1: Decoding Covid-19 with Genome Analysis

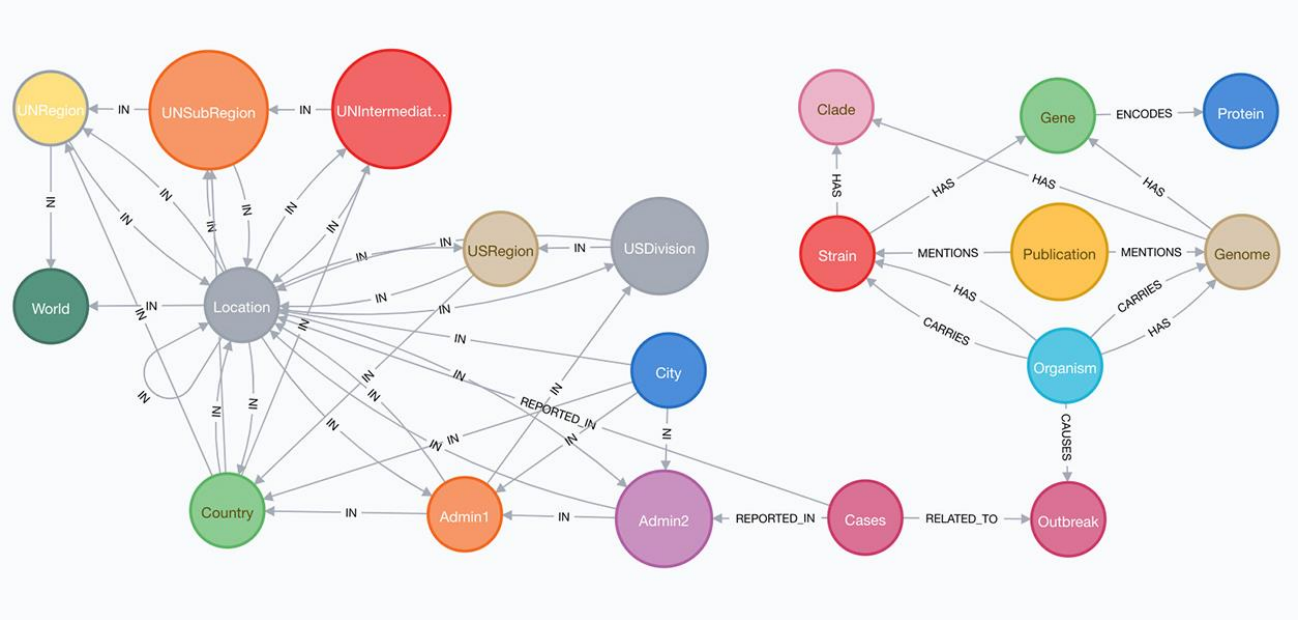
PROBLEM STATEMENT

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-ner

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.

WORK METHODLOGY

As CORD-NER-full dataset had 29500 rows, due to lack of computation power I had to work with only 10000 of it rows.

WORKFLOW OF MY NOTEBOOK:

1. Text Cleaning and Preprocessing
2. Article matching and Deep Cleaning
3. Topic Extraction (Latent Dirichlet Allocation)
4. T-SNE

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:

- Created keyword list to filter out required articled only.
- Filtered required articles and saved the index location in new dataframe.
- Used lemmatization and POS tagging technique to reduce ambiguity.
- Created word cloud to check frequently used words or phrases in the corpus.

Topic Extraction (Latent Dirichlet Allocation):

- Built LDA model to extract topics and coherence model to check the c_v coherence score to select appropriate topic number.
- Checked perplexity of my LDA Model.
- Extracted meaning full words for topics.
- LDA visualization.
- Checked topic per document with bar graph.

T-SNE:

- Generated document topic matrix.
- T-SNE clustering of LDA topics.

RESULTS

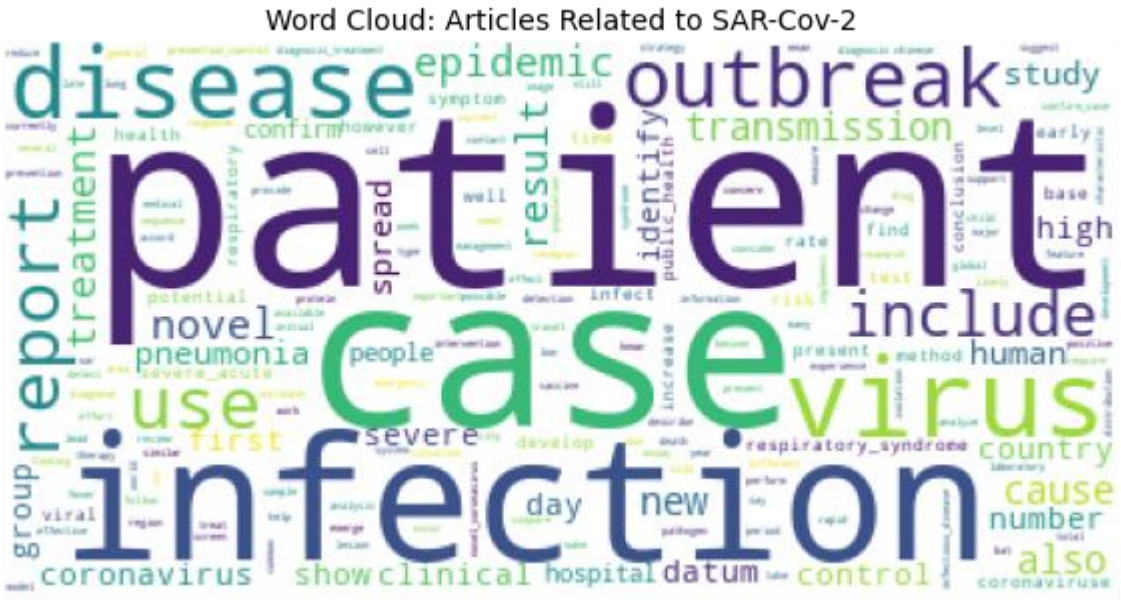
1. Filtered required article:

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

KEYWORD LIST –

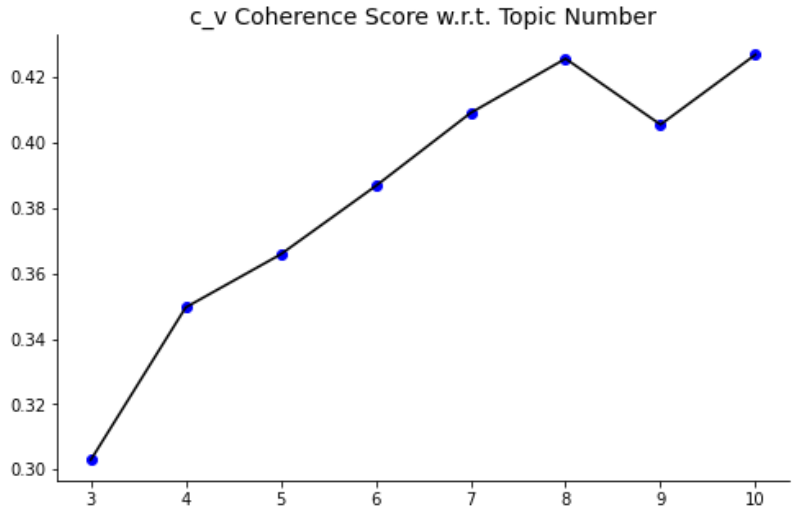
sars-cov-2, sars, cov-2, 2019-ncov, ncov, cov, covid19, covid, corona, coronavirus

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



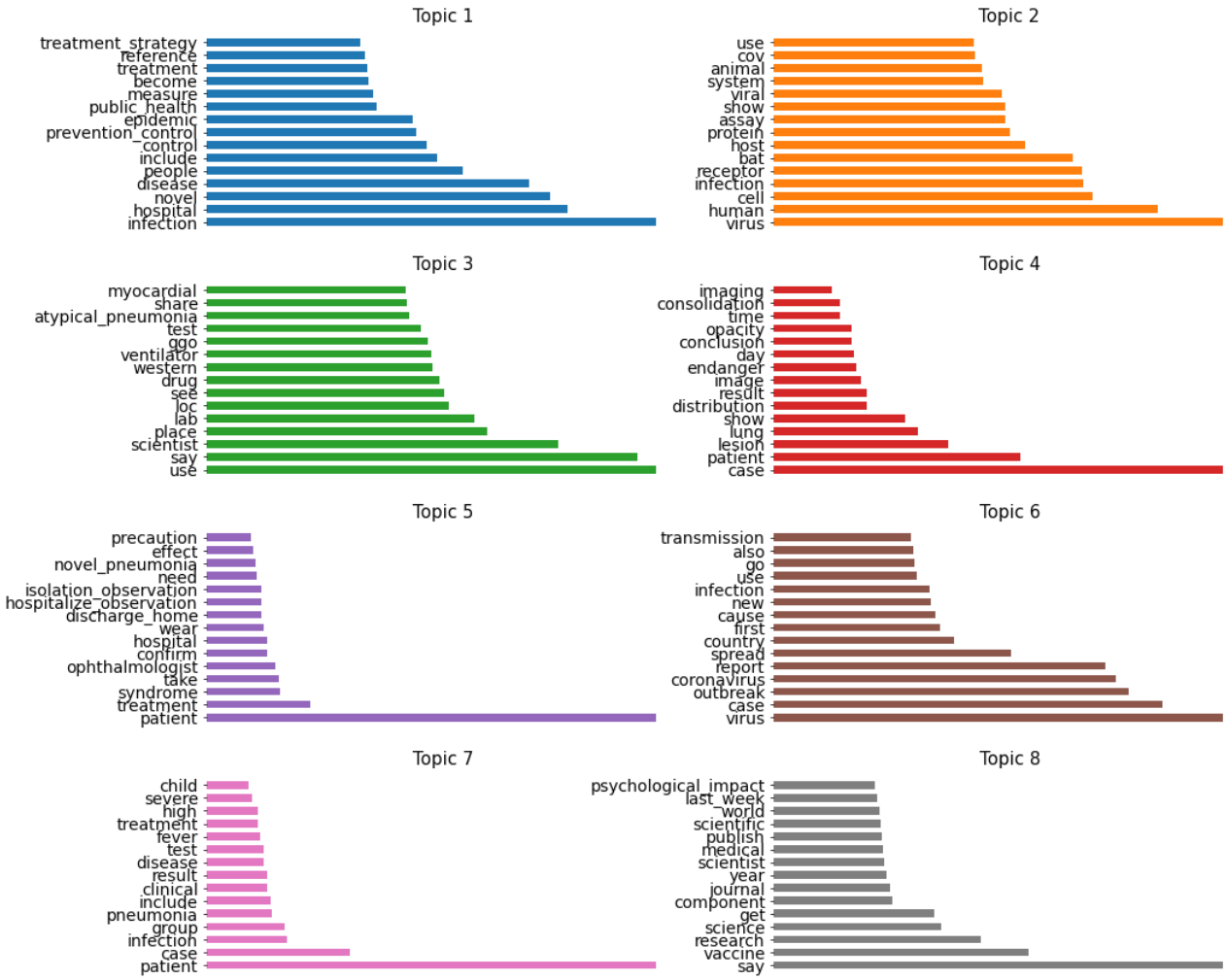
Words or phrases like 'patient', 'infection', 'disease', 'outbreak', 'virus', 'case', 'epidemic' occur frequently in this corpus.

3. c_v Coherence Score w.r.t. Topic Number:



The higher the c_v coherence score is, the more suitable the topic number should be. Hence, I choose 8 as the topic number for analysis.

4. Word per Topic:



1. First topic is talking about the hospital practices to ensure the public health by adopting public treatment strategies and infection prevention control measures. We can see words like 'infection', 'hospital', 'disease', 'prevention_control','public health','treatment strategies'.

2. Second topic is probably talking about evolution of virus inside human host. We can see words like 'virus', 'human', 'cell','host','protein','infection'.

3. Third topic talks about Scientists testing western drugs in lab for atypical pneumonia and myocardial. We can see words like 'Scientist', 'lab', 'test', 'ventilator', 'drug','western','atypical pneumonia','myocardial'.

4. Fourth topic probably talks about most of the patient cases orignated from distribution of lesions in lungs that were infected. We can see words like 'patient', 'case', 'lesion', 'infection', 'lung', 'distribution'.

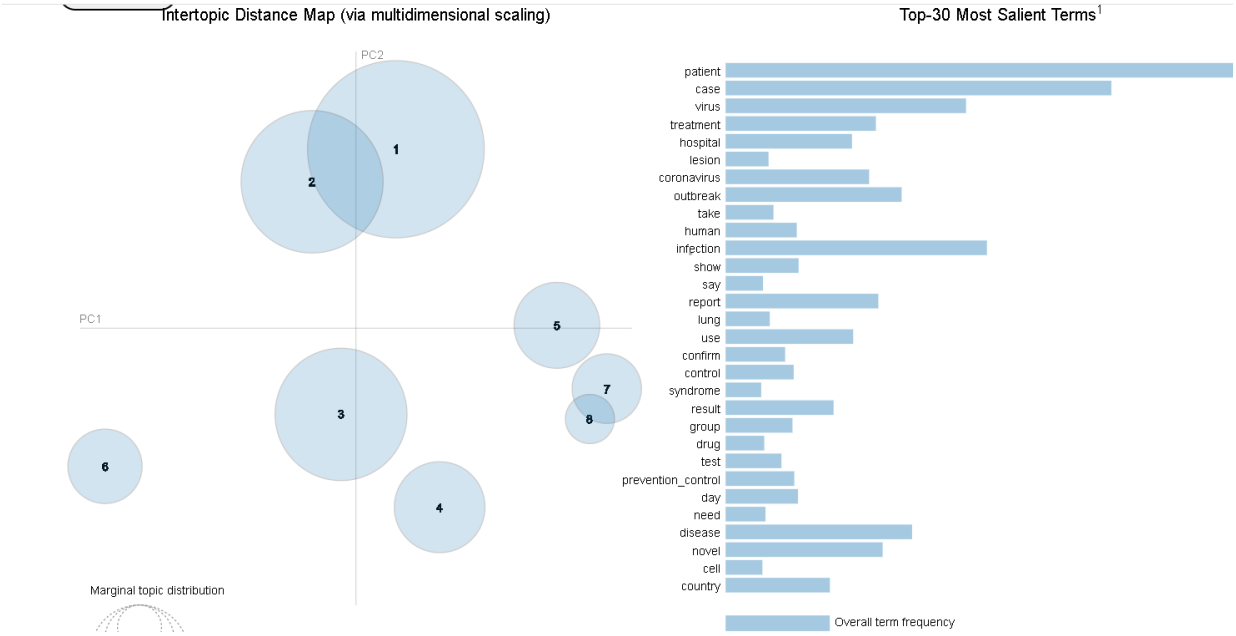
5. Fifth topic talks about observation on hospitalised COVID19 confirmed patients for syndrome, treatment, effect and precaution in isolation. We can see words like 'patient', 'treatment', 'syndrome', 'confirm', 'hospitalize_observation','effect','precaution'.

6. Sixth topic talks about rise in cases of COVID19 in the country. We can see words like 'country', 'case', 'transmission', 'outbreak', 'report', 'virus'.

7. Seventh topic probably talks about symptoms of COVID19 in a patient. We can see words like 'patient','case','fever','high'.

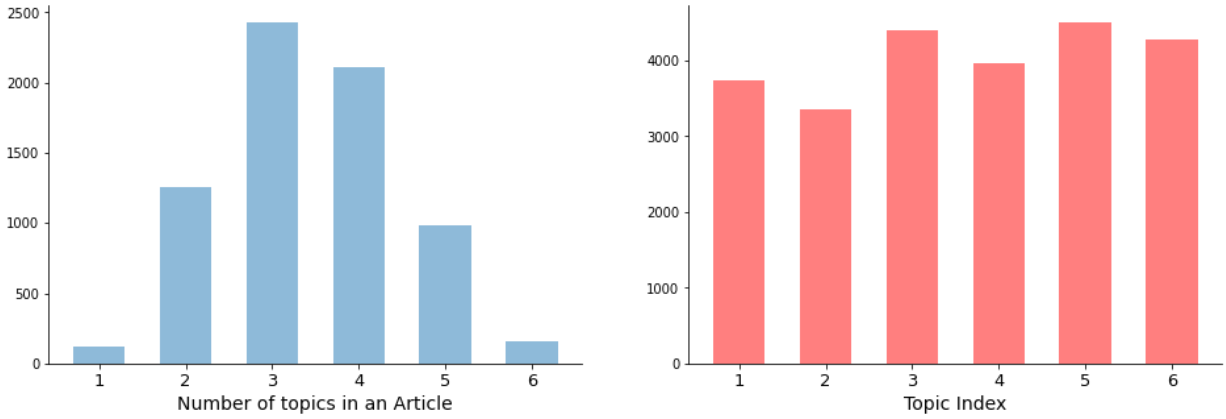
8. Eighth topic probably talks about psychological impact and research on vaccine by scientists in medical world and also about COVID19 cases in wild animals. We can see words like 'vaccine','scientist','research', medical','journal','wild_animal'.

5. LDA Visualization:



As we can see from the interactive figure above, topic 1 and 2 are very close as well as 7 and 8 (5 is also similar to 7). The other topics are separated appropriately.

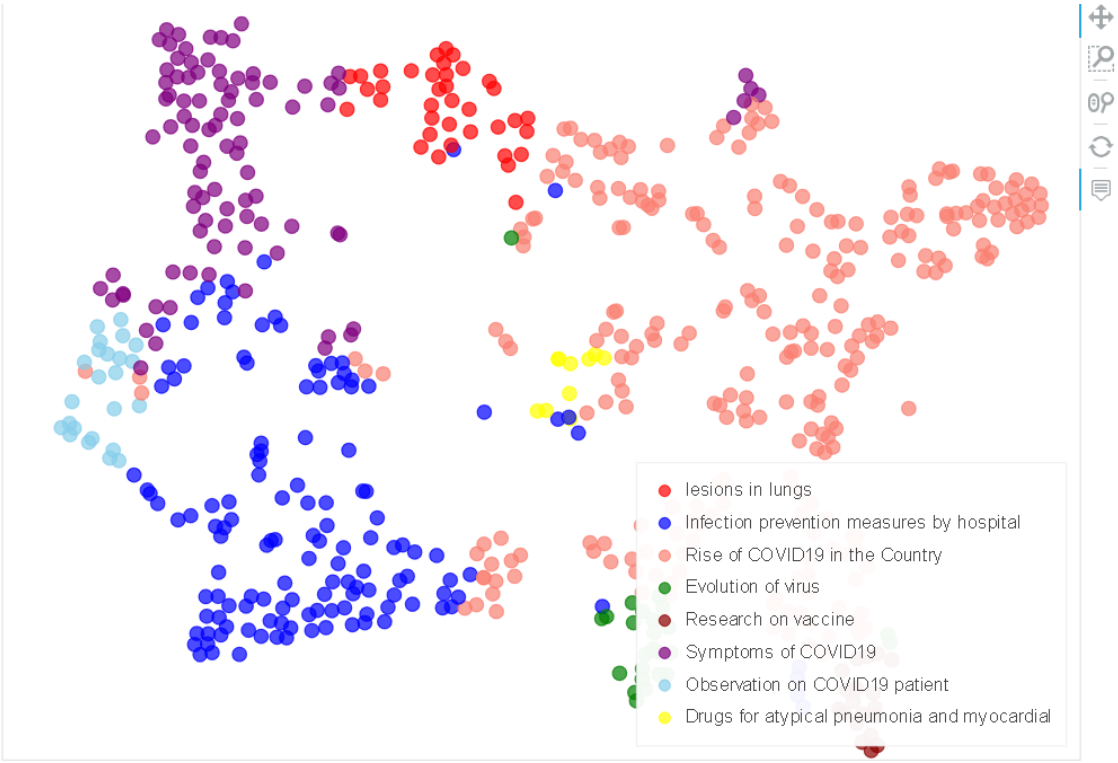
6. Topic per Document:

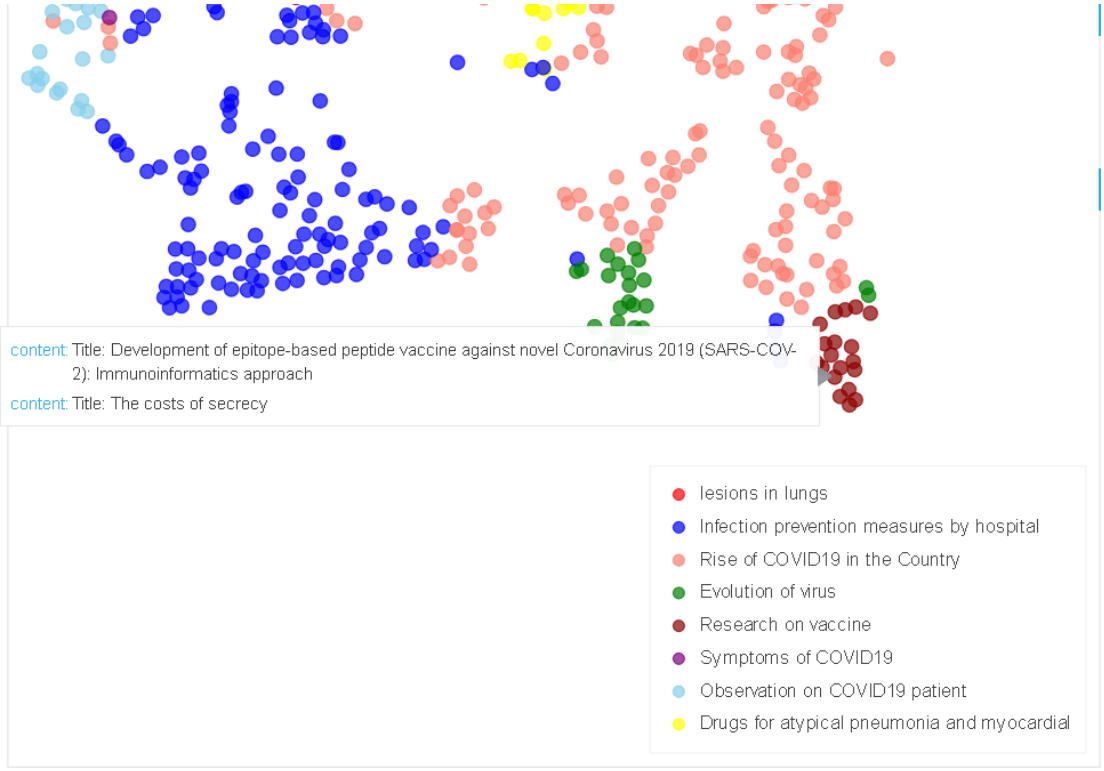


1.The left figure shows that there are very few articles who covered all the topics. 3rd and 4th topic are mostly present in many articles.

2.The right figure shows that occurrence frequencies of topics are not very close as we can see many articles covered 1st and 6th topic.

7. T-SNE clustering of LDA topics:





CONCLUSION

By the help of LDA topics and T-SNE clustering, we can easily identify the suitable article to get information from.
This will save time for researchers to go through many articles to search for required specific information.