Special Issues with Text-Related Data

- Text data is inherently messy.
- Firstly, computers can only process numbers, meaning that any text data needs some way to be converted to numbers.
- Some words that commonly appear in text have very little influence on the text. We call these **stop words**. (Such as “and”, “so”, “to”, etc.)
- A computer treats “cold”, “Cold”, “cold;” as separate objects, even though they are the same word.
- Additionally, there are rare words that may not appear very often but may have an influence on the text data.

Data Wrangling Completed

- Converting all words to lowercase
- Removing punctuation
- Counting the number of times a word appears and adding it as a feature to each post
- Removing rare words
- Adding an indicator to clarify the difference between “subreddit” as the feature and “subreddit” in the post
- Joining the training and test data
- Adding the word columns to the original data