Acquiring and Exploring the Data

The first step was to acquire the data. I used the PRAW library which is a Python wrapper for interfacing with Reddit's API. There was a limit in the amount of data I could extract per token so I was only able to extract 1000 posts per category at a time. To combat this problem I chose several categories to work with:

- **.top()**
- **.rising()**
- **.new()**
- **.hot()**

I explored the dataset in a few different ways, but the column I had most interest in was the 'body' column, which contains the meat of the sample document. Upon selecting this column, I applied a few analytical techniques to explore trends in the data.

My first strategy was to clean up the dataset to tokenize the text using regular expression tokenizer (regexptokenizer) and remove stopwords, lemmatize the text, and to exclude redundant words that did not contribute much to the overall context. I was also curious to see positionally tagged words. I used NLTK to achieve these steps.

The exclusion of certain words without much meaning to the overall context was an important step to eliminate outliers in the frequency distribution of common words because it is customary for certain subreddits to begin the document in certain ways. For an example, r/depression often started with "Anyone else feel like [...]", which may be an important feature during training, but did not help me explore the data. For any rows with missing values, I simply filled the gap with the .fillna() method. Then, I proceeded to acquire the frequency distribution for the 20 most common words used from each subreddit's datasets.

I was happy with the list of words I got back once the exclusions have been set. It was time to dig a bit further. I used NLTK's concordance finder to find sentences containing the word "want". This decision was based on purely intuition, and the only logic that backed this was my curiosity to find out what it is people with certain disorders 'want'? Then, I used a quadgram collocation finder to figure out which four words were commonly grouped together.

I repeated the above steps for all six datasets from the subreddits. Once completed, it was now time to glue them together into a master, labeled dataset which would then be converted into the correct input formatting for BERT.