**Name : Saptarshi Mukherjee**

**REPORT**

**1: Decoding Covid-19 with Genome Analysis**

**PROBLEM STATEMENT**

You are one of the researchers responding to the White House Office of Science and

Technology Policy centre’s call to conduct advanced research on Covid-19. A dataset that

represents the most extensive machine-readable coronavirus literature collection available for

data and text mining to date, with over 29,000 articles, more than 13,000 of which have full

text.

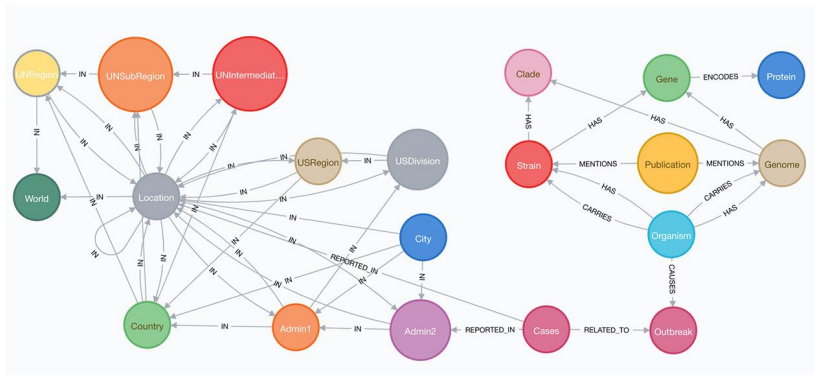
Using the CORD-NER dataset and Knowledge Graph, determine and map out the details of

the SARS-CoV-2 genome to assist understanding of the emergence, evolution and diagnosis

of this deadly virus.

Dataset used: <https://www.kaggle.com/code/xuanwangstat/cord-ne>

KNOWLEDGE GRAPH



**UNDERSTANDING FROM GIVEN KNOWLEDGE GRAPH: -**

* The location hierarchy from global to city levels is displayed on the left side of the basic

knowledge graph model.

* COVID-19 case counts are linked to host organisms, virus strains, genomes, genes, and

protein information, as well as articles that discuss the virus strains.

**WORKFLOW OF MY NOTEBOOK 1: (Covid19-CORD-NER)**

**Text Cleaning and Preprocessing:**

* First, imported 10000 rows from the json file (“CORD-NER-full.json”).
* Converted the json data to csv and imported it.
* Searched for articles published in the year 2020.
* Dropped articles with missing abstract.
* Dropped articles with duplicate abstract.
* In abstract column:
* Converted text to lowercase.
* Removed non-English words.
* Removed Stopwords.
* Removed words with single characters.
* Inverted index

**Article Matching and Deep Cleaning:**

* Developing a keyword list that would facilitate a search to locate only applicable articles.
* Took data from required website articles and saved the index location in a new data frame.
* It uses lemmatization and POS tagging in order to provide unambiguous information.
* The word cloud was created to realize which words or phrases were surprisingly recurring.

**Topic Extraction**

* The LDA model has been built to explore topics and the c\_v model features coherence measure.
* Checked perplexity of my LDA Model
* Thematically relevant sentences are huge technological bypasses.  
  LDA visualization.
* On each theme of the document, different bars will be used to illustrate the data visually.
* T-SNE:  
  Document topic matrix generated.  
  Multi-dimensionally clustering of LDA topic based on T-SNE.

**Semantic-Based Search:**

* Such data as those with duplicates or missing values.
* Dropping non-English articles.
* Spacy Parser and Tokenizer are a regex based library models that are task oriented in their nature.
* Sentence Tokenization
* Word2vec Training
* Ranking documents
* Storing the model and the dataframe as a variable is what we are saving here.

**WORKFLOW OF MY NOTEBOOK 2:**

**(COVID-19-CORD-NER- information-extraction- Question-Answering)**

1. NER Extraction from Text
2. Dependency parses
3. Question-Answering

**NER Extraction from Text:**

* spaCy based imports
* NER extraction using Spacy library
* Closer look at what spaCy is doing when it performs named entity recognition
* Finding same entity texts

**Dependency parses:**

* Encoding grammatical information by using spaCy's dependency visualizer.
* Identifying verbs + direct objects that are grammatically linked to a location.
* Identifying all the actions related to a single city, Wuhan.
* Top 10 Geopolitical Entities Related to "Origin", "Case", or "Transmission.

**Question-Answering:**

* Downloaded a transformer model that's already been trained on SQuAD from the

Huggingface model repository.

* Performed queries

**RESULTS**

**NOTEBOOK 1: (Covid19-CORD-NER)**

* **Filtered required article:**

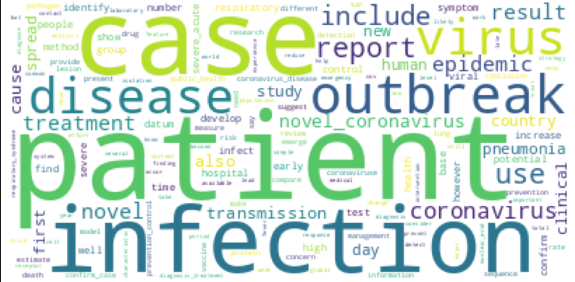
Among 10000 articles, found only 606 articles related to covid19 containing required

keywords.

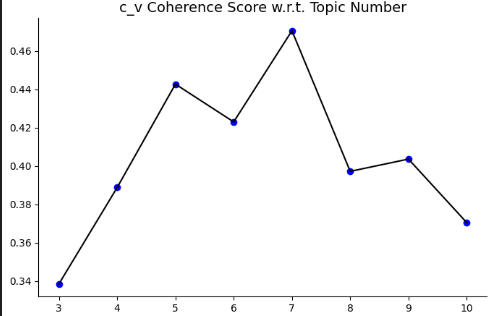
* **KEYWORD LIST –**

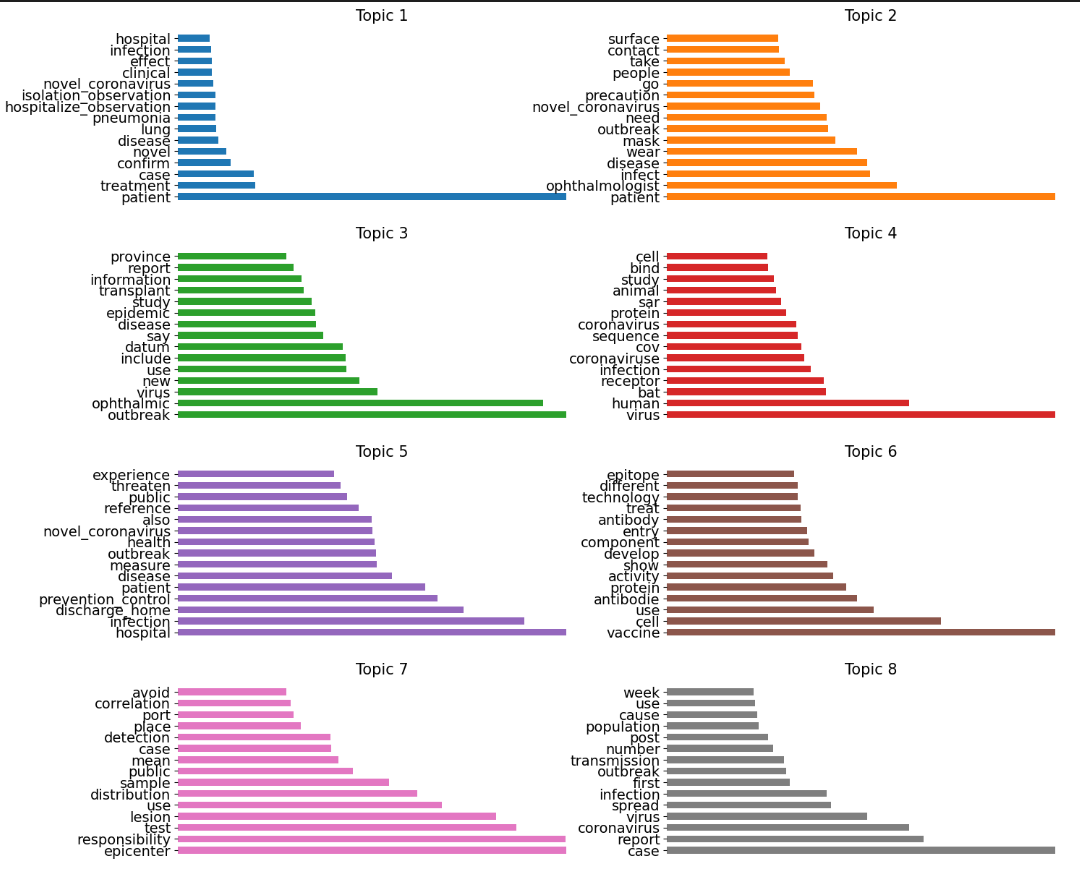
sars-cov-2, sars, cov-2, 2019-ncov, ncov, cov, covid19, covid, corona, coronavirus

* **Word Cloud: Articles Related to SAR-Cov-2:**



Words like 'infection', 'hospital', 'disease', 'prevention\_control','public health','treatment strategies' occur frequently in this corpus..





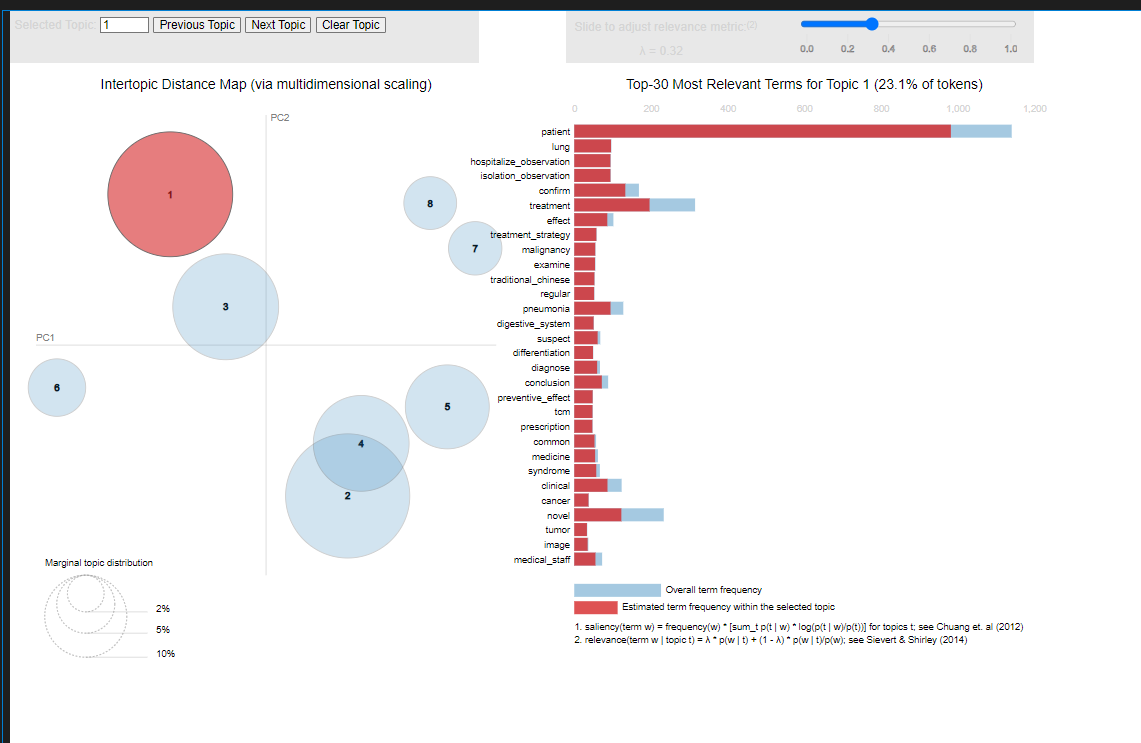
* The first topic highlights the role of hospitals in ensuring public health through the adoption of treatment strategies and infection prevention control measures.

We can see words like 'infection', 'hospital', 'novel', 'prevention\_control','public health','treatment '.

* The second topic delves into the evolution of the virus within the human host, touching upon aspects like viral infection at the cellular level. We can see words like,

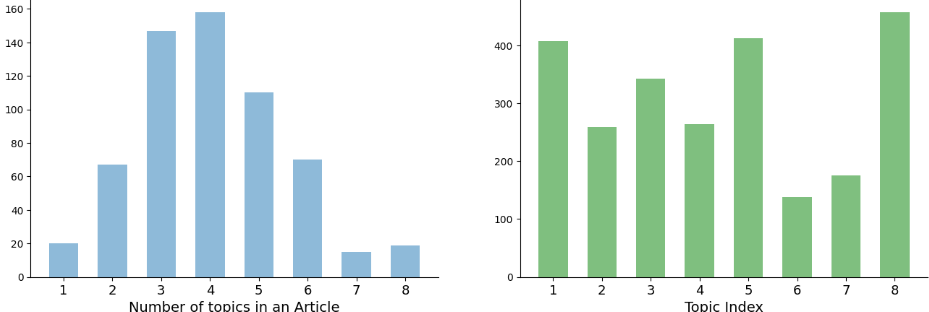
'human', 'cell','host','protein','infection' .

* The third topic discusses the efforts of scientists in testing western drugs in labs, specifically for conditions like atypical pneumonia and myocardial issues. We can see words like, 'lab', 'test', 'ventilator', 'drug','western','atypical pneumonia'.
* The fourth topic suggests that a significant number of patient cases originated from the distribution of lesions in infected lungs. We can see words like lesion', 'infection', 'lung'.
* The fifth topic talks about the observation of hospitalized COVID-19 confirmed patients, focusing on syndrome identification, treatment effects, and precautions taken during isolation. We can see words like 'syndrome', 'confirm', 'hospitalize\_observation'.
* The sixth topic addresses the rise in COVID-19 cases in the country, hinting at aspects like transmission, outbreak reporting, and the virus itself. We can see words like 'case', 'transmission', 'outbreak'.
* The seventh topic likely discusses the symptoms of COVID-19 in a patient, with a particular emphasis on fever. We can see words like,'case','fever','high'.
* The eighth topic seems to discuss the psychological impact of the pandemic, research on vaccines by scientists in the medical world, and COVID-19 cases in wild animals. We can see words like'vaccine','scientist', ‘medical','journal'.



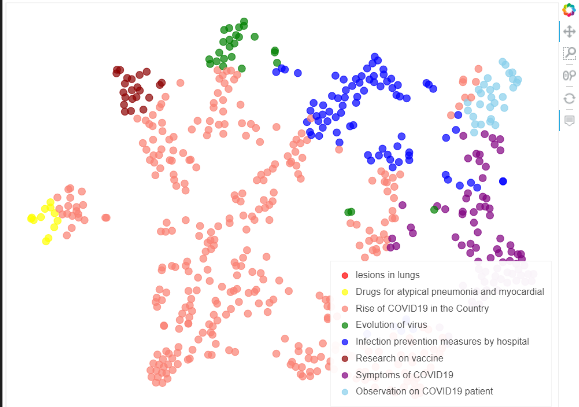
the interactive figure of Intertopic Distance Map (via multidimensional scaling) & Top-30 Most Relevant Terms for Topic 1

**Topic per Document:**

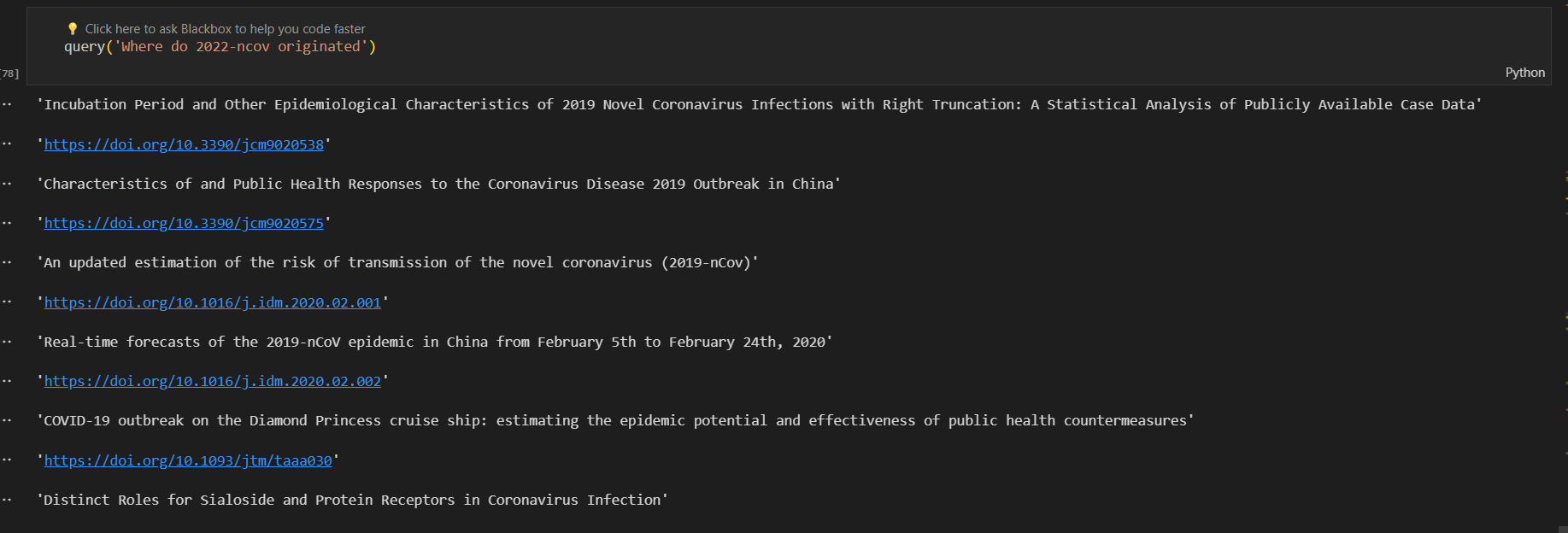
****

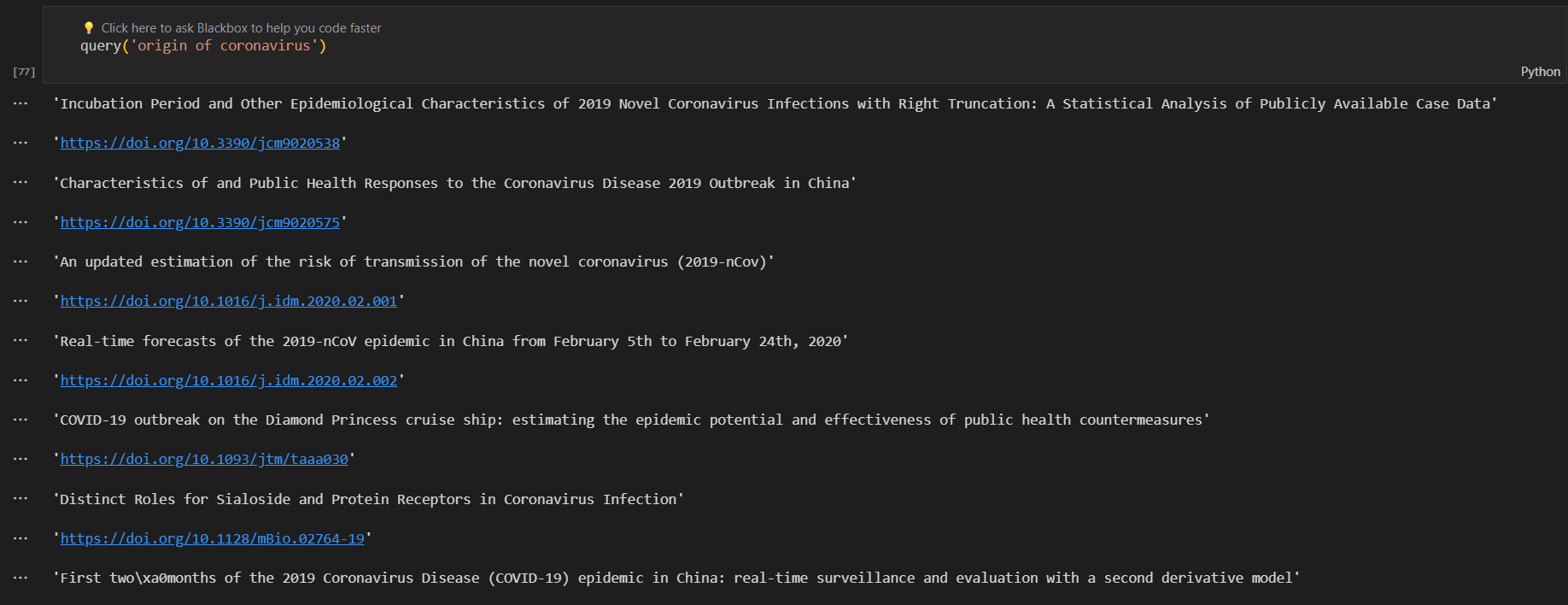
* Figure on the right hand side shows the fact that occurrences frequency of topics are not very close and they are far from each other since there are lots of article which covered the first and sixth topic.
* The left side of the figure tells us that only 1 or 2 of the articles had all the topics, which is very rare.3rd and 4th topic is generally pervasive in a lot of articles.

**T-SNE clustering**



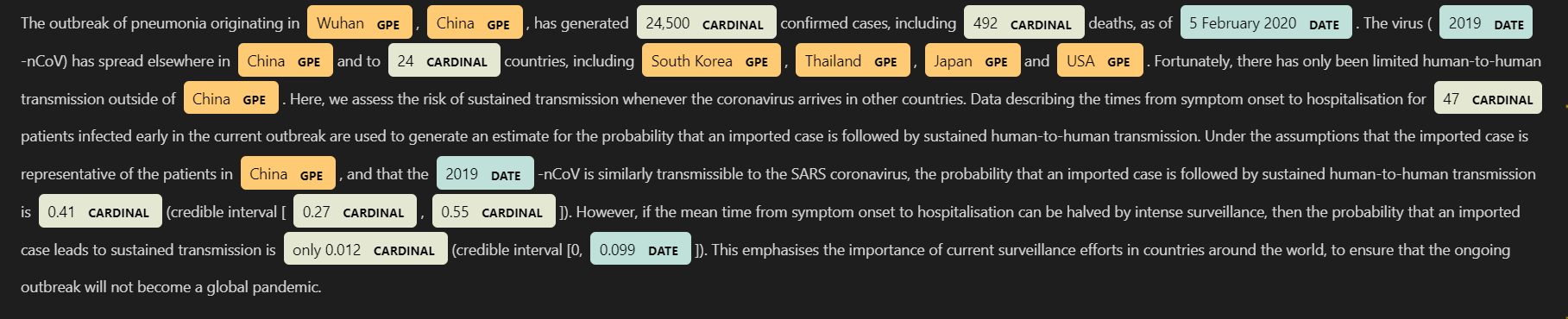
**Semantic-Based Search:**



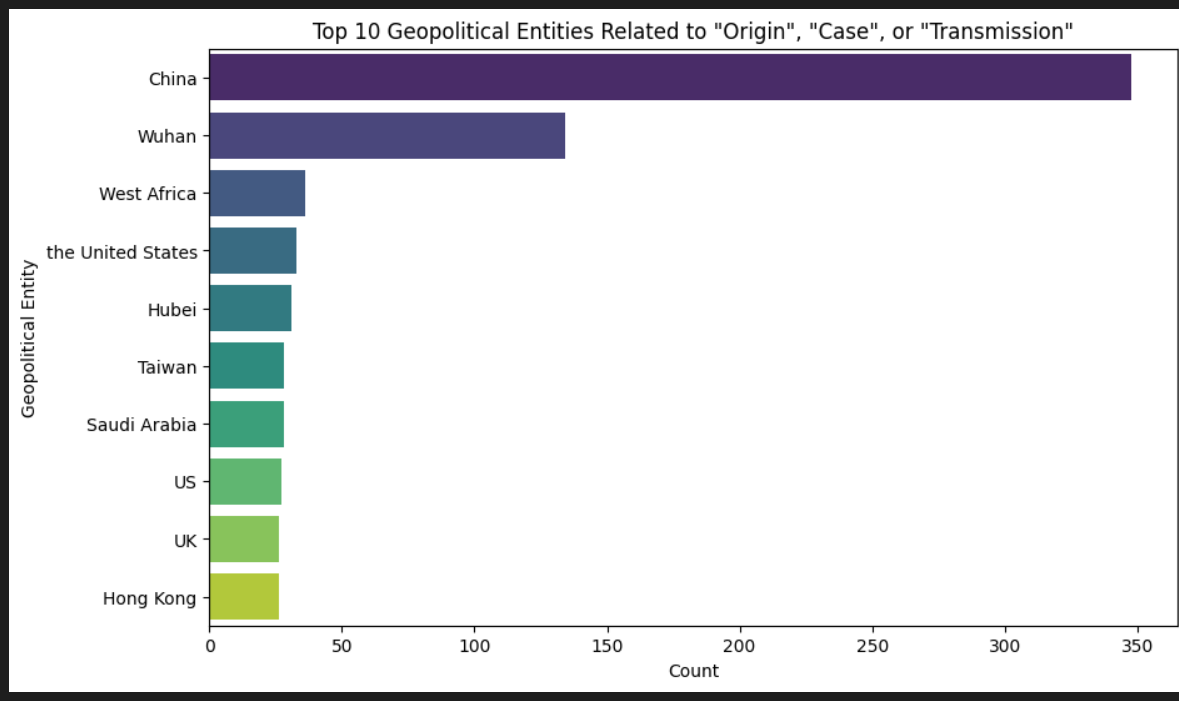


**MY NOTEBOOK 2: (COVID-19-CORD-NER- information-extraction- Question-Answering)**

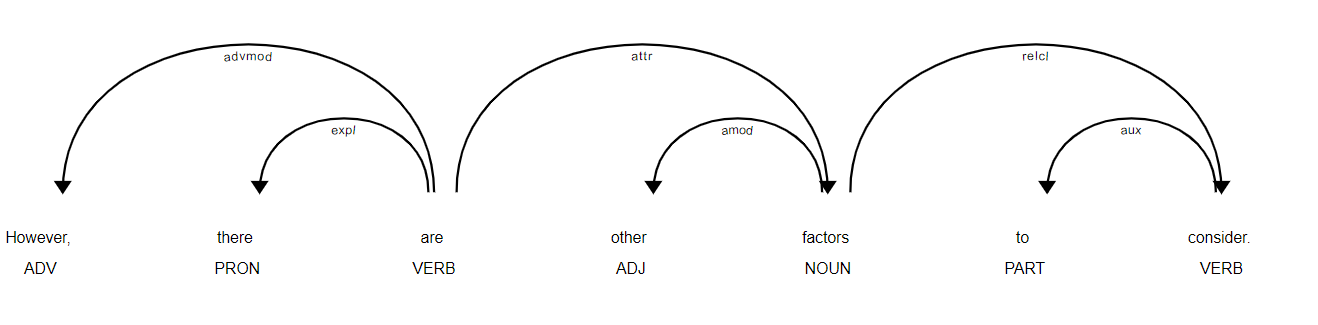
**NER Extraction:**

****

**Top 10 Geopolitical Entities Related to "Origin", "Case", or "Transmission :**

****

**Dependency parses:**

****

**Question-Answering:**



**CONCLUSION:**

By the help of LDA topics and T-SNE clustering, We can just as easily classify the proper articles to obtain the search result through a search engine with a semantic base, finding the most related links. Going back to the shop and picking the right kind of things. Besides, NER is beneficial for extracting data from pieces of writing; it can be used for many other purposes. Simultaneously, training on neural networks and ML models will be given. OpenAI's question-answering model serves a significant role in AI-driven virtual assistants. A journal will be made to process information retrieval using queries. This will help researchers to avoid the tiresome work of reading through many journals in search of a particular topic.