SANDIA REPORT
SAND2020-4510

Printed April, 2020

Sandia National Laboratories

Rapid Response Data Science for COVID-19

Alisa Bandlow, **Travis Bauer**, Patricia Crossno, Rudy Garcia, Lisa Gribble, Patricia Hernandez, Shawn Martin, Jonathan McClain, and Laura Patrizi

Prepared by Travis Bauer
Sandia National Laboratories
Albuquerque, New Mexico 87185
Livermore, California 94550

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology & Engineering Solutions of Sandia, LLC.

Available to DOE and DOE contractors from

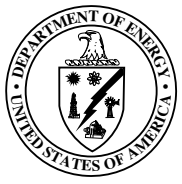
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@osti.gov
Online ordering: http://www.osti.gov/scitech

Available to the public from

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.gov
Online order: https://classic.ntis.gov/help/order-methods



ABSTRACT

This report describes the results of a seven day effort to assist subject matter experts address a problem related to COVID-19. In the course of this effort, we analyzed the 29K documents provided as part of the White House's call to action. This involved applying a variety of natural language processing techniques and compression-based analytics in combination with visualization techniques and assessment with subject matter experts to pursue answers to a specific question.

In this paper, we will describe the algorithms, the software, the study performed, and availability of the software developed during the effort.

ACKNOWLEDGMENT

We'd like to thank Justin Newcomer for helping with the initial stages of bringing the team together.

We'd like to thank Laura McNamara for human factors discussions and recommending a think aloud protocol.

CONTENTS

LIST OF FIGURES

1. INTRODUCTION

This report describes the results of a one-week effort to augment subject matter experts with natural language processing and compression-based analytics to make them more efficient at analyzing almost 30 thousand documents. This work will enable a user to navigate a space with respect to the document semantics. It is difficult for an analyst to communicate complex multifaceted research needs to a search engine. For example, through this study we realized that a query as simple as "stability of coronavirus" assumes a significant amount of background knowledge. A manual query into a search engine, based on our work with subject matter experts (SMEs), might include something like:

> (stability OR inactivation OR disinfection OR survival OR degradation OR "half life")
> AND (coronavirus OR "SARS 2" OR "SARS-CoV2" OR "SARS-COV-2") AND
> (sputum OR "throat swab" OR aerosol OR droplet OR phlegm OR mucus OR mucosa
> OR feces, OR stool OR urine)

Even a query this sophisticated would return many documents that are not relevant to the problem. Our subject matter experts discovered in the course of exploring the documentation that "stability" often referred to the state of patients and not the virus. It is more natural to present a system with examples of useful text and for the system to find related documents. Also, when displaying results, it may be easier to give the user a "big picture" overview of where their interests lie in a corpus rather than an ordered list.

Specifically, we developed a novel combination of compression-based analytics with dimensionality reduction and visualization techniques to assess their application to COVID-19 related problems. This resulted in a method to identify key documents for analysts to rapidly arrive at useful data for addressing COVID-19 research questions. The application of these unsupervised algorithms will help organize the information space to assist SMEs. We also assessed a variety of other technologies available at Sandia National Laboratories (SNL) to determine their applicability to this problem in future work.

Sandia is uniquely positioned to quickly bring together staff with expertise in biosecurity, public health, data science, and human factors. We have ongoing research in compression-based analytics and other information theoretic algorithms. These algorithms operate with minimal feature engineering. Consequently, it is easy to rapidly apply these algorithms in new domains. In other research, they have been applied to authorship identification, analysis of seismic activity, and partition discovery in binary files. In this new work, we will adapt them to biomedical and public health questions for COVID-19. We also have SMEs in biosecurity, public health, and human factors who facilitated the rapid adaptation of these capabilities into this problem domain.

For this work, our SMEs' research question was "Stability of SARS-CoV-2 in aerosol droplets and other matrices." Among the algorithms explored, t-SNE demonstrated the strongest capability of separating documents into clusters for exploration. Our SMEs were able to investigate this problem quickly when

they were augmented by the machine learning algorithms applied in this research. They determined that the scientific documents available were likely insufficient due to the lack of research available at the time of the study to answer their specific question.

2. RESEARCH GOALS AND RESULTS

2.1. EXECUTIVE SUMMARY

Key technical accomplishments include:

- We developed and ran two different search studies, one with two experts and one with three. The SME's addressed the research question "Stability of SARS-CoV-2 in aerosol droplets and other matrices," drawn from the DHS Master Question List for COVID-19[1]
- We studied whether a 2D representation of 29,000 documents, augmented with information about analysts' interests, can be used to help analysts efficiently navigate a large data set to quickly answer a specific question.
- We assembled a team of data scientists, engineers, a human factors expert, a virologist, a geneticist and public health expert, and a biosecurity and biodefense expert into a single team that proposed, organized, and executed a COVID-19 related project in one week.
- We explored several dimensionality reduction algorithms and compression-based analytic algorithms, integrating them in a way so SMEs could leverage them to study their question.
- We made a python library publicly available that can analyze document text, embed documents into a two-dimensional space, and let individuals explore them in Jupyter Notebooks. The scripts are designed to make it easy to work with the COVID-19 Open Research Dataset (CORD-19).

Key findings:

- Regarding the research question, "Stability of SARS-CoV-2 in aerosol droplets and other matrices," the SMEs determined that although there are many relevant articles, the documents in our dataset at the time of the study are likely insufficient to answer the question fully.
- Each SME was able to explore a document set of 29K documents in an hour (including a tutorial and explanation of the tool from the Human Factors and Data Science researchers) to come to this conclusion. The algorithms applied allowed them focus on the most relevant documentation quickly.
- The users' experiences during the second search study suggested that compression-based analytics can be used to augment a dataset in a way that the users found useful for navigating large data sets. However, these algorithms need to be better tuned for maximum effectiveness and quantitative studies need to be conducted to assess performance improvement.
- The agility and speed with which prototype interfaces can be developed for desktop computer systems now allow a team to respond and adapt to specific needs quickly in ways that large scale systems cannot. This kind of rapid adaptation is possible especially because of relatively recent

[1] https://www.dhs.gov/sites/default/files/publications/2020_03_18_mql_covid-19-sars-cov-2_-_cleared_for_public_release_0.pdf, page 4, "What do we need to know?"

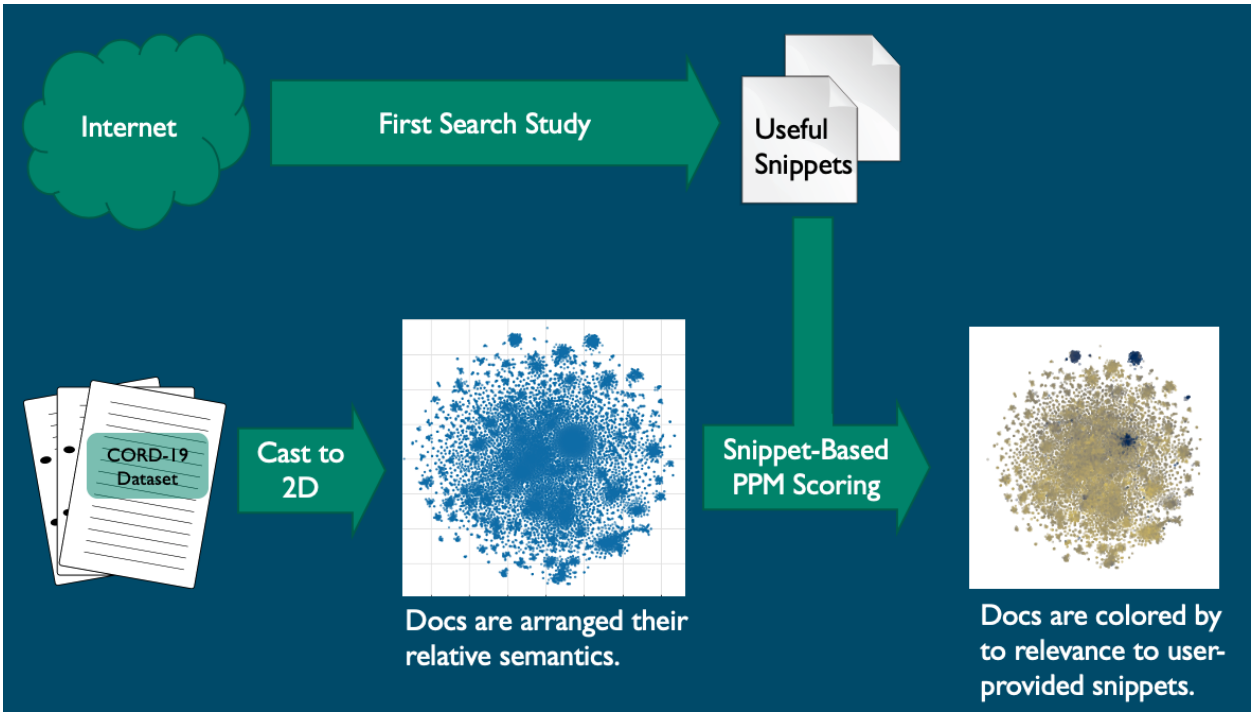


Figure 2-1. Creating an Augmented Document Space

advances in software libraries for visualization and interaction. It is possible to rapidly research and adapt to SMEs' needs as they work through a data set. These are exciting opportunities for addressing pressing problems that require rapid responses.

2.2. DETAILED SUMMARY

The key research question was whether we can use a combination of information theoretic and statistical natural language algorithms to augment subject matter experts studying a specific research question.

The research was conducted in two phases. The first phase involved arranging the the 29K documents contained in the CORD-19 dataset into a two-dimensional (2D) space for SMEs to navigate. The arrangement of the space was based on the relative semantics of the documents as measured by algorithms described in this document and available in the associated source code. The documents were then colored based on their similarity to snippets from other documents provided by the SMEs they considered helpful to answer their question of interest. This phase is shown in Figure 2-1.

In the second phase of the research the SMEs navigated this 2D space. The nature of the interface let them do two things. First, they could study documents that were rated as being similar to their snippets. Second, they could pivot. The flow for this second phase is illustrated in Figure 2-2.

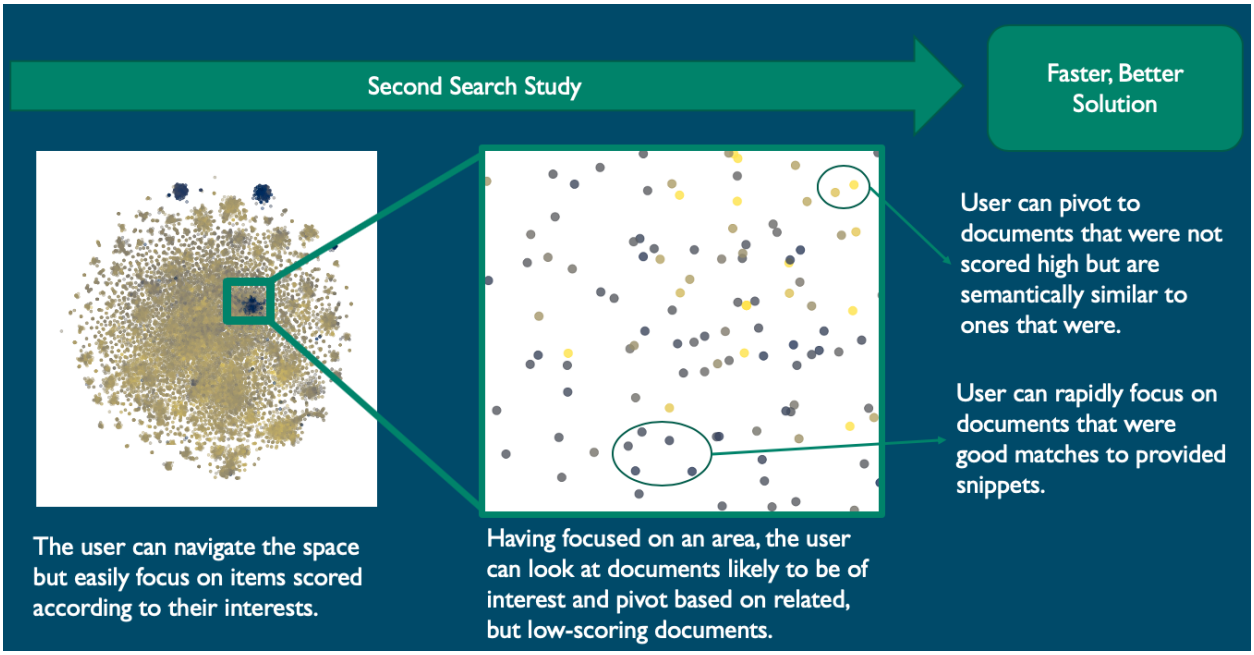


Figure 2-2. Using an Augmented Space

3. PHASE 1: CREATING AN AUGMENTED SPACE

3.1. CASTING DOCUMENTS TO 2D

The data used in this project was referenced in the "White House Call to Action to the Tech Community on New Machine Readable COVID-19 Dataset."[2]. This document set contained 29,315 documents at the time this study was conducted containing a variety of topics relevant for specific problems to the COVID-19 pandemic[1].

These documents were converted into a document/term matrix using Term Frequency/Inverse Document Frequency (TFIDF)[1]. TFIDF is a standard method for converting a document set into a vector space. Each term is assigned a weight in each document d. The weight is a function of both the number of times the term occurs in d and its frequency in the document set as a whole. A term will receive a higher weight in documents where it occurs frequently, but terms that occur frequently in the document set as a whole will generally get a lower weight. For this project, only terms that occurred in at least 20 documents were considered for inclusion in the feature space. This may have been too restrictive, as it reduced the possibility that small clusters of documents on specialized topics could form. The resulting sparse vector space contained 60,458 terms.

In order to display these results to a user, the documents were placed in a 2D space using three forms of dimensionality reduction, Singular Value Decomposition (SVD)[7], t-distributed Stochastic Neighbor Embedding (t-SNE)[11], and Uniform Manifold Approximation and Projection (UMAP)[12]. Given the timeline of the project, limited time was spent exploring the hyperparameters to find the best layout. This aspect of the project could be more thoroughly explored in future work. Figure 3-1 shows how the layout generated by each of the three algorithms for the hyperparameters were used. A full description of how each algorithm works is beyond the scope of this paper and would be better served by the provided references. However, a brief summary of each algorithm is provided below including some characteristics that are relevant for this project.

SVD generates a rank k approximation of a matrix by performing a matrix decomposition $X = U\Sigma V^T$. In this decomposition, Σ is a diagonal matrix. To create a reduced rank matrix with k dimensions, only the largest k singular values of Σ are kept and, if necessary, the three components are multiplied together[2]. In machine learning for text analysis, this has the benefit of uncovering latent information in relationships among terms. This means that in the reduced rank matrix, a document may have weights

[1]Note that other parts of this document refer to 33K documents. New versions of the dataset with more documents were released as this research was concluding.

[2]In some applications, the matrix components can be used directly, which can save space.

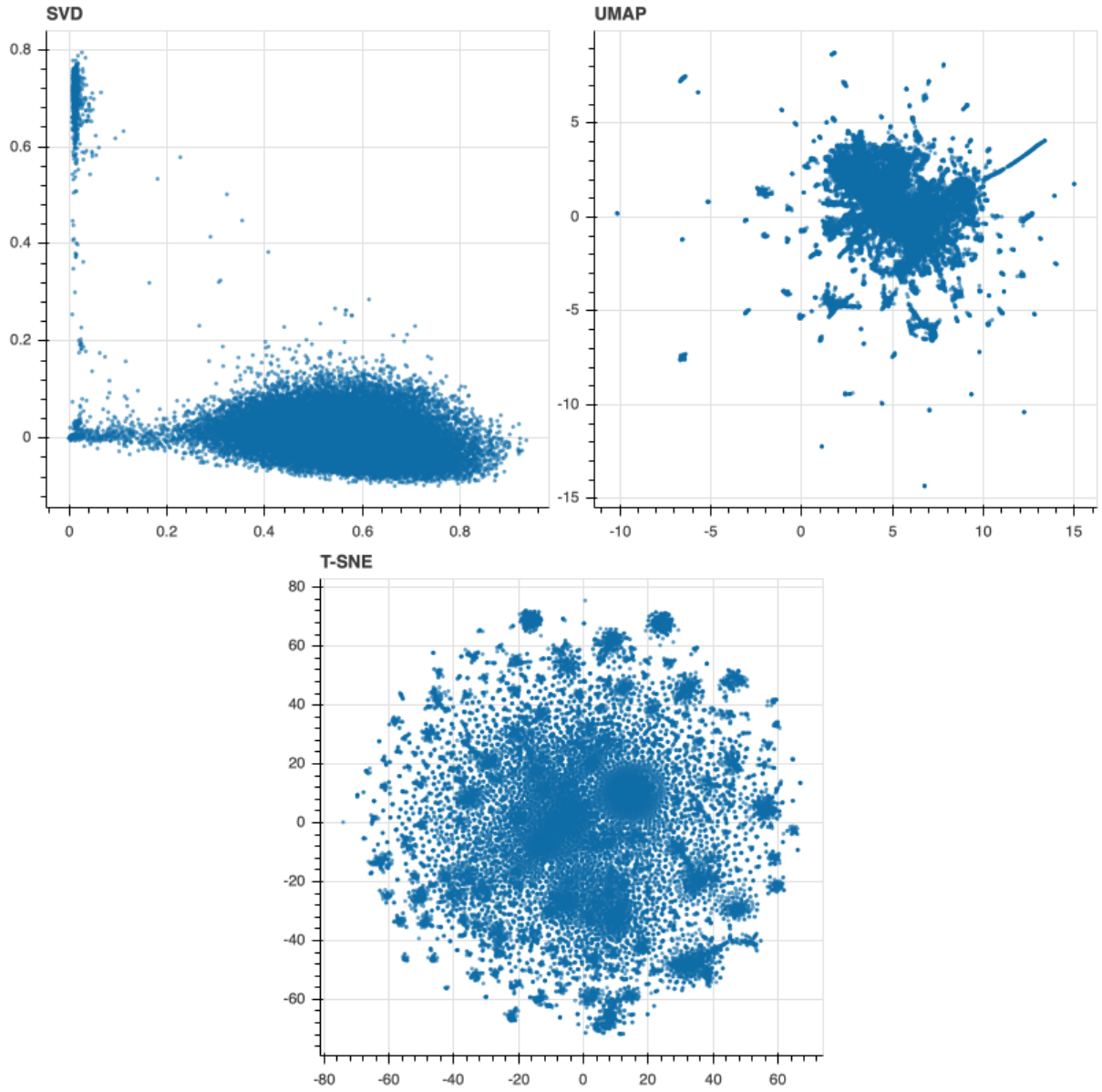


Figure 3-1. 2D Representations of the CORD-19 Dataset

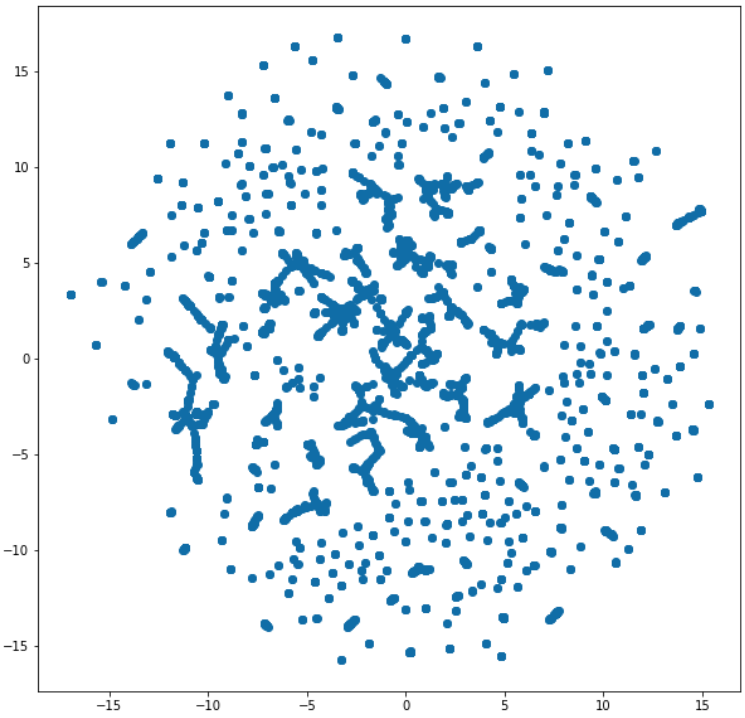


Figure 3-2. A UMAP-based Representation CORD-19 with num_neighbors=2

for a particular term even if the term never occurs in the document but is semantically related. Using SVD in this way with document/term matrices is called Latent Semantic Analysis (LSA)[6]. When used for LSA, it is more common to reduce the rank to approximately 300 and use the results for indexing and search. To embed the documents in a 2D space for visualization (as done in the present work), only the first two singular values are kept and, if necessary, the constituent parts can be multiplied to acquire a feature space with two dimensions. These two dimensions are treated as X and Y coordinates.

The top left diagram of Figure 3-1 shows the CORD-19 dataset cast into two dimensions pictured roughly as two elongated clusters. These clusters are meaningful, as will be described later. However, visually, this layout does not provide much differentiation for a user to explore; thus, this form was not chosen. The other layouts, UMAP and t-SNE, provided more differentiation and visual clustering.

UMAP is a popular method used for finding a lower dimensional representation of higher dimensional data. UMAP constructs a k-nearest neighbor graph of the data based on the distances among the points in the original space and then uses that graph to place the points in the lower dimensional space. A strong point of UMAP is that it does a good job of preserving global structure. For representing the overall topology, how well global structure is preserved can be critical. However, distances are also normalized locally. This means that the distance between two points might not mean the same thing from one part of the two dimensional space to another. For a use case such as ours, where we are more interested in helping the user find tight clusters and drive depth quickly, this can be a disadvantage. If the user is mostly interested in finding clusters, seeing the clusters visually and comparing densities across a 2D space might be more important than the accuracy of the overall topology.

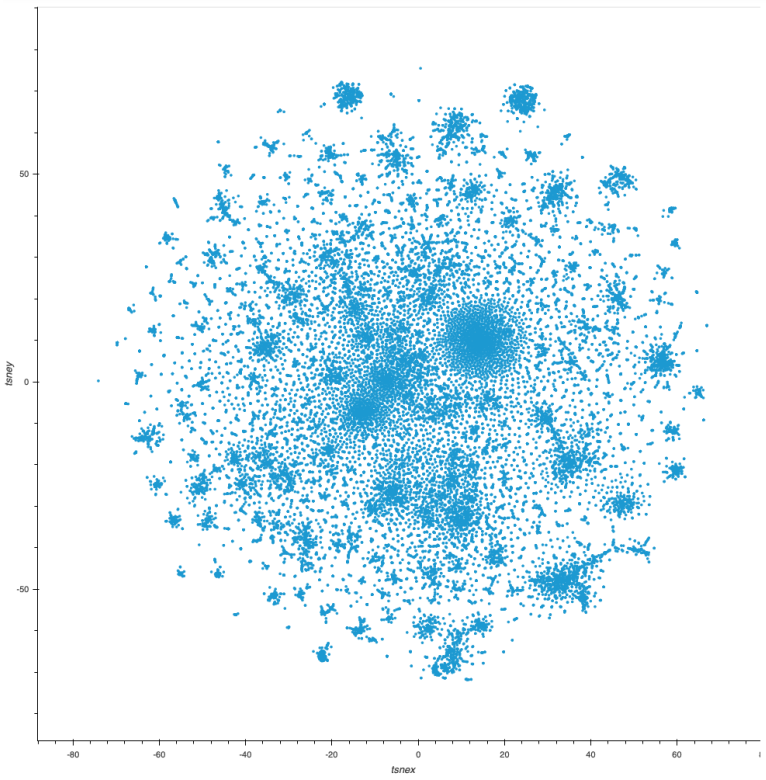


Figure 3-3. T-SNE based layout

The upper right layout in Figure 3-1 illustrates how UMAP arranged the vectors space using the default parameters[3] in the commonly used python implementation of the algorithm[4]. This is certainly suboptimal. Figure 3-2 shows the layout with num_neighbors reduced from 15 to 2. There are a large number of clearly visible clusters. By manipulating spread and min_dist, it will likely be possible to develop a better representation. This should be pursued in future work

The third algorithm, and the one chosen for this study was t-SNE. t-SNE interprets distances between points probabilities where larger distances are interpreted as lower probabilities. If point A is close to point B and far from point C, then conceptually the probability of B given A is high and the probability of C given A is low. Using either Principal Component Analysis (PCA) [9] or random values as an initializer, t-SNE iteratively moves points in the lower dimensional space to reduce the Kullback-Leibler divergence between the original and the lower dimensional representations of the data. This process can lead to a poor representation of the global structure of the space. For points that are far apart, it is less important where they are than simply the fact that they are far apart. An advantage of this technique is that the probabilistic interpretation can aid in uncertainty quantification.

The bottom plot in Figure 3-1 shows the layout produced by t-SNE. There are a variety of parameters that could be tuned to improved the performance here as well. However, because the algorithm produced clearly defined clusters for the parameters used, and further inspection showed that these

[3]Key parameters for our discussion are num_neighbors=15, metric=euclidean, min_dist=0.1, spread=1.0
[4]https://pypi.org/project/umap-learn/

clusters were semantically related was sufficient for this study. Figure 3-3 shows the same plot, but with the individual points were made smaller.

3.2. FIRST SEARCH STUDY

Each search session was conducted as a Skype meeting during "work at home" orders during the pandemic because of social distancing policies. This unique situation provided to conduct a study when the relevant parties are not co-located. The sessions were designed by our Human Factors expert based on best practices[8, 15, 14]. SMEs shared their desktop so an observer could view their screen and watch their activity. The Skype session was recorded with the SME's permission. The first part of the session recorded the SME performing a search on a given topic for 30 minutes. In the second part of the session, both the observer and the SME reviewed the search video, and the SME explained her search decisions and thoughts. The reason for not performing a "think aloud" search during the first part is because SMEs make snap judgments about articles; forcing the SMEs to articulate all their thoughts while searching would interfere with their ability to make these snap judgments. While a post-hoc explanation will not be entirely accurate, we decided that this overall design would capture the most realistic and natural search behavior.

All SMEs were given the same research question relevant to COVID-19/SARS-CoV-2. Each SME opened a new Word document to capture interesting information during their search. The SME was instructed to spend 30 minutes searching for articles relevant to answering the COVID-19 question using the search system or engine of their choice (Google Scholar, PubMed, Web of Science, etc.). They we allowed to use as many search systems or engines as needed. As the SME found relevant or interesting information in an article, she copy-and-pasted the text into a Word document. This Word document, one per user, became the snippets that were used to create scores for the document space.

At around 30 minutes, during a natural pause in the search, the SME was stopped. A break was taken while the Skype video was generated. Next, the Skype video was played back in the Skype window. The SME provided a post-hoc think aloud of her search by explaining why she selected search terms, how she decided which articles to view, and what content she was looking for in each article. This think aloud session was again recorded in Skype, providing a voice over commentary for the original search video. The think aloud session provided context to the text snippets collected in the Word document.

3.2.1. Observations

Our SMEs noted that because SARS-CoV-2 is a novel coronavirus, and we are in the early stages of the pandemic, they were not surprised to find a low number of scientific manuscripts. It takes time to run experiments, draft the manuscript, obtain peer review feedback and publish. Three SMEs performed the same keyword searches across multiple search engines (Google Scholar, PubMed, Web of Science) and found relatively similar results. While the 30-minute search time was an artificially short window, near the end of the session, no new, relevant results were found; thus the SMEs began to focus more on the references in the most relevant articles.

Relevance and Categorization: The SMEs were focused on finding articles with experiments and data. They also employed within-document search strategies to determine relevance. The SMEs typically scanned each document in the following order: abstract, figures and tables, the Results section, and the Methods section. The SMEs categorized the articles in a variety of ways, such as:

- completely irrelevant
  - wrong topic
  - opinion piece, which generally won't be about experimental data
- relevant (experiment on SARS-CoV-2 that covers one or more matrices and one or more environmental conditions)
- semi-relevant
  - overview/survey papers (references are useful to look up)
  - case studies
  - paper that meets some of the search criteria (e.g., a very good stability paper but about SARS-CoV-1 or MERS)
  - a paper that introduced another way of thinking about the problem (e.g., biosafety, one health perspective)

Keywords and Ontology: While the SMEs used a common and consistent set of search terms in their queries, the keywords they searched for in the titles and text were much more complex, heavily leveraging their domain expertise. The think aloud session allowed us to collect these keywords. A majority were synonyms that were not included in the queries. Examples include:

- COVID-19: coronavirus, SARS 2, SARS-CoV2, SARS-CoV-2
- stability: inactivation, disinfection, survival, degradation, half life
- case study: first case
- matrices: sputum, throat swab, aerosol, phlegm, mucus, mucosa, droplet, feces, stool, urine
- surfaces: copper, plastic, stainless steel, cardboard, gown, mask

Disambiguation: In one instance, the SMEs discovered that the keywords could have multiple meanings. Sometimes this difference could be ascertained from the title or abstract, but more often the SMEs had to skim the full body of text. For example, a search on "stability coronavirus" will return articles about the stability of the coronavirus (what they wanted) and the stability of patients who had contracted the coronavirus. Coronavirus refers to both SARS and MERS, so they had to determine if the paper was about SARS 2 (what they wanted) or SARS 1 or MERS, possibly influenza (wrong virus). Another disambiguation was whether the experiment focused on real-life environmental conditions (what they wanted) or biosafety.

3.3. SNIPPET-BASED PREDICTION BY PARTIAL MATCHING SCORING

A product of each search study was a document containing the SMEs text snippets from Search Study 1. The next step in this study involved using these snippets to score the CORD-19 data for the next study.

To create these scores, we used a compression-based analytic. Compression-based analytics use relative compression ratios in various forms to compute a distance between two items. These analytics are rooted in the idea of a normalized information distance [10] based on Kolmogorov complexity. Because Kolmogorov complexity is non-computable, the number of bits in a compressed piece of data can serve as a proxy. This has advantages over many other machine learning algorithms because it does not require significant feature engineering to create a vector space. It also tends to work well with a relatively small amount of data.

Compression-based algorithms have been shown to be high performing algorithms in authorship detection[17][16]. We have applied compression-based algorithms on a variety of different types of data, including the discovery of anomalous network traffic[19] and deception detection in natural language text[18].

Prediction by Partial Matching (PPM)[3][13] is a method for accumulating token probabilities for arithmetic coding[20]. These probabilities are accumulated during the process of compressing a piece of data and can be thought of as a model of the information content of the data. This model can then be used during the compression of new data items. If a new data item compresses well with the new model, that means that the original data and the new one shares information content.

foreach *snippet* s **do**
  Create a model m by compressing s using PPM ;
  foreach *document* d *in the CORD-19 dataset* **do**
    $Score_{d,m} \leftarrow LenCompressed(d, m)/Len(d)$;
  end
end

Figure 3-4. Algorithm for computing document scores

This method was used to score all the CORD-19 documents using the snippets from the first search study. The goal is to compute a score $Score_{d,m}$ for each document d compared to the model m for each snippet. If $LenCompressed(d, m)$ is a function that computes the number of bytes in a compressed form of d using model m, and $Len(d)$ is simply the number of bytes in the d, when we score each document against each snippet using the algorithm described in Figure 3-4.

Figure 3-5 shows the distribution of scores for the CORD-19 documents computed separately for the snippets from each of the two SMEs. A low score means the document is highly compressed and is thus shares information with the snippet. In other words, a low score means that the document may be of interest to the user. The possible scores range from $(0, 1])$. The distributions appear roughly normally distributed with a long tail on the right. The two vertical lines in each histogram indicates the 0.3 percentile, where 0.3%, or 87 documents, fall to the left. The second vertical line in each histogram indicates the 20th percentile. These markers were used for coloring the points the user saw when navigating the documents. However, examining some alternative scoring mechanisms provides some insight into how the documents cluster and are placed in their respective 2D plots.

Figure 3-6 shows a linear color mapping scheme using the Cividis256 palette in Bokeh[5]. This palette is a

[5]https://bokeh.org

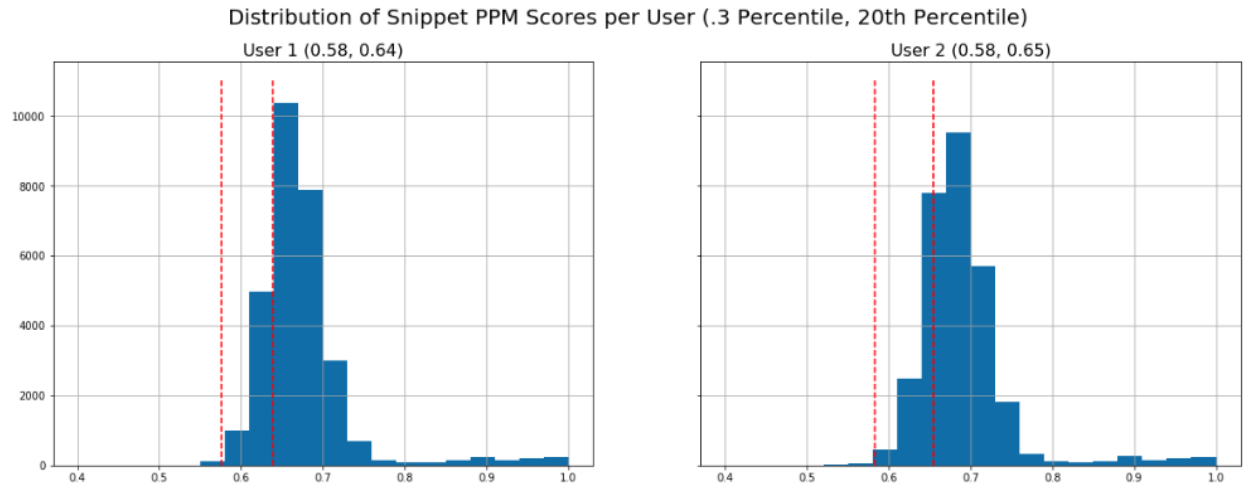


Figure 3-5. Distribution of the Snipped-based Scores for each User

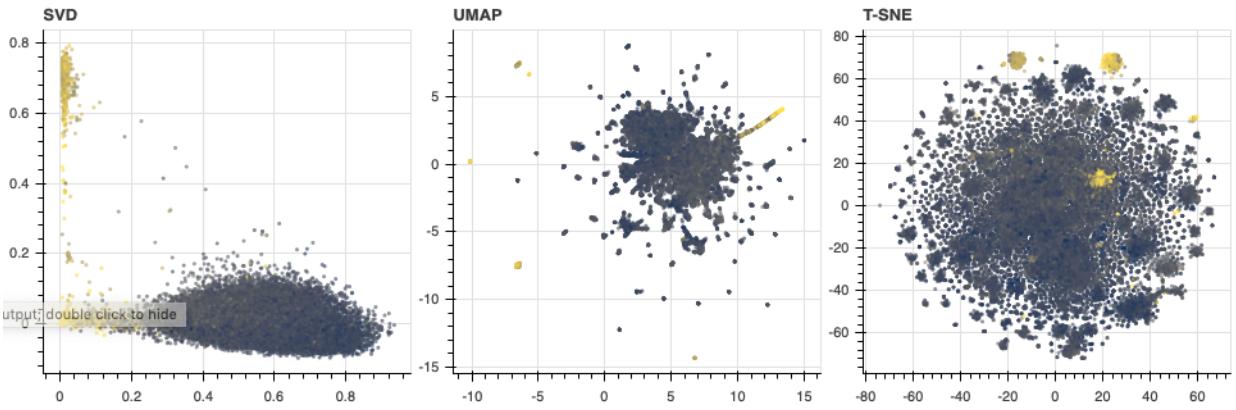


Figure 3-6. Extreme scores highlight non-English Documents

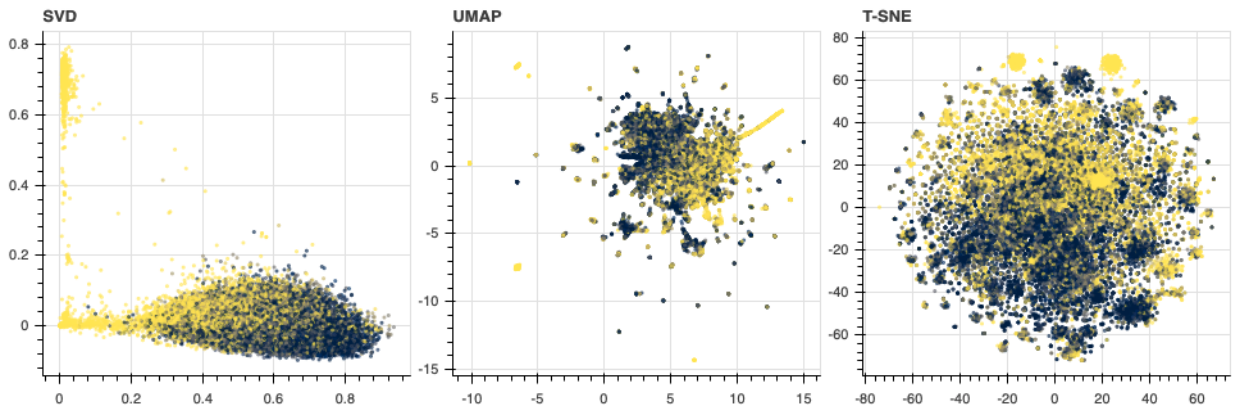
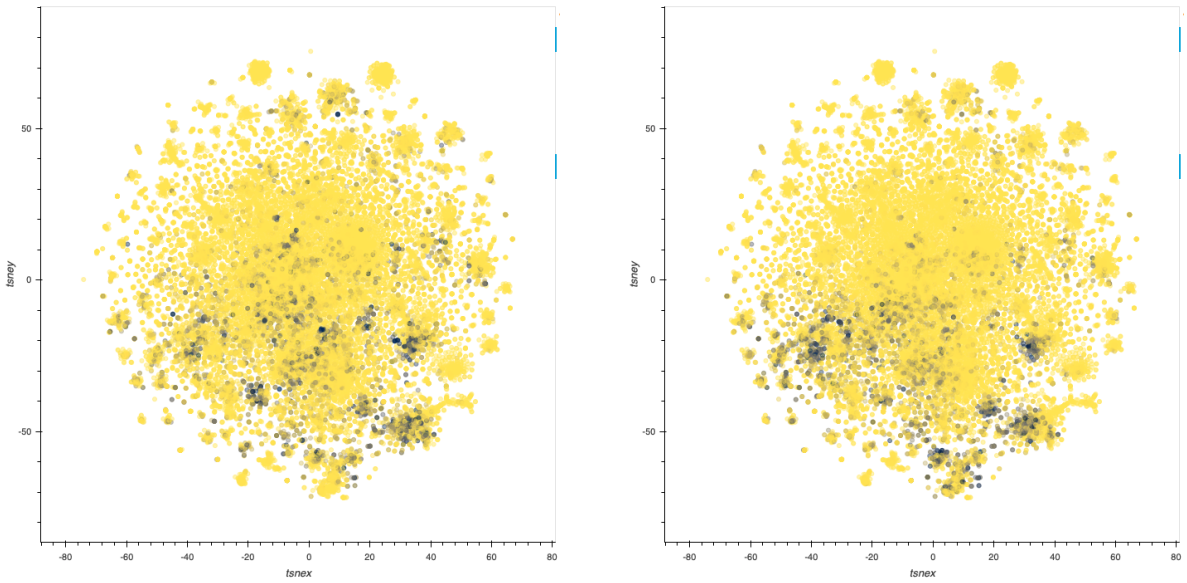


Figure 3-7. Documents Highlighted Based on the 25th to 75th Percentile

gradient from black to yellow, black representing low scores. The colormap provides high contrast for easier visibility. In addition, the use of just two colors on a gradient reduces ambiguity regarding the meaning. The same points are plotted, but using each of the three algorithms described earlier. What is immediately obvious is that most of the points are dark. This is because of the right skew in the scores. Most scores range between 0.4 and 0.7 for both users. The yellow points represent the small number of values at the high end. Upon closer inspection of these documents, it was found that most were non-English. SVD, UMAP, and t-SNE place these documents into one or more clusters. The compression scoring correctly identified these documents as most unlike the English language snippets collected by the SMEs. Note that for t-SNE, there are three regions. The region in the middle of the cloud of documents were generally in multiple languages.

Figure 3-7 colors points with a linear scale that ranges between the 25th and the 75th percentile. This means the bottom 25% of the documents are the darkest color (corresponding to the best match for the user), and the top 25% of the documents are the lightest shade of yellow possible. The remaining documents fall somewhere along the spectrum. For the UMAP and SVD plots, one can see that the space is arranged in a way that tends to separate the black and the yellow points. This suggests that an unsupervised arrangement of the points in two dimensions correlates with the supervised labeling and the documents that match the snippets tend to be close to one another. The arrangement for t-SNE, however, is more diffuse. This is consistent with fact that t-SNE does not preserve global structure as well as UMAP.

Coloring the documents on a scale ranging between the 25th to the 75th percentile still leaves the user unable to differentiate among a full quarter of the dataset, or over 7,000 documents. The decision was made to color only a small number of the best scoring documents enable the SME to quickly identify and examine the documents most relevant to their snippet. Figure 3-8 shows the appearance of the document cloud as shown to the SMEs. The range in this document cloud was from the 0.3 percentile to the 20th percentile. This means that approximately 87 documents were black and just over 23,000 of the 29,000 documents were fully yellow. Of course, the exact coloring for each SME was different, but they did share a few regions. Specifically, there are two regions in the lower right that have highly relevant documents. The regions to the left show more diversity. One SME commented that "I found



(a) Highlighting based on User 1 Scores **(b) Highlighting Based on User 2 Scores**

Figure 3-8. How the highlighted items looked to the users

the tool easy to use, graphically impressive, and overall fun to play with. I also found the color contrast to be extremely helpful."

4. PHASE 2: USING AN AUGMENTED SPACE

4.1. PRESENTING DOCUMENTS TO THE USERS

This project was only a week long and was not a tool building exercise. A key outcome of this first search study involved capturing SME feedback regarding the usability of our developed methodology, specifically whether it allowed them to find relevant information faster. Thus our next search required giving the SMEs the ability to interactively explore the document set, which was easily done. The advent of python[1], jupyter notebooks[2], and the holoviews[3] family of libraries make it easier than in the past to build interactive interfaces for SMEs. In just 55 lines of code (including blank lines for code readability), it was possible to define a user interface that would run in a Jupyter notebook and provide a user with the ability to zoom and pan through the 2D space. This interface is shown in Figure 4-1. SMEs also had the ability to hover over points and see document titles. By clicking on a single point, the user could peruse the beginning of the document displayed on the right margin of the page. This interface also supports a full text search with a well-equipped query language using a pure python search engine[4]. This drastically increased our ability to quickly allow our SMEs to interact with the output of our analytics. Although it was out of scope for the current effort, some of the recommendations from the users during the search sessions could have been implemented rapidly and experimented with immediately.

4.2. SECOND SEARCH STUDY

Each search session was conducted as a Skype meeting. SMEs shared their desktop so an observer could see their screen and watch their activity. The Skype session was recorded with the SME's permission. Each SME's text snippets document was used to score the documents in the CORD-19 dataset, creating a unique highlight of relevant documents. Each was given a quick tutorial of the interface. They were instructed to select the "coloring" related to her individual text snippets. They were then asked to zoom into a cluster of documents that were scored as relevant to the original search (text snippets). An example of what it looked like when zooming in on a sample set of scored documents is shown in

[1]https://www.python.org
[2]https://jupyter.org
[3]http://holoviews.org
[4]https://pypi.org/project/Whoosh/

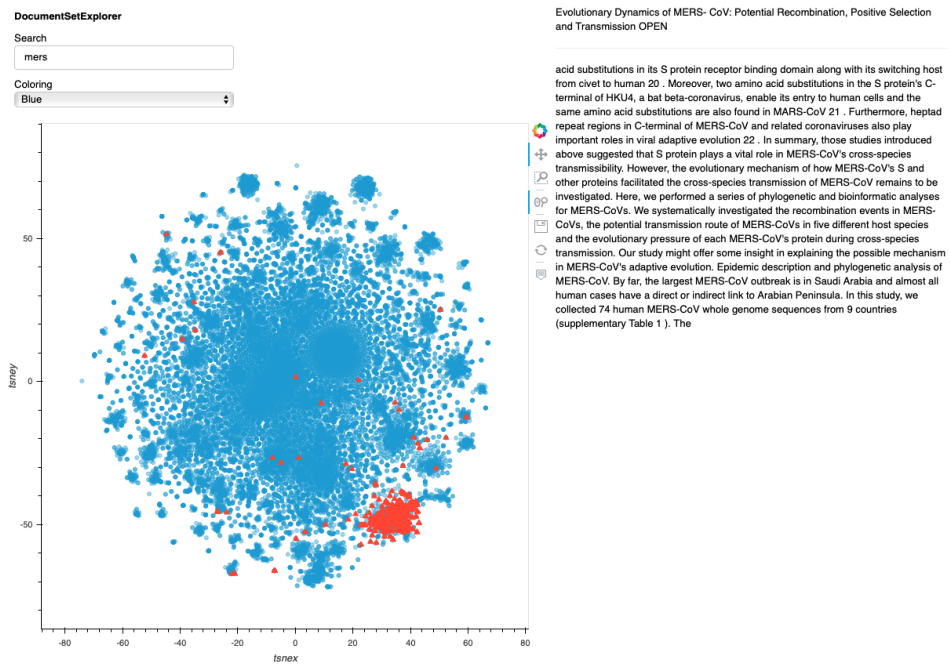


Figure 4-1. Interface Used by Analysts

Figure 4-2. Sample Zoomed in Region for User 1

Figure 4-2. We asked them to review a few documents and score them as either relevant to the research question or non-relevant and provide explanations as to why they were scored in that way.

4.2.1. SME 1

SME 1 investigated about six clusters. She said that the irrelevant documents she found were truly irrelevant. She designated only one cluster as “maybe” relevant.

At the end of her evaluation, we asked her thoughts on using the clustering and highlighting tool to explore documents. She was surprised that some of the “relevant” documents had not been scored as irrelevant. Regarding the low number of truly relevant documents, The SME said that we are at a disadvantage because ver few papers have been published on COVID-19 and SARS-CoV-2 at the time of this study. Thus, the best approach would be to investigate other coronaviruses, SARS-CoV (the virus causing SARS) and MERS-CoV (the virus causing MERS). The SME noted that if more time was available, she was confident she would be able to find most of the existing publications relevant to the question using our method.

When deciding which documents to examine further, she judged relevance based on the document title. She also relied on publication year, journal, institution and authorship to judge relevance, although this citation information was not displayed. When reading documents, the SME noted she missed seeing the publication tables and figures (which also were not displayed) stating "We are missing good data without them." Indeed, it was noted in the first search study, both SMEs focused on reading the abstract first followed by skimming documents for tables and figures to judge relevant. In the second search, the text body was presented as a long block with no breaks, which made it difficult to read. The SME reported a feature supporting within-document keyword search with highlighting would also be helpful.

The SME did not use the keyword search. However, when it was showed to her, she appreciated the search results highlighted on top of the data allowing her to see how they relate to the clusters. This would inform her search strategy.

Overall, the SME reported she believes our tool presents significant opportunity for analysts. She liked the visual nature of the tool, specifically the color contrast between all publications versus those most relevant to her search. She enjoyed "peeling back" the colors to focus on a subset of the most relevant papers for her analysis. Lastly, the SME liked the panning and zooming features and looked forward to using the tool in future search opportunities.

4.2.2. SME 2

The second SME investigated approximately three clusters. She evaluated many of the “relevant” documents as relevant and “maybe relevant”. (Note: Remember that each SME had a different scoring system based on their text snippets document.) She noted that she found a pair of documents she felt were very similar, yet one was scored as relevant and one was not. When she investigated the “irrelevant” documents, she found a lot of papers that had relevant sounding titles, but determined that they weren’t relevant when she skimmed the text. In a real search, she felt the tool would save her time by showing her which articles she can skip. After about three clusters, we asked her to try the keyword

search feature. She initially focused on clusters of red triangles (search results). She noted that one can not tell if a red triangle was originally scored as relevant or irrelevant.

Overall, she liked the graphical interface and thought it was useful. When she clicked on a document node, she noted the need to see the title, authors, year, journal and abstract. These fields help her judge the quality and relevance of the paper more quickly. As SME 1, when scanning a document, SME 2 first looked at the abstract, the tables and figures, results, and summary of methods sections, but again, the section descriptions and section breaks were unavailable in the summary displayed, forcing her to scroll through the main body of text. The SME noted she would like the ability to drag a dot into a list to keep track of what she's found and what was relevant, such as a favorites list. She also suggested a way to track which documents she already reviewed.

The SME liked zooming into a cluster and looking at titles, so she could intuitively understand how the documents were grouped. She requested a better understanding of the two graph axes and how the documents are clustered, as this would guide how she searches (i.e., what does it mean for a cluster to be in the lower right corner versus upper right?). She assumed that more black dots in a cluster meant she would find more documents relevant to her original search. She also requested the ability to see keywords that describe a cluster when she hovers over them rather than or in addition to article titles. Lastly, it would be nice to be able to toggle the irrelevant documents on/off.

5. OTHER APPLICABLE TECHNOLOGIES

Several different technologies were applied to the data used in this study. Even though these technologies were not used in the studies described in this paper, either could potentially be opportunities for transitioning the findings into more of a corporate, production environment.

5.1. SLYCAT RESULTS

We have ingested the results of the text analysis and SME testing into Slycat. Slycat is a web-based platform for analyzing large data sets interactively through a browser interface[5]. Slycat provides a visual analytics approach that blends analysis and visualization into an exploratory framework. Although Slycat was developed for analyzing ensembles of simulation data, the platform can also be used to visualize generic table data, in which the columns are variables and the rows are multivariate samples. Slycat is available open source on GitHub (https://github.com/sandialabs/slycat). An online version of our user manual can be found on Read the Docs (https://slycat.readthedocs.io/en/latest/manual/user-manual.html). Alternatively, we offer a faster path to build a local, single-user version of Slycat using Docker-compose (https://github.com/sandialabs/slycat/tree/master/docker/compose/slycat-compose).

For this work, we generated and ingested a data set that integrates the subset of the COVID-19 metadata table that has a full text description with several analysis results and scoring from SME testing. Figure 5-1 shows the full data set drawn using t-SNE analysis coordinates and color-coded by the first SME's score.

The point size in this rendering has been reduced to reveal interesting structures in the document groupings. Alternatively, we can scale the points up (to facilitate direct selection in the scatterplot) and filter the results to show only the upper end of the SME scores. This enables users to reduce the large number of documents to a targeted set, then interactively drill down and view the text of specific papers. Links to pdfs (text-only versions generated from the COVID-19 json files) are stored in a shared repository and can be interactively viewed by hovering over points in the scatterplot. A pdf viewer in Slycat brings up the text, which can be read. Multiple simultaneous viewers enable comparisons between sets of documents suggested by the analysis clustering as shown in Figure 5-2.

We have also added DOI links so that users can see the original versions of the papers (including graphs and figures that are not included in the text versions shown in Figure 5-2) through the publisher's websites. These links take the user to a separate browser tab outside of Slycat.

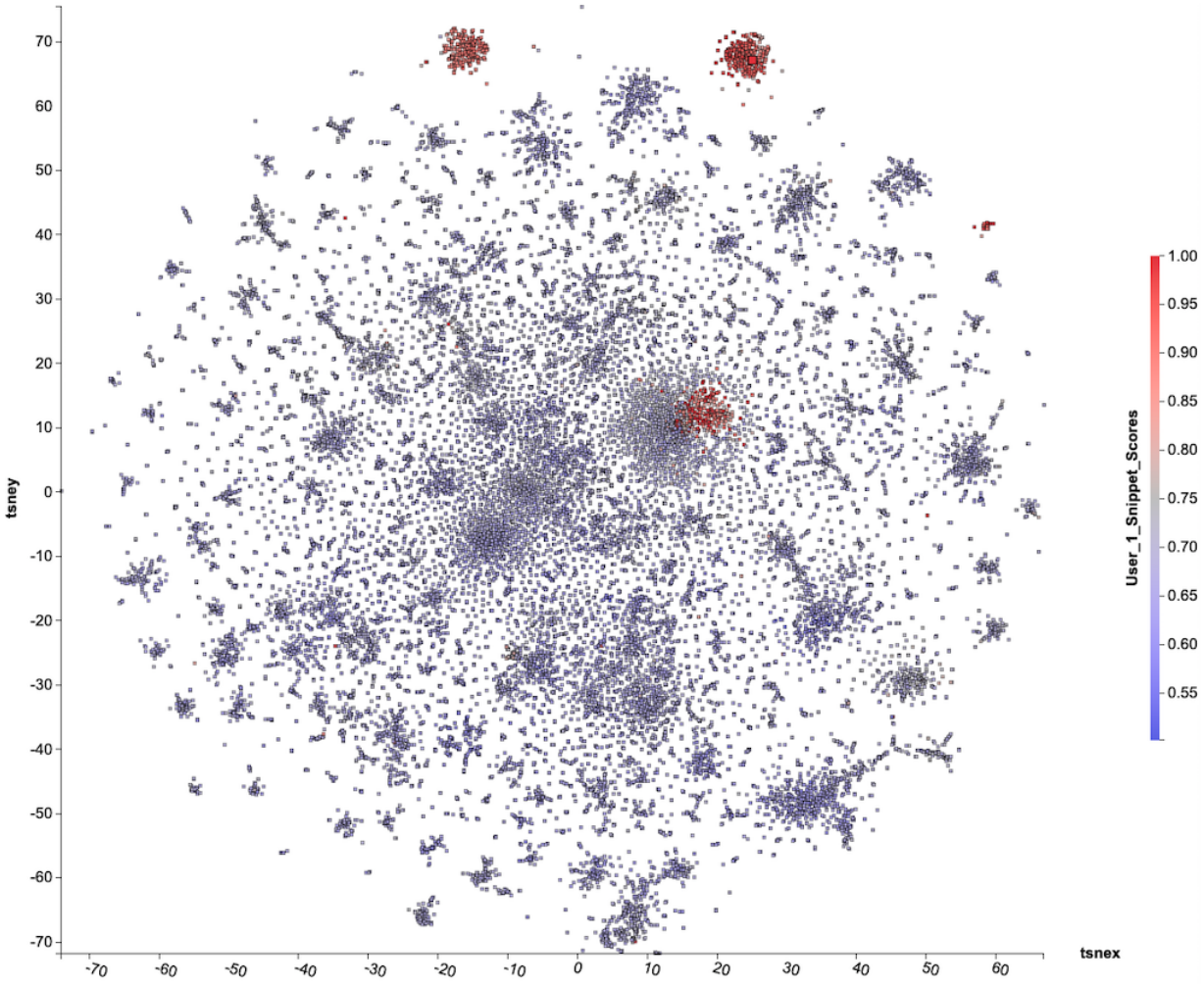


Figure 5-1. Slycat Display

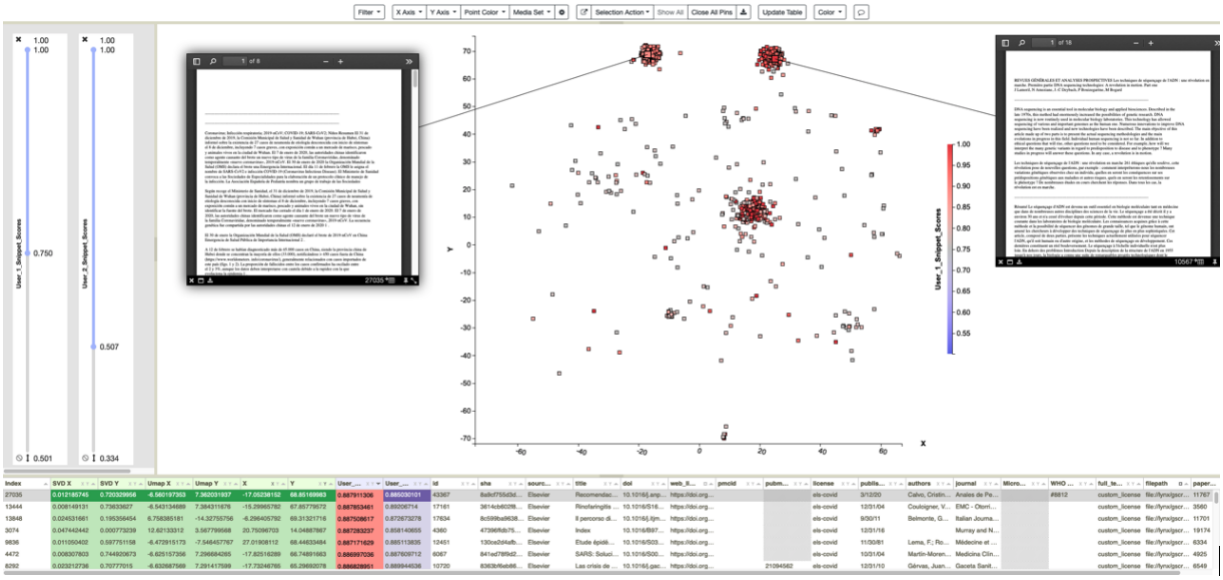


Figure 5-2. Slycat Full Interface

In future work, we would like to integrate the Solr query capability into this interface to enable keyword searching and document retrieval.

5.2. SEARCH AND DISCOVER WITHIN COVID-19 PUBLICATIONS

In order to support this LDRD, we needed a way to discover and search COVID-19 publications from the COVID-19 Open Research Dataset (CORD-19) easily and within Sandia. Our approach uses knowledge from previous (Big Data for Actionable Intelligence - BDAI) and existing efforts (Sandia Open Data Access – SODA), open source software (Apache Solr, Banana), and on-perm cloud computing (Azure Stack) infrastructures to provide this search and discovery capability. A screenshot of the COVID-19 Publications Dashboard is shown in Figure 5-4.

We'd like to acknowledge the following:

- BDAI (ALD 6300 funded effort for FY18 and FY19) – Provided experience and knowledge on open source technologies, data strategy, data architecture and data management.
- SODA (ALD 6300 funded effort – FY20) : Developing the opendata.sandia.gov portal to host data, applications and dashboards.
- Department 10799 – Provided support for the Apache Solr Cluster and Azure Stack infrastructure.
- Department 1461 – Provided Jupyter Notebook support for accessing external sites.

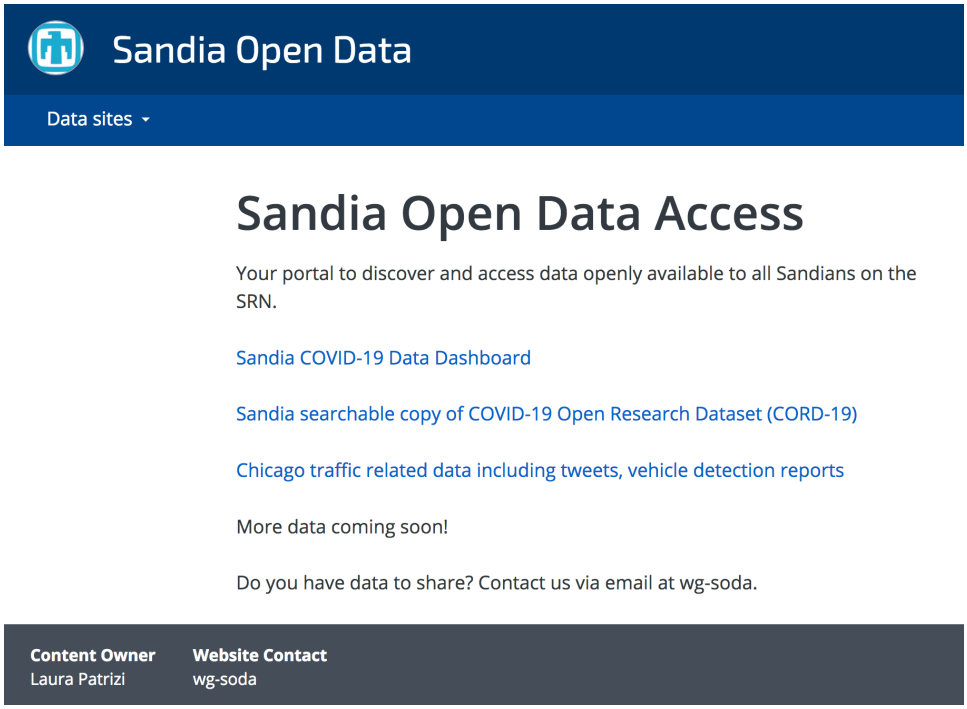


Figure 5-3. Open Data Homepage

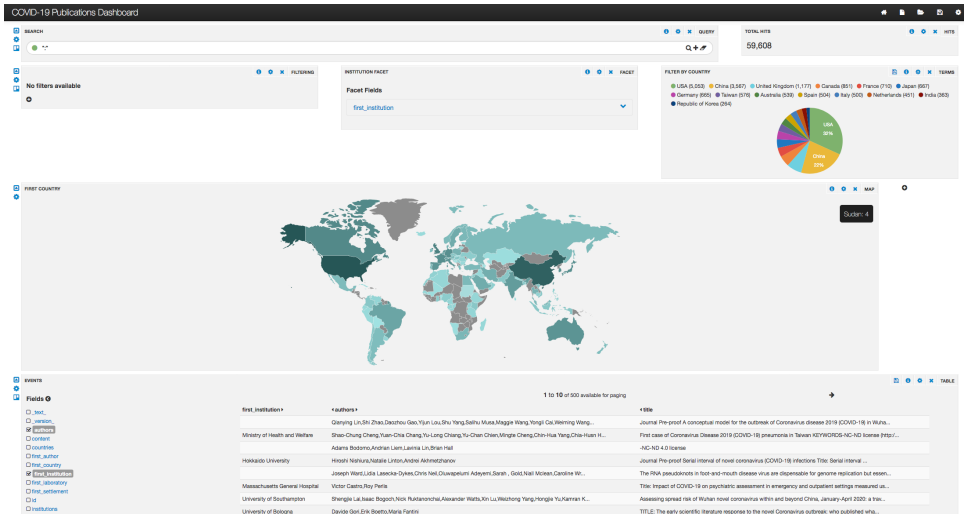


Figure 5-4. Soda Dashboard

1. Download all 33k publications from the Open Research Dataset to a Sandia Gitlab Repo.
2. Created an Apache Solr Collection based on the COVID-19 JSON schema.
3. Developed a Go script to take each JSON publications, curate the fields and index the resulting curated data into the Apache Solr COVID19_Pubs Collection.
4. Configured a Banana Dashboard to point to the Apache Solr COVID-19 Pubs Collection.
5. Placed the dashboard files on a web server which opendata.sandia.gov could access.
6. Updated the opendata.sandia.gov web page to reference the COVID-19 Publications Dashboard

Figure 5-5. Technical Approach for Loading SODA

6. DISTRIBUTION OF RESULTS AND SOFTWARE AVAILABILITY

Two piece of software described in this paper are available as open source.

Name	Description	URL
galen-view	Jupyter stand-alone dashboard software used in this paper's studies. Downloads, analyzes, and presents CORD-19	https://github.com/sandialabs/galen-view
Slycat	Web-based platform for analyzing large data sets interactively through a browser interface	https://github.com/sandialabs/slycat

6.1. GALEN-VIEW

Galen-view is the software used for the search studies in this project and has been published on github at https://github.com/sandialabs/galen-view.

This software downloads the latest version of the CORD-19 data, performs the TFIDF and t-SNE analysis of the documents, and creates the interactive visualization on a standalone computer.

This one week project was executed in response to the "Call to Action to the Tech Community on New Machine Readable COVID-19 Dataset" [1]. The call to action directed participants to submit entries to the Kaggle platform's related challenges. [2]. Our submission fits within the "What has been published about information sharing and inter-sectoral collaboration." [3]. Our entry there will enable one to run the code used in this study.

6.2. SLYCAT

Slycat is available on github at https://github.com/sandialabs/slycat. It is described more thoroughly in section 5.1.

[1] https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/
[2] https://www.kaggle.com/covid19
[3] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks?taskId=583

BIBLIOGRAPHY

[1] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*. Vol. 463. ACM press New York, 1999.

[2] *Call to Action to the Tech Community on New Machine Readable COVID-19 Dataset*. `https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/`. Accessed: 2020-03-16.

[3] John Cleary and Ian Witten. "Data compression using adaptive coding and partial string matching". In: *IEEE transactions on Communications* 32.4 (1984), pp. 396–402.

[4] *COVID-19 Open Research Dataset (CORD-19)*. Version 2020-03-27. Mar. 2020. DOI: `10.5281/zenodo.3715505`. URL: `https://pages.semanticscholar.org/coronavirus-research`.

[5] Patricia Crossno. "Challenges in visual analysis of ensembles". In: *IEEE computer graphics and applications* 38.2 (2018), pp. 122–131.

[6] Susan T Dumais et al. "Using latent semantic analysis to improve access to textual information". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1988, pp. 281–285.

[7] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions". In: *SIAM review* 53.2 (2011), pp. 217–288.

[8] Karen Holtzblatt and Hugh Beyer. *Contextual design: defining customer-centered systems*. Elsevier, 1997.

[9] Harold Hotelling. "Analysis of a complex of statistical variables into principal components." In: *Journal of educational psychology* 24.6 (1933), p. 417.

[10] Ming Li et al. "The similarity metric". In: *IEEE transactions on Information Theory* 50.12 (2004), pp. 3250–3264.

[11] Laurens van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE". In: *Journal of machine learning research* 9.Nov (2008), pp. 2579–2605.

[12] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[13] Alistair Moffat. "Implementing the PPM data compression scheme". In: *IEEE Transactions on communications* 38.11 (1990), pp. 1917–1921.

[14] K. Moran and K. Pernice. "Remote Moderated Usability Tests: How and Why to Do Them." In: *URL https://www.nngroup.com/articles/moderated-remote-usability-test/* ().

[15] Jakob Nielsen. "Thinking aloud: the# 1 usability tool. 2012". In: *URL https://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/*. (2012).

[16] Anderson Rocha et al. "Authorship attribution for social media forensics". In: *IEEE Transactions on Information Forensics and Security* 12.1 (2016), pp. 5–33.

[17] Efstathios Stamatatos. "A survey of modern authorship attribution methods". In: *Journal of the American Society for information Science and Technology* 60.3 (2009), pp. 538–556.

[18] Christina L Ting, Andrew N Fisher, and Travis L Bauer. "Compression-based algorithms for deception detection". In: *International Conference on Social Informatics*. Springer. 2017, pp. 257–276.

[19] Christina Ting et al. "Compression analytics for classification and anomaly detection within network communication". In: *IEEE Transactions on Information Forensics and Security* 14.5 (2018), pp. 1366–1376.

[20] Ian H Witten, Radford M Neal, and John G Cleary. "Arithmetic coding for data compression". In: *Communications of the ACM* 30.6 (1987), pp. 520–540.

DISTRIBUTION

Hardcopy—External

Number of Copies	Name(s)	Company Name and Company Mailing Address

Hardcopy—Internal

Number of Copies	Name	Org.	Mailstop

Email—Internal (encrypt for OUO)

Name	Org.	Sandia Email Address
Technical Library	01177	libref@sandia.gov

Sandia National Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.