Bryan Werth
2/24/2016

Reflection

Project Overview

For this project, I decided to work with a stand up comedy special transcript that I found online. Initially I wanted to use it to randomly generate knock knock jokes, but I decided that I would want more time to produce a program of that complexity. Instead, I analyzed the word usage in the transcript. To do this, I created a histogram using every word used by the comedian. I also found the average sentimentality of the special by using the sentiment function. By doing this, I hoped to learn more about the most common words used by a comedian, the word variety, and the positive or negative tone of the comedy.

Implementation

The first major part of implementation involved filtering the transcript that I found online. I knew I wanted just the words, so I split the text file to isolate the text from the html. I also stripped out most punctuation by replacing each character with an empty string and lowered any capital letters using the lower method. I also created a list of the words used in the transcript by stripping the text. I defined data and datalist as global variables for future use by all of the written functions.

Next, I defined a histogram using datalist and sorted it by word frequency. I defined a function called fill_hist that defined an empty dictionary, and used the get method to return the value associated with each word. After this, I defined another function, sort_hist, that turned the dictionary into a list and used the sorted function to sort based on the value associated with each word. Finally, I calculated the total number of words that were only used once and compared that with the total number of unique words used.

To find sentimentality, I defined a sentimentality function to find the sentiment polarity associated with the lines in the transcript. First, I split the text by line breaks, creating a list of each individual line. I looped through this list, checking for empty lines and lines with a sentiment polarity of zero. I averaged the remaining lines together to get an average sentiment polarity.

Results

The most interesting thing I discovered about the transcript was that Robin Williams used a lot of different words. In total sixty percent of the words used were only used once. The only words that were used repeatedly were curse words and transitional words such as the and and. This does make sense considering how much Robin Williams jumped around during his stand-up. I would expect the word usage to be very different for other comedians.

The sentiment of the stand-up was about what I expected it to be. Comedy can be negative and insulting sometimes, but in general, Robin Williams was not known for that. As such, I expected the sentiment to be somewhat positive. I found the sentiment polarity to be .58. This was a little lower than expected, but it made sense given the nature of comey.

Reflection

This project went pretty smoothly for me on average. I was a little disappointed that I had to back down from my initial plan, but I was interested in what I eventually came up with. I was confused about unit testing my program because it was so dependent on global variables. I wish

I had had the opportunity to talk to a professor about that. The process I took was very different for this project because there was not a predefined structure for the program. I probably could have been more organized in program flow and structure, but overall I am happy with what I was able to produce.