CS 410 Project Progress Report

Author/Captain: Rustom Ichhaporia (netID: rustomi2)

Progress Made So Far

So far, I have managed to figure out how to query the text data from both my Apple iMessage account and my Facebook Messenger account into a pandas DataFrame. For the purposes of the project so far, I have only been using the iMessages to simplify the analysis. I thought it might be more interesting to attempt to conduct a topic modeling study instead of the sentiment analysis I had previously planned, so I attempted the topic modeling approach using Latent Dirichlet Allocation. After preprocessing the text using NLTK, I used the gensim library to implement the topic modeling and plotted the clusters of most important words in word clouds. Unfortunately, this clustering did not seem to be very effective. I have come to learn that LDA might not be a good model choice for very short text documents, which is the case for my text messages.

Remaining Tasks

In lieu of that, I will either implement a different model for topic modeling, default back to my previous plan of sentiment analysis, or come up with an entirely new analysis. I will ensure that I dedicate a sufficient amount of time to the analysis to demonstrate effort no matter which approach I take.

Challenges

So far, I don't see any significant challenges being faced other than the poor choice of modeling". I think I may de-scope using text sources from multiple social media sites due to complexity with data parsing in different formats.