COVID-19 Research Articles Analysis and Visualization

Dongjin Choi Tina Yuting Guo Seil Kwon Isaac Thomas Rehg Andrew Wang Chaerin Yoo

The **CORD-19 Browser** is a **dashboard system** with an **automated pipeline** that summarizes COVID-19 research articles by incorporating three techniques: **clustering**, **citation networks**, and **Q&A system**.

Motivation

The production of vast amounts of Covid-19 research publications has made it harder for users to look for publications, and also hinders the process of exploring them.

This could lead to inefficiency in publication searching, making the process unnecessarily difficult. We concluded that visualization of the papers could help solve the problem.

Data

Covid-19 Open Research Dataset (CORD-19) is a text dataset of over 200,000 research articles' abstracts, body texts, and metadata downloaded from National Institutes of Health.
We sampled 10,000 articles for our project.
The size of dataset is nearly 200MB after processing.

Our Approach

1. Document Clustering

We represented documents using TF-IDF, reduced dimensionality using PCA and t-SNE, and clustered documents with k-means and LDA.

2. Graph Analysis

We created a citation network between the articles using NetworkX and ArgoLite. For analyzing the influential capacity of each papers, we focused on in-degree and VoteRank centrality metrics.

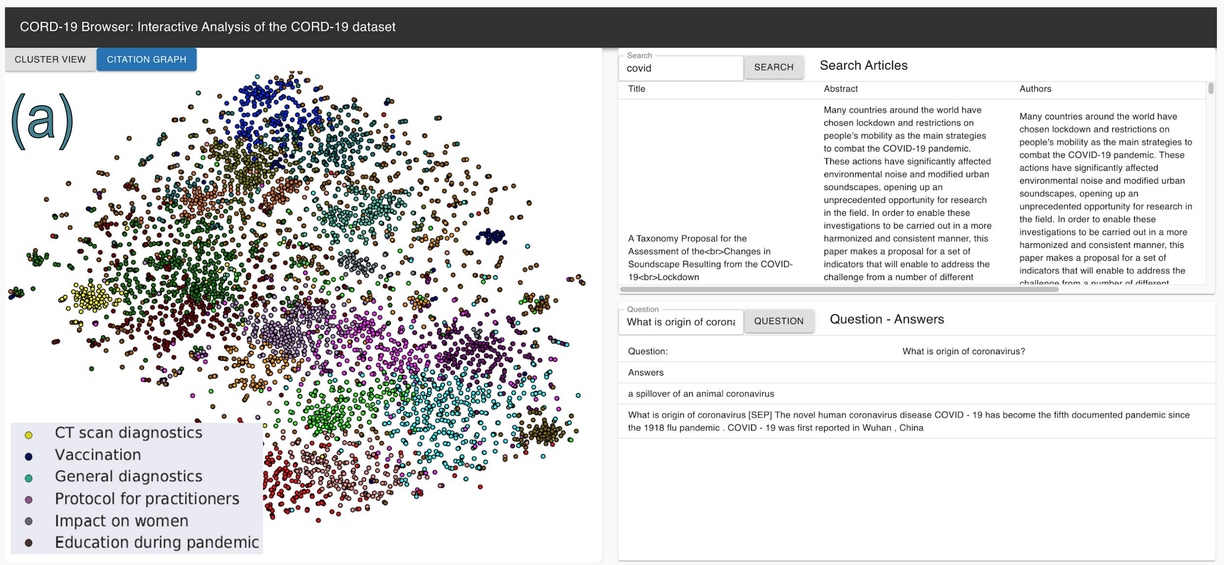
3. Question Answering

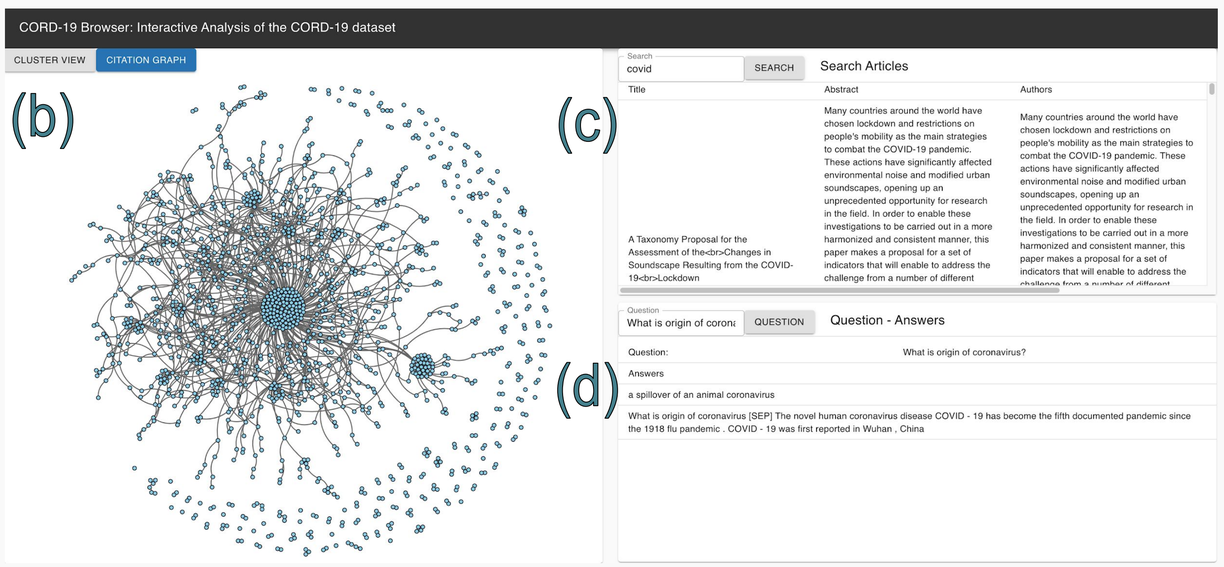
We used a deep-learning model from HuggingFace library which utilizes BERT model for automated question answering.

Experiments and Results

Comparing with LDA, k-means produced 20 different topics of cluster assignments with less overlap. Overwhelming number of articles referenced one paper[1] 156 times, and VoteRank was more efficient in determining influential papers, compared to degree centrality.

[1] Epidemiological and Clinical Characteristics of the Early Phase of the Covid-19 Epidemic in Brazil





Solution to the Ever Growing Information

As shown above, our solution provides users with clustered view(a) of the CORD-19 papers into different topics such as "Vaccination" or "Impact on women" and a comprehensive citation network(b) in one sight. When "covid" is searched In panel (c), we can see that all the covid related articles are searched. In panel (d), when we ask "What is the origin of coronavirus?" the answer "a spillover of an animal" is shown.

Evaluation

After surveying 26 people about their experience with our system, most of them gave a score of 4 on a scale of 1 to 5 for questions such as "I would use this system frequently" or "I found the Q&A tools very helpful."