April 17, 2017
CMDA 3654 T/TH

Project Part 1

1. Our team name is Twitch Scrapers, because our project revolves around scraping information from the website www.twitch.tv. The members of our groups 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.

   The goal of our project is to analyze the chat accompanying the VODs of streams for a variety of information. Some of the questions we intend to answer include:
   - What percentage of viewers are active in chat?
   - What's the average sentiment value of the chat messages in given channels?
   - How many messages per second can be expected given the number of viewers?
   - How many moments of excitement are in the VOD based on increased activity of chatters?
   - How many unique messages are chatted? (Many messages are copy pasted from other users)

3. The data our group has gathered so far is the result of making various API calls to the Twitch servers. Calls to the API return a list of JSON formatted data containing information including the content of the message, user who sent the message, timestamp, and other useful information. The calls only return messages within a 30 second time frame, so we used a loop to make calls for every 30 seconds of the video to get the chat messages from the entire duration. We then used this list and associated information to create a Pandas DataFrame of each JSON key and its value for every message in the chat. Once the DataFrame was created, we used the methods learned in previous assignments to clean the data, removing unnecessary characters and creating a new column of the cleaned chat strings. This DataFrame was then converted to a .csv file for submission and further analysis in other Jupyter notebook files.

4. Member contributions:
   a. Jacob Smith – Wrote the PDF, contributed to development of the idea and assisted with the methodology behind retrieving the data.
   b. Jordan Deang – Responsible for scraping the actual data using Python and formatting it into Pandas DataFrames and eventually .csv files.

c. Kevin Jiang – Analyzed the API for Twitch streams, scoped out data files for documentations related to Twitch, contributed to the development of ideas pertaining to the data being read into Python.