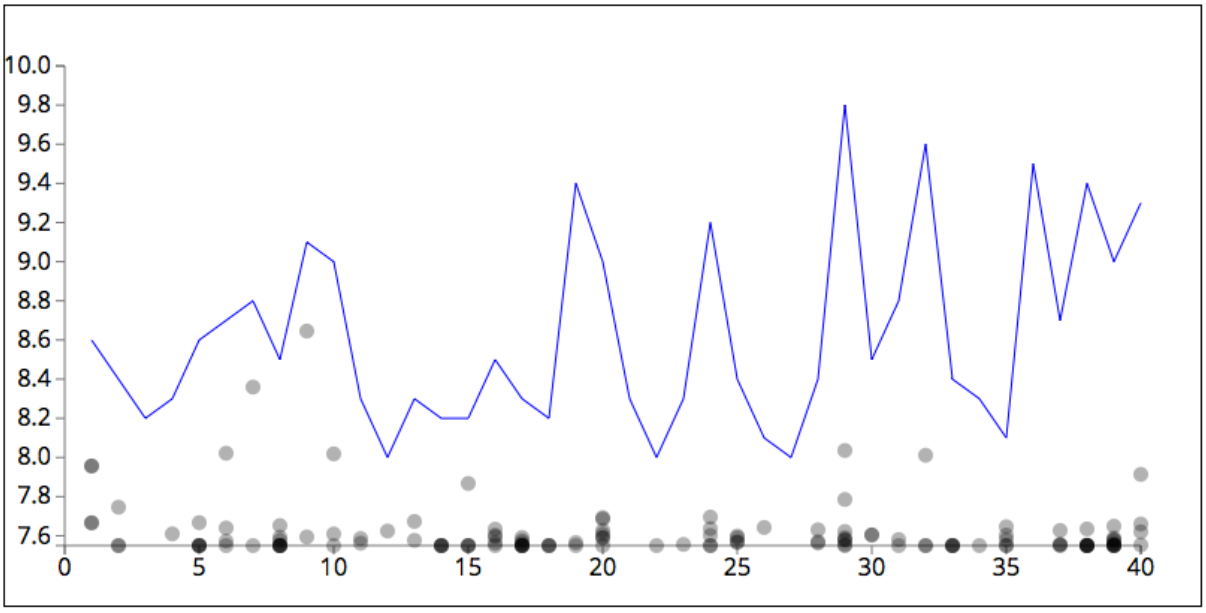
CS 3300 Project 1 Progress Report

Abraham Meller (abm79), Michael Huang (mxh2), and Mahak Goel (mg785)

2.27.15

Since our last meeting, our group has made significant progress on our project. Michael was able to write a parser which parses the Game of Thrones Wiki and classified the importance of a every character who died. The way the parser worked was by counting the number of occurrences of the character's name in all episode summaries up to their death, and dividing that value by the sum of the occurrences of character as a means of normalizing the data. We cannot just the number of occurrences of the characters name as their importance measure because there are important characters who died early in the show, but they have fewer name occurrences just because they were a part of the show for a less amount of time. We shifted from episode appearances to number of times their name was on the wiki summaries because characters who died early on would have a high number of episodes they appeared in to total episodes ratio, even if they weren't important. Our parser returns a json object with objects for each episodes which contain the names of the people who died and their importances.

Our visualization currently looks like this:



The x-axis has the episode number and the y-axis has the IMDB rating as well as character importance. The blue line shows the IMDB ratings and there is a dot for each character who died and the higher up the dot is, the more important the character was. There is definitely a correlation between the importance of the character who died and IMDB rating for that

episode because there are spikes in the rating for episodes in which there is a dot, a character who died, that is high on the graph, indicating that the character was important.

However, one clear problem we have is the scale for the character importances. There are a lot of characters that have been assigned zero importance because they don't appear in the summary. Also characters who died early tend to have higher importances because the ratio of occurrences of their name to occurrences of every name is higher, just because there are less characters introduced so far early on in the show, and later in the show many more characters have been introduced. For example, ned stark is the most important character according to our scales, but not necessarily the most important character in the entire show.

We now have to figure out how we can re-scale the character importance data so it's more accurate. We're trying to work with our existing measure, of occurrences of the characters name in the summary, and are also looking into other data online that shows the number of the character's lines in the show, and the character's total screen time as a measure of the character's importance. We are also working on making our visualization more aesthetically pleasing and more with the Game of Thrones theme. Other than that, we feel like we're in decent shape because we our visualization shows some interesting information and correlation and we just have to enhance it.