Jackson Dean

Final Project

5/1/17

**Introduction:** Few would doubt the enormous social impact of hip-hop music and rap, especially on younger generations in the United States. Unfortunately, due to media coverage and generalizations about the genre’s audience, many people believe that the genre has a negative social impact. For my project, I wanted to investigate the claim that rap contains mostly negative messaging. Knowing that it would be difficult to analyze entire songs for prosocial messages, I decided to focus on individual lines and to look at the sentiment polarity for each rather than looking at the meaning behind entire songs. Therefore, the goal of my project was to investigate whether lines in rap songs are averagely positive or negative on a polar scale. I hypothesized that an average of sentiment polarity scores for a collection of lines from rap lyrics will be greater than 0, indicating that the average line in rap has a positive sentiment polarity.   
**Background:** There is a significant body of research in this field. The most influential in my project was a thesis written by Terrell Green, “Rap Lyrics, Music Videos, and Themes of Sex, Violence, and Prosocial Behavior”. In this thesis, the author did an analysis of many rap songs, examining the messages the songs conveyed. He studied a variety of music videos and lyrics, specifically looking for messages about sex, violence and prosocial behavior. Potentially surprisingly to some, he found that prosocial behavior was actually the most common theme of the three. For this reason, I expected to find that most lines in rap songs would have a positive sentiment polarity. I also conducted some research about the best way to find sentiment polarity of text. I settled on VADER sentiment analyzer because it is built into NLTK and it was designed for tweets and therefore likely performs well with some slang common in rap songs. Lastly, I needed to find a source for my rap lyrics. I looked at a number of lyric websites, but many of them did not allow me to sort by genre or had a very incomplete list of lyrics. I finally settled on OHHLA.com, which exclusively has rap lyrics and a basic HTML format that I could parse with beautiful soup.   
**Design:** The first task that I attempted to resolve was gathering a large corpus of rap lyrics from a variety of artists, subgenres, and years. I wrote webscrapping methods to accomplish this. There was one method for each layer of the website. For example, the first method took in the URLs for pages that had groups of artists and returned a list of links to the artist pages. The next method took in those artists and returned the albums, then the next returned the songs and the final method returned the lyrics. I wrote the methods this way so that I could experiment with different artists, albums, and songs individually and to keep the code from being one difficult-to-read function. I wrote all of these lyrics to a text file so that I didn’t have to regather them every time, which was a time-consuming process because of the hierarchy on the site. For my analysis, I ran the text file through NLTK’s tokenizer and then the VADER sentiment analysis.

**Concerns:** The first major concern I had was finding a website that had a large variety of easily-accessible rap lyrics. Most lyrics websites are formatted in a way that makes them difficult to access with a python script, and the website that I finally settled on was easy to navigate but required going through many pages for each song which was time-consuming. I solved this problem by writing all the lyrics to a text file the first time the program ran, so I could easily access them for future experiments. Once I was able to collect the lyrics and run my analysis, I found that the sentiment analysis didn’t live up to my standards. Although it worked well on lyrics that had very obvious positive or negative vocabulary, much of the meaning in rap is inferred and hidden behind slang or pop culture references. The data demonstrated this phenomenon as the polarity scores were largely neutral rather than being positive or negative. The best way to deal with this issue would be to design my own sentiment analyzer specifically for rap lyrics. Overall, I am happy with my program, however, given unlimited time, I would like to design my own detailed sentiment analyzer and to test a variety of different variables. It would be interesting to look at the differences in sentiment polarity between songs from different subgenres or time periods. I suspect that another reason I found such a neutral sentiment polarity is because my corpus included very obscure songs, which may not reflect the genre as it influences popular culture.