**Design Doc | *Due April 25, 2020***

**Team Crestone Peak**

* Jason Engel, Zach Johnson, Natalie Garibaldi, Julian Beaupre, Yubing Yang

**Components**

**Dataset:** [CBC News Coronavirus/COVID-19 Articles](https://www.kaggle.com/ryanxjhan/cbc-news-coronavirus-articles-march-26)

* The data contains the authors, the title, the publish date, the description about the story, the main story, and the url.
* "Has the news media been overreacting or under-reacting during the development of COVID-19? What are the media's main focuses? How is the news correlated to public reactions or policy changes? You might find many insights with more than 3,500 CBC news articles."

**Amazon AWS:**

* An Amazon EMR cluster simplifies running big data frames, such as Apache Spark and Apache Hadoop, in Amazon Web Services (AWS). Its purpose is to help process and analyze data. WIth Amazon EMRs, it is easy to transform and transfer large datasets into/out of other AWS databases. An example of these databases is Amazon Simple Storage Service (S3). We will upload our data to the Leeds School of Business EMR cluster, hosted with Hadoop, on AWS, into an S3 bucket. Once uploaded, we can run our models on our data.

**Apache Spark:**

* Apache Spark is a fast and general-purpose cluster computing system. It is originally written in Scala and can be integrated with other programming languages such as Python, Java, R, and SQL. Within Apache Spark, there is PySpark, which is the Python API written in python. For this project, we will be using PySpark.

**Components Interface**

* Our dataset is first downloaded from kaggle and stored locally
* Then, using SCP we are able to copy the data into our Leeds Cluster
* From here, we can ssh to the cluster and copy the data to the S3 bucket via the Cluster’s permissions
* We can port forward to the server locally to access our notebook with a gui rather than command line, and we can submit our code/jobs via there

**Technical Requirements**

To successfully accomplish this project, there are several technical requirements that must be considered and completed. Technical requirements include:

* Merge/clean the dataset
* Ingesting the dataset to S3 bucket to easily run
* Impute values and dummy code if necessary for the pipeline
* Pipeline data and run spark job on cluster

**Functional Requirements**

Functional requirements can benefit and add value to a business. In the case of our data, potential “businesses” could be governments gauging the feelings of their citizens to make a decision on stay-at-home orders and lockdowns. These requirements include:

* Making predictions based on our research
* Computing the data and finding the target
* Deploying model and visuals

**Description of math**

Emotion over time

After our analysis and processing of the language, we will find our ‘score’ for each article and their given emotions. From here, we will look at how the data is distributed and scale or modify it if necessary. A value will be given so we can show the range of emotions for various categories in a time-series visual to see the progression of the writings.

**Description of models**

Use Sentiment analysis to interpret and classify emotions within the text of the articles. Give articles emotional ratings based on text. Natural Language Toolkit (NLTK) used to classify articles and parts of articles into a predefined sentiment. Tokenization used to split strings into smaller tokens to make the language accurately processed. Removing stop words within text to provide more accurate scores and minimize noise within the contents of the articles. Other noise that needs to be removed besides stop words are special characters, mostly used for punctuation. In order to execute the sentiment analysis, the tokens need to be converted to a dictionary. Once we have the dictionary the next step is to split the data into training and testing sets. From here we can build and test the models.

The sentiment analysis model used is based on the six stages of grief: denial, anger, bargaining, depression, acceptance, and finding meaning.