Alex Beshansky

MIS3640

November 2, 2016

Assignment Three: The Battle of the Wikipedia Pages; Boston Red Sox vs. New York Yankees

Project Overview:

The goal of this assignment is to determine how the Wikipedia pages for these two historic baseball teams, the Boston Red Sox and New York Yankees, were constructed. In sports, teams constantly go through high and low points over the course of their franchise. The techniques used looked at both word structure and repetition as well as sentiment analysis in hopes of understanding how these two teams are being portrayed through one of the most widely used resources online. These two teams are known for the strongest rivalry in all of sports.

Implementation:

In order to compare the two Wikipedia pages, I have developed a three system model that allows me to run the analysis on the three different criteria in an organized fashion. The three criteria that I am examining are; the Boston Red Sox Wikipedia age, the New York Yankees Wikipedia page and comparisons between the two pages. The first step was to organize each page individually in order to run the preliminary functions and collection data. After importing the pages and converting it to pickle file, the dictionary was created. Then, a simple analysis was run to determine basic criteria on the page. I then created a function to determine the highest frequency of words for each page. This would be used to determine the focus of the writer while creating the page. If one word come up often that would not have expected, or comprised of a major role in the history of the team, it would be determined. Another way of conducting this analysis was by identifying a set of key words that I would find as important to the subject, such as team names, World Series championships, and baseball terminology.

Next, sentiment analysis was run on each of Wikipedia pages. Through this, the goal was to analyze how the writer interpreted the history of the two teams. The New York Yankees are a team known for success, having won the most World Series Championships of any other team. They have been the most competitive team over the last few decades and continuously fight their way into the playoffs. The Boston Red Sox have not had as much fortune. In 2004, the Boston Red Sox snapped a drought after winning in 2004, and again in 2007 and 2013. The Yankees have only won in 2009 over the last fifteen years.

One decision I made was to remove all stop words from the list that shows the frequency of all the words. Although this list would not actually pull the first ten words by frequency, this change was to pull the first ten meaningful words by frequency. This allows the analysis to display the words that will tell the story rather than construct the grammar. For the purposes of this analysis, this decision would create stronger end results.

Lastly, I created the third file to run analysis to be able to compare the two pages, Analysis.py. This allows information to be gathered exclusively on the individual pages as well as a unique third page. The process is simplified because of the ability to import the two individual dictionaries into the third python file. This allows the file to not have to run every program, therefore preventing any slow processing that may result. This could have been run individually for one of the files, but would create unnecessary clutter.

Results:

Through the initial text analysis, basic structure was established to create a baseline. The Boston Red Sox Wikipedia article had 2,863 unique words for a total of 13,087, or unique word to word ratio to word of 4.57. The New York Yankees article had 2,425 unique words for a total of 11,057, or a word to unique word ratio of 4.56. These two are incredibly similar ratios. Further analysis would look at other Wikipedia pages to determine whether this event is a coincidence or is within the normal ratio.

Another form of analysis was examining a select few “key words” which we believed would be important to the two pages. The Boston Red Sox Wikipedia page features the Boston Red Sox approximately 225+ times, while mentioning the Yankees approximately 50+ times. On the other hand, the Yankees page features the Red Sox 40+ times while mentioning the Yankees 240+ times. Again, the similarities between mentioning themselves and their opponent are apparent. This brings truth the steady rivalry between the two teams.

The next two words also present interesting findings. The words “baseball” and “champion” are present one more time in the Red Sox’s page. Now “baseball” may be insignificant, but it is interesting to find “champion” more present for the Red Sox. They have won more World Series Championships recently, but overall, the Yankees have won far more. This analysis can provide that the Wikipedia pages are geared more towards recent events.

The next piece of analysis that was examined were words that appeared the same amount of times for each article. This analysis could not create significant findings but a couple were interesting. First, the word “Jacoby” appeared on both. Jacoby, referring to Jacoby Ellsbury, is a played drafted by and spent most of his time with the Boston Red Sox. He was then signed by the Yankees and is now currently playing for them. Jacoby spent most of his time with the Red Sox, as well as a much more successful career, so the original assumption was that he would have shown up more for the Red Sox. Another word, “2009,” is significant because it was the last time the Yankees won the World Series. The original assumption was to see this word appear more times for the Yankees.

The last analysis that was examined was sentiment analysis. The Boston Red Sox produced the following results: {'compound': 1.0, 'pos': 0.101, 'neg': 0.053, 'neu': 0.846} The New York Yankees produced these results: # {'neu': 0.837, 'neg': 0.065, 'pos': 0.099, 'compound': 0.9999}. This analysis does not show much because of the similarities between the two pages. Although the Yankees have had a more successful history, the Red Sox recent success has been enough to limit the differences. This analysis demonstrated that these two articles were written with a neutral and nonbiased point of view.

Reflection:

This project definitely had its challenges throughout the process. The pickle file did not create the most efficient process of analyzing the data, so sometimes it took a while to run the analysis. This analysis was able to obtain findings that would not have been originally perceived based on the common bias in writing, especially for sports. Both of these teams had strong and weak moments, which were not shown through this analysis of a primarily neutral attitude. Moving forward, other sports rivalries could be analyzed to determine attitudes of these Wikipedia pages and how this could have been affected. It was also difficult to use these new resources such as the Sentiment Intensity Analyzer. Moving forward, it would be important to start with a plan to examine multiple sources for a more accurate representation of data. This project has been an exciting and motivating experience to continue to make use of the resources of Python.