**Contents** :

1. [**Introduction**](#_Introduction:)
2. [**Dataset**](#_Dataset:)
3. [**Data Wrangling**](#_Data_Wrangling:)
   1. **Inverted index**
   2. **Lemmatization**
   3. **Part of Speech tagging**
4. [**Exploratory Data Analysis**](#_Exploratory_Data_Analysis)
   1. **Article matching using WordCloud :**
   2. **Topic modeling using Latent Dirichlet Allocation:**
   3. **Topic Extraction using SentenceBert + K-Means Clustering + TF-IDF**
5. [**Modeling**](#_Modeling:)
6. **Conclusion**

# [Introduction](#First_heading)

Coronavirus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The first known case was identified in Wuhan, China, in December 2019.[7] The disease has since spread worldwide, leading to an ongoing pandemic.Symptoms of COVID-19 are variable, but often include fever, cough, headache,fatigue, breathing difficulties, and loss of smell and taste. Symptoms may begin one to fourteen days after exposure to the virus. At least a third of people who are infected do not develop noticeable symptoms.Of those people who develop symptoms noticeable enough to be classed as patients, most (81%) develop mild to moderate symptoms (up to mild pneumonia), while 14% develop severe symptoms (dyspnea, hypoxia, or more than 50% lung involvement on imaging), and 5% suffer critical symptoms (respiratory failure, shock, or multiorgan dysfunction).[15] Older people are at a higher risk of developing severe symptoms. Some people continue to experience a range of effects (long COVID) for months after recovery, and damage to organs has been observed. Multi-year studies are underway to further investigate the long-term effects of the disease.

Given the large number of literature and the rapid spread of COVID-19, it is difficult for health professionals to keep up with new information on the virus. This is a difficult time in which health care workers, sanitation staff, and much other essential personnel are out there keeping the world afloat.

The goal of this project is to :

* Develop Natural Language Processing & other AI techniques to generate insights support of the ongoing fight against this infectious disease
* Can clustering similar research articles together simplify the search for related publications?
* How can the content of the clusters be qualified?
* Keep up with the rapid acceleration in new literature
* This tool was created to help make it easier for trained professionals to sift through many, many publications related to the virus, and find their own determinations

# Dataset

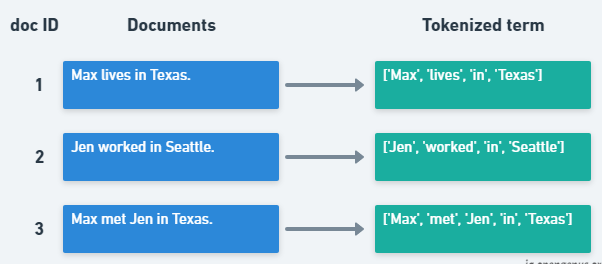
In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19). CORD-19 is a resource of over 500,000 scholarly articles, including over 200,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. This freely available dataset is provided to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease. There is a growing urgency for these approaches because of the rapid acceleration in new coronavirus literature, making it difficult for the medical research community to keep up.

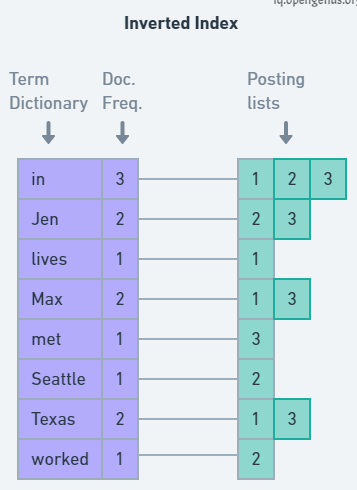
# Data Wrangling

CORD-19 has several years of data. So, the first step is to filter it to 2021 related papers. There are 669 articles, out of which 170 has missing abstract. So, removed the ones that are missing.

Apart from that, checked if there are any duplicates. Next, using NLTK package, removed non-English words, stop words (like ‘the’, ‘to’, ‘and’ etc.) and words with single characters.

Next, created an inverted index. Inverted index is a fundamental technology used commonly in search engine . Inverted index assigns each word with a list of document id and it helps retrieve the articles with keyword matching fastly. For example {'patient':[1,3,6]} means that the word patient occurs in the second,fourth and sixth document in the corpus. So, there are 8294 unique words in inverted index. For instance :

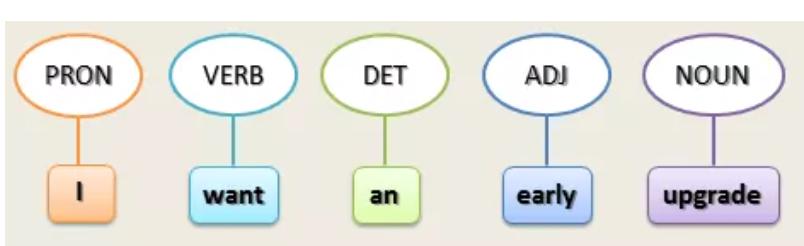




Lemmatization is another technique to doing things properly with the use of a vocabulary and morphological analysis of words. The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For instance:

am, are, is $\Rightarrow$ be  
car, cars, car's, cars' $\Rightarrow$ car

Part of Speech tagging ('NOUN','ADJ','VERB','ADV’)



Below are the steps followed:

* Parse the text from the body of each document using Natural Language Processing (NLP).
* Turn each document instance di into a feature vector Xi using Term Frequency–Inverse Document Frequency (TF-IDF).
* Apply Dimensionality Reduction to each feature vector Xi using t-Distributed Stochastic Neighbor Embedding (t-SNE) to cluster similar research articles in the two-dimensional plane X embedding Y1.
* Use Principal Component Analysis (PCA) to project down the dimensions of X to several dimensions that will keep .95 variance while removing noise and outliers in embedding Y2.
* Apply k-means clustering on Y2, where k is 6, to label each cluster on Y1.
* Apply Topic Modeling on X using Latent Dirichlet Allocation (LDA) to discover keywords from each cluster.
* Investigate the clusters visually on the plot, zooming down to specific articles as needed, and via classification using Stochastic Gradient Descent (SGD).

# Exploratory Data Analysis

Article matching using WordCloud : Sample data contains words related to risk like ['risk','risky','risks','risked','risking']



**Topic modeling**

Topic Models, in a nutshell, are a type of statistical language models used for uncovering hidden structure in a collection of texts. In a practical and more intuitively, you can think of it as a task of:

Dimensionality Reduction, where rather than representing a text T in its feature space as {Word\_i: count(Word\_i, T) for Word\_i in Vocabulary}, you can represent it in a topic space as {Topic\_i: Weight(Topic\_i, T) for Topic\_i in Topics}

Unsupervised Learning, where it can be compared to clustering, as in the case of clustering, the number of topics, like the number of clusters, is an output parameter. By doing topic modeling, we build clusters of words rather than clusters of texts. A text is thus a mixture of all the topics, each having a specific weight

Tagging, abstract “topics” that occur in a collection of documents that best represents the information in them.

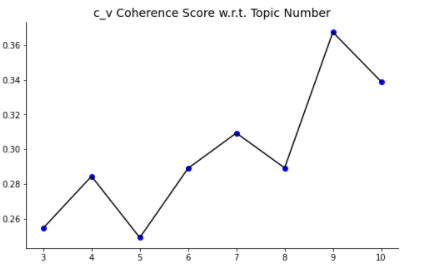
There are several existing algorithms you can use to perform the topic modeling. The most common of it are, Latent Semantic Analysis (LSA/LSI), Probabilistic Latent Semantic Analysis (pLSA), and Latent Dirichlet Allocation (LDA)

**1st method:**

Topic modeling using Latent Dirichlet Allocation: LDA model are available and fast to implement with the help of either two libraries (sklearn and genism).

* sklearn LDA pipeline: CounterVectorizer + LatentDirichletAllocation + GridSearchCV
* gensim pipeline: Dictionary + doc2bow + LdaModel + CoherenceModel

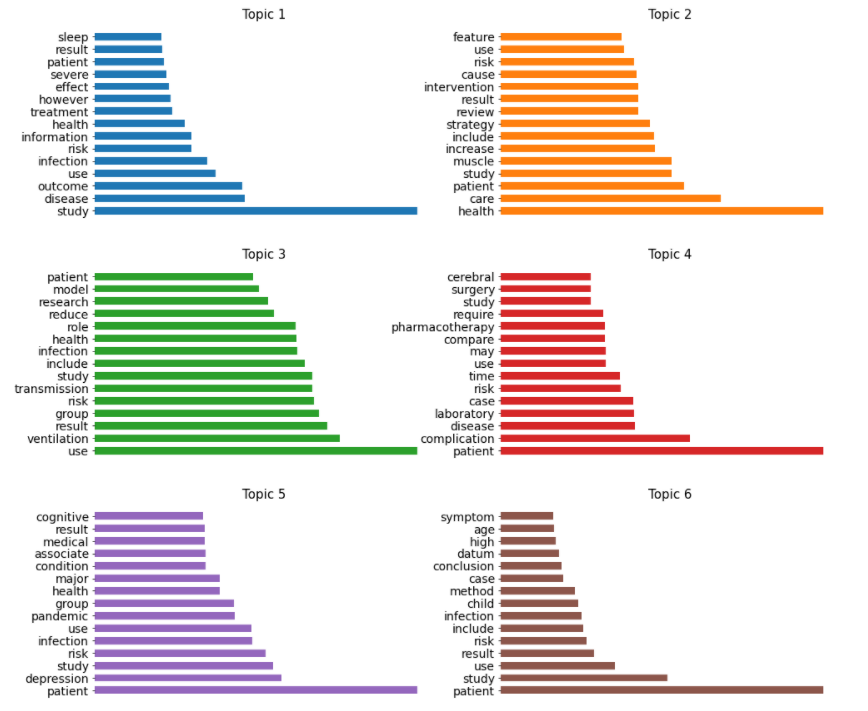
I also used Topic Coherence which is measures to score a single topic by measuring the degree of semantic similarity between high scoring words in the topic. These measurements help distinguish between topics that are semantically interpretable topics and topics that are artifacts of statistical inference.



The higher the c\_v coherence score is, the more suitable the topic number should be. However, the coherence score will vary if we run the iteration different times. Hence, I choose 6 as the topic number for analysis.

Perplexity (-7.46) is used as a metric for model evaluation

Below is the plot showing word per topic:



**Takeaway**

•The first topic is severity of disease. We can see the word like severe,patient,infection.

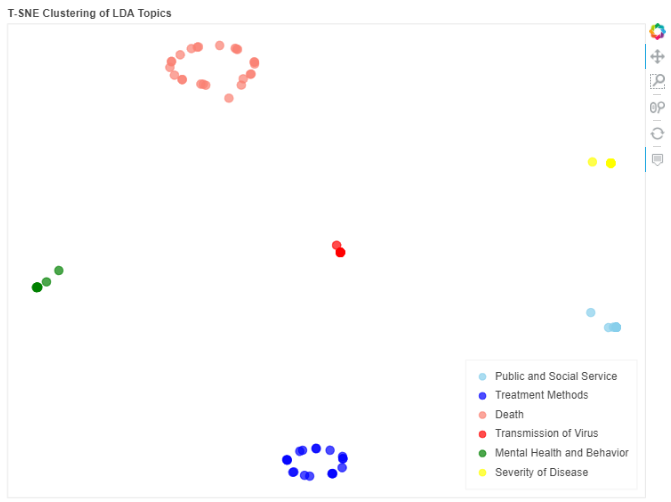
•The second topic is probably talking about the public service and work during pandemic. We can see the word like intervention,study,review,result.

•The third topic is probably talking about transmission dynamics of the virus. We can see the word like transmission,model,group.

•The fourth topic is probably talking about mortality/death. We can see the word like surgery,complication,treatment,

•The fifth topic is probably talking about mental health of people during pandemic. We can see the word like depression, pandemic.

•The sixth topic is probably talking about how it is affecting for different age groups. We can see the word like age,high,child,infection,risk,conclusion.



The label of each article is decided by the topic with highest probability. According to the potential risk factors that Task 2 mentions and the topic inference acquired from the word-per-topic part, I lable each topic with some 'real ideas'.

As we can see from the figure, there is not too much overlap between documents of different clusters(topics) but the clusters are very close with each other.

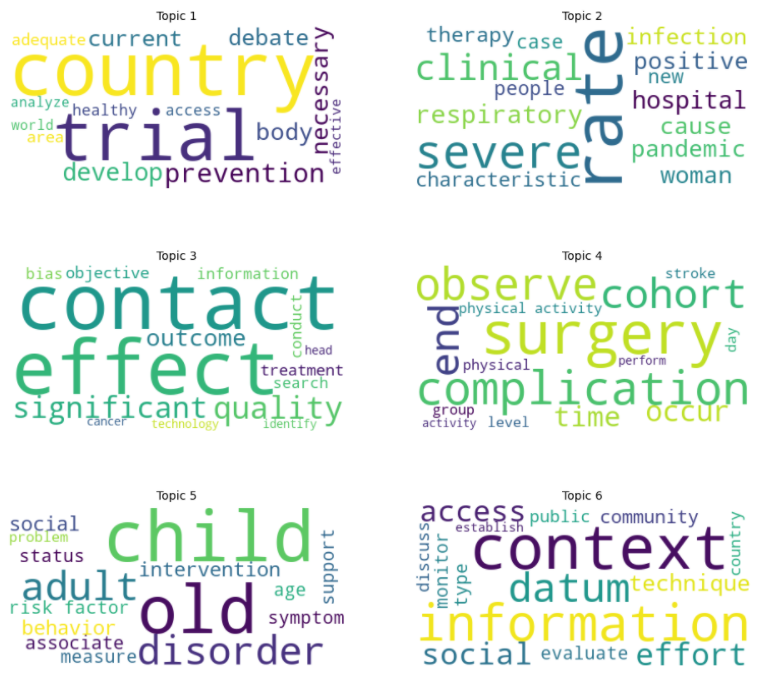
**Why Choose T-SNE?**

'T-SNE gives the impression that it has classified the data by bringing it to two-dimensions but in reality, it doesn’t reduce the dimensions. It is a visualizer, which tells how each class is distributed and is there any overlap between them'. An intuition of T-SNE from the article on Medium: PCA vs LDA vs T-SNE — Let’s Understand the difference between them

**2nd method:**

## Topic Extraction using SentenceBert + K-Means Clustering + TF-IDF

Leveraging BERT and TF-IDF to create easily interpretable topics. Although topic models such as LDA and NMF have shown to be good starting points, based on a [medium](https://towardsdatascience.com/topic-modeling-with-bert-779f7db187e6) article, it took quite some effort through hyperparameter tuning to create meaningful topic. The author of it invented a modified TF-IDF alogirhtm called 'class-based variant of TF-IDF' to help extract the topics from clustered documents.



**Takeaway**

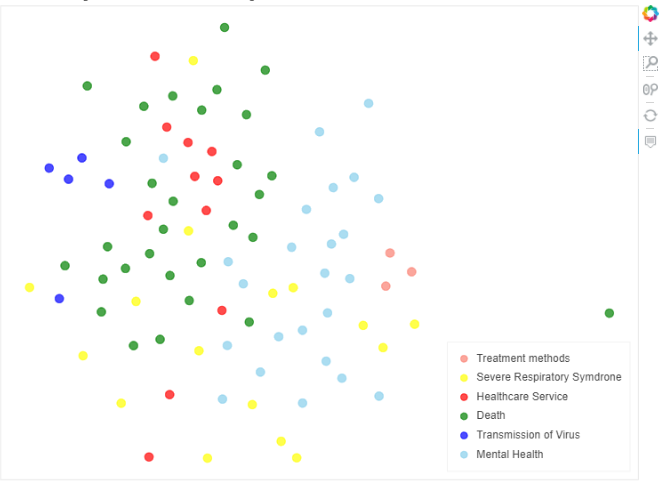
* The first topic is similar with topic 3 generated by LDA, talking about the transmission dynamic of the virus.
* The second topic is talking about severity
* The third topic is talking about research
* The fourth topic is similar with topic 4 generated by LDA, talking about complication, surgery
* The fifth topic is similar with topic 5 generated by LDA, talking about mental health
* The sixth topic is probably similar with topic 6 generated by LDA, talking about how it is affecting for different age groups.

# Modeling

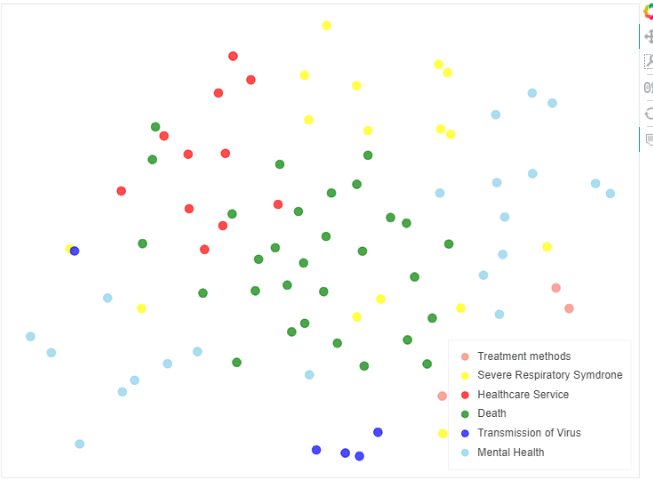
* Topic Modeling using Latent Dirichlet Allocation
  + Selected the number of topics (6) based on coherence score & perplexity
* Topic Modeling using BERT+ K-Means clustering +Term Frequency-Inverse Document Frequency
* Dimensionality Reduction with Autoencoder in Keras. I used this [blog](https://blog.keras.io/building-autoencoders-in-keras.html) as a guidance to build the autoencoder

Topic extracted using this is more specific (2) compared to the original (1)

1)



2)



# Conclusion

Topics extracted from two methods are similar, but the topics from the second method are more specific.

**LDA**

* Traditional and go-to algorithm for topic analysis
* Can extract real ideas from topics and cluster documents based on topics easily with this algorithm

Advantage : Can get relatively clear boundaries when clustering the documents

Disadvantage : real ideas from topics are somewhat ambiguous because it is not easy to tune the hyparameters of this unsupervised learning algorithm

**Bert and its modified versions**

* Outperform everything and achieve state-of-art results in all kinds of NLP tasks
* SentenceBert can give very good representation of the documents

Advantage : Real ideas from topics are more meaningful and interpretable

Disadvantage : Decision boundary of documents clustering is not clear