DS 5230 - Unsupervised Machine Learning and Data Mining
COVID-19 Literature Clustering

Rounak Bende
Khoury College of Computer Science
Northeastern University
Boston, USA
bende.r@northeastern.edu

Pareekshit Reddy Gaddam
Khoury College of Computer Sciences
Northeastern University
Boston, USA
gaddam.pa@northeastern.edu

I. Introduction

Given the large number of documents and the rapid spread of the CoronaVirus, it has been challenging for healthcare professionals to stay updated with the new information regarding the virus. This is where we thought clustering similar documents could be helpful. Since the problem is unsupervised as we do not have any labels for the clusters, we plan to use topic modeling and extract keywords to associate these clusters to relevant topics. Highly similar Publications will have the same label and will be plotted in the form of a cluster. Topic modeling will be used to find the keywords in the clusters. There is a need for these approaches to be applied and derive meaningful insights as the information about these topics is increasing rapidly, with many articles being written and published every day.

The remainder of the paper is laid out as follows: The second section reviews related work done on clustering covid related articles. The background material for this paper is described in Section 3. The proposed strategy, as well as the methods and algorithms we used, is described in Section 4. Section 5 describes the results of several algorithms based on the experiments we conducted. The study is concluded in Section 6 with a discussion on future work.

II. RELATED WORK

Several works related to clustering covid-related articles have been done in the recent past using various Partitioning, Hierarchical, and density-based clustering methods. Since the clusters are overlapping as a result of articles talking about relevant topics, techniques like Gaussian Mixture Models can perform well. A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. Gaussian Mixture Models can be used to perform soft clustering, which means that a particular data point can belong to multiple clusters at the same time.

III. BACKGROUND INFORMATION

The world had to endure a pandemic on an unprecedented scale in 2020. But the scientific research community was quick to respond and started to research this disease at a very fast pace. At one point according to a survey it is estimated that 23,000 papers were published from January 2000 to May 2020. This huge inflow of information left the researchers overwhelmed and difficult to keep up with fresh knowledge about the Corona Virus.

To tackle the COVID-19 pandemic, Georgetown University Center for Security and Emerging Technology, Microsoft Research, the Allen Institute for AI, IBM, and the National Library of Medicine, The Chan Zuckerberg Initiative created the CORD-19 dataset. [1]

The White House Office of Science and Technology Policy put out a call to action for all data scientists to help better understand the virus through analysis using data science techniques. Our project is now focused on using natural language processing, clustering, and dimensionality reduction techniques to produce a structured organization of the literature.

IV. PROPOSED APPROACH

We initially parsed the text from each document using Natural Language Processing. We preprocessed the data to remove stopwords and performed stemming and lemmatization. We used the langdetect package to identify the language of the articles and retained only those articles which were written in English for our further analysis. We initially used NLTK's Porter Stemmer and WordNet Lemmatizer for our preprocessing. But upon exploring we found specific libraries that work well on scientific data. We then made use of the scispacy package to perform our preprocessing tasks. This package also consists of a different set of stopwords which are more probable to occur in scientific documents. After performing preprocessing, we then vectorized each instance of a document into a feature vector using Term Frequency-Inverse Document Frequency. TF-IDF allows us to correlate each word in a document with a number that indicates how

important that word is in that document. [5]Then, documents with similar, relevant words will have similar vectors. After we have feature vectors, we performed dimensionality reduction as the dimensionality was high and used techniques like Principal Component Analysis to project down the dimensions of the data to several dimensions that will keep 95 percent of the original variance in the data. Dimensionality reduction was really important because we were handling data with a large vocabulary size. After reducing the dimensions, we performed clustering to identify similar documents in the corpus using Algorithms like K-Means, Gaussian Mixture Models, Spectral Clustering and density-based clustering models like DBSCAN. Density-based clustering did not perform as well as K-Means and Gaussian Mixture models in terms of the silhouette score. We then visualized the results using t-SNE along with the labels we got from K-means algorithm. We can compress our high-dimensional features vector to two dimensions using t-SNE. The body text can be plotted using the two dimensions as x,y coordinates. When downsized to 2D, t-SNE will try to preserve the relationships of the higher dimensional data as nearly as feasible. The similar articles will thus be in closer proximity to each other. As we were not certain about the number of clusters we used an elbow plot to find an optimal value for the number of clusters.

Once the clusters are formed, we used topic modeling using Latent Dirichlet Allocation to find keywords within each cluster so that we can associate these clusters to relevant topics. With LDA, each document can be described by a distribution of topics and each topic can be described by a distribution of words. LDA is a generative statistical model that allows sets of words to be explained by a shared topic.

We also performed a similar document suggestion task where we checked for the similarity between a given document and returned 'k' similar documents arranged based on similarity. We used different measures of similarity like cosine similarity, Euclidean distance and Jensen Shannon distance. We performed TF-IDF transformation to get two real-valued vectors and computed similarity among these vectors.

V. EXPERIMENTS

A. Dataset Description

The Kaggle CORD-19 competition dataset is brought to us by the White House and a consortium of leading research groups. This dataset consists of over 500,000 articles in several languages but mainly in English, consisting of information about COVID-19, SARS-CoV-2, and other related coronaviruses. This freely available dataset was made public for the research community to apply the recent advances in Natural Language Processing and other Machine Learning techniques to generate meaningful insights and aid the medical research communities. There is a need for these approaches to be applied and derive meaningful insights as the information about these topics is increasing rapidly, with many articles being written and published every day. This vast amount of data makes it difficult for the scientific and medical research communities to keep up with the ongoing publications.

B. Experimental Setup

The task we are going to perform as a part of this project is to cluster covid-19 articles. We have performed preprocessing on the content of the post and put it into a cleaner format by removing stopwords,punctuations and also performed lemmatization. Once we had the data cleaned, we used TF-IDF Vectorizer to convert these articles to get real-valued vectors. We then performed dimensionality reduction using Principal Component Analysis and plotted the data in lower dimensions using T-SNE. We then used several clustering algorithms on the data with reduced dimensions.

C. K-Means Clustering

We plotted the elbow plot to find the optimal number of clusters to use for clustering.

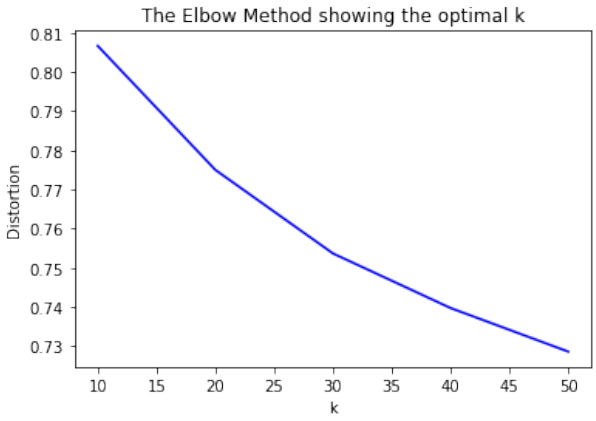

Fig. 1. Elbow Graph

TABLE I
K-MEANS CLUSTERING RESULTS

Metrics	Scores
Silhouette Score	0.07267476833829357
Davies Bouldin Score	1.3606865220591242

D. Spectral Clustering

Spectral clustering is a collection of algorithms for locating K clusters using a matrix's eigenvectors. This matrix is typically created using a collection of pairwise similarities between the clustering points. This process is known as similarity-based clustering or graph clustering.

TABLE II
SPECTRAL CLUSTERING RESULTS

Metrics	Scores
Silhouette Score	0.112864901023528897
Davies Bouldin Score	1.3288469631481714

E. Gaussian Mixture Models

The GaussianMixture object implements the expectation-maximization (EM) algorithm for fitting mixture-of-Gaussian models. It can also draw confidence ellipsoids for multivariate models, and compute the Bayesian Information Criterion to assess the number of clusters in the data. Unlike other models explored earlier, Gaussian Mixture Models perform soft clustering.

TABLE III
GAUSSIAN MIXTURE MODEL RESULTS

Metrics	Scores
Silhouette Score	0.14836118096961962
Davies Bouldin Score	1.285991163648109

F. Results

Comparing all models, Gaussian Models performed the best.

We visualized the data in lower dimensions using T-SNE with K-Means and GMM Labels as plotted below.

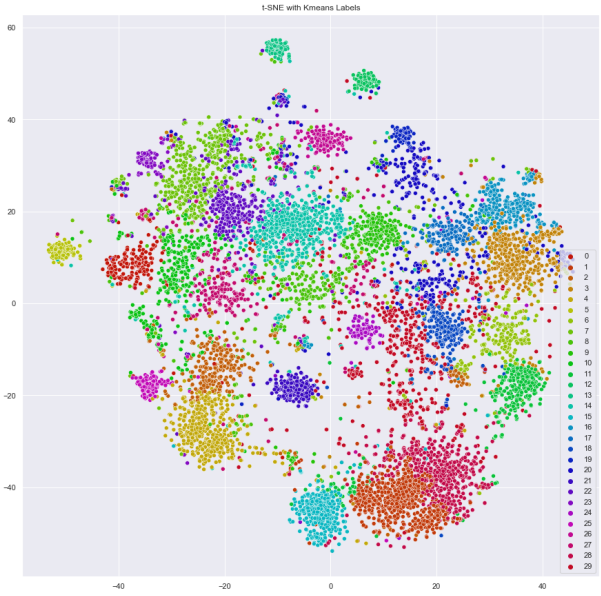

Fig. 2. T-SNE with K-Means Labels

We could observe an overlap between the clusters. To get more insights into what each cluster is talking about we performed LDA to identify keywords within each cluster and also plotted word clouds.

If we observe the Figure 4 wordcloud, we could see the keywords are mostly related to poultry, birds etc. One article from this cluster is:

- Domestic Poultry and SARS Coronavirus, Southern China
- Link:http://doi.org/10.3201/eid1005.03082

If we look at the second wordcloud it is talking about several viruses in general.One article from this cluster:

- Fever in travellers returning from the tropics

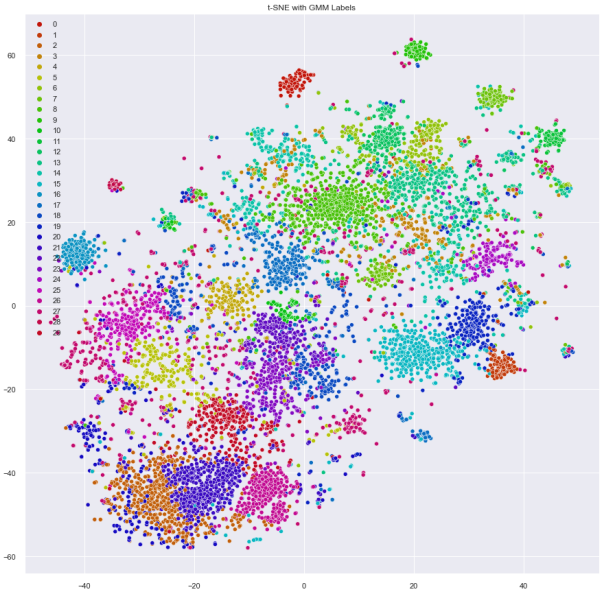

Fig. 3. T-SNE with GMM Labels

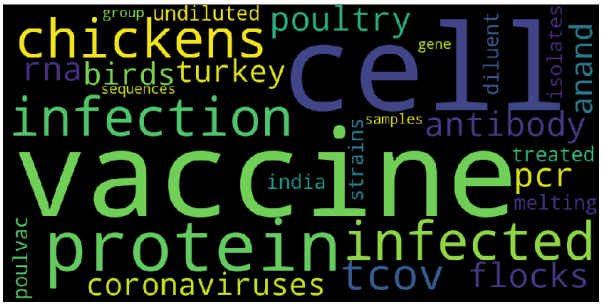

Fig. 4. First Word Cloud

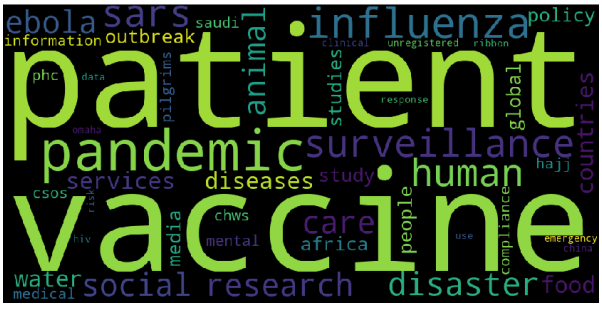

Fig. 5. Second Word Cloud

- Link:http://doi.org/10.1016/j.medcle.2019.03.013

Searching for a single key term may help in inadvertenly filtering out highly similar papers that use different phrasing. This would be really helpful considering the volume of information related to covid-19 and other viruses available right now.

We then implemented a document suggestion task to return 'k' similar documents based on euclidean distance, cosine similarity and Jensen Shannon divergence.

G. Jensen Shannon Divergence

The Jensen–Shannon divergence is a method of quantifying the similarity between two probability distributions in probability theory and statistics. It's also known as total divergence from the average or information radius. It is based on the Kullback–Leibler divergence, but with a few noticeable (and important) features, such as being symmetric and having a finite value.Jensen–Shannon distance is a metric that measures the square root of the Jensen–Shannon divergence [2].

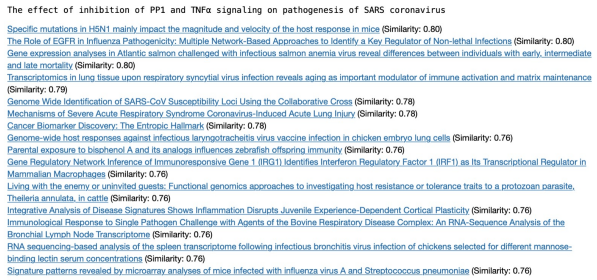

Fig. 6. "k" Similar Documents

The above figure shows the list of documents returned as a suggestion for the given document using Euclidean distance as a metric.

VI. CONCLUSION AND FUTURE WORK

In this project, we sought to cluster available literature on COVID-19 and reduce the dimensionality for visualization purposes in this research. Professionals can quickly find material related to a central topic by grouping the literature in this way. Every article is linked to a wider topic cluster, eliminating the need to actively search for relevant material. We performed LDA to identify the keywords in the clusters. Unsupervised methods may or may not group occurrences in a predictable manner. There is no correct answer for how the papers should be clustered.We observed that the clusters were overlapping. We plan to implement model-based overlapping clustering [3] in the future and also explore other evaluation methods [4]. In the future, we also intend to use Word2Vec to generate embeddings to provide as input to the models.

REFERENCES

[1] ALLEN INSTITUTE FOR AI, United States of America, accessed 4 May 2022, ¡https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge¿

[2] "Jensen–Shannon divergence" Wikipedia, Wikipedia Foundation, 25 April 2022, https://en.wikipedia.org/wiki/Jensen

[3] Q. Fu and A. Banerjee, "Multiplicative Mixture Models for Overlapping Clustering," 2008 Eighth IEEE International Conference on Data Mining, 2008, pp. 791-796, doi: 10.1109/ICDM.2008.103.

[4] M. K. Goldberg, M. Hayvanovych and M. Magdon-Ismail, "Measuring Similarity between Sets of Overlapping Clusters," 2010 IEEE Second International Conference on Social Computing, 2010, pp. 303-308, doi: 10.1109/SocialCom.2010.50.

[5] P. Bafna, D. Pramod and A. Vaidya, "Document clustering: TF-IDF approach," 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 2016, pp. 61-66, doi: 10.1109/ICEEOT.2016.7754750.