|  |  |
| --- | --- |
| **spring 2020 Project: COVID-19 Open Research Dataset Challenge (CORD-19)** | Abstract  In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19) at Kaggle. CORD-19 is a resource of over 59,000 scholarly articles, including over 47,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses |

**Introduction**

The current world is facing a serious problem of COVID-19. In this project we are using a large corpus of COVID-19 research to understand the pandemic better. The coronavirus disease 2019 (COVID-19) caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The disease was first identified in Wuhan, Hubei, China in December 2019.This is biggest outbreak which has caused widespread medical and economic disruption in recent history.

To fight with this pandemic across the globe medical researchers, healthcare workers, and even data scientists are working together. To get the solution for this problem, professionals need to go through the number of published articles at various sources. There is an increasing number of articles available related to COVID-19 pandemic and it is hard for the researchers and healthcare professionals to read each article. So, the need is increasing to make use of machine learning techniques available to help medical research community to keep up with the rapid acceleration in new coronavirus literature to understand the virus better. This project makes use of such methodologies on the given dataset to find answers to questions, and connect insights across, this content in support of the ongoing COVID-19 pandemic.

Our motivation to pursue this problem is to contribute to the world of medical research through our knowledge in AI and machine learning. We would like to leverage the knowledge we gained about supervised and unsupervised learning algorithms in this course and apply them as much as possible to this problem statement. Although the scope of this problem is a multi-step process, we would like to start off with the knowledge in hand and further pursue this problem by increasing our knowledge base.

As our data is not labelled, we have used two unsupervised algorithms: K-means Clustering and Hierarchical Clustering. The input to our algorithm is the abstract of the documents published in the timeframe of Covid 19 (2019 and 2020). We aim at producing an output where we can identify how many major clusters of information exists in this document base and what topics are these clusters addressing just by analyzing the abstract of these documents. This information could then be used as a starting point for more thorough labelling of clustered documents. This can be used by medical researchers who want to access documents to their area of interest without having to deep dive inside a collection of articles.

**Dataset Description**

In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19) at Kaggle. CORD-19 is a resource of over 59,000 scholarly articles, including over 47,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. [1]

**Link:** <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge>

The data provided on the Kaggle website contains a metadata excel which consists of information about the jsons and a collection of json files which contains the body of each document. We are only working with the metadata excel in our scope and hence analyzing the abstract of the documents and not the body.

Below are the columns of the metadata file:

(1) corduliid: The unique id generated for each paper.

(2) sha - The paper records have PDFs and the hash of the PDFs are in 'sha'

(3) source\_x: It shows which source the paper belongs. Metadata for papers from these sources are combined: CZI, PMC, Elsevier, WHO,BioRxiv/MedRxiv.

(4) title: It gives the title of the paper

(5) Various 'keys' are populated with the metadata:

- 'pmcid': populated for all PMC paper records

- 'doi': populated for all BioRxiv/MedRxiv paper records and most of the other records

- 'WHO Covidence': populated for all CZI records and none of the other records

- 'pubmed\_id': populated for some of the records

- 'Microsoft Academic Paper ID': populated for some of the records

(6) abstract: It contains the abstract of the paper.

(7) publish\_time: When the paper was published that date is stored in this column.

(8) authors: The author of the papers.

(9) full\_text\_file: PDFs were processed with fulltext ('has\_full\_text'=True)

**Preprocessing and cleaning**

In terms of data cleaning, the meta data excel sheet contains a lot of nulls for several columns. However, in our analysis, we are analyzing the textual content to see how useful it could be for research purposes. Hence the only features we are considering now is abstract.

**Duplicates:** We found several documents with duplicate abstracts. This could be because of researchers submitting their papers at multiple sources. Hence, we only considered documents with unique abstracts.

**Languages:** There were also a lot of documents present in Spanish, French, German, Italian, etc. Documents in other languages except English have been marked out of scope for this project. Hence, we have dropped these records.

**Null values:** There were also a lot of documents with no abstract values. Such documents would need further processing where their bodies could be analyzed. Hence for the scope of this project, we have dropped such records.

**Tokenization -** Tokenization is the process of segmenting text into words, clauses or sentences. In our dataset we will split the abstract of the article into individual words and the punctuation is also produced as separate words.

**Stemming -** Stemming reduces related words to a common stem. It is an optional process step, and it is useful to test accuracy with and without stemming.

**Stop words removal-** Stop words are commonly used words that are unlikely to have any benefit in natural language processing. These includes words such as ‘a’, ‘the’, ‘is’. So, finding and removing stop words.

**Learning Algorithms**

1. **K means Clustering –**

K-means clustering is a simple unsupervised machine learning algorithm that is used to solve clustering problems. It follows a simple procedure of classifying a given data set into several clusters, defined by the letter "k," which is fixed beforehand.

**The algorithm:**

1. K points are placed into the object data space representing the initial group of centroids.

2. Each object or data point is assigned into the closest k.

3. After all objects are assigned, the positions of the k centroids are recalculated.

4. Steps 2 and 3 are repeated until the positions of the centroids no longer move.

In the context of this project, we used TF-IDF for vectorizing the information in abstract since while working with text documents, we need a way to convert each document into a numeric vector. This process led to a creation of 4000+ features for the abstract feature. Hence in order to be able to apply some unsupervised learning models on this further, we used PCA to reduce the dimensions. While applying K- means clustering, we used the variance as 95% and further analysis showed that 8 clusters would be a good starting point.

1. **Hierarchical Clustering -**

Hierarchical cluster analysis or HCA is an unsupervised clustering algorithm which involves creating clusters that have predominant ordering from top to bottom. [2]. The algorithm groups similar objects into groups called clusters. The endpoint is a set of clusters or groups, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly like each other.

This clustering technique is divided into two types:

1. Agglomerative Hierarchical Clustering
2. Divisive Hierarchical Clustering

**The algorithm:**

1. Make each data point a single-point cluster → forms N clusters
2. Take the two closest data points and make them one cluster → forms N-1 clusters
3. Take the two closest clusters and make them one cluster → Forms N-2 clusters.
4. Repeat step-3 until you are left with only one cluster.

In the context of this project, we applied HCA using the linkage method as ward and centroid. We have used Agglomerative Hierarchical Clustering. The number of features as an input to the algorithm was coming from the dimensions reduced by PCA.

1. **Principal Component Analysis –**

Principal component analysis (PCA) is a statistical technique to convert high dimensional data to low dimensional data by selecting the most important features that capture maximum information about the dataset. The features are selected on the basis of variance that they cause in the output.

PCA process:

* Initialize the PCA class by passing the number of components to the constructor.
* Call the fit and then transform methods by passing the feature set to these methods. The transform method returns the specified number of principal components.

In the context of this project, we have used PCA to select the most important features after we create vectorized features for the abstract column using TF-IDF.

1. **Latent Dirichlet Allocation -**

Latent Dirichlet Allocation (LDA) is one of the approaches of topic modeling, which is an unsupervised machine learning algorithm, where each document is considered as a collection of topics and each word in the document corresponds to one of the topics.

So, given a document LDA basically clusters the document into topics where each topic contains a set of words which best describe the topic.The following are the steps to implement LDA in Python.

1. Import the dataset.
2. Preprocess the text data
3. Create Gensim dictionary and corpus
4. Building the Topic Model
5. Analyze the results
6. Dominant topic within documents

In the context of this project, we used LDA to view the major topics in the clusters formed by K-Means algorithm. We did not use this on the clusters from HCA since the topics were evident in those clusters on a visual word cloud.

**We used the following techniques:**

**Vectorization -** Machine learning algorithms most often take numeric feature vectors as input. Thus, when working with text documents, we need a way to convert each document into a numeric vector and this process is known as text vectorization. For this purpose in our dataset we are using TF-IDF.

TF-IDF stands for Term Frequency-Inverse Document Frequency which basically tells importance of the word in the corpus or dataset. TF-IDF contain two concept Term Frequency (TF) and Inverse Document Frequency (IDF)

Term Frequency (TF) is defined as how frequently the word appear in the document or corpus. As each sentence is not the same length so it may be possible a word appears in long sentence occur more time as compared to word appear in sorter sentence.

Inverse Document frequency (IDF) is another concept which is used for finding out importance of the word. It is since less frequent words are more informative and important.

**Dictionary and Corpus creation –**To perform topic modeling using LDA the two main inputs are the dictionary(id2word) and the corpus. Here we are using gensim library for building the dictionary and the corpus.In Gensim, the words are referred to as “tokens” and the index of each word in the dictionary is called “id”. Dictionary is nothing but the collection of unique word-id’s and corpus is the mapping of (word\_id, word\_frequency).

**Word Cloud** - Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud.

**Discussion**

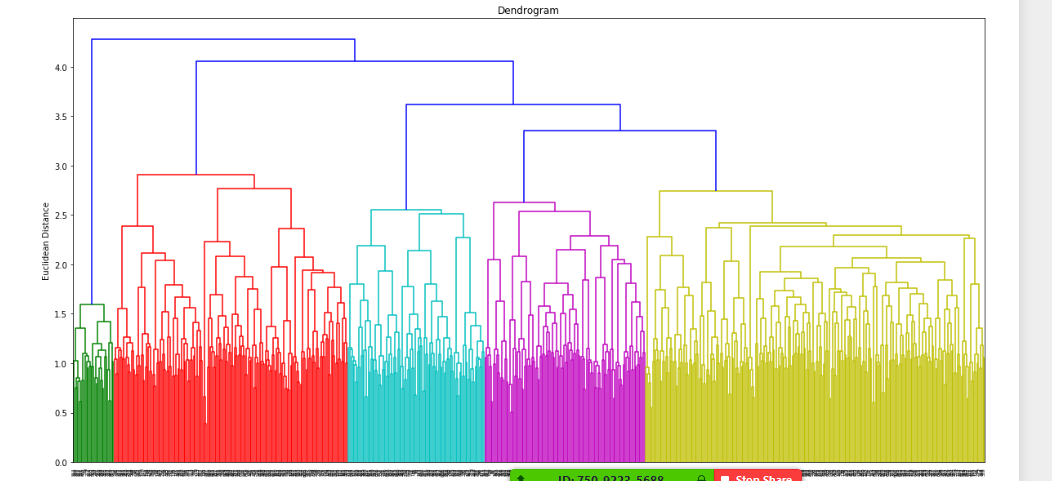
**Hierarchical Clustering**

We executed Hierarchical Clustering with the following combinations.

1. Choosing all the features after vectorizing using the TF-IDF method, we ran PCA on top of it. this resulted in about 700 features and the dendrogram showed evidence for 2 clusters.
2. Choosing 2\*\*12 top features based on TF-IDF scores, when we executed PCA, we ended up with 547 features and the dendrogram showed evidence for 3 clusters.
3. Both the above were executed with the n\_components for PCA (variance) as 95% and we used agglomerative clustering.
4. We then switched to reducing the number of features to 1000, and the PCA variance (n\_components) to 95%. This resulted in the dendrogram showing 4 clusters using agglomerative clustering.
5. Further we ran the same for variance in PCA as 90%, but this did not show a very good clustering, we were able to detect 3 clusters.
6. Now, we chose the top 500 TF-IDF scores, and PCA with 90% variance and this resulted in 255 as the number of reduced dimensions. The dendrogram showed 5 distinct cluster possibilities.

*X = vectorize(text,500)*

*pca = PCA(n\_components=0.90, random\_state=42)*

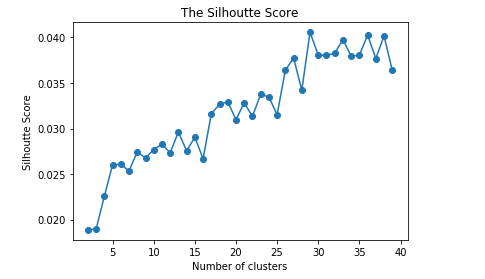
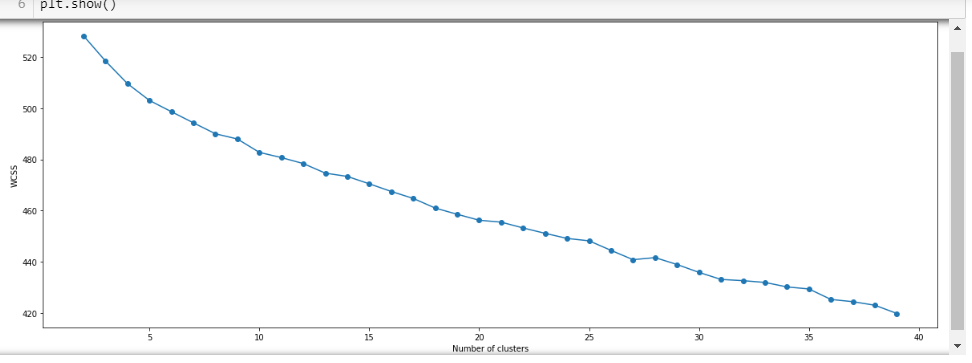


**K means clustering**

1. Switching to k means, with the choice of top 2\*\*12 TF idf scored features and a PCA variance of 95%, the elbow method and silhouette score showed an evidence of 8 clusters. However, the elbow curve itself did not show a significant bend.
2. Following the final test for Hierarchical clustering (6 from Hierarchical) for the top 500 TF-IDF scores, and PCA with 90% variance resulting in 255 as the number of reduced dimensions. The dendrogram showed 5 distinct cluster possibilities.

*X = vectorize(text,500)*

*pca = PCA(n\_components=0.90, random\_state=42)*

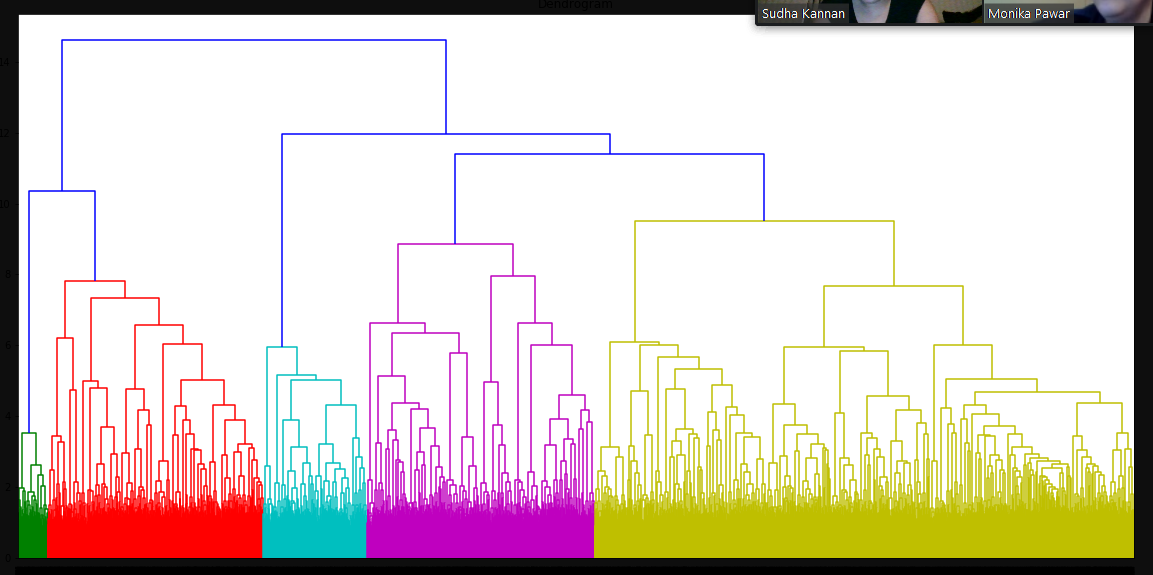


In k means, based on the above graph the candidates for the number of clusters and based on the elbow method could be: 8, 20, 27. Based on the silhouette score, the apt number of clusters should be around 28.

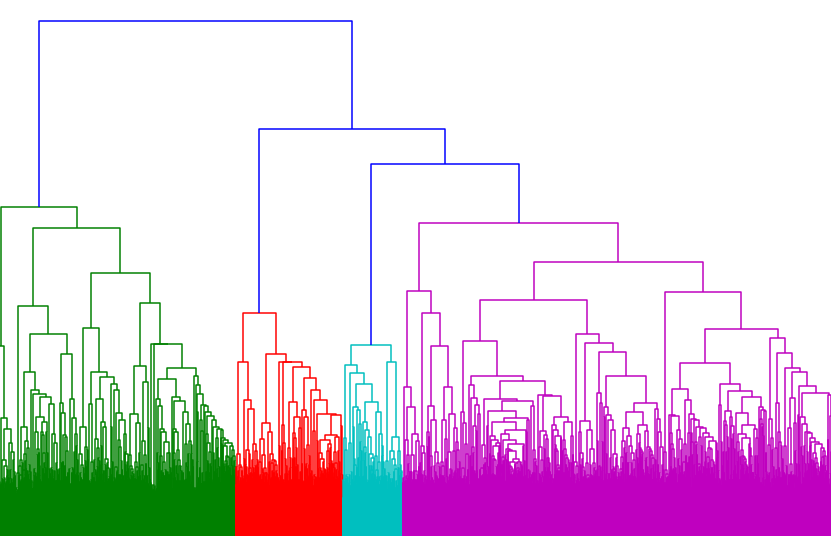
**Since Hierarchical and K-means are showing completely different number of clusters from the range of 2 to 30, we decided to increase the size of our training dataset. We included the abstracts of all the documents published in 2019 and 2020 and not just Nov and Dec 2019.**

**Switching back to Hierarchical:**

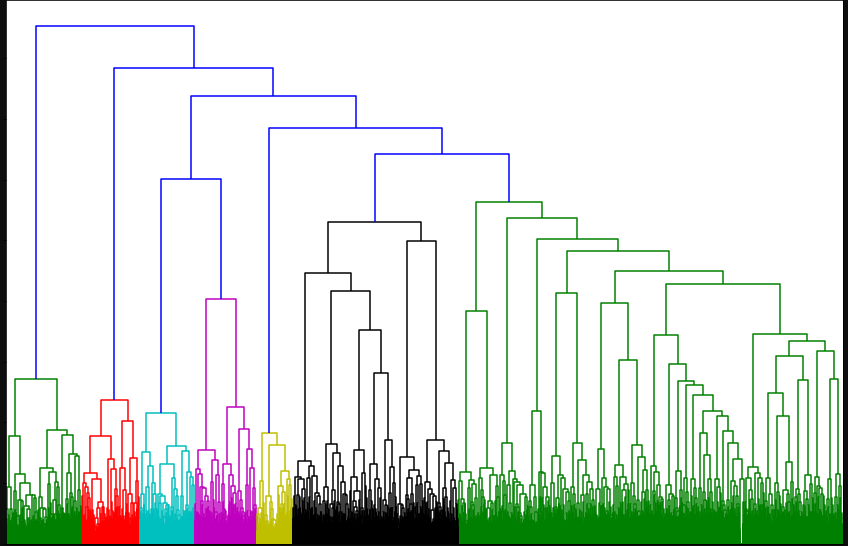
1. For top 500 TF – IDF scores, we found 368 significant dimensions using PCA and 2 clusters using HCA.



1. For top 2000 TF – IDF scores, we found 1190 significant dimensions using PCA and 2 clusters using HCA.



9.For top **50** TF – IDF scores, we found **40** significant dimensions using PCA and 2 clusters using HCA



All the HCA runs were using “Ward” as the linkage method. Results using centroid were not

readable. We always saw only 2 clusters forming from HCA when we increased our training dataset.

**Conclusion**

We tried multiple runs on the vectorized features generated from “abstract” using HCA and K means. When the number of top TF-IDF features chosen was more than 500 or 1000, HCA yielded around 5 clusters whereas K-means showed more than 25 clusters with a very low silhouette coefficient (~0.35).

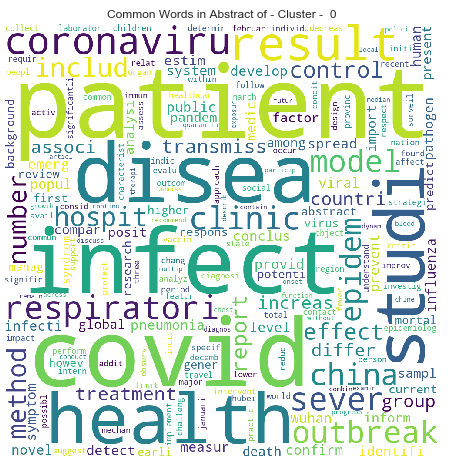
When we reduced the number of features to 50 (TF-IDF), the K-means algorithm showed a clear number of clusters in the elbow method whereas HCA showed 2 clusters. Since 50 features is a very low number to analyze, we think that we should go for at least 1000 features as an input to PCA and use the clusters derived from clustering for further analysis.

During this analysis, we also increased our dataset to incorporate all documents published in 2019 and 2020.

**With this, for 1000 TF-IDF vectors, PCA resulted in 688 significant dimensions, agglomerative HCA showed presence of 2 significant clusters. K – Means did not show a significant bend in the elbow method- 8, 10, 15, 17 clusters being candidates for further analysis, and the Silhouette coefficient suggested to create around 18 clusters. We pursued creating 10 clusters for ease of analysis at this stage.**

### Clustering alone is not enough to gain any useful insights from this analysis. We should also pair it with word cloud or LDA (Latent Dirichlet Allocation) so that we can check what are the main words and topics brought out by the algorithm.

### So, after applying Agglomerative Hierarchical clustering what is present inside these 2 clusters?



Common things: both clusters talk about infection.

First cluster (0) shows that the main words addressed in the documents of this cluster is about Covid infections, clinics and hospitals, epidemic and outbreak, patients with respiratory infection, transmissible, China, Wuhan, health treatment, studies.

Second cluster (1) shows that the main words addressed in the documents of this cluster are – protein, viral studies, immunity, vaccine, analysis, studies about effect on humans, antibody strain.

### After applying K means clustering: what is present inside these 10 clusters?

1. Main topics for Cluster number: 0[(0, '0.025\*"covid" + 0.025\*"china" + 0.021\*"wuhan" + 0.015\*"infect" + 0.013\*"coronaviru"'), (1, '0.030\*"coronaviru" + 0.021\*"infect" + 0.020\*"diseas" + 0.020\*"covid" + 0.018\*"china"'), (2, '0.031\*"covid" + 0.023\*"china" + 0.016\*"coronaviru" + 0.016\*"wuhan" + 0.013\*"outbreak"'), (3, '0.022\*"china" + 0.021\*"covid" + 0.020\*"report" + 0.015\*"infect" + 0.014\*"coronaviru"'), (4, '0.026\*"provinc" + 0.025\*"china" + 0.024\*"wuhan" + 0.024\*"hubei" + 0.016\*"number"')]

First cluster (0) talks about the outbreak of covid-19/ corona virus in china and the number of infections in Wuhan, Hubei.

1. Main topics for Cluster number: 1[(0, '0.017\*"activ" + 0.016\*"immun" + 0.015\*"viral" + 0.013\*"respons" + 0.012\*"infect"'), (1, '0.025\*"infect" + 0.016\*"activ" + 0.014\*"respons" + 0.014\*"express" + 0.013\*"induc"'), (2, '0.028\*"infect" + 0.013\*"express" + 0.013\*"viral" + 0.012\*"replic" + 0.010\*"protein"'), (3, '0.020\*"infect" + 0.020\*"antivir" + 0.017\*"activ" + 0.013\*"viral" + 0.012\*"effect"'), (4, '0.020\*"immun" + 0.019\*"express" + 0.013\*"infect" + 0.013\*"pdcov" + 0.010\*"respons"')]

### Second cluster (1) talks about protein, virus and the immunity response to the virus. There are keywords pointing to PDCov and the immunity for it in infected patients. As per CDC’s website is abbreviation for Porcine Deltacoronavirus Infection and Transmission in Poultry, United States. As per CDC [21], in 2009, porcine deltacoronavirus (PDCoV) was detected in fecal samples from pigs in Asia, but its etiologic role was not identified until 2014, when it caused major diarrhea outbreaks in swine in the United States.

1. Main topics for Cluster number: 2[(0, '0.027\*"model" + 0.021\*"infect" + 0.019\*"epidem" + 0.016\*"diseas" + 0.015\*"number"'), (1, '0.040\*"model" + 0.022\*"infect" + 0.016\*"estim" + 0.014\*"predict" + 0.014\*"covid"'), (2, '0.025\*"number" + 0.025\*"estim" + 0.023\*"covid" + 0.021\*"epidem" + 0.021\*"china"'), (3, '0.019\*"model" + 0.018\*"covid" + 0.016\*"transmiss" + 0.015\*"infect" + 0.014\*"measur"'), (4, '0.033\*"covid" + 0.023\*"countri" + 0.014\*"model" + 0.012\*"itali" + 0.011\*"outbreak"')]

Third cluster (2) talks about number of infected people due to the pandemic in China and also talks about a model built on infected numbers. There is a link to Italy in this cluster as one of the countries where there was an outbreak and this model was perhaps used there. The topic suggests that the model was used to predict and estimate the numbers in China.

1. Main topics for Cluster number: 3[(0, '0.020\*"protein" + 0.014\*"receptor" + 0.009\*"entri" + 0.009\*"glycoprotein" + 0.007\*"structur"'), (1, '0.032\*"protein" + 0.019\*"antibodi" + 0.012\*"infect" + 0.012\*"human" + 0.011\*"neutral"'), (2, '0.053\*"protein" + 0.013\*"infect" + 0.013\*"viral" + 0.011\*"express" + 0.009\*"interact"'), (3, '0.032\*"protein" + 0.017\*"structur" + 0.011\*"viral" + 0.010\*"interact" + 0.010\*"function"'), (4, '0.032\*"protein" + 0.017\*"structur" + 0.013\*"coronaviru" + 0.011\*"spike" + 0.010\*"studi"')]

Fourth cluster (3) This cluster talks about a spike protein on COVID-19, spike glycoprotein, which could be used in the development of vaccines for human. It also talks about proteins and antibody interactions with human, perhaps how the protein structure of the virus functions when in contact with humans. More analysis is required to interpret this clearly.

1. Main topics for Cluster number: 4[(0, '0.022\*"virus" + 0.015\*"viral" + 0.014\*"human" + 0.012\*"speci" + 0.010\*"genom"'), (1, '0.038\*"strain" + 0.016\*"sequenc" + 0.012\*"genom" + 0.009\*"vaccin" + 0.008\*"chicken"'), (2, '0.022\*"sequenc" + 0.021\*"genom" + 0.016\*"strain" + 0.015\*"sampl" + 0.014\*"virus"'), (3, '0.031\*"sequenc" + 0.013\*"genom" + 0.010\*"studi" + 0.009\*"viral" + 0.008\*"infect"'), (4, '0.020\*"genom" + 0.018\*"sequenc" + 0.012\*"human" + 0.011\*"coronaviru" + 0.011\*"strain"')]

Fifth cluster (4) talks about a specific virus genom and sequence data. It also talks about studies on different strains of the virus in a chicken and studies of the genom sequence data on a human.

1. Main topics for Cluster number: 5[(0, '0.029\*"health" + 0.012\*"covid" + 0.012\*"diseas" + 0.011\*"countri" + 0.010\*"public"'), (1, '0.023\*"health" + 0.016\*"covid" + 0.014\*"diseas" + 0.011\*"public" + 0.010\*"infect"'), (2, '0.026\*"health" + 0.019\*"covid" + 0.011\*"pandem" + 0.008\*"patient" + 0.007\*"infect"'), (3, '0.030\*"covid" + 0.013\*"emerg" + 0.011\*"social" + 0.010\*"public" + 0.010\*"infect"'), (4, '0.032\*"health" + 0.021\*"diseas" + 0.014\*"public" + 0.009\*"outbreak" + 0.009\*"commun"')]

Sixth cluster (5) talks about social and public keywords but not about distancing, one interpretation of this could be that public and social actives are leading to emerging [or emergency- to be analyzed further] health conditions and patient and countries infected due to pandemic.

1. Main topics for Cluster number: 6[(0, '0.051\*"vaccin" + 0.023\*"immun" + 0.014\*"develop" + 0.014\*"respons" + 0.012\*"infect"'), (1, '0.045\*"vaccin" + 0.016\*"infect" + 0.010\*"protein" + 0.010\*"immun" + 0.009\*"epitop"'), (2, '0.027\*"vaccin" + 0.017\*"immun" + 0.016\*"respons" + 0.011\*"specif" + 0.010\*"antibodi"'), (3, '0.042\*"vaccin" + 0.018\*"immun" + 0.010\*"respons" + 0.010\*"antigen" + 0.009\*"antibodi"'), (4, '0.071\*"vaccin" + 0.012\*"diseas" + 0.011\*"develop" + 0.009\*"studi" + 0.007\*"strain"')]

Seventh Cluster (6) talks about vaccine development and the protein, antibodies structure of the vaccine to increase immunity response. It also talks about EpiTOP which is a QSAR approach for ligands binding to several related proteins. This cluster also has documents with information on antigen which is a toxin or foreign substance which induces an immune response in the body, especially the production of antibodies.

1. Main topics for Cluster number: 7[(0, '0.049\*"patient" + 0.030\*"covid" + 0.019\*"sever" + 0.018\*"diseas" + 0.014\*"associ"'), (1, '0.042\*"patient" + 0.028\*"covid" + 0.014\*"lesion" + 0.014\*"hospit" + 0.012\*"diseas"'), (2, '0.065\*"patient" + 0.029\*"covid" + 0.017\*"group" + 0.017\*"sever" + 0.017\*"clinic"'), (3, '0.034\*"patient" + 0.028\*"covid" + 0.015\*"infect" + 0.011\*"clinic" + 0.009\*"studi"'), (4, '0.054\*"patient" + 0.022\*"infect" + 0.015\*"covid" + 0.009\*"diseas" + 0.009\*"treatment"')]

Eight cluster (7) talks about patients of covid-19 having severe condition of disease who are hospitalized, may be having lesions which is small lesions on feet of coronavirus patients, possibly indicating a new COVID-19 symptom

1. Main topics for Cluster number: 8[(0, '0.013\*"clinic" + 0.011\*"diseas" + 0.009\*"studi" + 0.008\*"treatment" + 0.007\*"review"'), (1, '0.011\*"infect" + 0.009\*"studi" + 0.007\*"treatment" + 0.007\*"diseas" + 0.006\*"increas"'), (2, '0.014\*"diseas" + 0.009\*"infect" + 0.007\*"develop" + 0.007\*"studi" + 0.006\*"human"'), (3, '0.010\*"studi" + 0.009\*"diseas" + 0.008\*"infect" + 0.006\*"pathogen" + 0.006\*"human"'), (4, '0.016\*"infect" + 0.011\*"patient" + 0.010\*"studi" + 0.010\*"respiratori" + 0.008\*"result"')]

Ninth cluster (8) shows topics about studies conducted on the treatment of this disease (covid 19) on infected humans. It has documents that describe how the infection was developed in humans and studies on the pathogens in people who were infected. Also, has studies on the respiratory system of the infected patients.

1. Main topics for Cluster number: 9[(0, '0.024\*"infect" + 0.015\*"respiratori" + 0.012\*"elisa" + 0.011\*"sampl" + 0.010\*"viral"'), (1, '0.029\*"respiratori" + 0.029\*"children" + 0.023\*"infect" + 0.017\*"pneumonia" + 0.015\*"influenza"'), (2, '0.021\*"infect" + 0.018\*"detect" + 0.017\*"respiratori" + 0.016\*"influenza" + 0.015\*"sampl"'), (3, '0.028\*"respiratori" + 0.024\*"infect" + 0.021\*"detect" + 0.020\*"posit" + 0.019\*"sampl"'), (4, '0.037\*"detect" + 0.027\*"assay" + 0.013\*"sensit" + 0.013\*"specif" + 0.013\* "develop" )]

The last cluster (9) has documents on studies samples of the virus on respiratory infections. It talks about ELISA (enzyme-linked immunosorbent assay) which is a plate-based assay technique designed for detecting and quantifying substances such as peptides, proteins, antibodies and hormones. Other names, such as enzyme immunoassay (EIA), are also used to describe the same technology. The documents are also related to respiratory infections in children like pneumonia and influenza and studies on such positively infected samples. The last topic in this cluster is about detection and assay. Assay is an analysis done to determine the presence of a substance and the amount of a substance.

**Statements of Individual contribution**

**Monika:**

1. Initial dataset research
2. Exploratory data analysis

* URL clicks, Null values,
* Counts and distribution of data,
* Abstract words distribution,
* Word clouds for covid 19 publications in 2019 2020

1. Data cleaning
2. PCA and K means
3. LDA
4. Documentation 50%
5. Conclusions and Analysis

**Sudha:**

1. Initial dataset research
2. Draft approach to load json files implemented
3. Exploratory data analysis

* Graph to show publications by year for all years
* Word clouds amended to show SARS, MERS, COVID 19

1. Data cleaning
2. PCA and Hierarchical
3. LDA
4. Documentation 50%
5. Conclusions and Analysis

**References**

1. https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge
2. <https://www.kdnuggets.com/2019/09/hierarchical-clustering.html>
3. <https://www.kaggle.com/pjoshi5/covid-19-project-medical-tasks>
4. <https://www.kaggle.com/blind0ne/somantic-brain-inspired-clustering-and-querying>
5. <https://www.kaggle.com/maksimeren/covid-19-literature-clustering>
6. <https://www.ecmconnection.com/doc/data-mining-and-document-mining-0001>
7. <https://www.kaggle.com/maksimeren/covid-19-literature-clustering>
8. <https://www.ecmconnection.com/doc/data-mining-and-document-mining-0001>
9. <https://www.kaggle.com/dawnsepehr/cord-19-task-4>
10. <https://www.kaggle.com/jonathanbesomi/a-qa-model-to-answer-them-all>
11. <https://www.kaggle.com/huascarmendez1/cord-19-vaccines-mining>
12. <https://rstudio-pubs-static.s3.amazonaws.com/79360_850b2a69980c4488b1db95987a24867a.html>
13. <https://www.kaggle.com/huascarmendez1/cord19csv>
14. <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/discussion/138849>
15. <https://pypi.org/project/gensim/>
16. <https://www.kaggle.com/hvillacorte/covid-19-vaccines-and-therapeutics#q5>
17. <https://plotly.com/python/distplot/>
18. <https://towardsdatascience.com/unsupervised-nlp-topic-models-as-a-supervised-learning-input-cf8ee9e5cf28>
19. https://www.kaggle.com/mrmimic/opinions-extraction-tool-several-examples
20. <https://towardsdatascience.com/nlp-extracting-the-main-topics-from-your-dataset-using-lda-in-minutes-21486f5aa925>