COVID-19 Literature Clustering

Maksim Ekin Eren, Nick Solovyev, Charles Nicholas, Edward Raff
meren1@umbc.edu, sonic1@umbc.edu, nicholas@umbc.edu, edraff1@umbc.edu
Malware Research Group, CSEE Department, UMBC

Objectives

Given the large number of scientific papers regarding the rapidly spreading COVID-19, it is difficult for a health professional to keep up with the research community. Can we apply what we learned from finding similarities among malware specimens, on a large scale [1][2], to a different domain:

- Can we cluster similar articles via Unsupervised Learning to make it easier for health professionals to find relevant research articles?
- Is there an underlying topic in each cluster that we can discover via Topic Modeling?

Approach

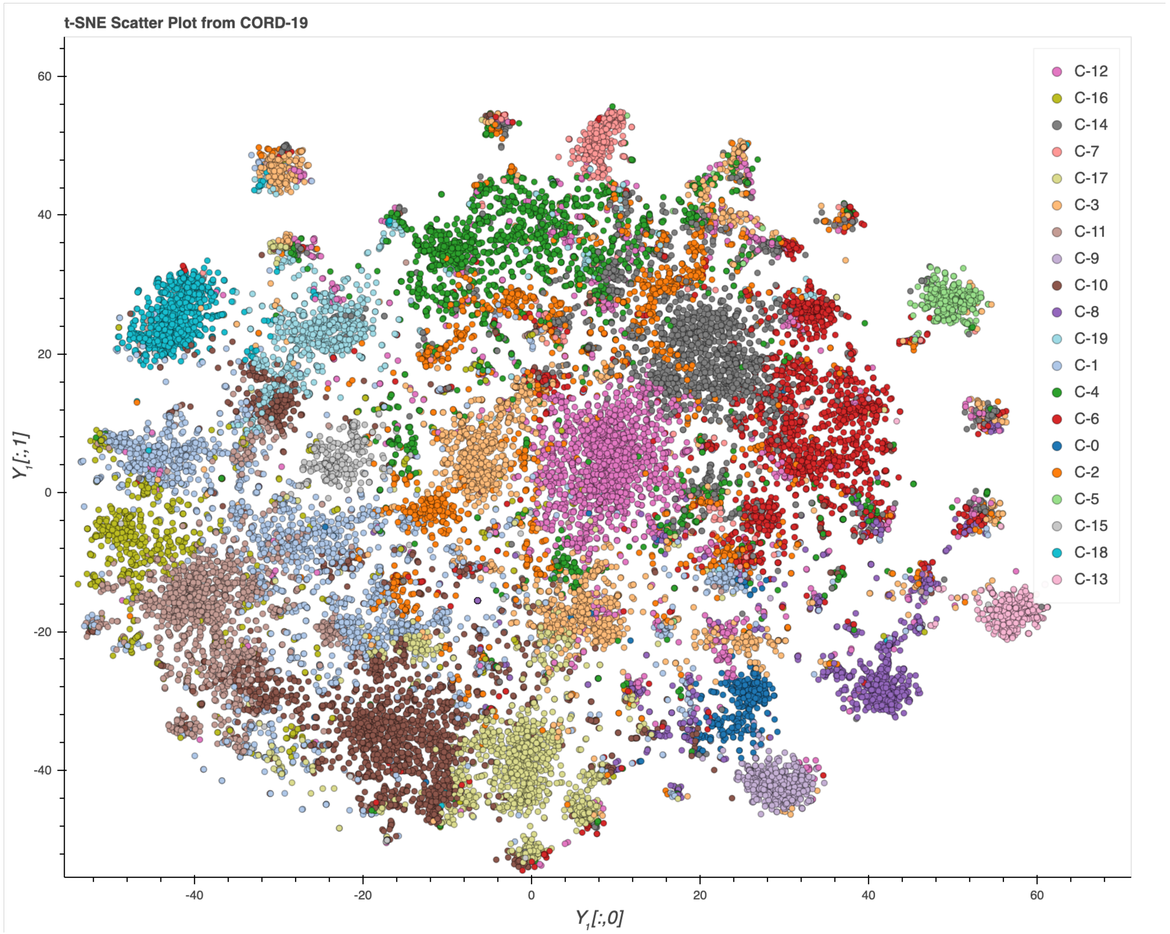
- Parse the text from the body of each document using Natural Language Processing (NLP).
- Turn each document instance d_i into a feature vector X_i using Term Frequency–inverse Document Frequency (TF-IDF).
- Apply Dimensionality Reduction to each feature vector X_i using t-Distributed Stochastic Neighbor Embedding (t-SNE) to cluster similar research articles in the two dimensional plane X embedding Y_1.
- Use Principal Component Analysis (PCA) to project down the dimensions of X to a number of dimensions that will keep .95 variance while removing noise and outliers in embedding Y_2.
- Apply k-means clustering on Y_2, where k is 20, to label each cluster on Y_1.
- Apply Topic Modeling on X using Latent Dirichlet Allocation (LDA) to discover keywords from each cluster.
- Investigate the clusters visually on the plot, zooming down to specific articles as needed, and via classification using Stochastic Gradient Descent (SGD).

Dataset Description

"In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19). CORD-19 is a resource of over 44,000 scholarly articles, ..., about COVID-19, SARS-CoV-2, and related coronaviruses..." [3]

Plot

At first glance, we can see clear clusters appearing on the scatter-plot of the t-SNE Y_1 embedding. Also, we can see that the labels we got by k-means, here represented by different colors, seems to be clustering together.



Discover the Clusters on the Interactive Plot: http://www.bit.ly/covid19-clustering

Results

To further investigate the clustering results, manual analysis is necessary. For example:

- We can see that all articles related to Middle East Respiratory Syndrome (MERS) are collected in Cluster 18.
- Cluster 16 has articles revolving around the economic and social impact of corona-virus.
- Cluster 15 includes topics about air and virus particles while housing a sub-cluster of articles that focus on face masks and filters as a topic.

Articles will classify to the clusters found by k-means with $F1$ score of 88.14% using SGD.

References

Edward Raff and Charles Nicholas.
An alternative to ncd for large sequences, lempel-ziv jaccard distance.
In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 1007–1015, New York, NY, USA, 2017. ACM.

Charles Nicholas.
Mr. shakespeare, meet mr. tucker.
In *High Performance Computing and Data Analytics Workshop*, September 2019.

Kaggle.
Covid-19 open research dataset challenge (cord-19).
https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge, March 2020.
Allen Institute for AI in partnership with the Chan Zuckerberg Initiative, Georgetown University's Center for Security and Emerging Technology, Microsoft Research, and the National Library of Medicine - National Institutes of Health, in coordination with The White House Office of Science and Technology Policy.

QR to the Interactive Plot

SCAN ME

April 3, 2020 - v1.0