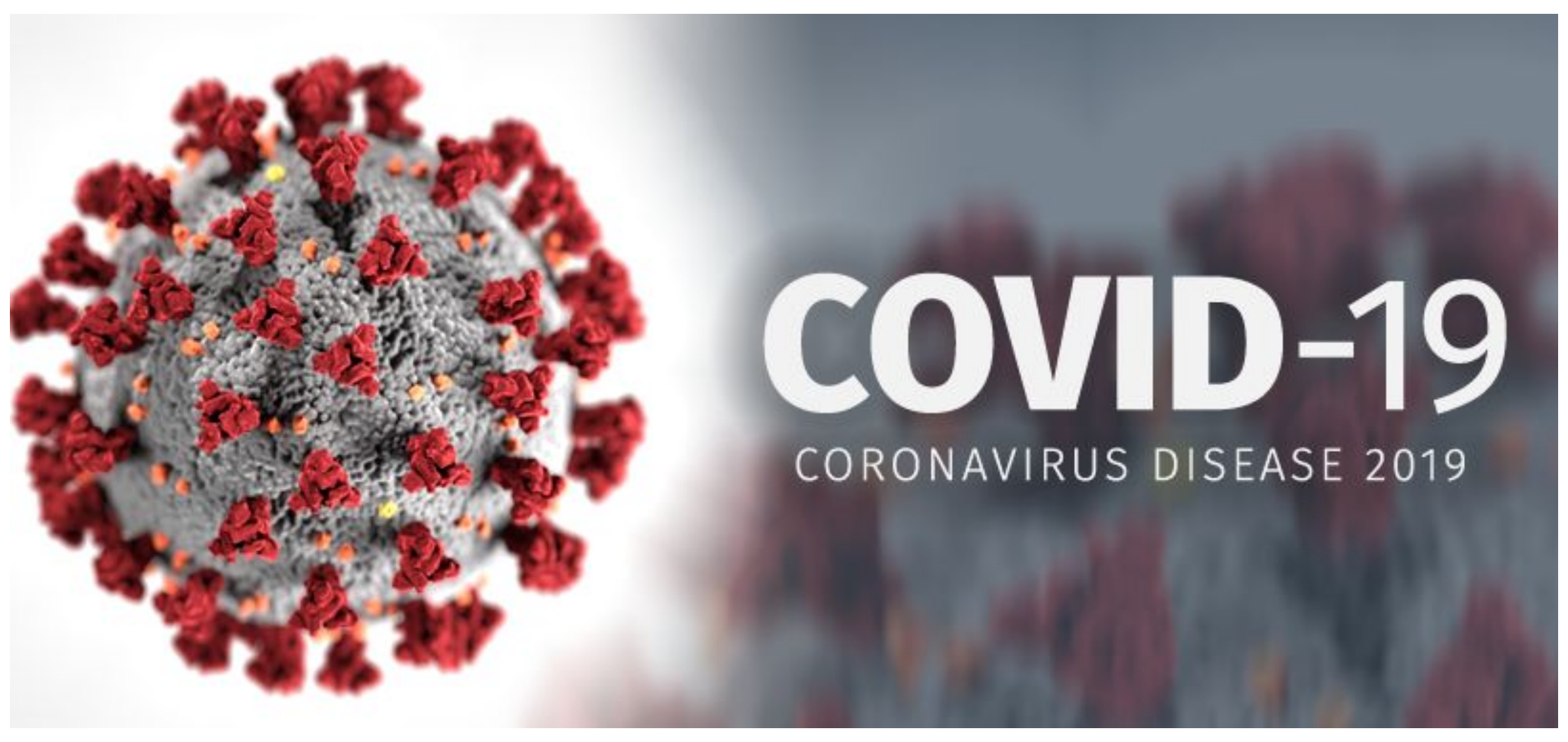
Data N Depictions

DnD
Team #94

Scott Cardinal, Shad Hopson, Keegan Nesbitt,
Tina Nguyen, Erhan Posluk, Annmarie Thomas



Introduction & Motivation

The rapidly growing amount of research on COVID-19 is humanly impossible to stay up to date on without any assistance. Researchers, policymakers, and others can't quickly find and review research important to their roles, as the volume is simply too great and new content comes out too rapidly. In our implementation, we propose an approach that takes the text of the research, extracts topics, and relates the topics with each other and with the individual papers. This is achieved through a combination of Natural Language Processing, similarity calculations, and graph processing, implemented on AWS and locally. Using this output, we present the user with a way to intuitively and interactively navigate through the sea of COVID-19 research for the content most relevant to them.

Data

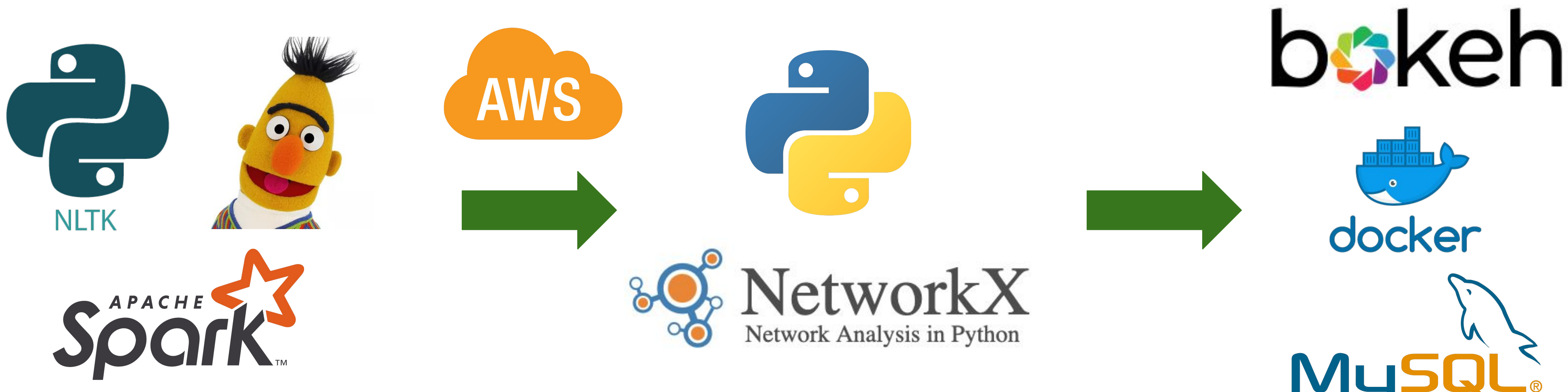
COVID-19 scholarly articles from the CORD-19 Dataset

Data Components:
- Research paper abstracts
- Research paper full-text (excluded from our scope)
- Authors
- Publication date
- Journal

Format:
Combination of JSON and CSV downloaded from Kaggle website

Total Documents: ~477,000
Usable Documents: ~300,000

Approach Overview



Text Preprocessing
Topic Extraction

Graph Processing
Topic Similarity

User Interface

NLP Preprocessing

Prior to doing any meaningful processing, we needed to do some cleaning of the data to get the raw text into a cleaner, more consistent format.

Steps Taken:
1. Duplicate and bad record removal
2. Consistent letter casing
3. Replace contractions with full words
4. Remove noisy characters (numbers, single letters, symbols, etc.)
5. Tokenization - convert sentences to individual words
6. Lemmatization - change words to their roots
7. Stop Word removal

'Eevee, the quick brown fox, jumped over the lazy dog.'

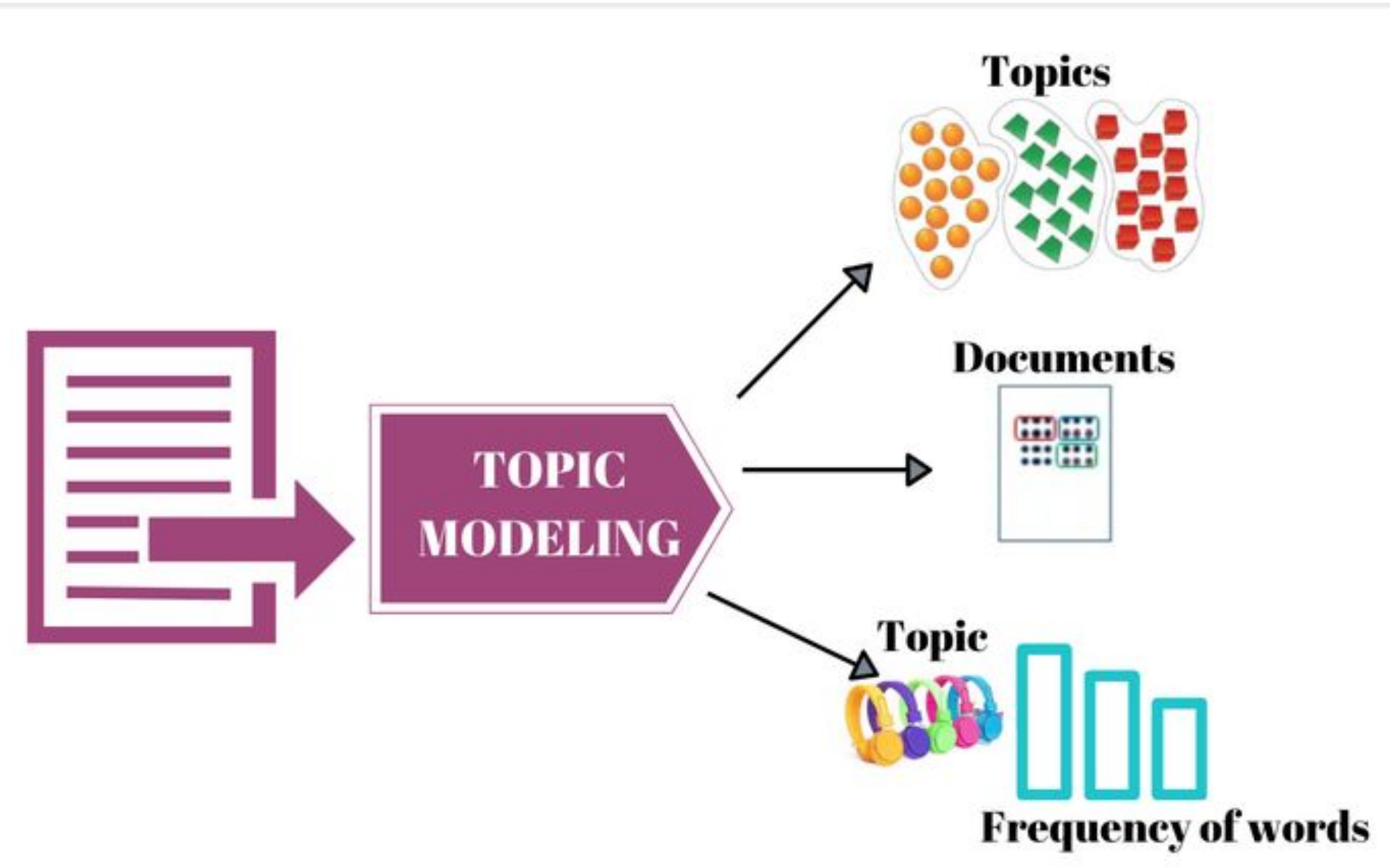
['eevee', 'quick', 'brown', 'fox', 'jump', 'over', 'lazy', 'dog']

Processing of full data set took ~2 hours

Utilized Apache Spark to distribute applying NLTK operations to each document via UDFs and SparkSQL functions

Executed distributed code on an AWS EMR cluster consisting of 5 m5.xlarge nodes

Topic Extraction



Showing the raw text of the papers is not feasible for a large corpus, so we need to try to distill the papers into a more limited number of topics. To do this, we'll also need to employ approaches that convert human readable text into vector representations. Each of these topics can be described by a weighted collection of terms.

Processing of ~100,000 documents set took:
BERT: 3.5 hours
LDA: 1 hour

Executed code on an AWS EC2 m5.2xlarge instance with 100GB of EBS storage

BERT (Bidirectional Encoder Representations from Transformers)	**LDA** (Latent Dirichlet Allocation)
• In the BERT model, we are using BERT to transform the documents into numerical representations called embeddings, such that within the space of the embeddings similar words will be close together. BERT comes pre-trained on a huge corpus, and uses this to provide general context on the english language. • Representative words given for topics: Topic 1: teacher, learning, teach, teaching Topic 2: mrna, replication, genome, ribosome Topic 3: graph, algorithm, neural, semantic	• In the LDA model we use LDA to extract representative words of clusters made with TFIDF and the Gaussian Mixture model based on the probability of the word being in a document in that topic. • In general LDA tended to have more general topics than in the BERT model. • Representative words given for topics: Topic 1: health, covid, disease, research Topic 2: student, online, learn, medical Topic 3: health, mental, depression, patient

Amazon
EC2

Topic Modeling & Clustering

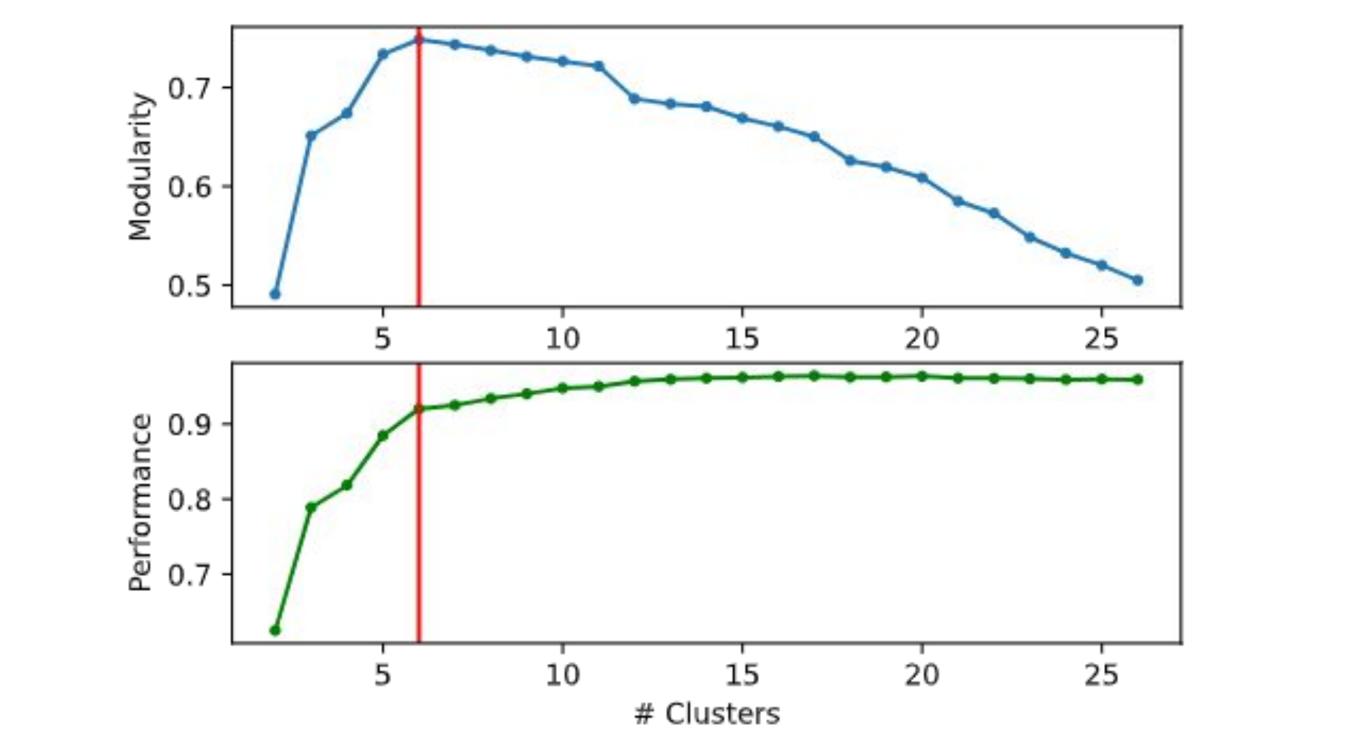
After topic extraction, each document has a probability of association with each topic. The majority of documents, however, do not fall cleanly into only one topic. Multiple topics could potentially be associated with individual papers.

The goal was to derive associations **between** the extracted topics based on their co-occurrence within individual documents. Similarly, associations between documents could be determined by correlation of their probability of association with each extracted topics as a compositional profile of probabilities.

The method used was adapted from weighted gene co-expression network analysis (WGCNA). Rather than raw correlation, the topological overlap within a network *derived* from those correlations was used to generate a weighted, undirected graph of both topics and papers. The Girvan-Newman algorithm for network community detection is used to identify potential topic clusters. Optimal number of clusters is determined from balancing community modularity and and cluster performance measures.

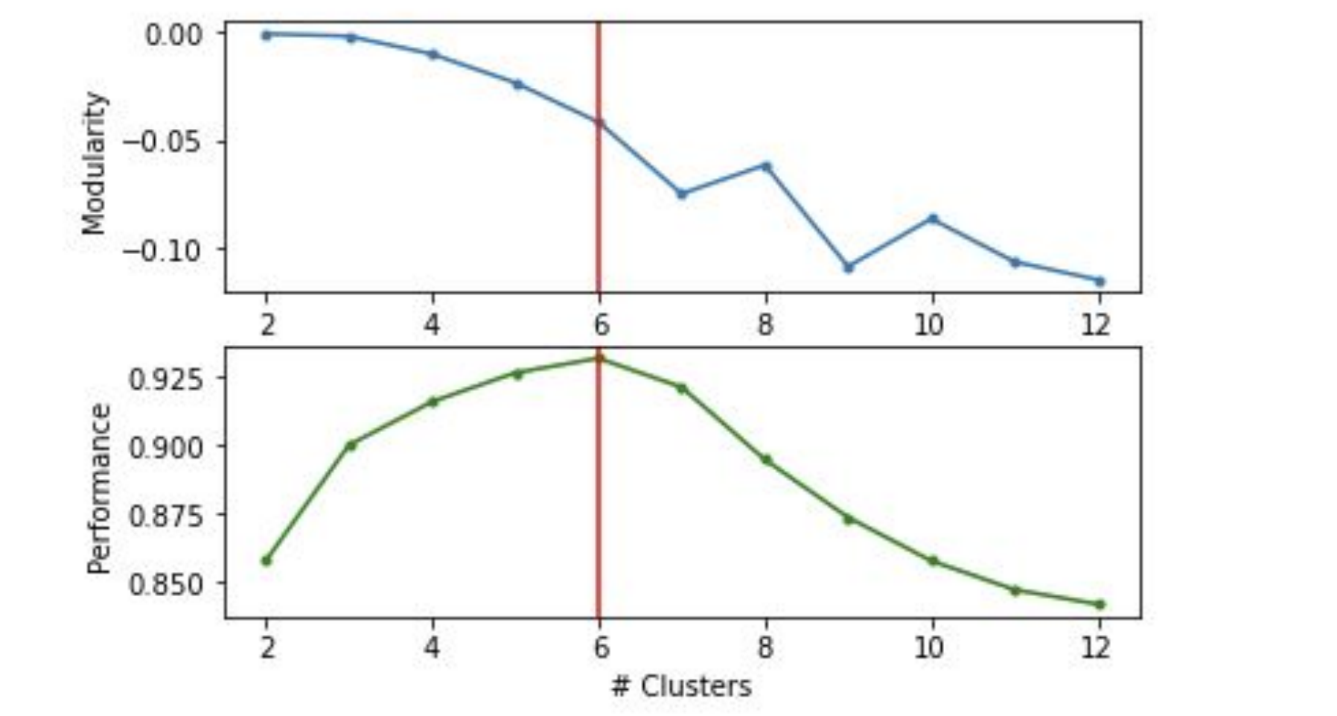
BERTopic

Girvan-Newman (edge_betweenness_centrality_equal_with_weight): 5k sample



LDA

Girvan-Newman (edge_betweenness_centrality_equal_with_weight): 5k sample

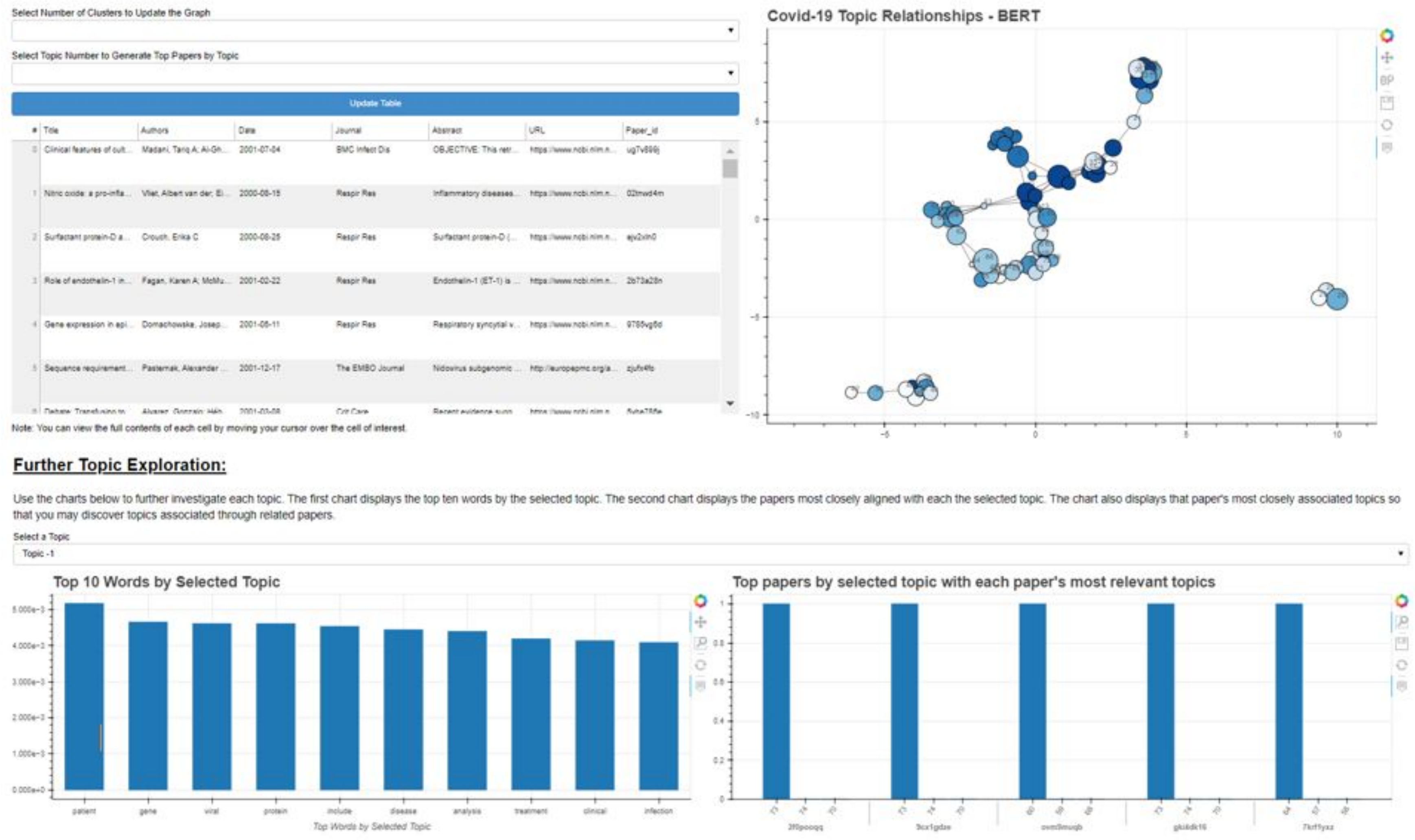


User Interface

The UI is created with Bokeh connected to a data backend using MySQL. The UI starts with a network graph displaying clusters of topics. Once a user selects a node, this filters a data table with the top five papers associated with that highlighted topic.

For further data exploration, there are two bar charts where the user can either view 1) the top ten words within that topic and their probability of being in that topic 2) a grouped bar chart displaying the top 5 topics each paper is associated with and their probability.

This provides multiple ways for a user to explore topics without being overwhelmed by a network graph because it allows them to see the connections between the topics more clearly.



Select Number of Clusters to Update the Graph

Select Topic Number to Generate Top Papers by Topic

Update Table

#	Title	Authors	Date	Journal	Abstract	URL	Paper_id
0	Clinical features of cult...	Madani, Tariq A; Al-Gh...	2001-07-04	BMC Infect Dis	OBJECTIVE: This retr...	https://www.ncbi.nlm.n...	ug7v899j
1	Nitric oxide: a pro-infla...	Vliet, Albert van der; Ei...	2000-08-15	Respir Res	Inflammatory diseases...	https://www.ncbi.nlm.n...	02tnwd4m
2	Surfactant protein-D a...	Crouch, Erika C	2000-08-25	Respir Res	Surfactant protein-D (...	https://www.ncbi.nlm.n...	ejv2xln0
3	Role of endothelin-1 in...	Fagan, Karen A; McMu...	2001-02-22	Respir Res	Endothelin-1 (ET-1) is ...	https://www.ncbi.nlm.n...	2b73a28n
4	Gene expression in epi...	Domachowske, Josep...	2001-05-11	Respir Res	Respiratory syncytial v...	https://www.ncbi.nlm.n...	9785vg6d
5	Sequence requirement...	Pasternak, Alexander ...	2001-12-17	The EMBO Journal	Nidovirus subgenomic ...	http://europepmc.org/a...	zjufx4fo
6	Debate: Transfusing to...	Alvarez, Gonzalo; Héb...	2001-03-08	Crit Care	Recent evidence sugg...	https://www.ncbi.nlm.n...	5vba78fe

Note: You can view the full contents of each cell by moving your cursor over the cell of interest.

Covid-19 Topic Relationships - BERT

Further Topic Exploration:

Use the charts below to further investigate each topic. The first chart displays the top ten words by the selected topic. The second chart displays the papers most closely aligned with each the selected topic. The chart also displays that paper's most closely associated topics so that you may discover topics associated through related papers.

Select a Topic
Topic -1

Top 10 Words by Selected Topic

Top papers by selected topic with each paper's most relevant topics