TopEx: Topic Exploration of COVID-19 Corpora

Results from the Biocreative VII Challenge Track 4

Amy L. Olex[1,2], Evan French[1,2], Peter Burdette[1], Srilakshmi Sagiraju[1], Thomas Neumann[3], Tamas S. Gal[1,3], and Bridget T. McInnes[2]

[1] C. Kenneth and Diane Wright Center for Clinical and Translational Research, Virginia Commonwealth University, Richmond, VA, USA; [2] Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA; [3] Massey Cancer Center, Virginia Commonwealth University, Richmond, VA, USA

***Abstract*—TopEx is a Natural Language Processing (NLP) application developed to facilitate the exploration of topics and keywords in a set of texts through a user interface that requires no programming or NLP knowledge, thus enhancing the ability of non-technical researchers to explore and analyze textual data. The underlying algorithm groups semantically similar sentences together followed by a topic analysis on each group to identify the key topics discussed in a collection of texts. Implementation is achieved via a Python library back-end and a web application front-end built with React and D3.js for visualizations. TopEx has been successfully used to identify themes, topics, and keywords in a variety of corpora, including COVID-19 discharge summaries and tweets. Feedback from the BioCreative VII Challenge Track 4 concludes that TopEx is a useful tool for text exploration for a variety of users and tasks.**

Keywords—natural language processing; topic analysis; user interface; covid-19

I. Introduction

In this digital age, research in any field inevitably requires the analysis of large data sets, including textual data. It is important to become familiar with any new set of data prior to running an analysis (1), and this is especially true with textual data. However, there are few applications that allow one to process textual data without requiring programming skills or knowledge of Natural Language Processing (NLP) techniques. For example, Python packages, such as PyLDAvis (2) and LDAExplore (3), provide functions to perform a topic analysis on a set of texts as well as visualize the results; however, these tools require a knowledge of Python programming (1). There are applications that utilize graphical user interfaces, but many of these require a subscription (ATLAS.ti), are difficult to install/customize (4,5), or are no longer available (6–8). Thus, the analysis of unstructured text documents still pose challenges to many, hence there is a need for an easy-to-use, programming-free topic exploration tool that can be utilized by researchers from any domain.

In this work we present TopEx, an NLP tool that allows for the automated and easy exploration of topics in a set of text documents. TopEx is domain agnostic and is designed to allow processing and exploration of niche corpora, such as those associated with COVID-19. With a user-friendly web interface and interactive graphical display of results, TopEx removes the barrier of having to learn a programming language or NLP techniques in order to explore topics present in a set of text documents, augmenting the type of research non-technical researchers can perform.

II. Algorithm

The NLP pipeline implemented in TopEx is composed of three primary steps (Fig. 1), and the algorithm is described in detail in Olex et. al. (9). Briefly, TopEx assumes each sentence expresses one topic, and aims to group sentences based on their similarity. To do this, TopEx first normalizes sentences (Sentence Normalization), which includes removing contractions and uninformative words, identifying parts of speech, reducing words to their base form, and converting all characters to lowercase. Next, sentences are converted into a numerical representation (Sentence Representation) by generating a Term Frequency-Inverse Document Frequency (TF-IDF) matrix (10), which is created using the input set of texts (aka corpus). The TF-IDF is first used to identify each

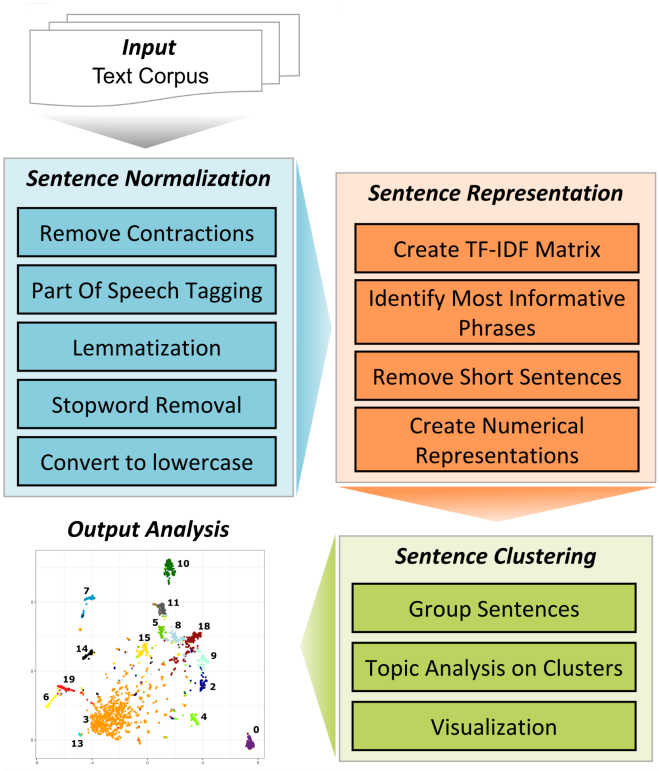

Fig. 1. TopEx NLP Pipeline.

sentence's most informative phrase, then a reduced form of the TF-IDF matrix is used to convert the phrase into a numerical vector that represents the whole sentence. Finally, clustering is performed on the numerical sentence vectors (Sentence Clustering) to group similar sentences together, and a topic analysis using Latent Dirichlet Allocation (LDA) (11) is run on each cluster to identify main topics. By default, the TF-IDF matrix is created using the input corpus, however, users can optionally import custom background corpora to add additional domain-specific context to the analysis. Results are visualized using word clouds and a UMAP (Uniform Manifold Approximation and Projection) (12) plot, and can be saved in a delimited text file for easy browsing of sentence clusters.

III. Implementation and Availability

TopEx is implemented in two parts: 1) as a Python 3.x library that can be downloaded from PyPi (https://pypi.org/project/topex) using the "pip install topex" command, and 2) as a web application with a graphical user interface (TopExApp). The TopEx Python library allows technical users the ability to integrate TopEx analyses into their NLP pipelines, as well as access the latest features and functionality. TopExApp is a web application with a graphical user interface that removes the high technical barrier traditionally associated with NLP and empowers researchers of any domain to directly utilize NLP tools without technical assistance. The TopExApp front end is built with React and D3.js is used for data visualizations The back-end consists of a Python Flask API, which wraps the TopEx Python library. TopExApp is hosted on a web server at topex.cctr.vcu.edu for general public use. For users exploring documents containing protected health information (PHI), or otherwise sensitive data, a local Docker image can be built from the code on GitHub (https://github.com/VCUWrightCenter/TopExApp). Additionally, source code for the TopEx Python library is also on GitHub (https://github.com/VCUWrightCenter/TopEx).

IV. User Interface and Results Exploration

The TopExApp web interface (Fig. 2) allows users to easily define their input corpus in multiple formats from the "Load Data" tab: a set of text files, a pipe-delimited file containing the text of one document per line, a MEDLINE formatted file, or by entering keywords to search PubMed for relevant abstracts. Users can customize the clustering pipeline by changing default settings in the "Parameters" tab, however, the default settings should produce decent results for most analyses. Results are visualized as a UMAP reduction and a set of word clouds. UMAP is a dimensionality reduction algorithm that takes a large numerical matrix and reduces it to two dimensions so that it can be easily visualized (12). UMAP reductions are plotted as a scatter plot (Fig. 3) where each dot represents a sentence that was clustered, and dots that are closer together indicate sentences that discuss similar topics. The color of the dot refers to the cluster that each sentence was assigned to. Users can explore the data by hovering over points to see 1) the full text of the sentence, 2) the most informative phrase in that sentence, and 3) a list of terms characterizing themes present in that sentence's cluster. In addition, word clouds, one for each cluster, correlate the frequency of each word in a cluster to the size of the bubble (Fig. 3, insets). Word clouds give the user insight into which words or topics may be dominating a particular cluster. The raw data behind the visualizations can be downloaded using the Import/Export tab and can be used in external programs such as R to generate high quality, customizable figures.

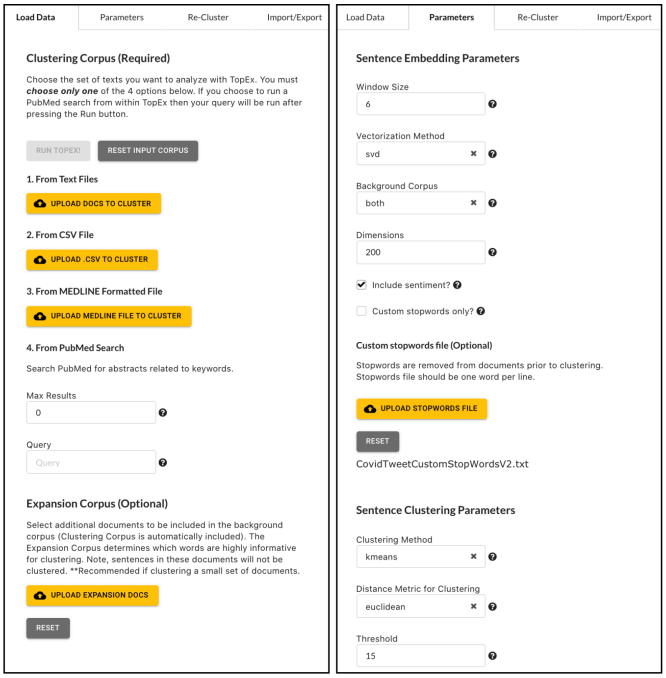

Fig. 2. Screenshots of TopEx tabs and menu items.

V. System Performance

As TopEx was built to be an exploratory tool, there are no specific benchmarks that can give an estimate of its performance other than the interpretability and relevance of the results. However, for the initial use case we did create a manually classified set of responses to assess whether or not TopEx could recreate those clusters. This revealed that TopEx performed better on topics that had a distinct vocabulary (9).

VI. Example Use Case: Evolution of COVID-19 Pandemic Tweets

TopEx is a versatile tool and can be used with any type of textual data. Table I provides a list of current use cases, both for internal operations (processing discharge summaries) and research. For this work we provide another example use case that utilizes TopEx to explore the evolution of topics in a subset of COVID-related tweets during the year 2020. Briefly, we randomly sampled 2000 English language tweets from the 22nd day of each month starting on March 22, 2020 from the COVID-19 Twitter Chatter dataset collected by Banda et al. (13). Fig. 3 contains the results generated from clustering this subset of COVID tweets for March 2020 (Fig. 3A) compared to December 2020 (Fig. 3B) with topics of various clusters highlighted. Topics in March revolved around stopping the spread of COVID, staying home, COVID testing, and the

TABLE I. CURRENT TOPEX USE CASES

Text Type	Use Case
Reflective Medical Writings	Identify common challenges experienced by medical students. Manuscript is under review.
COVID-19 Discharge Summaries	Identify key phrases and terms associated with COVID-19 patients to develop better rule-based queries using an in-house NLP system at VCU Massey Cancer Center.
Government COVID-19 Communications	Identify how mitigation strategies implemented in South Korea changed over time during the first wave of the COVID-19 pandemic. Manuscript is in preparation.
COVID-19 Tweets	Assess changing topics of community interest during the pandemic.

media's reaction to then President Donald Trump. In December the topics shifted to the new UK variant, vaccinations, and the COVID relief bill. In both months we see a cluster reporting new cases and the death toll. This cluster appeared in every month analyzed from March to December, which demonstrates that TopEx is able to identify consistent themes occurring throughout the year, as well as transient issues of interest to a community.

VII. BIOCREATIVE USER FEEDBACK

TopEx was submitted to the BioCreative VII Challenge Track 4: COVID-19 Text Mining Tool Interactive Demo and was tested by a variety of users. As the system developers, we provided four target-audience users (three non-technical clinical researchers and one grant funding organization representative), and the Track 4 organizers solicited the research community for additional participants. Users were sent a tutorial to complete and then asked to test out the system with some of their own data. Feedback was obtained from a survey of 30 questions with 14 requiring a numerical rating and 16 asking for unstructured text feedback. Table II lists the questions that required a numerical rating and the average score TopEx received from all users except one who was looking for a gene disease association tool rather than topic trends. TopEx received a total of seven responses with two who indicated they are directly involved in COVID-19 research. Two users were from government organizations, one from a patient organization that funds research grants, and the remaining four from an academic setting. Users indicated their field of work includes biocuration, informatics, and grant funding. Prior to this challenge, only two users had used topic analysis systems in the past, which included topic modeling using LDA and Chalklabs services. The following sections include a summary of the remaining results from tested functionalities, bottlenecks, usefulness of results, and whether they would consider using the system in the future.

A. *Tested Functionalities*

Functionalities tested by the users included uploading data in the different supported formats, modifying the parameters, clustering, visualizing, and exporting results. Specifically, two users tested out the CSV upload feature. Four users tested the PubMed search feature with queries like "Craniosynostosis AND gene AND variant", "Short QT syndrome", "COVID-19 AND gene AND variant", "(ncRNA or miRNA) AND Alzheimer", "long covid", and "Glioblastoma". In addition, users explored the visualizations by zooming in and hovering over the plots to see individual sentence information, and one user tested the re-clustering functionality.

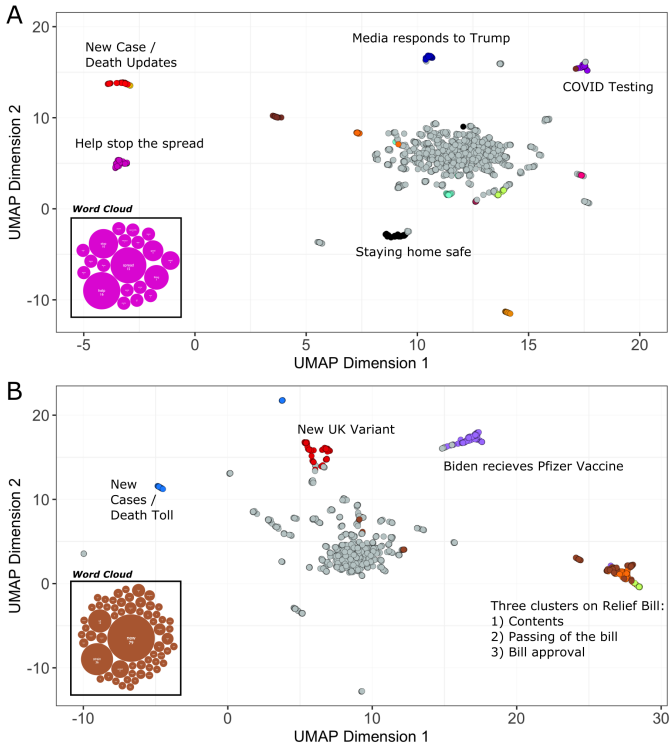

Fig. 3. UMAP scatter plots and example word clouds from TopEx results for tweets from A) March 2020 and B) December 2020. Scatter plots were generated in R from the coordinate text file output by TopEx.

B. *Positive Feedback and Usefulness of Results*

The majority of users (five out of seven) specifically indicated that they liked the interface because it was "*very clean*", "*easy to use and navigate*", "*intuitive and relatively easy to follow*", and "*the look and feel are very appealing*". Additionally, users indicated "*it is very quick in returning clustering*", useful for "*identifying trends in the data*" and for obtaining a "*quick grasp to understand topics*". One user noted that the hovering feature for viewing the details of a specific data point in the visualization was helpful, and another appreciated the multiple options for loading input data, particularly the PubMed search option. Overall, five out of seven users indicated that the returned results and program outputs were in a useful format.

C. *Bottlenecks and Suggested Improvements*

Four out of seven users identified bottlenecks when using the system. The most frequent was due to poor error handling when loading documents that were formatted improperly, followed by a lack of clarity on the meaning and purpose of

the many customizable parameters provided by TopEx. Other users also reported long runtimes for large queries, specifically large numbers of manuscript or grant abstracts, which limited the utility of TopEx for use in their work.

The most frequently suggested improvement was to provide lay descriptions of the parameters for non-specialist users. While it was noted by one user that the parameter explanation is extensive in the user documentation, it was suggested some of this material be posted directly on the site to aid users in navigating the parameter refinement process. Additionally, clarifying the formatting requirements for CSV upload was suggested as several users were unable to get this functionality to work. Other improvements include allowing users to enter stopwords into a text box without having to upload a separate file, including a pagination feature for the uploaded documents so the scroll bar is not so long for large inputs, referencing the source document name (i.e. PMID or file name) on the details drill down and in downloaded results, higher quality export of images, the ability to submit jobs and have the results emailed to you, incorporating named entity recognition, and other user interface improvements.

D. Overall Impression and Future Use

While a few users had difficulties, those that were able to use the system found it easy to navigate and intuitive with results being output in a helpful format for exploration. The majority of users indicated they would either definitely or possibly consider using TopEx for their own research in the future. Those that would not use TopEx included one person who was looking for a different type of tool to do gene disease associations, and a second who experienced very long run times. TopEx scored an average of 3.3 for overall impression (Table II). When asked if they would recommend TopEx to a colleague involved in COVID-19 research, all but two users rated TopEx with a score of 5 or greater (highest score of 8, lowest score of 1). Thus, the overall impression was that they liked the clean interface, found the system simple to navigate, and that the tool would be useful in future research, including COVID-19.

Future uses of the system included analysing different types of textual data. The users described various domains in which the system could be useful at analysing different types of data including thematic gene lists for curation, topic analysis of grant funders and interview transcripts. A summary of suggested use cases for TopEx is provided in Table III.

TABLE II. USER SURVEY QUESTIONS AND TOPEX AVERAGE

Question	Rating Rubric	TopEx
I think that I would like to use this system frequently.	1=Strongly Disagree 5=Strongly Agree	3.2
I found the system unnecessarily complex.		2.3
I thought the system was easy to use.		3
I think I would need support from the developer to be able to use this system.		3.3
I found the various functions of the system well integrated.		3.8
I thought there was too much inconsistency in this system.		2.3
I would imagine that most people would learn to use this system very quickly.		3.2
I found the system very cumbersome to use.		2.8
The system has met my expectations.		3.2
I felt very confident using the system.		3.5
I needed to learn a lot of things before I could get going with the system.		3.3
How easy was it to format and input data into this tool?	1=Not at all easy 5=Extremely easy	3
Please rate your overall impression with the system.	1=Very Negative 5=Very Positive	3.3
How likely is it that you would recommend this system to a colleague performing COVID-19 related research?	1=Not at all likely 10=Extremely likely	6

VIII. IMPROVEMENTS AND FUTURE WORK

Our goal for TopEx is to enable non-technical researchers easy access to NLP analyses, thus, the user feedback from the BioCreative Challenge has been very informative and helpful in identifying and prioritizing future improvements. Based on user feedback, immediate future work will include adding on-site documentation to aid users in navigating the various analysis parameters at their disposal. Additionally, improving the functionality and documentation surrounding the CSV input format is vital to alleviate immediate usability concerns.

Other future work will include parallelization to improve run times on larger documents, investigating a submission base framework for very large corpora, providing a text box so users may input stopwords within the application instead of uploading a text file, and including the document name or PubMed PMID in the on-screen and downloaded output so users can re-identify documents or articles for additional research. Further work will also include upgrades to the user interface, implementation of different levels of clustering to include paragraph or document summarization along with sentence level clustering, as well as improved visualizations, and temporal analyses.

TABLE III. SUGGESTED USE CASES FOR TOPEX

Text Type	Use Case
PubMed Abstracts	Identify main themes in a set of queried abstracts from PubMed.
Grant Summaries	Identify topics addressed in a set of grants that need to be assigned to reviewers.
Publications	Identify thematic gene lists for manual curation from a collection of publications.
Interview Transcripts	Analysis of transcripts of interviews in social behavioral work for common themes.
Open-ended Survey/Blog Responses	Assessing themes or topics addressed in open-ended survey responses or topic-focused blog posts.

IX. Conclusions

TopEx is a novel, domain agnostic, NLP tool that provides a user-friendly interface for non-technical users to explore topics present in a set of texts. It has already been shown to be useful in navigating reflective writings from medical students (9, second manuscript under review), COVID-19 discharge summaries, and tweets (Section VI). TopEx was submitted to the BioCreative VII Challenge Track 4 and has been evaluated by a diverse group of users. The overall impression from users is that TopEx is easy and intuitive to navigate, provides useful output, and they would consider using TopEx in their future research. In conclusion, end-users have indicated that TopEx is a user-friendly NLP tool that facilitates the exploration of topics in a set of texts, and enhances the ability of non-technical researchers to explore and analyze text data.

Acknowledgment

The authors would like to thank Sean Kortola, Aiden Meyers, and Suzanne Prince for work in implementing the first prototype of TopExApp for their senior year capstone project as students of VCU's Computer Science Department. Additionally, the development and maintenance of TopEx is supported by CTSA award No. UL1TR002649 from the National Center for Advancing Translational Sciences. Its contents are solely the responsibility of the authors and do not necessarily represent official views of the National Center for Advancing Translational Sciences or the National Institutes of Health.

References

1. Horn F, Arras L, Montavon G, Müller K-R, Samek W. (2017) Exploring text datasets by visualizing relevant words. arXiv:170705261 [cs]. [cited 15 Feb 2021]. Available: http://arxiv.org/abs/1707.05261

2. Sievert C, Shirley K. (2014) LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Baltimore, MD. Association for Computational Linguistics; pp. 63–70. doi:10.3115/v1/W14-3110

3. Ganesan A, Brantley K, Pan S, Chen J. (2015) LDAExplore: Visualizing Topic Models Generated Using Latent Dirichlet Allocation. arXiv:150706593 [cs]. [cited 17 Feb 2021]. Available: http://arxiv.org/abs/1507.06593

4. Alexander E, Kohlmann J, Valenza R, Witmore M, Gleicher M. (2014) Serendip: Topic model-driven visual exploration of text corpora. *IEEE Conference on Visual Analytics Science and Technology (VAST)*. pp. 173–182. doi:10.1109/VAST.2014.7042493

5. uwgraphics/SerendipSlim. UW Graphics Group; 2021. Available: https://github.com/uwgraphics/SerendipSlim

6. Yang Y, Yao Q, Qu H. (2017) VISTopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling. *Visual Informatics*.1: 40–47.

7. Luo D, Yang J, Krstajić M, Ribarsky W, Keim D. (2012) EventRiver: Visually Exploring Text Collections with Temporal References. *IEEE Trans Vis Comput Graph*. 18: 93–105.

8. Havre S, Hetzler B, Nowell L. (2000) ThemeRiver: visualizing theme changes over time. *IEEE Symposium on Information Visualization 2000 INFOVIS 2000 Proceedings*. pp. 115–123.

9. Olex AL, DiazGranados D, McInnes BT, Goldberg S. Local Topic Mining for Reflective Medical Writing. (2020) *AMIA Joint Summits Transl. Sci. Proc.*. 459–468. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7233034/

10. Roelleke T, Wang J. (2008) TF-IDF uncovered: a study of theories and probabilities. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM; pp. 435–442.

11. Blei DM, Ng AY, Jordan MI, (2003) Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3, 993–1022.

12. McInnes L, Healy J, Melville J. (2018) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:180203426 [cs, stat]. [cited 16 Aug 2019]. Available: http://arxiv.org/abs/1802.03

13. Banda JM, Tekumalla R, Wang G, Yu J, Liu T, Ding Y, Artemova E, Tutubalina E, Chowell G. (2021) A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—An International Collaboration. *Epidemiologia* 2, 315–324. https://doi.org/10.3390/epidemiologia2030024