April 24, 2017
CMDA 3654 T/TH

Project Part 2

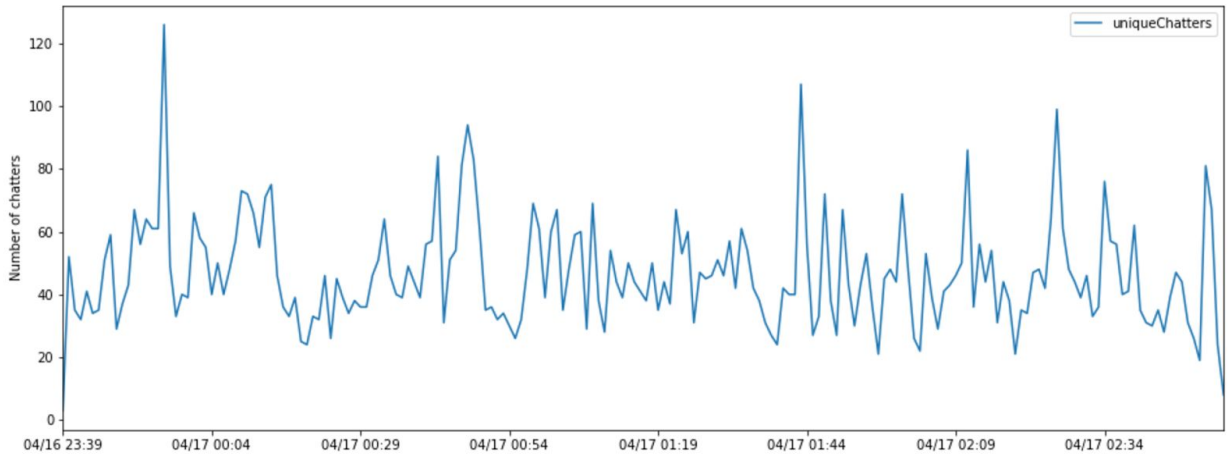
1. Our team name is Twitch Scrapers, because our project revolves around scraping information from the website www.twitch.tv. The members of our group are Jacob Smith, Jordan Deang, and Kevin Jiang.

2. Twitch is a website that hosts “channels” of users streaming themselves playing video games. Each stream is accompanied by a built-in chat, where viewers can communicate with the broadcaster as well as other viewers. Some streams are watched by thousands of people at any given time, resulting in a massive amount of data being produced in the form of chat messages. Twitch provides an open source API for developers to use to make calls against in order to retrieve information from the website using languages like Python.

   Changes from Part 1:
   Due to the nature of the website, it would be very difficult to distinguish between conversations of chatters. Each chat message is independent of all others, and there isn’t any mechanic for linking or checking if messages are directly related. For the time based questions from Part 1, we chose to operate upon intervals of 1 minute time frames. For instance, instead of asking “What percentage of viewers are active in chat”, we would ask “What percentage of viewers are active in chat between 10:30pm and 10:40pm”. This will allow us to more easily produce time based visualizations of our findings. To find these values, we will convert the list of every user who chats a message into a set, thus removing duplicate names and creating a unique chat user list. Then, we will create sets of users in given timeframes and compare that to the total number of chat users.

3. Listing of initial findings
   All of these findings are based around one VOD, a stream by the user ClintStevens. We wrote our code in such a way that other VODs can easily be imported and analyzed to find the same information found below.

   Number of unique chatters in the Clint VOD: 1459
   - To determine this number, we created a list using the “from” column of the csv, which contains the username of the person sending each message. We then converted the list to a set, removing any duplicate names from the data set and found the length of this set. We also created a column that stored the minute of each message, then grouped by this value and graphed the number of unique chatters in each minute of the VOD, shown below:

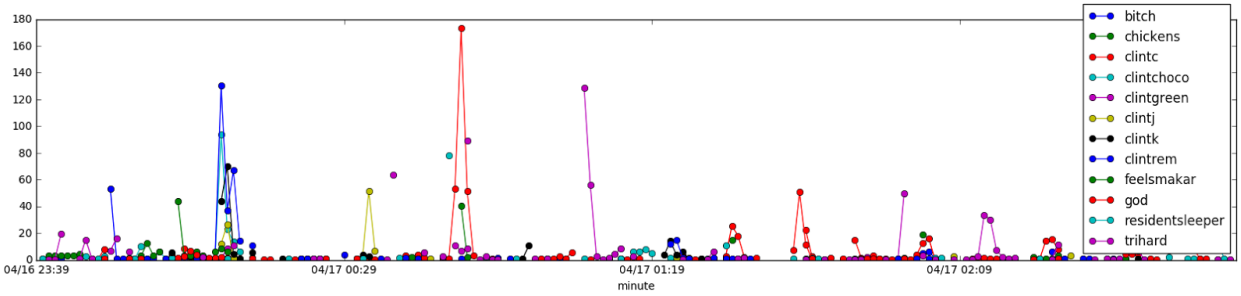


- Originally tried unsuccessfully to iteratively count the number of unique chatters
- This same file contains the code to analyze two other vods, both from formatted Twitch csv files created by running the processing notebook script.

Number of unique chat messages in the Clint VOD: 8491

- We first read in the “clint” csv data into the program and then removed the irrelevant punctuations. Next, we removed any duplicate chat messages that were not unique to the data set by using the set data structure to perform this operation. And lastly, we outputted the number of unique messages which resulted in the amount displayed above.

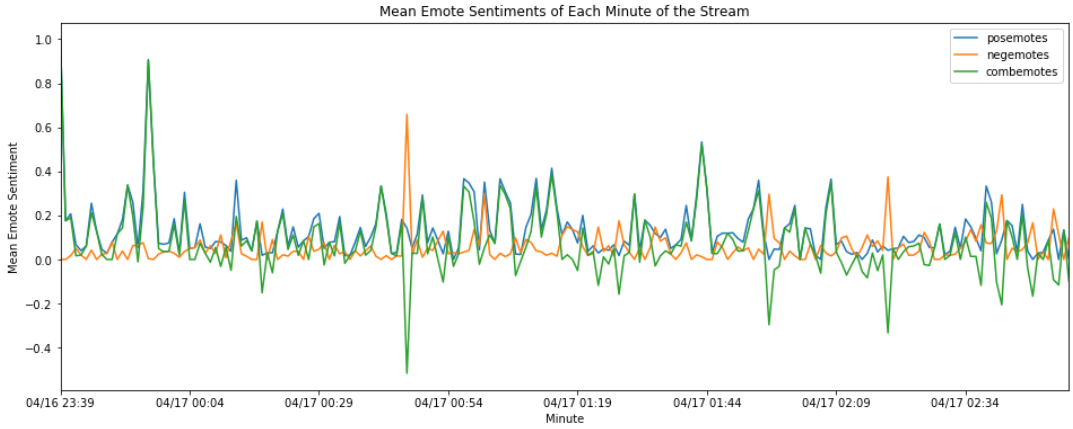
A broad generalization of the words used in the stream visualized as shown below:



Sentiment Analysis on Clint VOD:

- We conducted sentiment analysis on the processed messages of each minute in the VOD using the positive and negative lexicon used in the class. This showed that chat was mostly negative. Most negative sentiment came from use of the word “cancer,” and much of the positive sentiment came from people saying they love Clint. A time series of the sentiment shown in the figure below.

- We also created our own lexicon of positive and negative emotes since emotes are a large part of Twitch chat. This showed that emotes were generally positive. This contradicted some of our results from the basic sentiment analysis. The time series of the data is shown in the figure below.



- Computing the correlation between the basic sentiment analysis and our emote analysis shows that there is a very slight inverse correlation. This leads us to believe that the vocabulary used by Twitch chat is very complex and may require a combination of a normal lexicon and emotes to determine sentiment. However, the moments of high emote sentiment seemed more closely correlated to exciting moments in the VOD rather than the normal lexicon.

4. Contributions
    a. Jacob Smith - Wrote up the changes from Part 1, created the activeChatters document that analyzes chat user activity.
    b. Jordan Deang - Performed sentiment analysis on Clint Steven's VOD. Created a emote lexicon to analyze a correlation between emotes and sentiment.
    c. Kevin Jiang - Implemented the unique-messages file that analyzes the number of unique chat messages in the recorded stream; created the TD-IDF matrix that represents the relationship between the words used for each messages with the corresponding dates, thereby determining its importance; also included various other minor analysis data points located in the td-idf-twitch file such as boring events during the given VOD timeframe.