Final Report of Unsupervised-NLP-project

Problem statement:

We are develop text and data mining tools that can help the medical community develop answers to high priority scientific questions. The CORD-19 dataset represents the most extensive machine-readable coronavirus literature collection available for data mining to date. This allows the worldwide AI research community the opportunity to apply text and data mining approaches to find answers to questions within, and connect insights across, this content in support of the ongoing COVID-19 response efforts worldwide. There is a growing urgency for these approaches because of the rapid increase in coronavirus literature, making it difficult for the medical community to keep up.

Data 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 500,000 scholarly articles, including over 200,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. This freely available dataset is provide to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease. There is a growing urgency for these approaches because of the rapid acceleration in new coronavirus literature, making it difficult for the medical research community to keep up.

Accessing the Dataset

- We have made this dataset available on Kaggle. Watch out for periodic updates.
- The dataset is also host on AI2's-Semantic-Scholar, and you can search the dataset using AI2's new COVID-19 explorer.

Algorithms:

- Problem Understanding.
- Dataset Exploration and Cleaning.
- Natural Language Pre-processing.
- Unsupervised learning algorithms.

Tools:

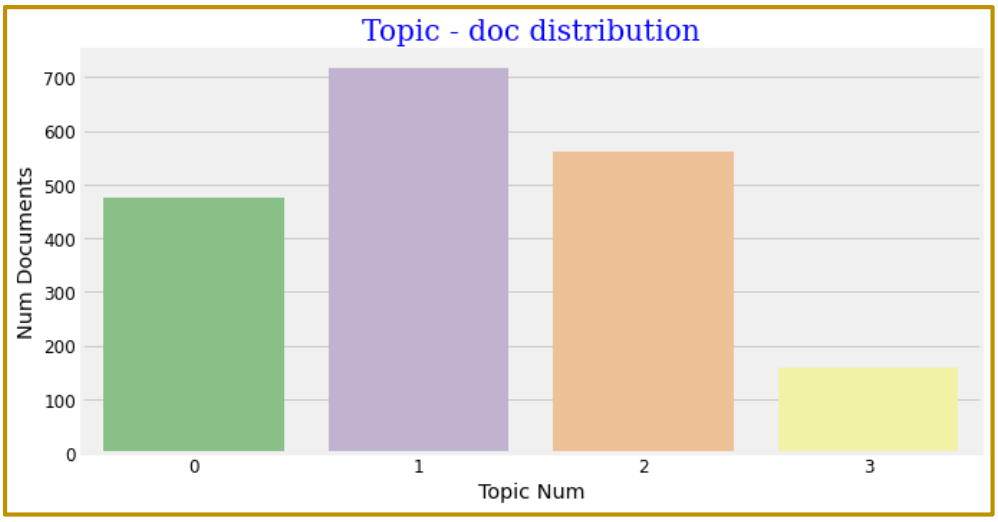
Technologies	Libraries
Python program language	Panda , nltk , sklearn
Jupiter Notebook	NumPy , seaborn,sklean
Microsoft Office	MatPlotLib , wordcloud, pyLDAvis

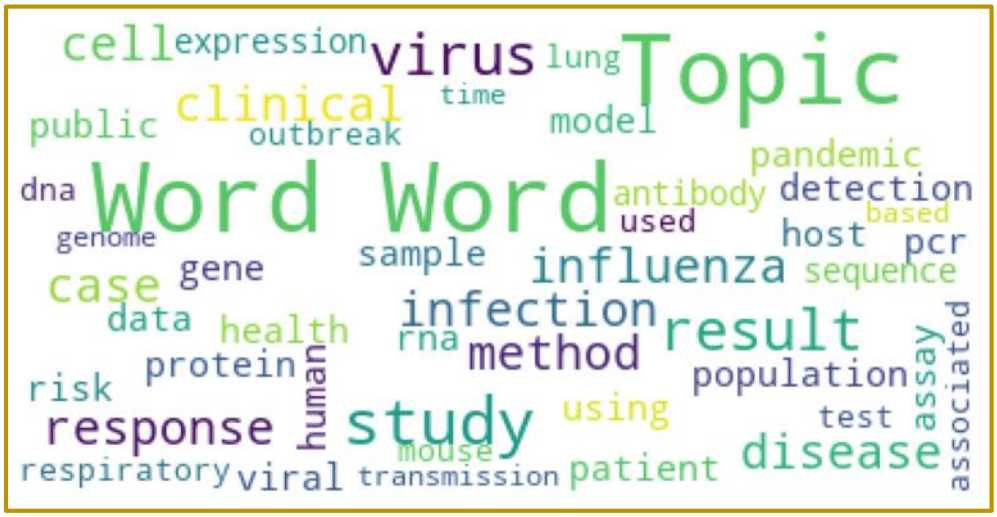
MVP Goal:

In this project we will find related articles by using Topic modelling. Here I am using, LSA , NMF , CorEx, and LDA. All this topic models are used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model.

Communication:

Charts





Presentation snips:

Presentation is [hear]

Team members:
Tahani Alqahtani ,Amani Albalawi.

Instructor:
Mohammad El Deeb , Kinza Waqar