**Capstone 2**

The world today has been greatly affected by COVID-19. The virus was discovered in 2019 and appears to have an East Asia origin. The timing of the virus’ inception (roughly around the Chinese new year) made its global spread a lot faster, and catastrophic for global economies.

The effects of this pandemic have ranged from mild inconveniences to cities and towns to complete lockdowns and hundreds of thousands in cities such as New York. In the United States alone there have been over 1.5 million confirmed cases with some 5 million cases worldwide at the time of this project.

As the world continues its contingency measures, namely social distancing and enforcing the wearing masks whenever out in public, most people are not sure what the future will look like. Countries that were infected early on and had gotten the spread under control are not facing a second wave. Some countries like the United States are still in the midst of the first wave. This means that the future is still very much uncertain, but what we do know is that for things to return to normal, there must be a vaccine. But while we wait for a vaccine it is possible that there could be other measures that we take as a people to protect ourselves and clues from battling other viruses could point us to faster vaccine development.

This project could be used by virtually any government, particularly a government with scientists who are in the race to develop a vaccine and developing better ways to protect people while we wait. Private companies who are developing products for help in the battle with COVID-19 would also be benefitted from the findings of this study.

The dataset was acquired from Kaggle and is currently being used as a competition of sorts for Data scientists and Artificial intelligence enthusiasts to build models and gain insights.

The data was arranged in several folders which included the archive folder, bioarchive, common use folder, custom license folder, and noncommon use folder. Each folder held data with information that was thematically similar to the others.

The individual articles were in json files with each file containing elements such as the title, metadata, the body, references used, and the bibliography. This meant that the first step of wrangling involved separating the body text from each file and removing the citations that were used inside the body. This was done iteratively and then the data was stored in a dataframe. From here, the data was preprocessed using the nltk and gensim libraries. These were used to remove stopwords (common words that appear in all the articles but do not offer any real information), remove special characters such as periods, questions marks, quotations marks, and the like.

After this, the documents were analyzed to see the most common words used in the form of 1-grams, 2-grams, and 3-grams. These were visualized using a graph. The documents will then be processed using a tf-idf to see what the most valuable words were as well as undergo an Latent Dirichlet Allocation to get a better idea about the topics that are in the corpus.





