**Requirement Document**

**Unstructured Data Modelling/Analysis**

Yuehua Yang, Tanhya Chitle, Jerry Manning, Bhanuprasad Tudy, Patrick Durkin

**Dataset:** The dataset we use is the CDC news with COVID-19 related articles since March 26. It includes 4609 rows and 7 columns. The seven categories show the authors, title, publish date, description and text information of each article. It also includes the url link for each article.

From the data, we aim to explore the text dominated fields to come up with analysis on patterns in the data. We also have an idea to merge this dataset with stock market data to understand the influence of news on the stock market.

Thus, the other dataset we want to use is the stock market data with the similar time period. The data is for the past 4 months with stock prices for each day the market is open. It includes

**Functional Specifications:**

Given that most of our data is text data, the word count function from pyspark could be a good starting point for us to do a basic analysis and understand the entire dataset in a brief way. Building on the word count, we aim to dig deeper into text analysis using sentiment analysis and word embeddings.

Using sentiment analysis we can obtain a comparison bar graph of the positive, negative and neutral words used and also make word clouds of the most frequently used words overall and in the specific sentiment. This would help us understand the overall sentiment balance of the dataset along with information on main words related to those sentiments. This sentiment analysis can further be compared to the stock market data to predict the effect of the positive and negative sentiment over the stock market.

By using word embeddings, we will create a vectorial representation of the word corpus with a model like Word2Vec using Bag of words or Skip gram. Using these word embeddings, we can create a relational network of words to understand similar and most related words. We can also find out the main topics the media has focused on. The model can also be extended to combine with a machine learning analysis to classify news into specific categories like sentiment or any other category we can come up with while processing and analysing the data.

Further, we can visually compare different countries on how actively they are fighting the disease using the sentiment analysis results on what countries are more populated with positive or negative sentiments.

In conclusion, we think this is an extremely rich dataset which can lead us to uncover interesting patterns on the media’s take towards the current global crisis. This being said, since this is a primarily text based dataset, we are concerned about the extent to which we will be able to clean the data to get accurate and precise results.