ETL Project Summary

For our group’s ETL project, we have chosen to aggergate NCAA Women’s Lacrosse Data. Specifically, we’ve chosen to gather and summarize the available game play data for the US Naval Academy’s (NAVY) Women’s lacrosse program.

While there are locations on the web where certain data is available, there is no central repository or API available to extract and analyze off of, thus justifying the necessity of this project.

To accomplish this, first we gathered the game schedule from navystats.com (Data Source #1), which redirects to <http://www.statbroadcast.com/events/statbroadcast.php?t=1&gid=navy>.

We used Splinter / Chromedriver to web scrape the HTML and gather the schedule information.

Next, we wanted to scrape the box scores and play by play information from NCAA.org. In order to do so, we needed to dynamically build the URLs for the game information from the schedule data we scarped. This was slightly more challenging than anticipated, as the NCAA.org used somewhat arbitrary aliases for certain programs to build the URLs.

In order for us to accomplish this, we had to subsequently integrate data from a CSV file where we created a table that allowed us to translate the aliases to the required values and generate the correct URLs for the box score and play by play data. The CSV file serves as Data Source #2 for the extraction purpose of the project.

With the URLs built, which was a major translation effort, as we had to determine whether navy was the “Home” or “Away” team, we were finally able to extract the box score and play by play data from NCAA.org.

After scraping the data using Splinter / Chromedriver / beautiful soup, we compiled it into a Pandas Dataframe as the foundation for future analysis. We gave each match a Game ID, and performed several minor transformation tasks to create a clean working dataset.

Finally, we saved the data out to SQL (sqllite) as our final data loading repository. SQLLite was chosen because we liked the name and it seemed like a reasonable choice at the time, but realistically, we could have dumped this stuff anywhere and as long as we were able to retrieve it later, it wouldn’t have made a major difference. I’m sure this matters more than I’m insinuating here when building applications for commercial and professional purposes, but for the purposes of our project we just picked the easiest one that would satisfy the requirements and went with that.