Serena Chen

Youtube Text Mining

Project Overview

I decided to mine the most recent comments off the most popular videos on Youtube. Youtube comments have a reputation of not being particularly well thought out, so I wanted to see if the statistics reflected how the comments' reputation. As a result, I looked at all the comments in English, and looked at how polarized they were, what words were most commonly used, and how many words were misspelled/were not real words in the dictionary. After starting the project and realizing how many non-English comments there were, I decided also to see how many comments of different languages there are.

Implementation

To read in the comments, I used the Youtube API. I grabbed the list of top videos in each possible category (the top videos for all categories are music videos), and went through the first 20 parent comments of each. If the parent comments had replies, I would log the replies as well. Even though I didn't do much with the difference between categories, I kept them all in separate files, which is why I have ~70 files.

I read these comments in in Analysis.py, and separated the comments by language using the langdetect module. For polarity, I used the sentiment analyzer, which returns a value between -1 and 1, where -1 is most negative, and 1 is most positive. I simply added up all the sentiments for the english words and divided it by the number of words for the average. I did something similar for extremity, except I added the absolute value of the polarity in order to get how strongly Youtube commenters tend to express. For word frequency, I fed in all the words to a dictionary. Easy Peasy. For the number of misspelled vs. correctly spelled words, I compared each word against the word lists from the Moby Project.

Results

The results were pretty surprising to me; according to my analyses, Youtube comments aren't nearly as bad as I thought they were. The first thing that surprised me that messed with my data was the amount of comments in languages other than English. According to my results, a good 1500 comments or so were not in English out of ~18000, with comments that had no alphabetic characters defaulted to English. Out of the comments that were in English, 7.5% of them were misspelled, which is a lot lower than I anticipated. The results may be skewed by words that are misspelled, but still are valid words in the English dictionary, and grammar mistakes, which Youtube comments are also known for.

The polarity and extremity of the comments were also surprising. I didn't expect anything in particular from the average polarity since some comments would be positive and some would be negative.However, I was expecting the polarity to be greater than .5, with people who have either very positive opinions or very negative opinions, but the extremity was only around .2.

The most common words in Youtube comments were pretty generic; most of them were the regular common English words. There were a lot of first person pronouns; ‘I’ is the second most common word, and ‘my’, ‘me’, ‘we’, and ‘I’m’ all made the top 100 most common words. There were definitely lots of opinion words (‘so’, ‘good’, ‘nice’, ‘cool’, ‘best’, ‘much’, ‘ever’, ‘great’) and words relating to the Youtube platform (‘video’, ‘song’, ‘movie’, ‘watching’). The only two words in this list that aren’t English words are ‘lol’ and ‘u’. Finally, the 100th most common word used in Youtube comments is ‘first’, showing Youtube culture and people’s sense of Internet superiority.

Reflection

I think it’s cool that I got to implement many different ways of doing text analysis. I feel like I didn’t do a lot of deep text analysis; I spent most of my time trying to figure out what comments were relevant (English words), and spent more time on deeper analysis like levenshtein distance, although with such a big data set, I’m not sure how I could use that kind of data in a meaningful way. I also feel like my analysis doesn’t give me a good picture of what the Youtube comments are really like. I’m not sure what to analyze for to further clarify Youtube comments; the picture I have right now shows very generic statistics without any details.