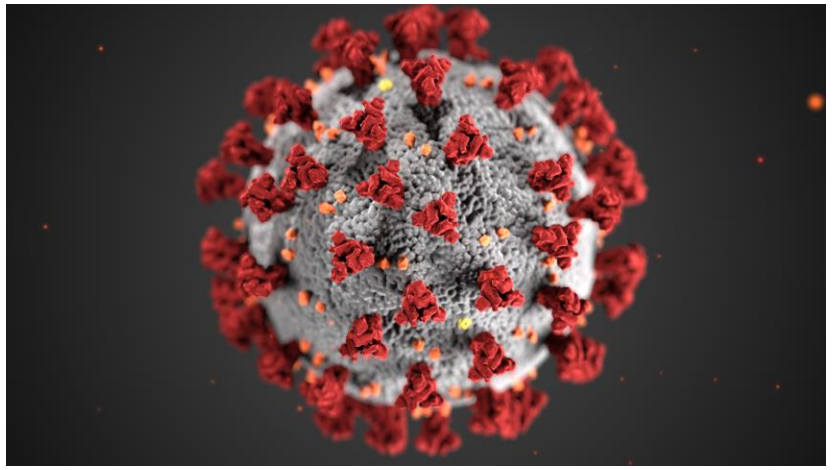
COVID19 - Open-Source Research Article Dataset

The focus of this project will be to analyze the COVID19 open-source dataset of nearly 500,000 research articles on the topic of therapeutics.

We aim to derive insights regarding therapeutics and tools to assist in the research process by providing navigation in the wide plethora of scholarly articles.

INTRODUCTION

- Coronavirus disease 2019 (COVID-19) is a contagious disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus. The first known case was identified in Wuhan, China, in December 2019. The disease has since spread worldwide, leading to an ongoing pandemic
- The scientific community has been extensively researching different aspect of COVID19 including the use of therapeutics.
- Therapeutics - the branch of medicine concerned with the treatment of disease and the action of remedial agents.
- Some fake news regarding COVID19 treatment (UV, bleach, alcohol etc.)



Remdesivir

Hydroxychloroquine

Convalescent Plasma

DATA EXPLORATION

- Over 500,00 scholarly articles
  - Notable Features: Title, Abstract, Publish Date, Authors, Journal
  - Heavily text-based data
- Few null values in dataset except in Journal feature
  - Drop journal feature
- Filter based on therapeutic keywords [1].
  - Over 321,000 scholarly articles which contain at least 1 keyword in its abstract
  - Over 5,800 contain at least 1 keyword in its Title
- Majority of papers published in 2020-2021

Figure 1: Heat Map of Null Values

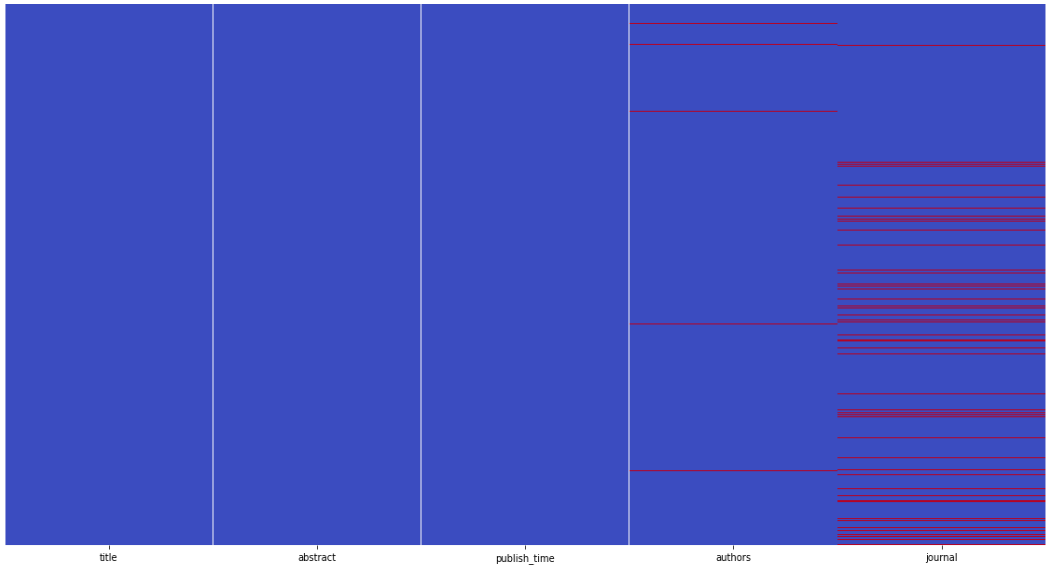
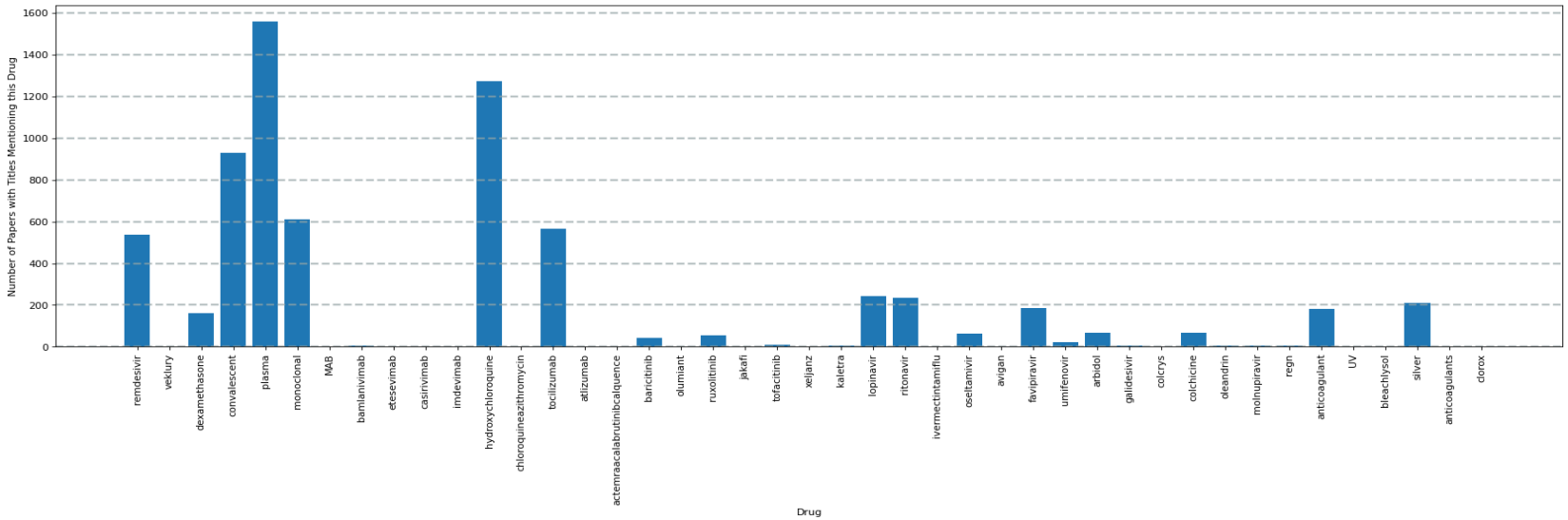


Figure 2: Visualizing Therapeutic Mentions in Title of Research Articles



- Top 10 Most Mentioned Therapeutics
  1. Convalescent Plasma
  2. Hydroxychloroquine
  3. Monoclonal Antibodies
  4. Tocilizumab
  5. Remdesevir
  6. Lopinivar
  7. Ritonavar
  8. Favipiravir
  9. Anticoagulants
  10. Dexamethasone

FEATURE ENGINEERING

- Word counts of abstract, and title explored
  - Normally distributed
  - Long tail on the abstract for some papers with exceptionally long abstracts
- Text data (title, abstract) pre-processed
  - Remove stop words
  - Make all text lower case
  - Remove unnecessary symbols
  - Remove unnecessary punctuation such as /\[]() etc.
  - Remove HTML, convert HTML to ASCII
  - Remove any URLs
  - Lemmatize and Tokenize
- TFIDF Vectorization
- Dimensionality Reduction via PCA
  - 737 Features
- K Means Clustering
  - Optimal K selected to be 20
  - SSE slope plateaus after this point

Figure 3: Title Word Count

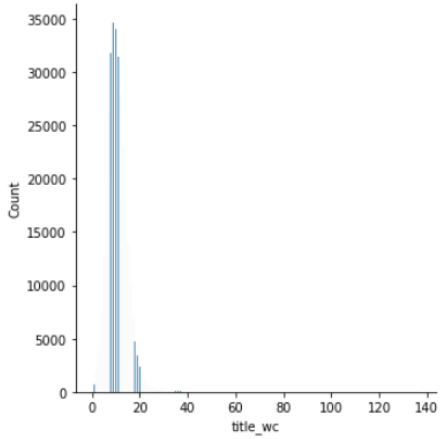


Figure 4: Abstract Word Count

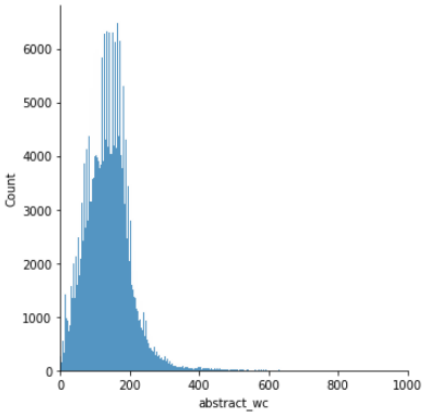
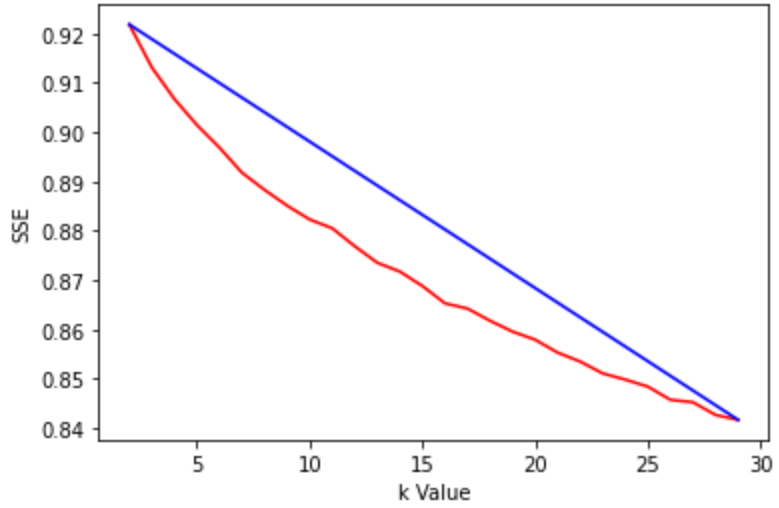


Figure 5: Elbow Method using SSE for kMean Clustering



MODEL RESULTS

- Visualize clusters
  - Dimensionality reduction to 2 dimensions using t-SNE
  - Perplexity = 100 provided best clustering results
  - Colors appear to match the number of clusters as determined by kMeans
- Now that we have clusters, let's analyze topics for each cluster
- Example cluster topics below from Word Clouds
  - General topics can be assigned to each cluster
- Prediction Model Built with Clusters from k-means as Labels
  - XGBoost
  - Test Accuracy = 90%

Figure 6: k=20 Clustering of Therepeutic Related Articles Embedded via t-SNE (2-dimensions)

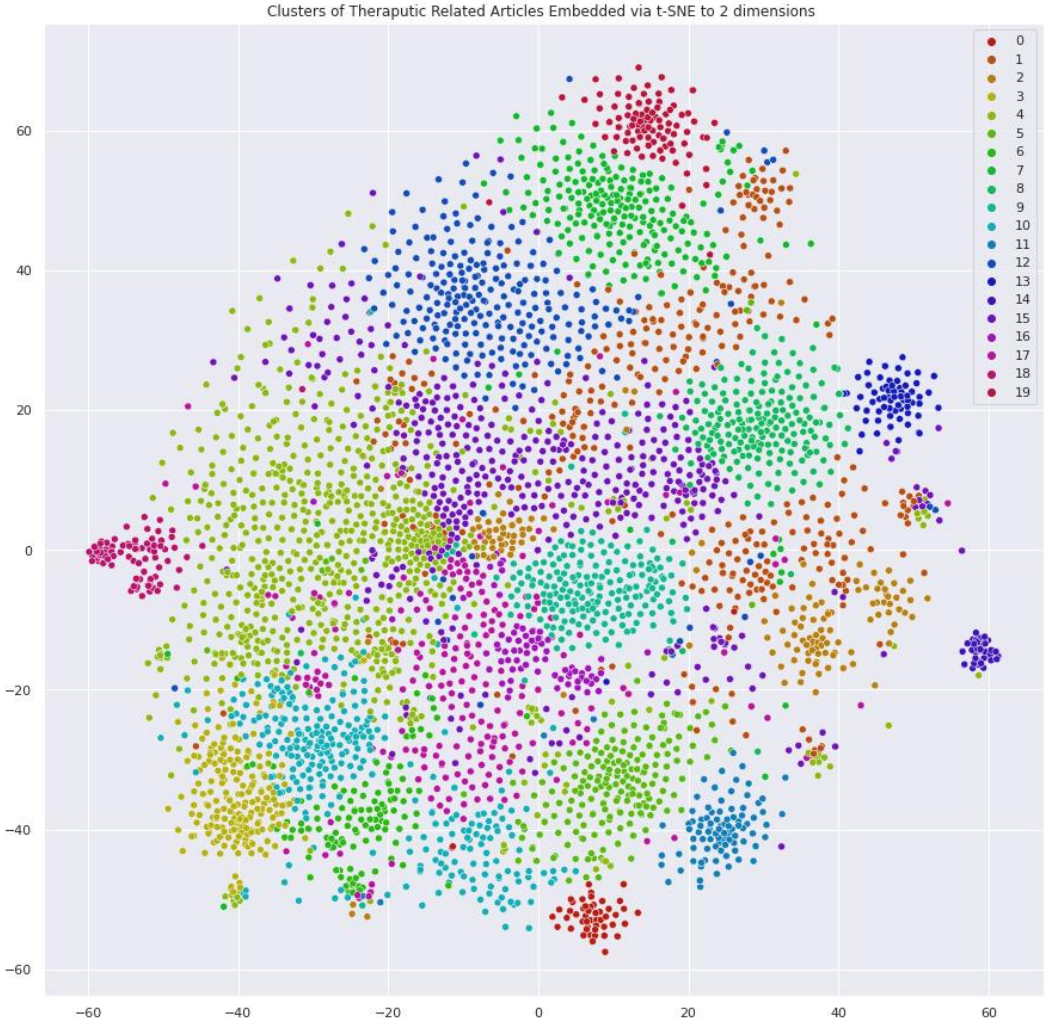


Figure 7: Cluster 1 Word Cloud

Figure 8: Cluster 9 Word Cloud

Figure 9: Cluster 10 Word Cloud

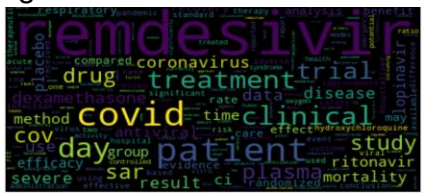


Figure 10: Cluster 3 Word Cloud

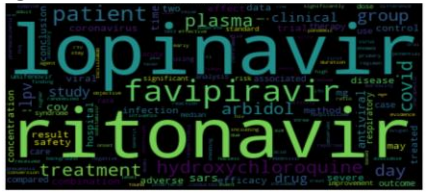
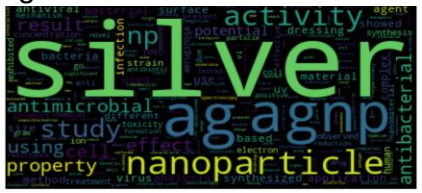


Figure 8: Cluster 8 Word Cloud

Figure 8: Cluster 19 Word Cloud



INSIGHTS & RECOMMENDATIONS

Insights

- Approximately 321,000 of these research papers contain mention of a therapeutic used to treat COVID19.
- Majority (>95%) of the research papers were published in 2020 and 2021 with very few papers before then.
- Convalescent Plasma is the most studied therapeutic.
- Hydroxychloroquine is the second most studied therapeutic which could be due to its publicity spouted by former President Donald Trump.
- Some therapeutics were correlated in their presence in reports which suggest they are related, or comparative studies have been conducted already.
- There appears to be approximately 20 distinct topics (some are related) in research of therapeutics.
- A supervised machine learning (XGBoost) model can be created to accurately predict the topic of a research paper. This can be utilized in order to find relevant articles in the field to aid researchers.
- Many COVID19 publicized on the internet have no research behind them
  - Veklury, bamlarivinab, etesevimab, casirivimab, imdevimab, etc.
  - Including many non-medical treatments such as UV and bleach which are mostly fake news on social media
- Cluster 15 is French research articles

Policy Recommendations

- Public health officials should investigate the usefulness of the following therapeutics for COVID19 (Remdesivir, Dexamethasone, Convalescent Plasma, Monoclonal Antibodies, Hydroxychloroquine, Tocilizumab, Lopinavir, Ritonavir, Oseltamivir, Favirpiravir). These have been most thoroughly studied and so conclusions might be able to be made whether they should be approved for use by doctors to treat COVID19 until vaccines can be rolled out.
- Public health officials across the world can use this project as a tool to determine the credibility of non-traditional treatments for COVID19 (UV, Bleach, Clorox alcohol etc.) as well as other therapeutics (hydroxychloroquine). If the therapeutics do not appear in keyword searched across all research papers they are likely not approved treatments so the public should be notified not to use them as to prevent more harm like in the case of using bleach or hydroxychloroquine.
- A therapeutic that has been commonly mentioned by Donald Trump is hydroxychloroquine despite concerns from the academic community that there is no evidence to suggest it's effectiveness. Journal articles from cluster 1,7,9 and 16 can be studied to draw conclusions by Public Health Officials.
- Research scientists if provided with a research article can input the abstract into the model to find relevant papers which have been identified by clustering. This can help with the research process to help tackle COVID19.
- The Government (NSERC) should continue to fund research for COVID19 therapeutics as they are critical for preventing serious illness and death due to COVID19 before vaccinations can be developed, manufactured and distributed.
- Public health officials should make public announcements or a website which summarizes current summarized research results (in laymans terms) for public information. Especially for treatments which are known to be non-effective or harmful.

REFERENCES

- [1] - https://www.goodrx.com/blog/coronavirus-treatments-on-the-way/

- [2] - https://www.nytimes.com/interactive/2020/science/coronavirus-drugs-treatments.html