* **Did you work with others on a team to create your project? If so, please add them (max 2) below.**

Abhishek Jonnavittula

Artur Bakhtriger

Crystal Song

* **Describe in a few paragraphs the problem you are trying to solve and the technical solution you implemented.**

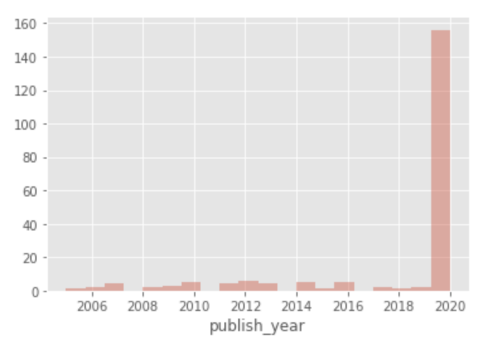
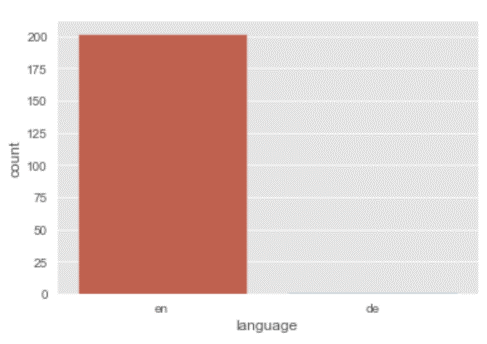
**Overall Objective:** Our goal is to use natural language processing and other techniques to identify a subset of articles that will provide us insights and information on non-pharmaceutical interventions for Covid-19 and other related viruses in the CORD-19 dataset, which is a resource of over 195,000 scholarly articles.

Specific characteristics of our solution are given below:

* + Efficiently and effectively “read” through the dataset and narrow down the scope to key documents that contain information on non-pharmaceutical interventions.
  + Further analyze the subset and systematically generate and identify key words and high frequency words
  + Create a document term matrix by applying LDA (Latent Dirichlet Allocation) to create an optimal number of “topics” to cluster the documents for further investigation
  + Analyze top frequency words within each topic, and associate them with target questions (please refer to “Task Details” on [Kaggle](https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks?taskId=587) for the specific target questions) to generate insights
  + Additional analysis was also carried out to compare LDA with Non-negative Matrix Factorization (NMF) in terms of topics generated on the same data set
  + Provide summaries and guidelines for audiences to leverage our results for further study

**Detailed Approach:**

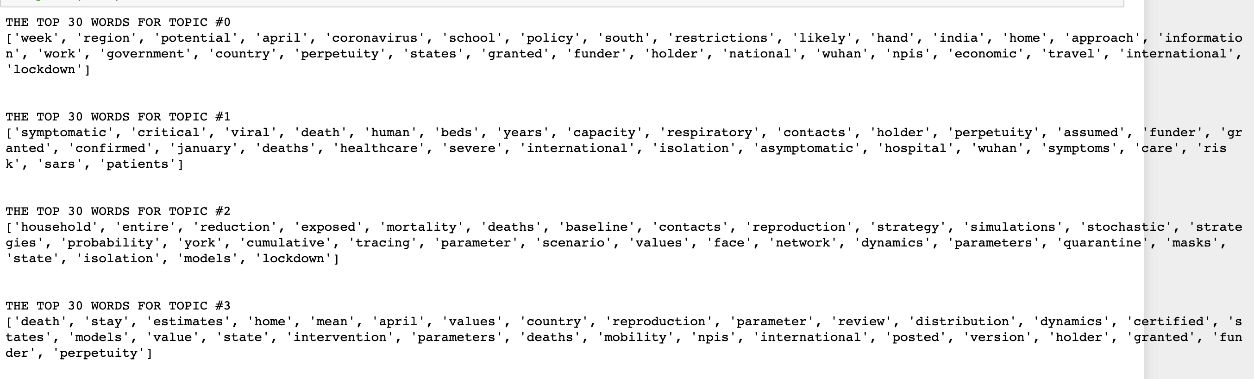
* + **Create** panda **dataframe** over the metadata.csv file; For the purpose of our analysis, only the following columns from the metadata.csv file are selected – sha, title, abstract, publish\_time and pdf\_json\_files
  + **Drop rows** with missing values in key fields, i.e. title, abstract and pdf\_json files.
  + **Convert** all values in the title and abstract fields to **lowercase**, to ensure consistency and to reduce duplication. Having values in the title and abstract fields is important to the initial feature engineering section of our solution. We noticed that publications of the type pdf\_json are the only ones that have an abstract. The total document count after these steps is 55,365
  + In order to arrive at a focused data set, we used a **keyword search** on the abstract to filter out non-relevant documents. For the purpose of our analysis the following key words were used – ‘non-pharmaceutical’ and ‘npi’. This reduced our document corpus to 203. It is important to note that different key words can be used to refine the corpus as needed
* Using **data visualization** techniques, we were able to get better understanding of our data set. Key learnings are that most of the publications were created between 2018 to 2020 and that the documents are predominantly written in English.

* To ensure that the results from our topic modeling exercise are meaningful, only **English** publication created after 2018 are retained. Our final corpus now contains 157 documents
* A **word cloud** created on the abstracts from 157 produced interesting results

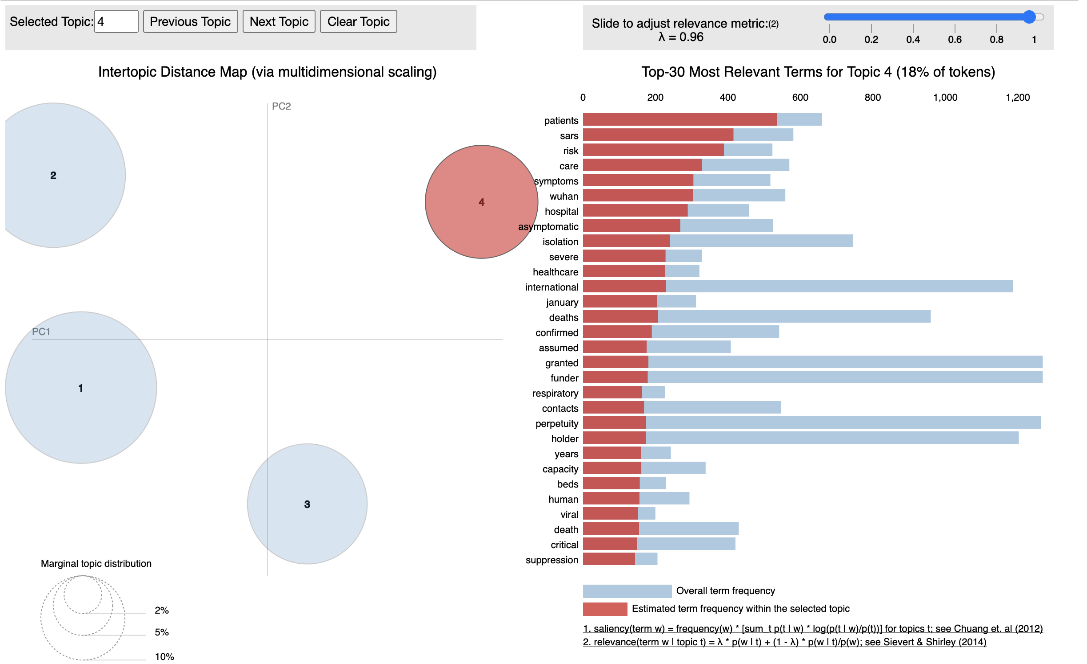


* A **list object** containing the body text elements from each of the 157 documents was then created
* A **document term matrix** was then created using the above list and **LDA** was used to segregate the document data into 4 topics. The number 4 was determined through trial and error.



* A **document-topic matrix** was then created to enable easy **correlation** between topics and associated documents.
* Documents in each topic were then reviewed to better understand the topic and tag them to the specific questions.
* A **pyLDA visualization** was also used to aid the discovery process

(\*Note: the numbers on the bubbles indicate the size of the cluster, and they are different from the index of the topic. 1,2,3,4 on the distance map corresponds to topic [3], [2], [0], [1] respectively.)



* Additionally, the topic modeling exercise was repeated using **Non-negative Matrix** **Factorization** to further validate our topics
* **What are the pros and cons of your approach?**

We have broken out the pros and cons of our approach according to the steps of our analysis:

Initial Data Cleaning

One of the initial steps we have done to identify relevant publications in the 195,000 document corpus. This was done using a key word search on the following terms- “non-pharmaceutical” and “npi”

**Pros:**

1. Quickly narrowed down the number of documents upon which the NLP algorithms will be executed on; reducing the time for analysis and usage on computing power
2. For a dataset that is already constructed in a way to present articles on a selected topic (Coivd-19, Sars and other coronaviruses), using more specific and directional keywords is helpful to ensure the results are targeted, specific and directly relevant for our study
3. Since the terms “non-pharmaceutical” and “npi” are rather standardized verbiage used by the academic community, we think the risk of omitting a large number of relevant articles is relatively low
4. Given the LDA and NMF are efficient algorithms, the solution can be executed without GPUs

**Cons/Improvement Opportunity:**

1. After apply key word search, inclusion of only recently published articles, and filtering out documents in foreign languages, we are left with 157 documents, which is a very small subset of the dataset, leading to potential risks of leaving out important documents that may be helpful for our study.
2. We only used the documents with abstracts available, and the key word search was done on abstracts only. A potential improvement opportunity is to conduct the same process over the entire body text for better inclusion.

Clustering using the LDA vs. NMF Methods and Other Potential Alternatives

We have chosen the LDA method to conduct topic modeling and have done additional analysis using the NMF method for comparison. However, there are other alternatives that can be used, and a combination of various methods can also be applied.

**Pros:**

1. Using both LDA and NMF, we were able to compare the results and choose the one that provides more insights and relevancy

**Cons:**

1. Determining the number of topics using the LDA algorithm can be quite arbitrary sometimes. Through trial and errors, we have chosen n = 4; but there may be a more optimal n that can be used to generate better results

Data Analysis & Insights Generation

**Pros:**

1. Using word cloud, matplotlib, seaborn and pyLDA, we have created visuals to present our results in a more user-friendly and engaging format
2. Before finalizing on the number of topics we wanted to generate, we have tested a few different scenarios and analyzed the result to decide on clustering into 4 topics. This iterative approach ensures that our topics are focused and not overly segmented

**Cons:**

1. Our analysis is relatively simple and may not provide direct answers to the targeted questions we are interested in addressing
2. Our team has no prior public health or healthcare background, making it hard to judge the quality of the documents or conduct more professional level analysis and research

* **What are your findings on the given dataset?**

After many iterations and experiments, we have applied the LDA method to create 4 clusters/topics on the subset of documents we identified as relevant to our research topics. Based on the key words and the titles of the documents within each topic, we have summarized the key themes of the four topics as following:

**Topic [0]:** Overall analysis of non-pharmaceutical measures and behaviors, such as masks, and the effectiveness and impacts. This topic focuses on **public response and impacts in the society.** Compare to Topic #3, topic 0 focuses more on presenting current restrictions and policies rather than on future trends and predictions. **(Topic 3 on the Intertopic Distance Map)**

**Topic [1]:** The **impacts** of NPIs on the results and pressure on the **healthcare systems and the healthcare demand**, with references to specific studies in certain countries and other similar diseases such as **SARS and Influenza**. **Topic 4 on the Intertopic Distance Map)**

**Topic [2]:** **Statistical modeling and research** results for specific Covid-19 related topics. The documents from this topic is a lot more **quantitative** than the other topics, and mostly introduce complex models/forecasts as the main subject of the documents. **(Topic 2 on the Intertopic Distance Map)**

**Topic [3]:** Observational and modeling studies on npis in the socioeconomic context, such as lockdown, social distancing, etc.. This topic focuses on **conducting comparisons and scenario analysis** on various strategiesand predicting **future outcomes, effects and trends. (Topic 1 on the Intertopic Distance Map)**

To further prove our findings; we have generated an intertopic distance map to show the inter-topic relationship, and how relevant they are in the overall article dataset. Shown on the map, topic 1 and 4 (which is actually topic[3] and topic [1]) are the most distant, while topic 1 and 2 (which is topic [3] and topic [2]) are the closest.

Based on our analysis on the key words (see above for the topic summaries), topic [2] and topic [3] both feature a lot of modeling and statistical research, which explains why they are close in distance with key words such as “models; parameters; distributions; etc.”. Topic [1] focuses on the short-term impacts of NPIS on the healthcare system while topic [3] focuses on long term predictions and comparisons on different strategies.

* **How might these findings help the fight against the COVID-19 virus?**

Given the audience’ specific needs and goals, he/she can leverage our results to quickly sort through over 190,00 documents to find the subset of documents that may be interesting and conduct further segmentation and analysis to address specific questions. For example, below we have picked out a few documents that are helpful in answering some specific question, by first narrowing down to the topic level, and then narrowing down to specific documents

* + **Rapid design and execution of experiments to examine and compare NPIs currently being implemented. DHS Centers for Excellence could potentially be leveraged to conduct these experiments.**

Topic 3;

Document 36: *The effect of multiple interventions to balance healthcare demand for controlling Covid-19;*

Document 106: *Centralized and decentralized isolation strategies and their impact on the Covid-19 Pandemic Dynamics. Some key points are:*

* 1. Adoption of physical distancing measured for a long time may reduce the number of infections by more than 92% by mid-2020
  2. In this article, modeling results also show the correlation between adopting quarantine and the duration of pandemic
  + **Methods to control the spread in communities, barriers to compliance and how these vary among different populations.**

Topic 1 presents in-depth studies of how different methods and policies have been used in different world regions. Documents in this topic can be used as a research source to investigate how the methods’ impacts vary among different populations

* + **Models of potential interventions to predict costs and benefits that take account of such factors as race, income, disability, age, geographic location, immigration status, housing status, employment status, and health insurance status.**

Topic 2:

*Document 43: Estimate of the development of the epidemic reproduction number rt from Coronavirus SARS case data and implications for political measures based on prognostics:*

Some of the costs that come along with interventions including economic burden, collateral damages such as increased frequencies of suicide.

Topic 3:

Document 107:

**Reginal boundaries** and **age cohorts** are the two main sources of heterogeneity in the diffusion process. Both dimensions can shape policy interventions aiming at containing the epidemic.

* + **Research on why people fail to comply with public health advice, even if they want to do so (e.g., social or financial costs may be too high).**

Topic 0;

*Document 125: universal Masking is urgent in the Covid-19 pandemic: Empirical Validation and Policy Recommendations*

*Document 16: To Mask or Not to Mask*

*Document 37: Covid-19 Datasets: A Survey and Future Challenges*

Topic 3:

Document 107:

An interesting fact regarding lock down is that “a too early lockdown delays the start of the epidemic without attenuating its severity”, meaning that an early lock down buys time but it postpones the problem without mitigating its severity. This may be a reason why people are resistant to certain intervention measurements that were put in place too early on.