CORD-19: NLP

Aditya Sharma (112676654)

Mukul Javadekar (112961738)

Project Objective:

CORD-19 is a COVID-19 Open Research Dataset, it is a collection of over 47,000 scholarly articles, including over 36,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. Our objective includes:

- Collecting data from the dataset from Kaggle (here we have used .csv file).
- Cleaning the data by removing all the unrequired things from the dataset i.e. commas, punctuations, numbers, non ASCII characters etc. Storing data in corpus and document-term matrix from.
- Doing exploratory data analysis (EDA) over the data to summarize the main characteristics of the dataset in form of top words and word clouds.
- Using Topic Modelling to get the topics and label papers with the topics.

Contributions:

Aditya Sharma:

- Arranging the data in corpus and document – term matrix form
- Creation of Word could and bar plot
- Topic modelling
- Presentation & report

Mukul Javadekar:

- Collecting the data set from Kaggle
- Cleaning the data
- Visualizing the result
- Presentation & Report

Dataset:

This dataset is freely available on Kaggle 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. This dataset can be found on below link - https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

Significance:

With the rapid increase in the coronavirus literature there is a need to apply data analysis techniques which can help researchers get proper insights from the data. 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.

Techniques and Tools:

- Data Gathering: Obtained from Kaggle
    - Data has 19 columns, of which we are interested in cord_id and abstract
- Data Cleaning: Get data in clean standard format for analysis
    - Corpus: a collection of written texts, especially the entire works of a particular author or a body of writing on a particular subject using pandas
    - Document-term Matrix: It is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms. So here we clean text by removing all numbers, letters not belonging to a-z, non-ASCII characters and next line using regular expressions. Tokenize it by word and put in matrix form using scikit-learn and count vectorizer (remove stop-words)
- Exploratory Data Analysis (EDA): Summarize the data
    - Word count – Aggregate word count in document-term-matrix by this we can get the top words
    - Bar plot – To visualize the top words using matplotlib

- Word clouds – A word cloud is a popular visualization of words typically associated with Internet keywords and text data. Here we use it to visualize the top word in each paper after removing the common words such as virus, infection etc. and using word cloud.
- Topic Modelling: Input is document term matrix as each topic has a set of words where order doesn't matter. We use gensim here which use Latent Dirichlet Allocation (LDA) for topic modelling. We used nltk for tagging.
- LDA learns topic mix in every document and word mix in every document. We start with a random number of topics; it goes through every word and randomly assigns it to a topic. It further updates the topic for every word based on how frequently that topic occurs in that document and how often the word occurs in the topic overall. Does in multiple iteration.
- nltk (Natural language tool kit): The Natural Language Toolkit (**NLTK**) is a platform used for building Python programs that work with human language data for applying in statistical natural language processing (NLP). It is used to filter out just the nouns or filtering nouns and adjectives for topic modelling.

Results and Conclusion:

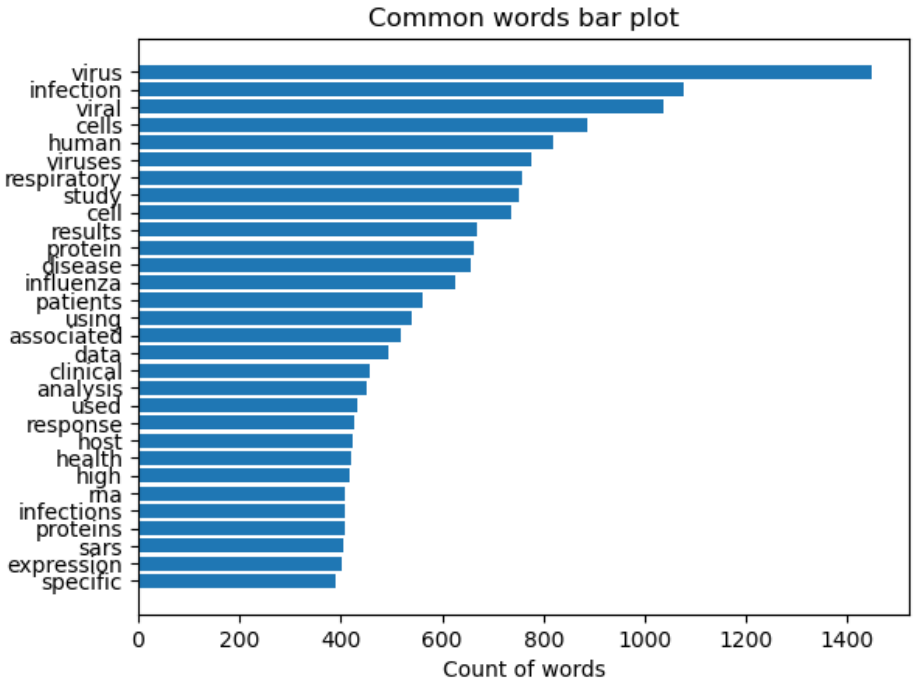

fig1. Top 30 words of 5000 items

From the above bar plot we can see that “virus” word has been appeared most number of times while “specific” is appeared least number of times.

fig2. Word cloud for several papers

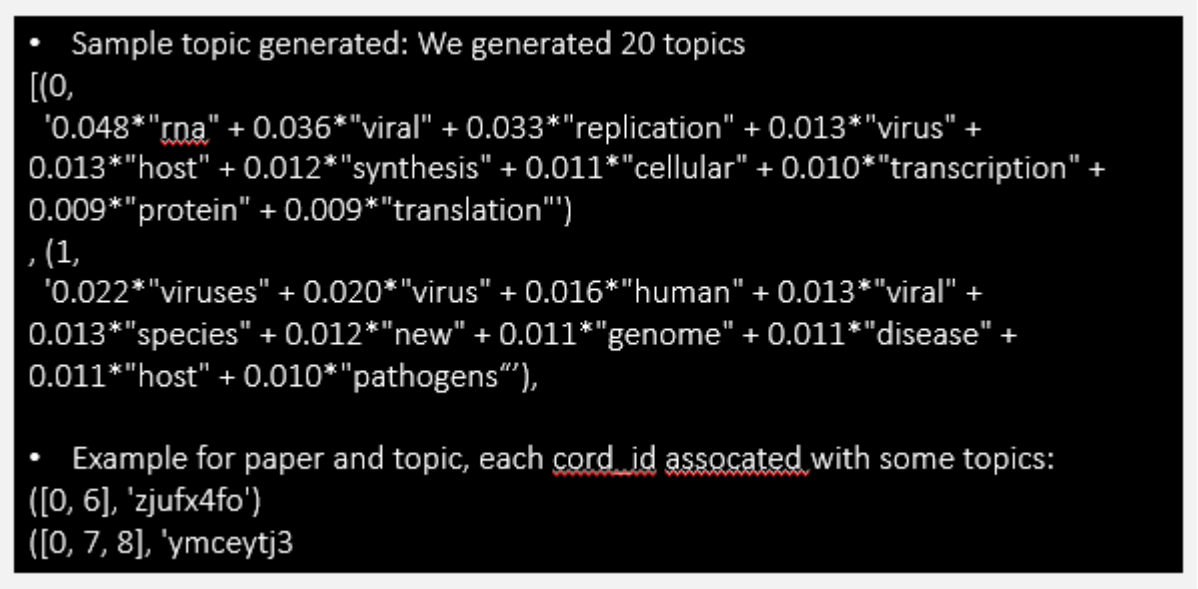

fig 3. Topic modelling

In this fig we can see that we generated 20 topics via topic modelling out of which we have stated 2 topics. As for the second one, we have shown the examples of two cord_ids and their associated topics.

Challenges and Future Work:

- While cleaning data, adding stop words, finding topics, domain related vocabulary and knowledge will be useful as large amount of data.
- Few topics can be selected and analysis on papers belonging to that topic can be performed
- The current analysis is one of general process followed for NLP although to assert relevance of the results domain related knowledge is required.

References:

1. https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge
2. https://en.wikipedia.org/wiki/Document-term_matrix
3. https://www.techopedia.com/definition/30343/natural-language-toolkit-nltk
4. https://matplotlib.org/3.2.1/api/_as_gen/matplotlib.pyplot.bar.html
5. https://www.machinelearningplus.com/nlp/gensim-tutorial/