**Short Paper**

**Analysis of COVID-19 Research Papers**

It is difficult for health professionals to keep up with latest information on the virus. By using clustering for visualization, the research papers can be represented by a scatter plot. In this plot, research papers of nearest similar topics will be plotted near each other. In this project we are trying to analyze the COVID-19 research papers , find the keywords and cluster them to find research papers of similar kind and we have plotted a graph based on the scores. We have used jupyter notebook and python to work on our project.

**Dataset:**

Our Dataset consists of columns of data related to Title, License, Abstract, Publish Time, Authors, Journal, Release Year, ID’s.

**Libraries Used:**

Numpy, Pandas, NLTK, OS, Pickle

**Data Cleaning:**

As we already know data cleaning is the process of removing incomplete or missing data. Our first step was to drop the null values. We did it using dropna. We did this first because it can adversely affect the performance and accuracy of any machine learning algorithm.

**Good Paper Impact Data:**

For a good and relevant paper we have use the following parameters to measure. Score, readers count, accounts. Here score means how has the research outcomes that we share with the platform and interactions with each other members and the reputation of peers.

**Textual Analysis:**

For textual analysis we applied the concepts we learnt in text mining and bigdata. Removed the non-essential words like what. is, a. Tokenization and Stemming, Lemmatization.

**Data Clustering:**

For data clustering we used the unsupervised learning technique for algorithms for the analysis of title. We plan to use the Principal Component Analysis. It helps us finding most significant features in a dataset. We have used vectorization to get some distinct features out of the text for model to train on. With this we plan to extract the common keywords from the cluster and plot a scatter plot.

Chart, bar chart

Description automatically generated

Chart, scatter chart

Description automatically generated