Project Proposal
Data Mining and Clustering of COVID-19 Literature

Team Members:

Abhilash Hemaraj	hemaraj.a@northeastern.edu	001305283
Niyati Chopra	chopra.ni@northeastern.edu	001443559

We will assign individual responsibilities as we go along with the project and work collaboratively.

Problem Description:
With the currently rising widespread cases of COVID-19 across all the countries in the world, there is also an increasing rise in the number of literary studies(research papers, scholarly articles, studies, etc.) that is being published everyday. This is why there is a growing need for the applications based on text mining, natural language processing, unsupervised machine learning and data mining techniques. The increase in the number of these literary studies is making it hard for the medical community to keep up.
The input and output is explained more in the datasets section.

The main aim of the project is to use advanced unsupervised learning methods such as Dimensionality reduction(PCA) and clustering techniques(Kmeans, DBSCAN) to parse and segregate the text corpus into different concentrations. Our secondary goals include performing Exploratory data analysis and build interactive visualizations

Algorithms:
Since the main goal of the project is clustering, we are going to use the popular clustering methods K Means and DBSCAN. The main reason for using K Means is that we have the freedom to specify how many number of clusters we want. And DBSCAN can be used as a point for comparing the two models and determine which one is more feasible and outputs sensible clusters.

The amount of text data in the corpus is vast, and fitting a model against such high dimensionality will not be feasible. This calls for use of dimensionality reduction techniques such as Principal Component Analysis, Singular Value Decomposition.

Dataset:
The dataset is available in the form of json files. These files are going to be parsed using python scripting and stored in the form of dataframes. The data is majorly textual in nature. It contains the titles, abstracts and body descriptions of sixty thousand literary studies related to COVID 19. We plan to use a smaller subset of about 10,000 abstracts due to computational limitations.

Once the data has been parsed and stored in the appropriate formats, text processing will be performed using string manipulation, regular expressions and other natural language processing techniques such stemming and lemmatization. This is done to make sure that the data consists of only readable text and not any filler words or other unwanted entities such as urls, etc.

Libraries and tools:

The libraries that will be used consist of python libraries like Matplotlib, Numpy, Pandas, Seaborn and Plotly (for EDA and visualizations), Scikit-learn (for machine learning) and NLTK (for text processing).

Results:

Our main goal is to be able to get distinct clusters that accurately represent each sub section of the literary studies. For example, one cluster representing all literature related to a vaccine, another representing literature on the long term health impacts of COVID, etc. We plan on comparing permutations of clustering techniques being used with the dimensionality reduction techniques being used to get the most optimal combination.

The main challenge we might face is while converting the text data into usable format to apply various machine learning tasks on it. This can be solved by performing different cleaning, preprocessing and feature engineering techniques for text mining.
Another challenge we might face is to prevent overfitting or overestimating the number of clusters. This we will solve by using techniques taught to us in class (like SSE and the elbow method).

References:

The data set has been taken from a Kaggle Competition titled CORD-19 Challenge.

- The link to the dataset: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge.
- https://dl.acm.org/doi/10.1145/3395027.3419591
- https://www.kaggle.com/phyothuhtet/document-clustering-self-organizing-map-kmeans