Project Overview:
I focused on using Reddit to gather data, specifically focusing on the subreddit r/Coronavirus. I organized the data into lists and dictionaries to make it easier to manipulate and change as needed. I hoped to get a better understanding of how seriously people are taking this pandemic and what kind of language is being used surrounding this topic. I was able to create a dictionary that shows the words being used by people on Reddit and the frequency at which they are being used. I also ran a sentiment analysis on the comments section of a select group of submissions to see if the language being used was positive and supportive, or more negative and argumentative.

Implementation:
Looking at the code from a high level perspective the majority of my work was initially just organizing and cleaning the data I pulled from Reddit. This involved creating functions to clean the data to remove punctuation, fix capitalization, etc., as well as breaking down the data into usable data sets. For example, I wanted to look at the frequency with which words were being used in the titles of Reddit Submissions. I did this by pulling the top 300 submissions from the r/Coronavirus subreddit.

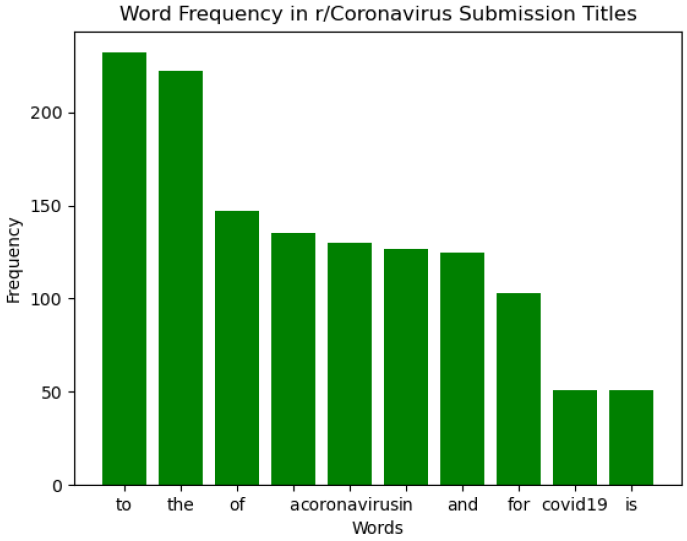
Originally I had stored the data so that each item in the list was a title. I decided to further break down the titles into individual words to make it easier to count their frequencies when creating the dictionary. When creating the visualization for the dictionary, I decided to only focus on the top 10 most frequent words, as using data for all ~1000 unique words in the titles would have been very hard to visualize well. The top 10 most frequent words were stored in a list with tuples. Each tuple contained the word and the frequency of use within titles. I decided to convert this list back into a separate dictionary just for the top 10 words because it was easier to use this method of storage and organization to create the x and y axis for the graph. The x-axis is just the dictionary keys and the y-axis is simply the dictionary values. This was much easier for me to understand and work with than the original list with tuples.

For the sentiment analysis, I referred to other projects I have done in past classes as well as resources I found online. I chose to use the TextBlob module as opposed to the Natural Language Toolkit because it was something I was more familiar with. To run this analysis I decided to focus on the comments because I thought it would give a better indication of people's feelings and reactions towards the Coronavirus. I first had to pull the comments from individual submissions and then compile them into a list that would process all the words and come to a consensus about how overall positive or negative that group of joined comments was.

Results:
Resulting from this project the word frequency analysis showed that the majority of frequently used words were common words like "the", "are", and "is". Not surprisingly, in the top 10 most frequent words was also "coronavirus" with ~130 mentions. I organized the printout to show the top 10 most used words with the keys and values from the dictionary I created printed beside each other in list format. The total number of words was ~6500 and the number of different words was ~2100. These numbers will fluctuate a bit because the code pulls new data each time it is run.

```
Words with 10 Highest Values:
Keys: Values
--------------
to  : 232
the  : 222
of  : 147
a  : 135
coronavirus  : 130
in  : 127
and  : 125
for  : 103
covid19  : 51
is  : 51


The total number of words is 6514.
The number of different words used in these posts is 2100.
```



The sentiment analysis showed overall negative sentiment with polarity of almost 0.05 and subjectivity of about .443. These numbers will fluctuate a bit as the code pulls new data each time it is run. Polarity is measured on a scale of -1 to +1. This means that the sentiment in terms of positive or negative was surprisingly neutral if not slightly positive. I expected the polarity to be much more negative when taking into consideration how scared a lot of people seem to be, how aggressively the media is reporting on what is now being called a “pandemic”, and how angry a lot of people have become with regard to how leadership is dealing with the issue. The words being used were mildly subjective on a scale of 0 to 1, but leaned more towards objectivity. Most comments were not as emotionally charged and opinion-based as I was expecting them to be when taking into consideration that comment threads are usually users voicing their individual opinions on the matter.

```
Sentiment(polarity=0.04885338303424477, subjectivity=0.4427792047631729)
```

Reflection:

Overall, I think this project went pretty well. It took me some time to figure out what data I wanted to pull, and then once I had it, figuring out what to do with it. Initially, I was only going to do the word frequency analysis for the project, but after completing that portion I felt like there was more work that could have been done. I would have liked to run a text similarity analysis as well, but was not sure how to implement it using reddit data. I think the addition of the sentiment analysis made the project a bit more interesting, although I did not get to do everything I would have liked with that either. I think if I had a better understanding of the different python modules going into this project would have saved a lot of Googling time, and given me a better direction in terms of what I wanted to achieve with this project.

During this project, I gained a better understanding of how to use APIs and the process of text mining. I think I did a good job of creating functions to simplify the code and make it easier to understand, although I could have definitely created more functions to streamline and optimize the process further. I can imagine there are much simpler ways to accomplish what I did for this project with a lot less code. With that said, I think my code could be used for other subreddits besides r/Coronavirus with very few changes, and could analyze other aspects of submissions besides the titles without much modification.

I could have improved this project by filtering out common words in titles like “the” and “is” for the word frequency analysis so that I could gain a better understanding of what the titles of these submissions were actually about. My visualization could have also been cleaner in terms of labeling and overall aesthetic, and I could have produced a few more to represent different aspects of the data and create a more complete picture. For the sentiment analysis, a lot of the code I used was taken from internet sources and modified to fit my needs. I did not fully understand all the code I was using, and therefore, could not fully modify the code to do everything I wanted for the sentiment analysis. I would have liked to create more visualizations for the sentiment analysis as well.