Using NLP to analyze biomedical literature related to the COVID19 pandemic

Liang Chen and Murat Melek

Abstract

COVID-19 has had a once in a generation impact on the World both in human and economic costs. Consequently, the number of research publications on the virus has increased exponentially. This paper outlines a new approach to research paper clustering by utilizing doc2vec algorithm and BM25 similarity measure.

1 Introduction

Since the first reported death due to Coronavirus (COVID-19) on January 11, the pandemic has claimed the lives of 644,832 worldwide and 146,460 in the United States. The heavy human and economic toll has led to an implosion of research across the world. Figure 1 shows the number of publications accumulated on COVID-19 Open Research Dataset.

Total number of publications shown in Figure 1 visualizes the difficulty of the task of staying on top of the vast range of biomedical articles, reviews and reports. This paper outlines a suggested procedure to visualize clustering of the publications on COVID-19 by taking advantage of recent advances in natural language processing.

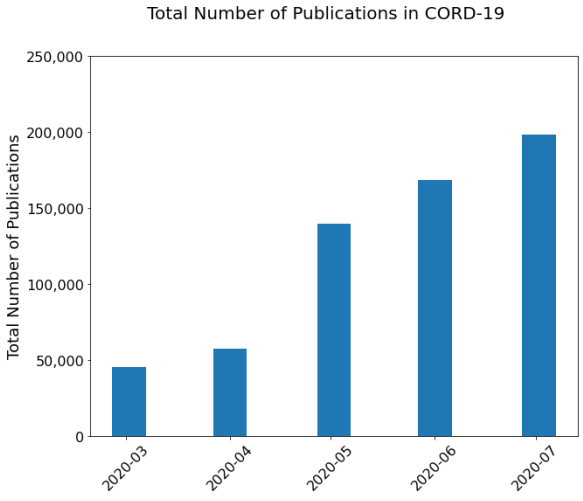

Figure 1: Number of Total Publications

The motivation of this paper is to examine whether clustering algorithms could be used to simplify the search for related publications in this ever expanding database. How can the content of the clusters be qualified? We will use clustering for labelling in combination with dimensionality reduction for visualization, and represent the collection of literature using a scatter plot. This plot will cluster publications of highly similar topics that share keywords.

2 Dataset

This paper utilizes COVID-19 Open Research Dataset (CORD-19). On March 16, 2020, CORD-19 was released by a collaboration between several institutions including the Allen

Institute for AI of MIT, Microsoft and White House. In the past four months the corpus which is curated and maintained by the Semantic Scholar team at Allen Institute for AI has grown to over 200,000 scholarly articles.

The dataset includes SPECTER document embeddings for each CORD-19 paper. The embeddings have a dimension of 768.

The site also provides json files that contain full text parses of a subset of CORD-19 papers. Semantic Scholar team was able to gather full text parses of papers that are available under an open access license. As of the date of the submission of this paper, full text of 65,992 publications obtained through PubMedCentral (PMC) are in the dataset. Full text of another 90,367 papers have been collated by parsing pdfs from several sources.

PMC and PDF parses result in two sets of full text JSON documents with a total size of 12 GBs.

3 Methodology

We used Approximate Nearest Neighbors Oh Yeah (Annoy) library of Python to map papers based on their Specter Embedding Vectors. By doing so, Annoy allows us to represent each publication as a vector in 768-dimensional space. This operation is undertaken once and the mapped structure is saved as an annoy file (.ann) to be later used by our algorithm to calculate Angular distances between vectors for similarity calculations.

Le and Mikolov (2014) proposed the use of *paragraph vectors* to construct fixed length document vectors from paragraphs with variable lengths. We used this approach to create fixed length document dense vectors based on the CORD-19 metadata. The document vectors were saved as a parque (.pq) file.

Document vectors dataframe contains CORD-19 publication id, cluster number, and 1d and 2d vector representations of the publication.

cord_uid	cluster	x	y	1d	2d
le0ogx1s	5	30.580741	8.813468	30.580741	[30.580741180972026, 8.813467785816274]

Figure 2: A row of Document Vector

We used the Okapi BM25 ranking algorithm to find similarity between the articles. Given a query consisting of a set of keywords (q) and document (d), similarity score is calculated as follows:

$$score(q,d) = \sum_{i=1}^{|q|} idf(q_i) \cdot \frac{f(q_i,d)\cdot(k_1+1)}{f(q_i,d)+k_1\cdot(1-b+b\cdot\frac{|d|}{avgdl})}$$

where

$idf(q_i)$ is the inverse document frequency weight of the query term q_i;

$$idf(q_i) = log\frac{N-df(q_i)+0.5}{df(q_i)+0.5}$$

k_1 and b are free parameters;

avgdl is the average document length;

|d| is the length of the document in words;

and $f(q_i)$ is the number of times q_i appears in d_i

Metadata has been cleaned by removing punctuation, dropping missing values and lowering all letters. Cleaned metadata was tokenized using NLTK package of Python

Python's rank-bm25 library was utilized for calculating similarity scores. Using the library, we have read and indexed the cleaned and tokenized paper abstract and text metadata.

Clusters were generated by kmean algorithm and dimensional reduced to the 2-D space by t-Distributed Stochastic Neighbor Embedding (t-SNE).

4 Results and discussion

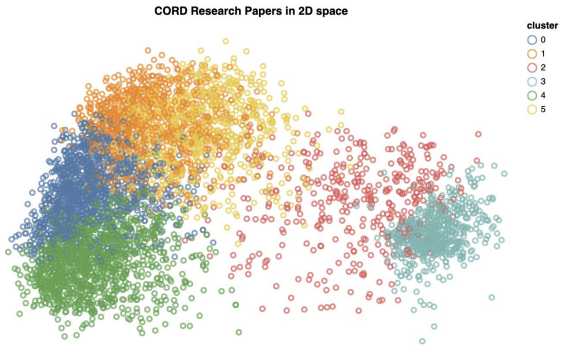

Figure 3: CORD-19 Dataset in 2D Space

We visualized tagged literature clusters using the Python's Altair package. The clusters provide insights into general topic areas of research papers (Fig 3).

We further visualized the keywords of abstracts of these research papers using the python wordcloud package. For instance, Cluster 0 includes those papers focusing on development of vaccines against the coronavirus; Cluster 1 includes papers that report clinical studies; Cluster 2 and cluster 3 includes papers studying severe acute respiratory syndrome; cluster 4 contains papers that study the coronavirus in at molecular levels, including RNA and protein; cluster 5 contains epidemiology studies that focus on the public health model of the global pandemic (Fig 4).

Clusters 2 and 3 include papers studying Severe Acute Respiratory Syndrome (SARS). Wordclouds for these clusters highlight the short and long form of SARS.

5 Conclusion

Literature clustering offers a bird's eye view to gauge publication trends in the space. The goal is to produce tools to identify relevant publications. We have presented a new approach to research paper clustering in this paper that utilizes doc2vec algorithm and BM25 similarity measure. Future work may include the use of this approach on diverse types of document sets to establish its robustness.

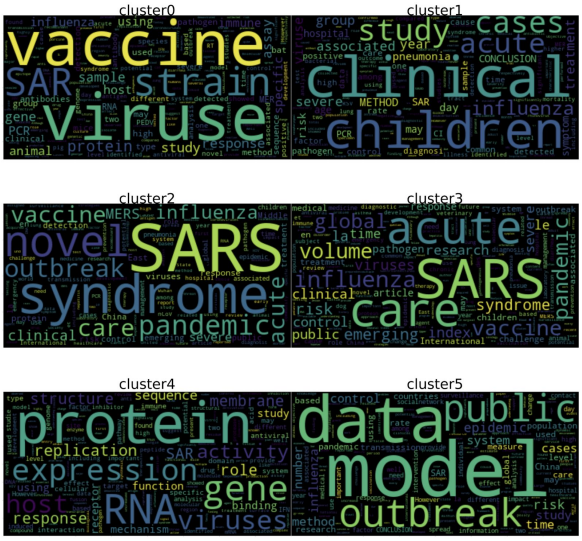

Figure 4: WordCloud of Clusters 0 through 5

References

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex Wade, Kuansan Wang, Nancy Xin Ru Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni and Sebastian Kohlmeier. 2020. *CORD-19: The COVID-19 Open Research Dataset*. Proceedings of the Workshop on NLP for COVID-19 at ACL 2020, Association for Computational Linguistics https://arxiv.org/abs/2004.10706

Stephen Robertson and Hugo Zaragoza. 2009. *The Probabilistic Relevance Framework: BM25 and*

Beyond. Foundations and Trends in Information Retrieval. 3, 4, *333–389*.

Annoy, https://github.com/spotify/annoy

Quoc Le and Tomas Mikolov. 2014. *Distributed Representations of Sentences and Documents*. Proceedings of the 31st International Conference on Machine Learning, Beijing, China.