Developing a Coronavirus Academic Paper Sorting Application

Christina A. van Hal, Xiaoqian Jiang, Luyao Chen, Yan Chu, Robert D. Jolly, Yaobin Lin, Jitian Zhao, Kang Lin Hsieh

***Abstract*—**The COVID-19 Literature Summary App, now live on the university website, was created for the primary purpose of enabling academicians and clinicians to quickly sort through the vast array of recent coronavirus publications by topics of interest. Multiple methods of summarizing and sorting the manuscripts were created. A summary page introduces the application function and capabilities, while an interactive map provides daily updates on infection, death, and recovery rates. A page with a pivot table allows publication sorting by topic, with an interactive data table that allows sorting topics by columns, as wells as the capability to view abstracts. Additionally, publications may be sorted by the medical topics they cover. We used the CORD-19 database to compile lists of publications. The data table can sort binary variables, allowing the user to pick desired publication topics, such as papers that describe COVID-19 symptoms. The application is primarily designed for use by researchers but can be used by anybody who wants a faster and more efficient means of locating papers of interest.

***Keywords*—**COVID-19, literature summary, information retrieval, snorkel.

I. Introduction

THE COVID-19 pandemic caused by the novel 2019-nCoV virus has highlighted a problem that a lot of data scientists have been grappling with during the past decade. There is so much information and data out there, it is no longer efficient or reasonable to expect a single expert in a field to read all of the papers, or even all of the abstracts of the papers, of a specific problem in the field [1]. After the virus began to spread at a rate unprecedented in 2020, there was a figurative explosion of papers on the topic. A simple search of the word “coronavirus” on Google Scholar leads to several hundred thousand papers. Suddenly, research articles are flooding major academic publication portals at an unprecedented speed, and every university had multiple teams churning out whatever COVID-19 research they could. This is not necessarily a bad idea, as the desire to discover a vaccine, drug, or preventive measure was universally shared. However, the sheer amount of information that was and is being generated was impossible for any person to sort through in a reasonable amount of time. The announcement of a world pandemic by the World Health Organization (WHO) in early March [4] was followed by numerous publications, and by late March 2020, there were hundreds of scholarly papers on this topic on a single site for research papers [1]. Even the discovery of the vaccine has not slowed the deluge of papers being published, the papers that one can get with a simple search of the word “coronavirus” now number over a million in Google Scholar.

The first issue that arose when deciding a path forward for this project was whether to use supervised or unsupervised machine learning algorithms to sort the coronavirus publications. The advantages of supervised machine learning are the ability to define the classes of topics and the results. The disadvantage of supervised learning is that it can be time-consuming to develop. Project personnel has to label the data, as well as supervise the process [6]. The advantages of unsupervised learning are that the machine can discover its’ patterns since no labels are given. It is also less cumbersome in terms of the time required to look at the enormous datasets [6]. In the end, a weak supervision machine learning algorithm was chosen, since the whole point of the project was to save researchers’ and clinicians’ time. This was agreed upon for several reasons, but primarily because it was desirable to have specific keywords included in the search for the abstract text sorting.

II. Methods

For sorting the papers into categories, a machine learning system called “Snorkel” [11] was used to sort our data into various categories. In this case, the data were text data from the abstracts of thousands of papers. In Snorkel, some of the labeling functions were used to create quick text data sorting into groups based on keywords. A similar project is underway with medical records at the U.S. Department of Veterans Affairs [5]. Essentially, the abstract text was pulled from CORD-19, and then Snorkel was used to combine experts’ knowledge for semi-automatic label the whole dataset for further information pipeline to process. The result was the output of a table with binary variables for each category, which allows for easy sorting and of the papers, including links to the papers. This could be accessed in two ways from the application. The first is the pivot table, which allows us to sort papers by the columns in our pivot table. The second is the data table with binary variables (more on that later). It should be noted that while the user interface is interactive, the underlying data are pre-computed to save time for the users. The results the application users are shown are pre-computed with Snorkel. While fairly quick and efficient (relatively speaking), this processing would quickly consume computing space and cause considerable server slowdown if multiple calls were placed on the server at one time, thus the pre-computation of our results.

C. A. van Hal is with the Rensselaer Polytechnic Institute, Troy, NY 12180 USA (corresponding author, phone: 832-459-6026, e-mail: vanhac2@rpi.edu).

X. Jiang and L. Chen are with the University of Texas Health Science Center at Houston, Houston, TX 77030 USA (e-mail: Xiaoqian.Jiang@uth.tmc.edu, Luyao.Chen@uth.tmc.edu).

Open Science Index, Computer and Systems Engineering Vol:15, No:12, 2021 publications.waset.org/10012376.pdf

The map data are fetched by a command that runs every couple of hours, which limits the strain on the servers as well. Fig. 1 gives a general overview of the application "pipeline," showing how the computer and the researcher interact to create the data in the background before it is shown in the application. Snorkel is a method that is based on Python [7]. Python is a general computer language commonly used for mathematical processing [8].

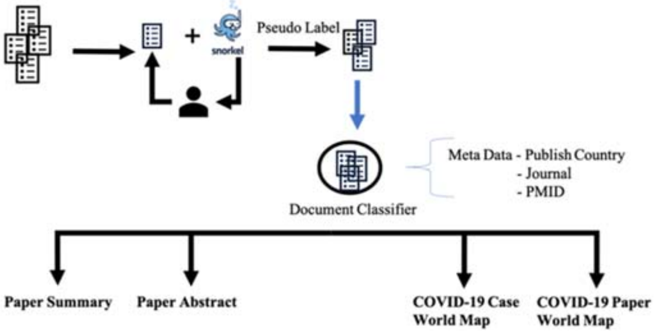


Fig. 1 The general overview of the COVID-19 literature mining application

The application itself (in this case, the "application" refers to the user interface) was written in the coding language "R," and the extended portion of that language meant for application building "R Shiny". The computer languages used here were designed for the easy computation and visualization of data [9]. "R" and "R Shiny" are languages that use a lot of packages with very specific functions to create certain kinds of visualizations [9], [10]. "R Shiny" was designed primarily to be able to create a user interface with "R" [10]. Most of the base code is "C," however, some of it is in other languages, such as JS [9], [10]. There was also some use of JS, HTML, and CSS, which allowed for better map formatting and interactivity. The backend of this is built on an R Shiny-server, which was used for ease of analysis. The data extraction for the interactive map with the number of cases by country was also built in R. All paper/text data came from CORD-19 [2]. In the backbone of the application, we integrated expert knowledge and several machine learning tools to build a semi-automatic updating pipeline. In this pipeline, experts and other application users can generate, examine, and modify annotation rules for generating a high-confidence dataset to train document classifiers. The document classifiers can generalize the pattern from the high-confidence dataset to another dataset and then retrieve other relevant information or articles. The application itself then calls upon the data, which are stored on the server after processing, and R and R Shiny make up the user interface. The data tables created in R and R Shiny, allow for user interactivity, the feature that was considered paramount for the researchers involved in this project. This approach makes it easy to update the application – you need only to update your data in the server and reload the application. The ease of update was also considered an extremely important feature due to the pace of the research taking place. The data frame will sort the data live for the user, which allows for more flexibility in the options the user can select.

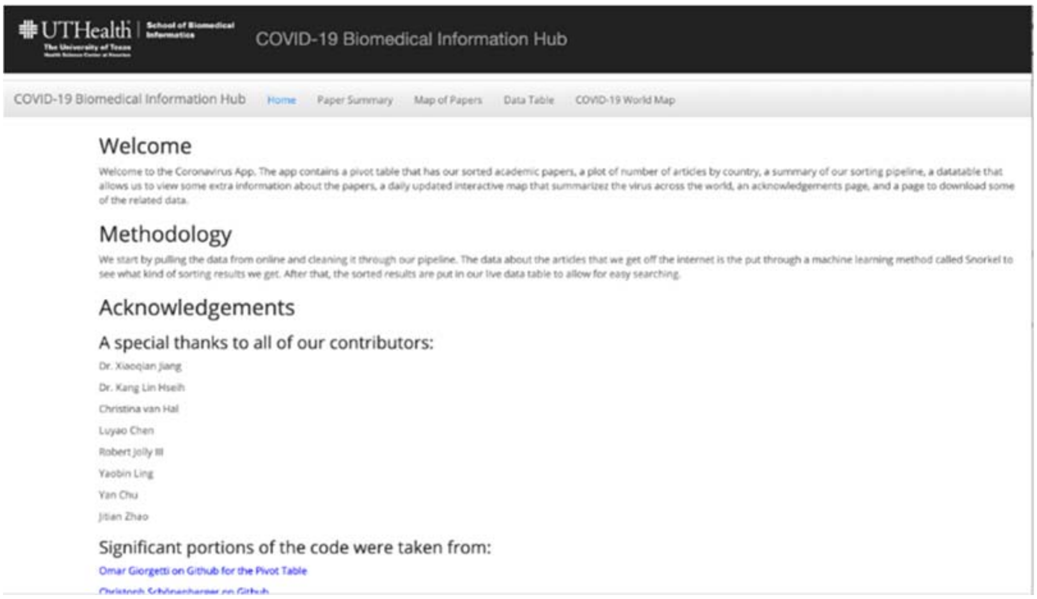
III. Results

The main focus of this project is to build a portable, transferable, and labor force efficient information retrieval pipeline for further usage. In current information retrieval works, generation, a gold standard, is a time-consuming and labor-intensive task. Using the gold standard dataset to optimize the document classifiers is also another goal. In the backbone of the app, experts can start from a small dataset to summarize the context patterns, word patterns within abstracts for building labeling rules. They can also utilize other relevant metadata information, including country, name of institution, and name of the journal, for building constraints for each labeling rule. Experts can apply other NLP tools such as topic modeling to examine the performance of rules. All labeling rules are applied with Snorkel. Snorkel can compromise the conflicts among rules and aggregate results into probability [7]. Experts can examine the output and set the threshold for further evaluation. After evaluation, the high confidence datasets will be put into Document Classifiers for optimization. The final model can be used to retrieve other relevant information from another independent dataset. The final information will be presented by the app.

From the researchers' perspective, they want to quickly identify the highest relevant information for them to read. This is why we believe that effective sorting is another key point in this project. Effective sorting of the data and the ability to call on that sorting quickly was the essential endeavor of this project. For example, users might want only to view papers that are about COVID-19 symptoms. A column was established that allows the user to filter papers by the presence of symptoms mentioned in the abstract or the lack thereof, reducing the need for researchers to read through abstracts manually. Another example is if the epidemiologist in question is interested in studying the spread of COVID-19, they can sort with a button on the app page. The data frame and pivot tables allow for easy sorting of the data. The app results are promising. Snorkel seems to have sorted the papers fairly well, as far as the researchers can tell and the graphics of the application are fairly quick in their generation. If researchers are not satisfied with the current results, Snorkel will allow them to generate their results and, so long as formatting was carefully observed, researchers could still use the code for the front end of the application with their preferred data in their server.

Some screenshots can demonstrate portions of the app. In Fig. 2, the front page of the app, we have a brief welcome message and the names of the creators of the app, along with a very brief summary of the methodology we used. This is also the page that allows for the researcher to download our data, which are all publicly available, though ours is cleaned. Fig. 3 is the abstract summary page and shows a sample pivot table. Other pivot tables can be formed by selecting more options; however, a simple example was desired due to coloration and size limitations. The pivot table can also be used for more visual-based results, such as a heatmap. Fig. 4 is the summary of current COVID-19 infection, death, and recovery statuses for each country. This gives a general overview of the spread of the virus and the current status of most countries with regard to infection and death rates. Fig. 4 should not be confused with Fig. 5, which is a map of the number of papers from each country. The creators of the app elected to keep the paper numbers separate to avoid confusion and several programming issues due to the structure of the data. Fig. 6 is our final figure. This shows the data table portion of the application. As you can see, there is a clear focus on the abstracts and text display in this part of the app. In full-screen viewing, this allows for much easier viewing than in the pivot table. It is a straightforward system to deliver a quick overview of thousands of papers to keep researchers up-to-date.

This app allows the user to search for multiple variables at once in both the data frame and the pivot table. The data frame was included for ease of viewing the abstracts. It allows for a more expanded view of the text. The pivot table allows for the addition and deletion of columns if you have a specific focus area. The pivot table also allows for some basic stats to be generated on-screen about the data, so it seemed prudent to include both. This includes basic statistics and a heatmap. The addition of the interactive world map and the map of papers serve to give the user an overview of what is going on in terms of the overall coronavirus and publications involving the virus. It also allows the viewer to see where the majority of the research is coming from. This may give them an idea of which country's research they wish to look at first. It might allow for a focus on the newly emerging regional variants of the virus that have been the recent focus of the news as the research emerges. In future modifications of the application, we hope to be able to link the map Fig. 5 to our data table, which would allow for ease of access in a very user-friendly way. The application is accessible to the public [3] and uses the CORD-19 database [2].


UTHealth | School of Biomedical Informatics
The University of Texas Health Science Center at Houston
COVID-19 Biomedical Information Hub
COVID-19 Biomedical Information Hub
Home
Paper Summary
Map of Papers
Data Table
COVID-19 World Map

Welcome

Welcome to the Coronavirus App. The app contains a pivot table that has our sorted academic papers, a plot of number of articles by country, a summary of our sorting pipeline, a datatable that allows us to view some extra information about the papers, a daily updated interactive map that summarizez the virus across the world, an acknowledgements page, and a page to download some of the related data.

Methodology

We start by pulling the data from online and cleaning it through our pipeline. The data about the articles that we get off the internet is the put through a machine learning method called Snorkel to see what kind of sorting results we get. After that, the sorted results are put in our live data table to allow for easy searching.

Acknowledgements

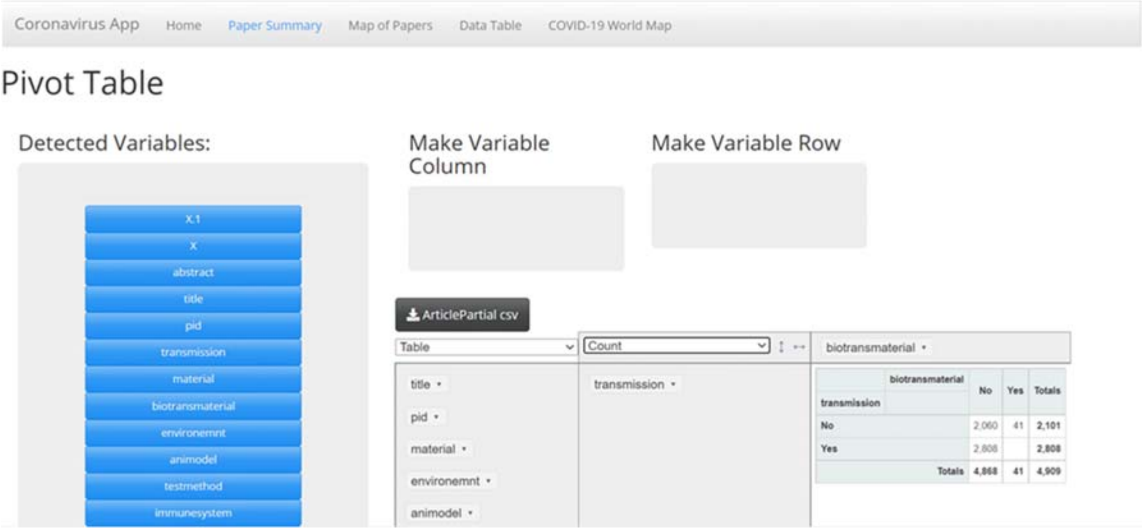
A special thanks to all of our contributors:

Dr. Xiaoqian Jiang
Dr. Kang Lin Hseih
Christina van Hal
Luyao Chen
Robert Jolly III
Yaobin Ling
Yan Chu
Jitian Zhao

Significant portions of the code were taken from:

Omar Giorgetti on Github for the Pivot Table
Christoph Schönenberger on Github

Fig. 2 The Front Page of the Application



Pivot Table

Detected Variables:

Make Variable Column

Make Variable Row

Table

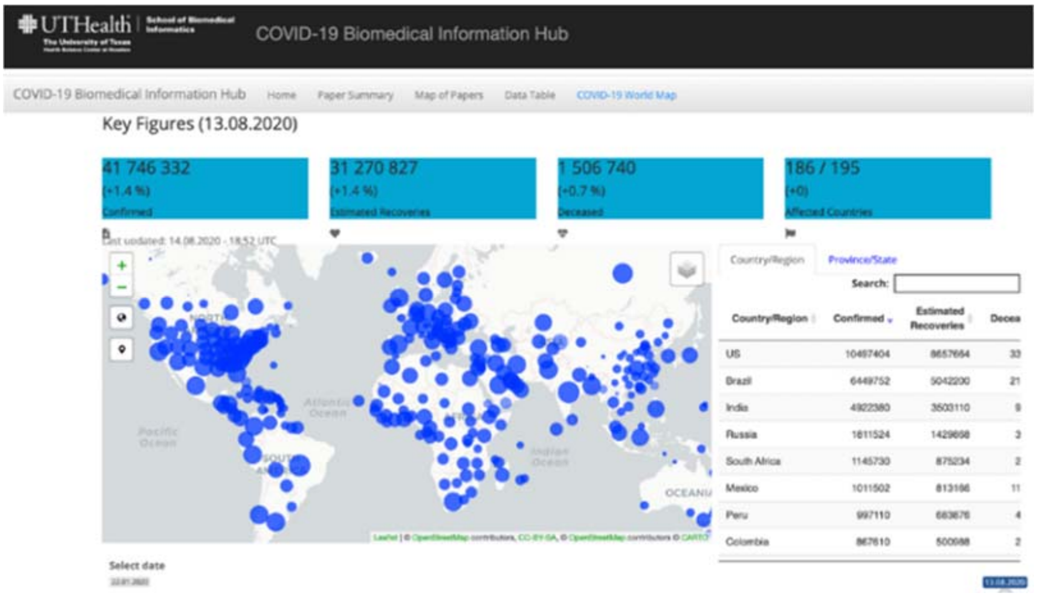
Count

transmission

biotransmaterial

transmission	biotransmaterial	No	Yes	Totals
No		2,060	41	2,101
Yes		2,808		2,808
	Totals	4,868	41	4,909

Fig. 3 Abstract Summary Page



COVID-19 Biomedical Information Hub

COVID-19 Biomedical Information Hub Home Paper Summary Map of Papers Data Table COVID-19 World Map

Key Figures (13.08.2020)

41 746 332 (+1.4 %) Confirmed

31 270 827 (+1.4 %) Estimated Recoveries

1 506 740 (+0.7 %) Deceased

186 / 195 (+0) Affected Countries

Last updated: 14.08.2020 - 18:52 UTC

Country/Region Province/State

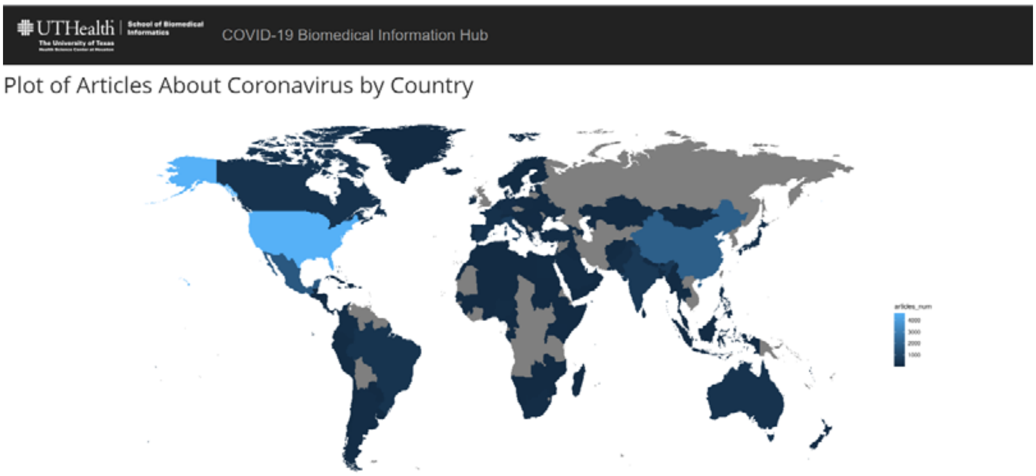
Search:

Country/Region	Confirmed	Estimated Recoveries	Decea
US	10497404	8657664	33
Brazil	6449752	5042200	21
India	4922380	3503110	9
Russia	1611524	1429698	3
South Africa	1145730	875234	2
Mexico	1011502	813166	11
Peru	997110	683676	4
Colombia	867610	500088	2

Select date

UTHealth The University of Texas
School of Biomedical Informatics
+
−
NORTH
Atlantic Ocean
Pacific Ocean
SOUTH AM
AFRICA
Indian Ocean
OCEANIA
Leaflet | © OpenStreetMap contributors, CC-BY-SA, © OpenStreetMap contributors © CARTO
22.01.2020
13.08.2020

Fig. 4 Summary of the Worldwide State of COVID-19

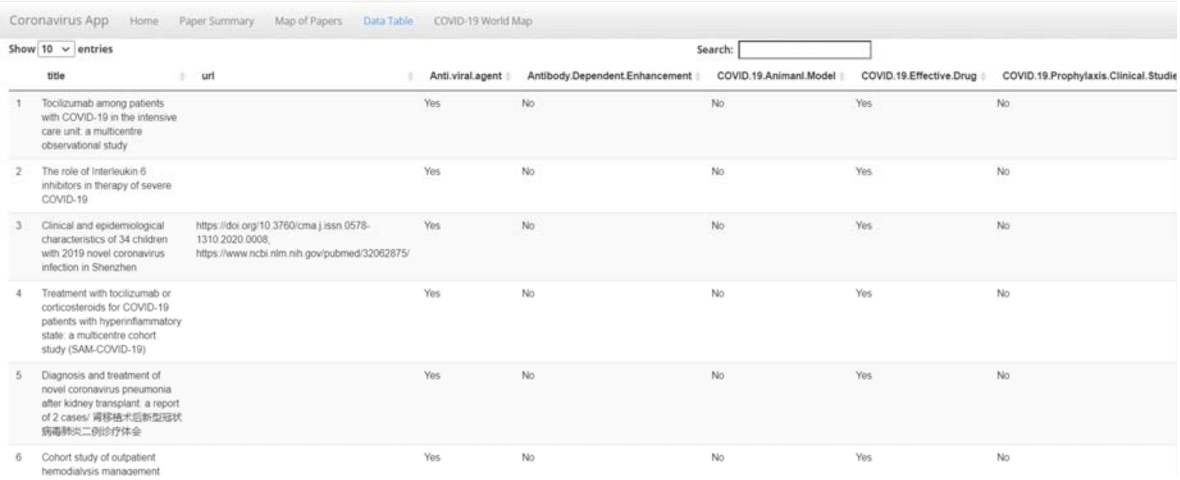


COVID-19 Biomedical Information Hub

Plot of Articles About Coronavirus by Country

UTHealth The University of Texas
School of Biomedical Informatics
articles_num
4000
3000
2000
1000

Fig. 5 Plot of Number of COVID-19 Articles by Country



Coronavirus App Home Paper Summary Map of Papers Data Table COVID-19 World Map

Show 10 entries Search:

	title	url	Anti.viral.agent	Antibody.Dependent.Enhancement	COVID.19.Animanl.Model	COVID.19.Effective.Drug	COVID.19.Prophylaxis.Clinical.Studie
1	Tocilizumab among patients with COVID-19 in the intensive care unit: a multicentre observational study		Yes	No	No	Yes	No
2	The role of Interleukin 6 inhibitors in therapy of severe COVID-19		Yes	No	No	Yes	No
3	Clinical and epidemiological characteristics of 34 children with 2019 novel coronavirus infection in Shenzhen	https://doi.org/10.3760/cma.j.issn.0578-1310.2020.0008; https://www.ncbi.nlm.nih.gov/pubmed/32062875/	Yes	No	No	Yes	No
4	Treatment with tocilizumab or corticosteroids for COVID-19 patients with hyperinflammatory state: a multicentre cohort study (SAM-COVID-19)		Yes	No	No	Yes	No
5	Diagnosis and treatment of novel coronavirus pneumonia after kidney transplant: a report of 2 cases/ 肾移植术后新型冠状病毒肺炎二例诊疗体会		Yes	No	No	Yes	No
6	Cohort study of outpatient hemodialysis management		Yes	No	No	Yes	No

Fig. 6 Data Table Summary of COVID-19 Papers

IV. Discussion and Limitations

We developed an application to enable the sorting of thousands of coronavirus papers. In the *Coronavirus App,* users can access our user interface to sort through the papers until they get sufficient papers on the topic of their interest. We expect this effort will assist researchers and clinicians in staying abreast of coronavirus research by reducing the papers to a more manageable number for reading purposes.

The design of the application also plays an important role in aiding the users in their search for topic-specific papers. The landing page gives an overview of the application, data, and the creators of the application. It was created in response to the relatively few options for searching papers in the databases themselves. Most databases that are encountered have very few filtration options that are based on the research topics of the papers themselves. Google Scholar, for instance, allows only sorting by date, relevance to the search text, patents, and the number of citations in other publications. Our application's second page has a pivot table that allows the sorting of the paper by multiple characteristics, and the user can customize the filter by themselves (e.g., users can understand how many papers are discussed with vaccine and antibody-dependent enhancement). This is useful for specific inter-correlated topics. Additionally, we have a page that gives a geographical overview of where the papers were written, which can help contextualize the research. For example, researchers can use this map to infer which countries across the world are the highest in infections for COVID-19. The page with the data table allows the user to sort the papers by their binary categories. There is also a page that gives the user a general overview of death, recoveries, and infections in each country, which can be useful in comparing countries' relative infection rates.

The application has several limitations, however. First, the loading speed is a little slow in real-time. The data in the server are not all automatically updated and may not be current up to the minute. Some of the variables may overlap, and, of course, our sorting may not be completely accurate. The application also does not rate the "value" or "importance" of the articles. The articles are all equally likely to show up, despite differences in sample size or limitations of that particular set of experiments. Another challenge might be integrating more article sources. The formatting for that might be more involved in terms of combining the text data. Overall, the general pipeline idea could be expanded to other topics with far too many research papers to ever feasibly read, however, this has certain caveats. The Snorkel sorting still requires some supervision and not an insignificant amount of time to use [7]. Certainly, it cuts down on time spent making the example dataset, but this still needs human involvement for it to work. The application user interface can be easily modified for other projects, provided the researchers are familiar with "R". The data itself is freely available, even if the papers are not. Both the back and front-end of the application could be improved, as is discussed below. However, the tool is still an excellent one if sorting COVID-19 papers is the desired outcome.

For continuing to improve this application, we consider applying some NLP technologies to mitigate the limitations. The first direction is combining ontology and weak supervision machine learning methods to rank the relevance of abstracts. Ontologies are a group of pre-defined control vocabularies that can use to retrieve a specific topic from documents. Based on this idea, we can consider each ontology as a classifier for determining a specific topic. Also, ontology is very specific to a given topic and should not affect other ontologies. We can consider each ontology-based classifier is independent of the other. If a signal of a given ontology is more intense than others, we can determine that this article has greater relevance to a specific topic than others. The potential limitation of the previous idea is each document can have multiple topics, which lead to conflicts of ontology. The weak supervision framework, Snorkel, is a skill in solving this problem. Snorkel will give a high-rank for a document because of low proportion conflicts between ontologies and vice versa. Thus, we can get a relative rank for a given of documents. Another potential benefit of using this framework is humans can update ontology and realize how the update affects the relevance. This can avoid black-box machine learning and reduce the re-training time for current NLP technologies. Additionally, there could be some improvements to our visualizations. A button that allowed researchers to click on a country in one of our maps would link up to a collection of the articles from that country on either the data table, the pivot table, or both. This would be difficult to implement in terms of making it real-time feasible, but doable. The paper summary map and the COVID-19 World Map could have also been combined. However, the researcher team felt it prudent to keep the maps separate, as combining the deaths and papers numbers may have been confusing or misleading. There might be ways to improve the speed of the data tables or the pivot tables, as well. Though pre-generating every option for each table seemed to be considered, it was rejected in favor of real-time generation, both due to time constraints and the fact that it may not have improved overall efficiency. The map for the number of papers and the paper data should have likely came from the same data pull from the server. The pivot table visualization had some other minor issues as well; namely, the code in the package had to be hand-edited to fit ADA requirements, given that the package source code was completed some time ago and updates to it have ceased. The pivot table package also has base code in an extinct computer language, making it incredibly difficult to modify in the modern-day.

Overall, the application worked as designed and sorted coronavirus articles by keywords in their abstracts. There are a few places for improvement, however, the application is more than just a nifty research tool for use in the ongoing research on COVID-19. It could also be considered a proof of concept of sorts. The pipeline could easily be used for other kinds of research and the researchers hope that it be considered. The pipeline shows a relatively easy, relatively researcher-friendly way to sift through the thousands and thousands of articles on any number of topics. With a few clicks in the source code of the user interface, it could also be edited to whatever the researcher desires to show a visual of. In essence, the creators are not just proposing a way to sift through coronavirus papers but a way to help sift through academic papers on nearly every subject.

V. Conclusion

In conclusion, an application was created that displays Coronavirus articles sorted by abstract keywords. The application is meant to assist researchers by providing them with articles relevant to their topic area of interest and removing the papers that are not related. The application allows the user to interact with their display, allowing them more flexibility in their search options than most databases would traditionally provide directly. While the application is primarily designed for researchers and clinicians, it is available for use by anyone interested in using it to get a quick grasp of the state-of-the-art research about the pandemic.

Acknowledgment

XJ is CPRIT Scholar in Cancer Research (RR180012), and he was supported in part by Christopher Sarofim Family Professorship, UT Stars award, UTHealth startup, the National Institute of Health (NIH) under award number R01AG066749.

K.H. is supported by Computational Cancer Biology Training Program Fellowship CPRIT Grant No. RP170593.

References

[1] Brainard, J. (2020). Scientists are drowning in COVID-19 papers. Can new tools keep them afloat? American Association for the Advancement of Science. Retrieved from https://www.sciencemag.org/news/2020/05/scientists-are-drowning-covid-19-papers-can-new-tools-keep-them-afloat

[2] CORD-19 Database (Computer Software). Retrieved from https://www.semanticscholar.org/cord19.

[3] Coronavirus Paper Sorting Application (Computer Software). Retrieved from https://safeapp1.uth.edu/app/01_covidapp.

[4] Cucinotta, D., & Vanelli, M. (2020). WHO Declares COVID-19 a Pandemic. Acta bio-medica: Atenei Parmensis, 91(1), 157–160. https://doi.org/10.23750/abm.v91i1.9397

[5] Ratner, A., Back, S., Ehrenberg, H., & Ré, C. (2017). Snorkel and The Dawn of Weakly Supervised Machine Learning. Stanford Dawn. Retrieved from https://dawn.cs.stanford.edu/2017/05/08/snorkel/

[6] Russell, S.J., Norvig, P. (2010) Artificial Intelligence: A Modern Approach, Third Edition, Prentice Hall ISBN 9780136042594.

[7] Snorkel Team. (2020, June 27). Snorkel: The System for Programmatically Building and Managing Training Data. Snorkel. https://www.snorkel.org/

[8] Python Source Releases (Version 3.8.5) (Source Code). Retrieved from https://www.python.org/downloads/source/.

[9] R-Source (Version 4.0.1) (Source Code). Retrieved from https://github.com/wch/r-source.3

[10] Shiny (Version 1.5.0) (Source Code). Retrieved from https://github.com/rstudio/shiny.

[11] Snorkel (Version 0.9.6) (Source Code). Retrieved from https://github.com/snorkel-team/snorkel.