistical inference (Feature Ranking/Information gain).

5. State in what way you intend for your project to be larger in scope than either of the homework assignments.
1. **Larger in pure size (Rows/Columns)**
2. **Larger in terms of sources (Many different CSV files needing to be merged together).**
3. **More specific insights to be drawn - Much heavier, more granular analysis.**

6. Based on your plans, you may want to start collecting data.
Initial data is already available to us, although we may try to integrate some form of web scraping once we get further along.

Data can be found at the below link:
https://github.com/chadwickbureau/baseballdatabank/tree/master/core