1: Decoding Covid-19 with Genome Analysis

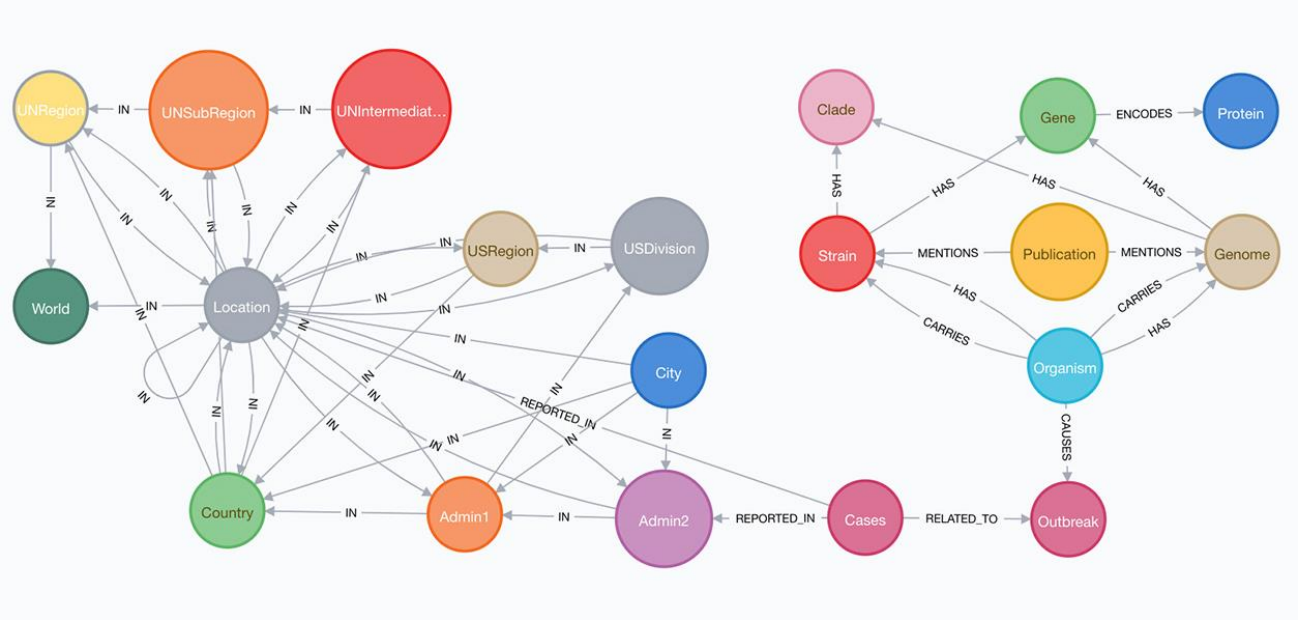
PROBLEM STATEMENT

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: (Covid19-CORD-NER)

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

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.

Semantic-Based Search:

- Duplicates and Null values.
- Dropping non-English articles.
- Spacy Parser and Tokenizer
- Sentence Tokenization
- Word2vec Training
- Ranking documents
- Saving the model and the dataframe

WORKFLOW OF MY NOTEBOOK 2: (COVID-19-CORD-NER-information-extraction-Q&A)

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.

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)

1. Filtered required article:

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

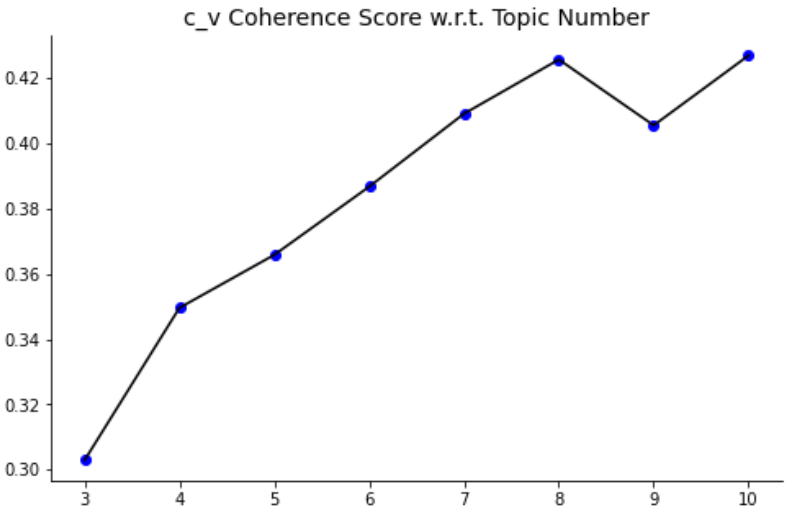
KEYWORD LIST –

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

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

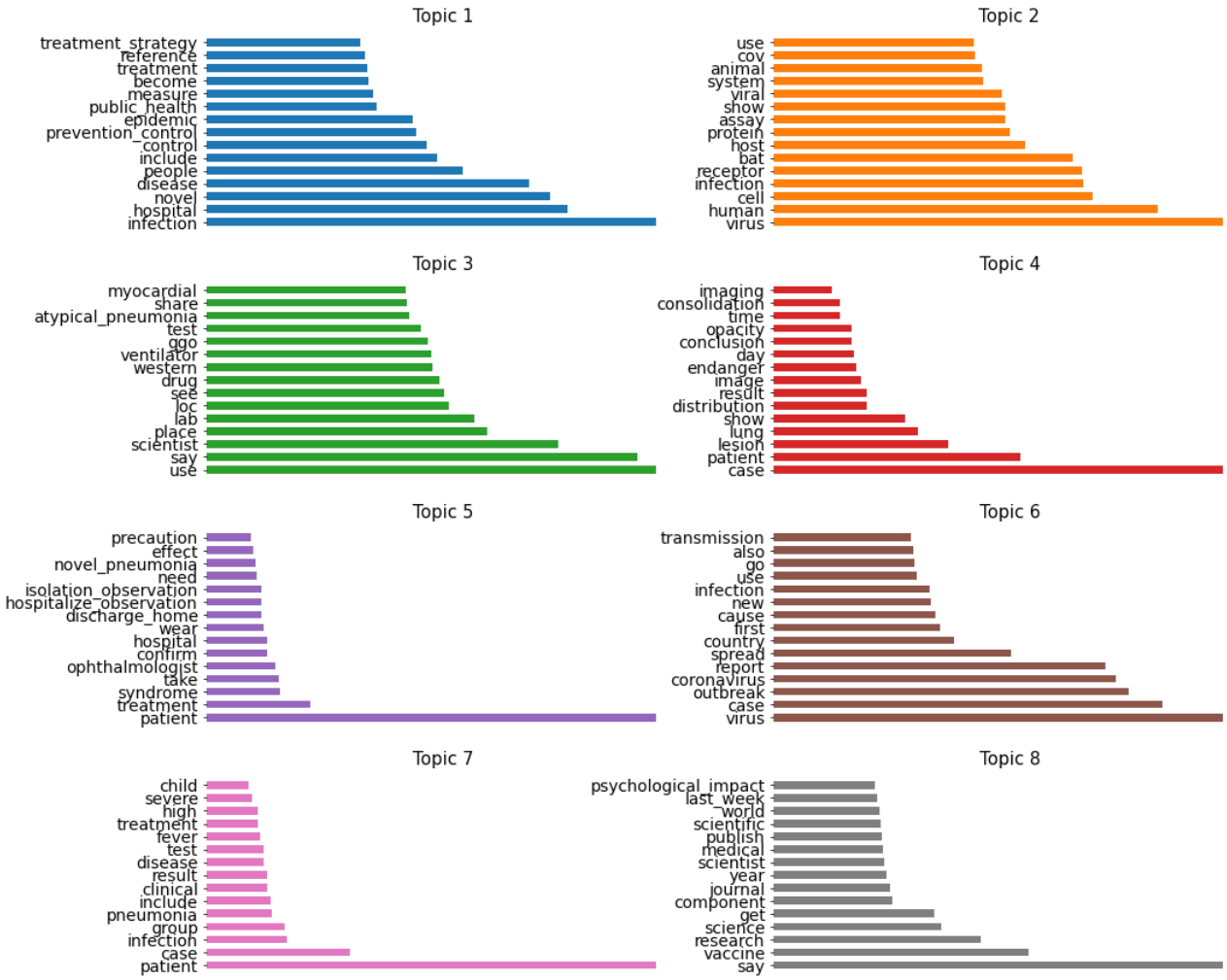
Words or phrases like 'patient', 'infection', 'disease', 'outbreak', 'virus', 'case', 'epidemic' occur frequently in this corpus.

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'.

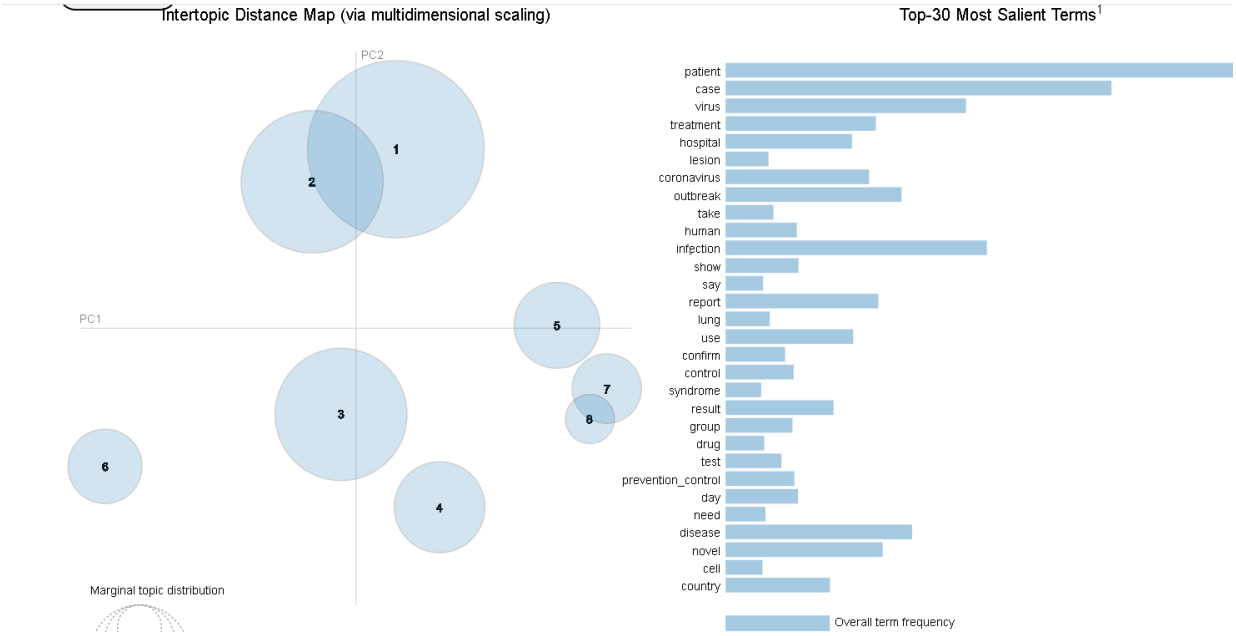
5. Fifth topic talks about observation on hospitalised COVID19 confirmed patients for syndrome, treatment, effect and precaution in isolation. We can see words like 'patient', 'treatment', 'syndrome', 'confirm', 'hospitalize_observation','effect','precaution'.

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'.

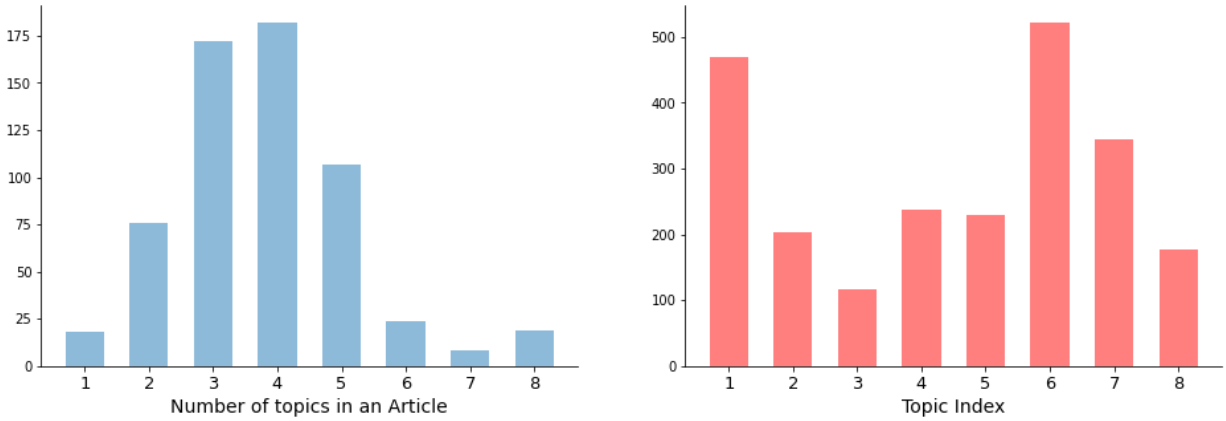
8. Eighth topic probably talks about psychological impact and research on vaccine by scientists in medical world and also about COVID19 cases in wild animals. We can see words like 'vaccine','scientist','research', medical','journal','wild_animal'.

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.

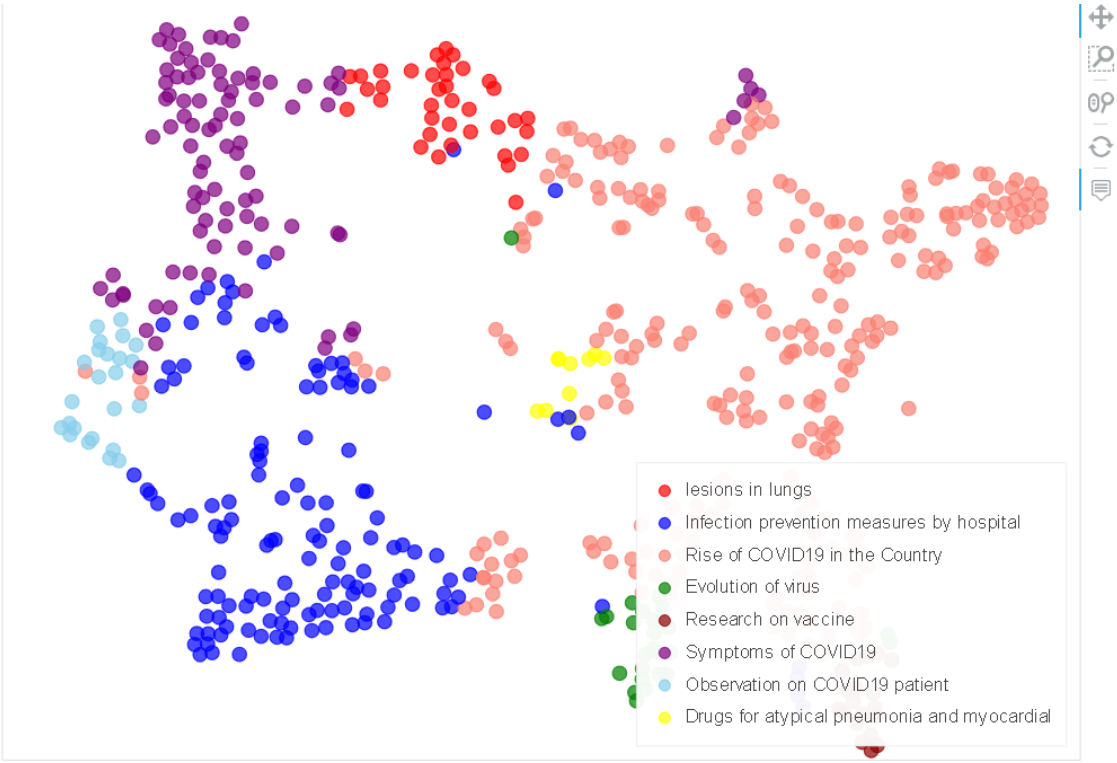
6. Topic per Document:

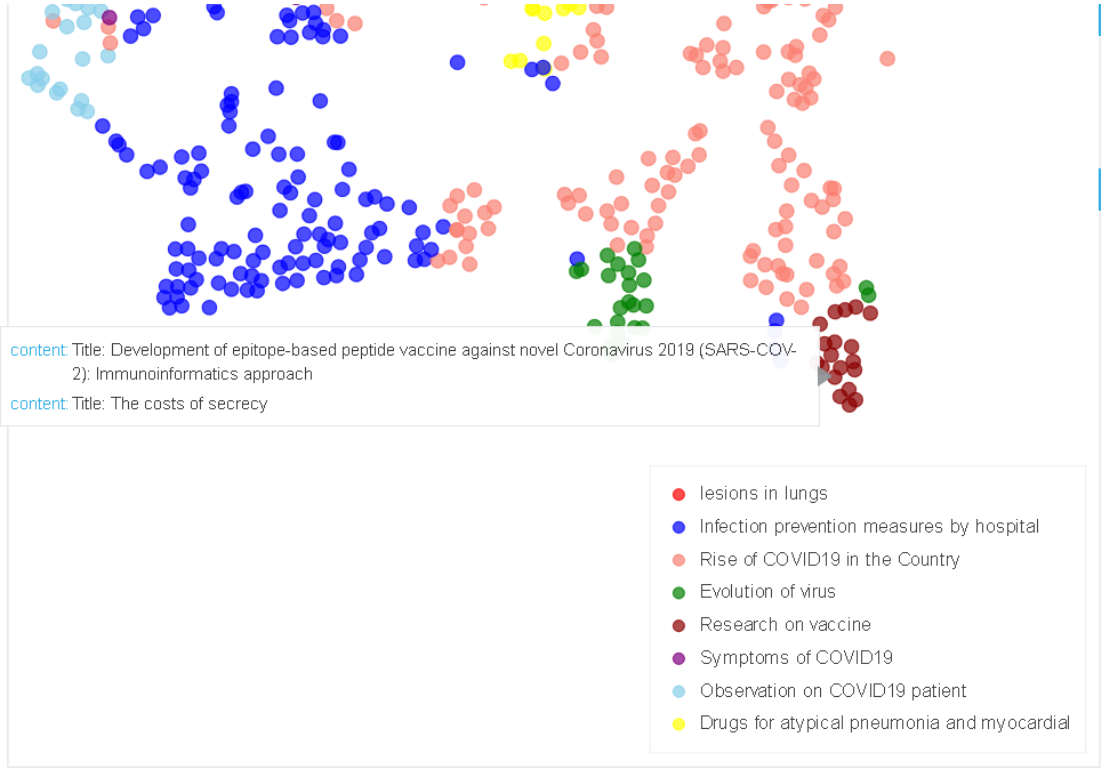


1.The left figure shows that there are very few articles who covered all the topics. 3rd and 4th topic are mostly present in many articles.

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:





8. Semantic-Based Search:

What do we know about virus origin, genetics and evolution?

```
query('origin of coronavirus')
```

```
'Case of the Index Patient Who Caused Tertiary Transmission of Coronavirus Disease 2019 in Korea: the Application of Lopinavi
tment of COVID-19 Pneumonia Monitored by Quantitative RT-PCR'

'https://doi.org/10.3346/jkms.2020.35.e79'
```

```
query('covid19 genetics', top_matches =9)
```

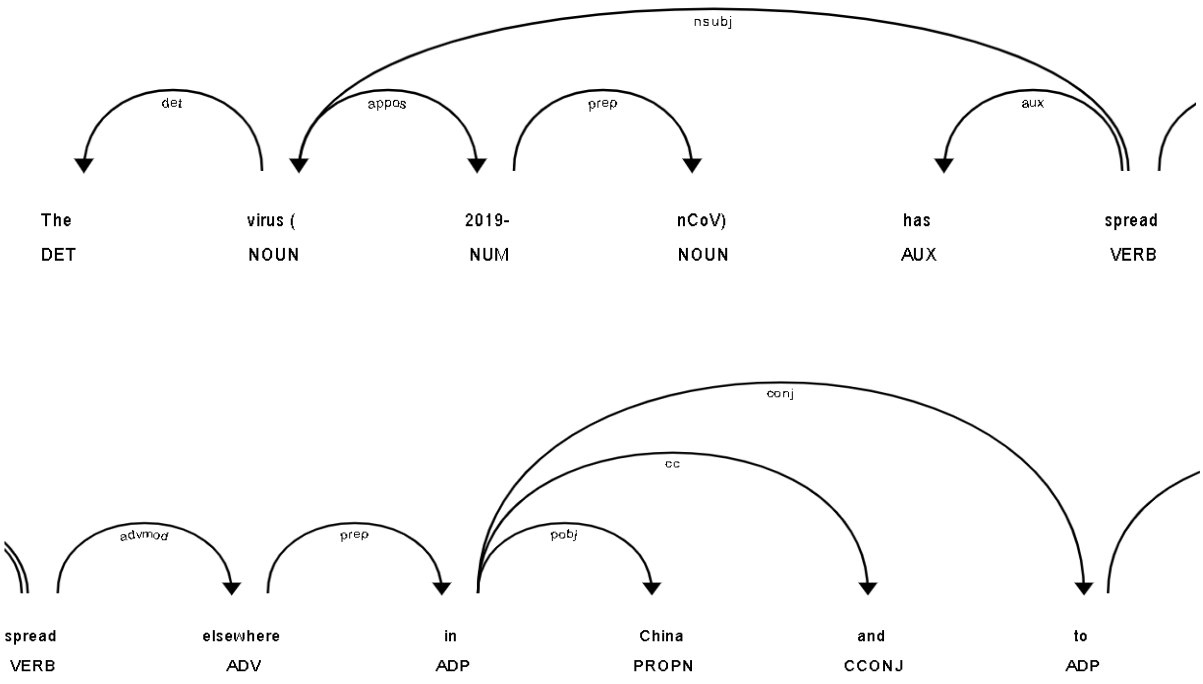
```
'Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes'

'https://doi.org/10.1093/bioinformatics/btaa145'
```

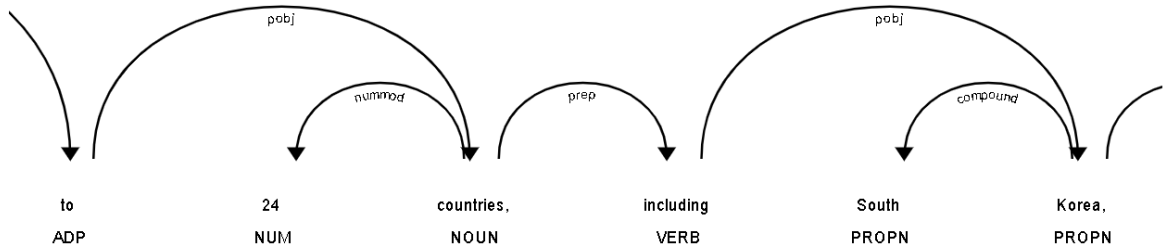
NOTEBOOK 2: (COVID-19-CORD-NER-information-extraction-Q&A)

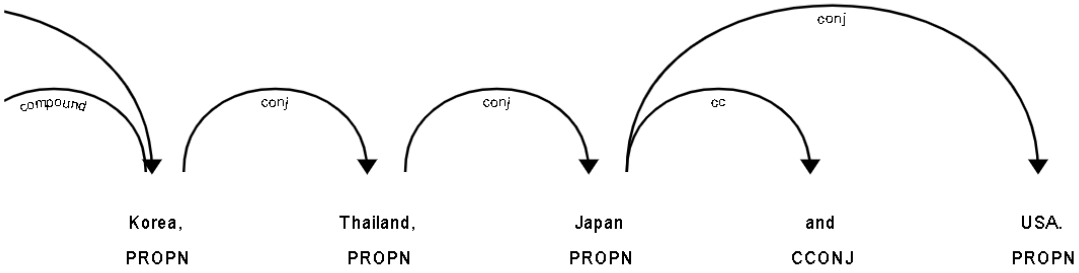
1. NER Extraction:

confirmed cases, including 492 CARDINAL deaths, as of 5 February 2020 DATE . The virus (2019-nCoV DATE) has spread elsewhere in China GPE and to 24 CARDINAL countries, including South Korea GPE , Thailand GPE , Japan GPE and USA GPE . Fortunately, there has only been limited human-to-human transmission outside of China GPE . Here, we assess the risk of sustained transmission whenever the coronavirus arrives in other countries. Data describing the times from symptom onset to hospitalisation for 47 CARDINAL patients infected early in the current outbreak are used to generate an estimate for the probability that an imported case is followed by sustained human-to-human transmission. Under the assumptions that the imported case is representative of the patients in China GPE , and that the 2019 DATE -nCoV is similarly transmissible to the SARS coronavirus, the probability that an imported case is followed by sustained human-to-human transmission is 0.41 CARDINAL (credible interval [0.27 CARDINAL , 0.55]). However, if the mean time from symptom onset to hospitalisation can be halved by intense surveillance, then the probability that an imported case leads to sustained transmission is only 0.012 CARDINAL (credible interval [0 CARDINAL , 0.099]). This emphasises the importance of current surveillance efforts in countries around the world, to ensure that the ongoing outbreak will not become a global pandemic.

2. Dependency parses:







3. Question-Answering:

[25]:
```
QA_input = {
    'question': "Where do 2019-ncov originated?",
    'context': doc.text
}
res = hugg(QA_input)

print(res)
```

```
{'score': 0.0001072749073500745, 'start': 184, 'end': 189, 'answer': 'China'}
```

[26]:
```
QA_input = {
    'question': "How 2019-ncov spread?",
    'context': doc.text
}
res = hugg(QA_input)

print(res)
```

```
{'score': 0.007484417874366045, 'start': 171, 'end': 209, 'answer': 'elsewhere in China and to 24 countries'}
```

[27]:
```
QA_input = {
    'question': "How corona virus evolved?",
    'context': doc.text
}
res = hugg(QA_input)

print(res)
```

```
{'score': 2.3563768536405405e-07, 'start': 844, 'end': 848, 'answer': 'SARS'}
```

[28]:
```
QA_input = {
    'question': "How dangerous is corona virus?",
    'context': doc.text
}
res = hugg(QA_input)

print(res)
```

```
{'score': 0.00035880590439774096, 'start': 374, 'end': 396, 'answer': 'sustained transmission'}
```

[29]:
```
QA_input = {
    'question': "Which city the first case originated from?",
    'context': doc.text
}
res = hugg(QA_input)

print(res)
```

```
{'score': 7.412996637867764e-05, 'start': 41, 'end': 46, 'answer': 'Wuhan'}
```

[30]:
```
QA_input = {
    'question': "How to prevent from corona virus?",
    'context': doc.text
}
res = hugg(QA_input)

print(res)
```

```
{'score': 0.00011674649431370199, 'start': 1076, 'end': 1096, 'answer': 'intense surveillance'}
```

CONCLUSION

By the help of LDA topics and T-SNE clustering, we can easily classify the suitable articles to get information from and Semantic based search engine with help get the links of most appropriate articles to go for. NER is also helpful to extract information from articles and can also be trained in neural networks and ML models. Question answering model used in the notebook will help extracting various information through queries.
This will save time for researchers to go through many articles to search for required specific information.