# Documentary Social Dilemma Re-Tweets Data

By Mike Scheibel

**Data set source:**

I found this data using Tweepy and Twitter API. I searched for re-tweets from the official account @SocialDilemma for the Documentary on Netflix. After many references from friends, I decided to watch this movie. It was very interesting and wanted to gather others opinions about the documentary through Re-Tweets.

**Data set description:**

The data was scraped from Twitter using Tweepy form the Twitter API exercise. I then wrote the data to a csv file named ` SocialDilemma\_tweets.csv` and it has 3 columns: The full text of the re-tweet; when it was created, and user screen name of the person who did the re-tweet. Reading through some of the text it was interesting. I think the limitations for the data might be the need for more columns of info like location.

**Descriptive statistics:**

Total Tokens: 3635

Unique Tokens: 1611

Average Token Length: 6.11

Lexical Diversity: 0.44

Top 10 words: ('social', 63),

('watch', 62),

('media', 44),

('people', 44),

('like', 43),

('us', 37),

('watched', 36),

('documentary', 29),

('watching', 29),

('see', 28)

I might want to go back to cleaning and pull out the links and other account handles that I removed from my first run of cleaning.

**Interesting questions:**

What is the location of some of these followers who re-tweeted?

What is the overall sentiment of the film, and if there is a even divide, what keywords differ between sentiment groups?

See if there are any common hashtags or emojis in the text?

Gather the hyper links in the text and maybe scrape information from the pages?

**Approach for a single question:**

What is the overall sentiment of the film, and if there is a even divide, what keywords differ between sentiment groups?

I would use the Sentiment Analysis and treat the re-tweets almost like a movie review and see if they tend to be more positive or negative. I would first try to find the polarity using WordNet and finding synonyms and antonyms. If I did not get the desired results, then I would turn to Naïve Bayes approach.