Objective & Approach

Student: Yuchao Wu (1000651984)

The objective of the study is to learn about the different ways of transmissions of CONVID-19 in the papers. By clustering ways of transmissions, hopefully targeted and corresponding actions can be taken by our healthcare system, government and industries to prevent the spread and fight against CONVID-19.

1 Data Cleaning

- **Strip HTML tag**
- **Remove punctuation**
- **Fix contraction**
- **Remove special characters**
- **Remove stopwords**

2 Exploratory Data Analysis

- Define the bag words related to ways of transmissions
- Descriptive analysis: countplots

3 Modelling

- Top N grams to expand the bag of words
- K-means clustering
- Hierarchical clustering

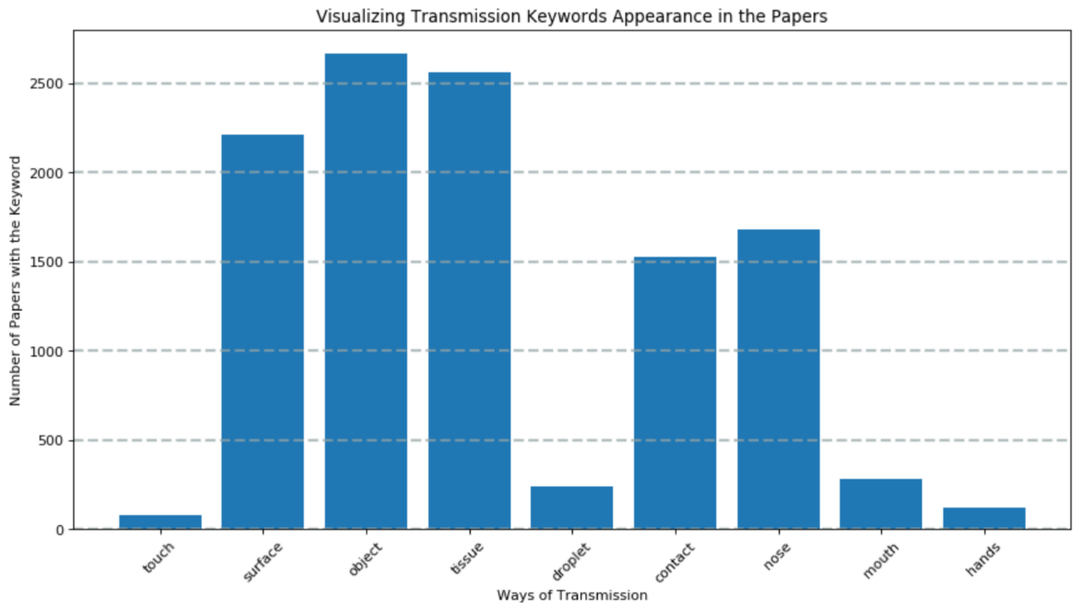
4 Insights on Policy & Guidance

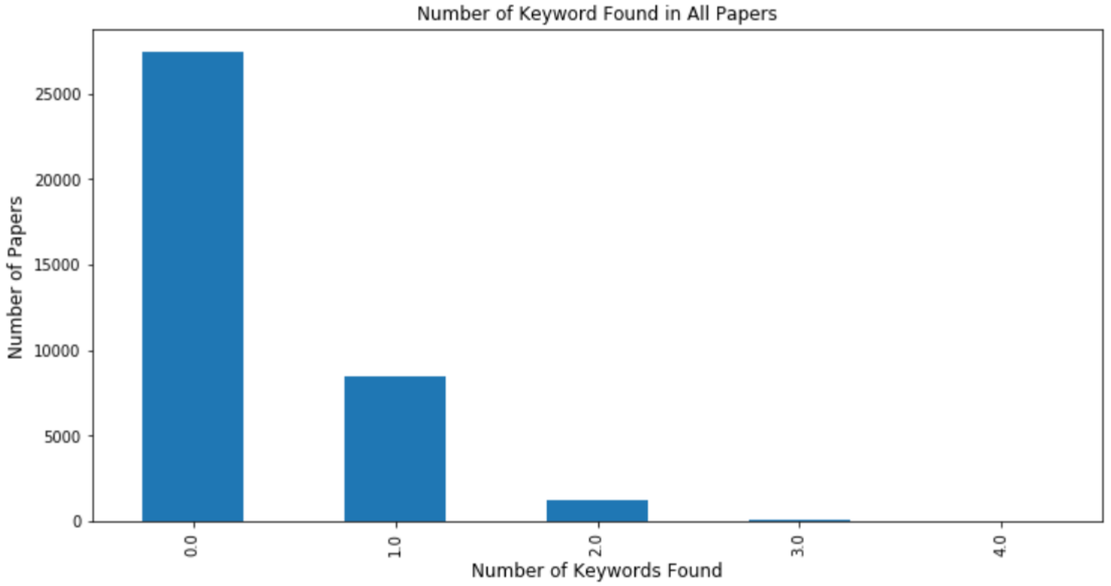
- Insights on Policy and Guidance based on clustering results

Exploratory Data Analysis

Bag of Keywords Related to Ways of Transmission

['touch', 'surface', 'object', 'tissue', 'droplet', 'contact', 'nose', 'mouth', 'hands']





- Out of the proposed 9 keywords, 3 keywords have appeared for over 2000 times in the corpus. Most of the keywords have got significant appearances in the papers, meaning that the study direction is also of interest to the researchers.

- Some of the keywords do appear in the same papers, which probably means that some keywords in nature are closer to each other and are possibly considered to be in the same bigger cluster of ways of transmissions.

Modelling – Fuzzy Logic to Find More Keywords

Since the bag of words defined through research online might not cover all ways of transmissions mentioned in the papers, a deep dive into the papers to find out if there are further keywords mentioned is necessary. The purpose of doing this is to cover more types of transmissions.

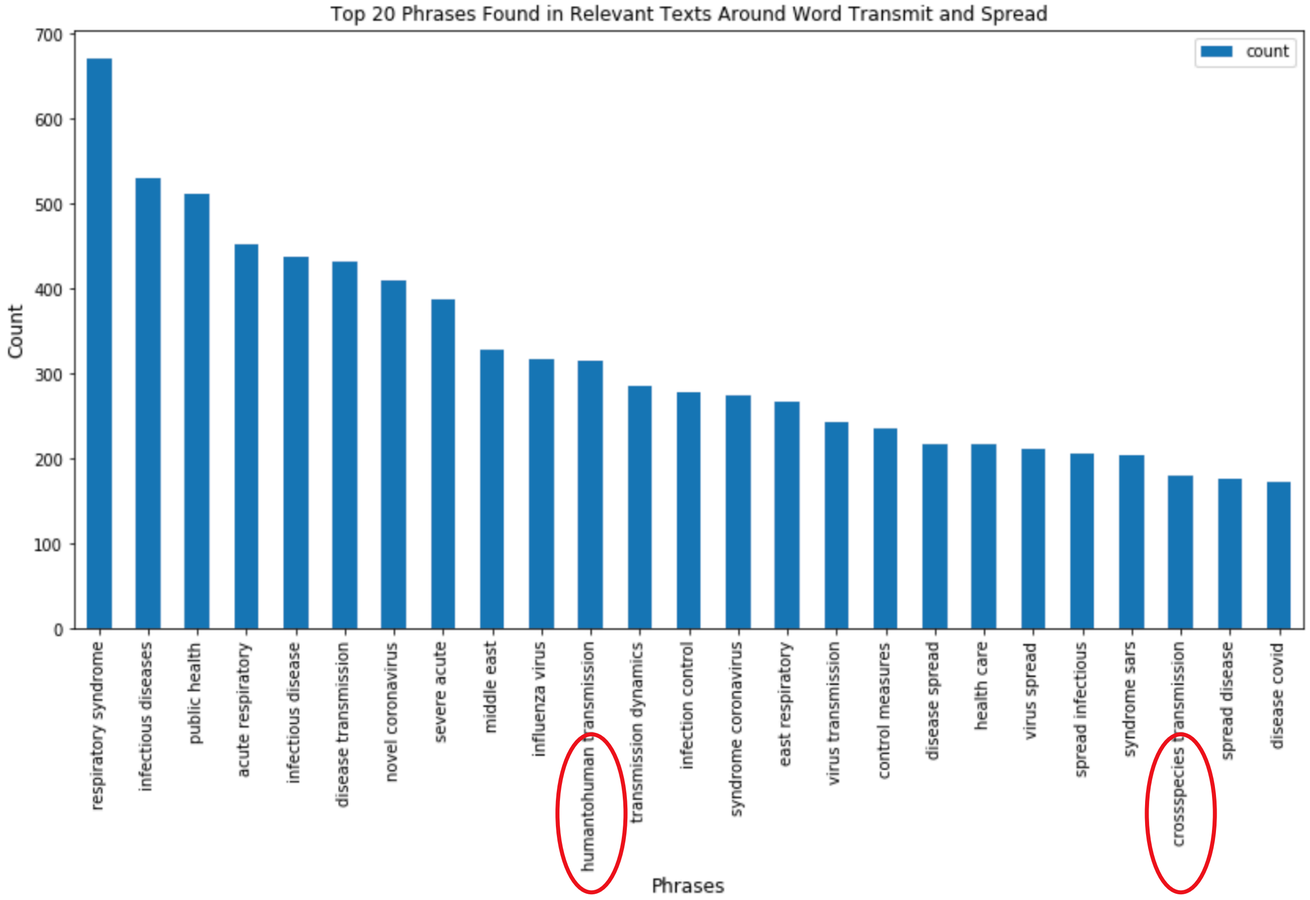
1. Naturally papers containing **'transmit', 'transmission' or 'spread'** should be relevant to the study of interest.

2. Applied **fuzzy logic** to filter for 15 words before and after these words might help find more types of transmission related keywords. e.g.

```
fuzz.ratio('transmission', 'transmissionsthis')
```

```
83
```

3. Digged deep into the top 30 phrases from the relevant texts, 4 new keywords are added to the bag of words: [‘humantohuman’, ‘crossspecies’, ‘airborne’, ‘nosocomial]



Modelling – Define Distance Matrix and Hierarchical Clustering

K-means Clustering

```
Cluster: 1
----------------------------
transmission dynamics
humantohuman transmission
crossspecies transmission
zoonotic transmission
viral spread
risk transmission
electron microscopy
transmission electron
united states
interspecies transmission
----------------------------
```

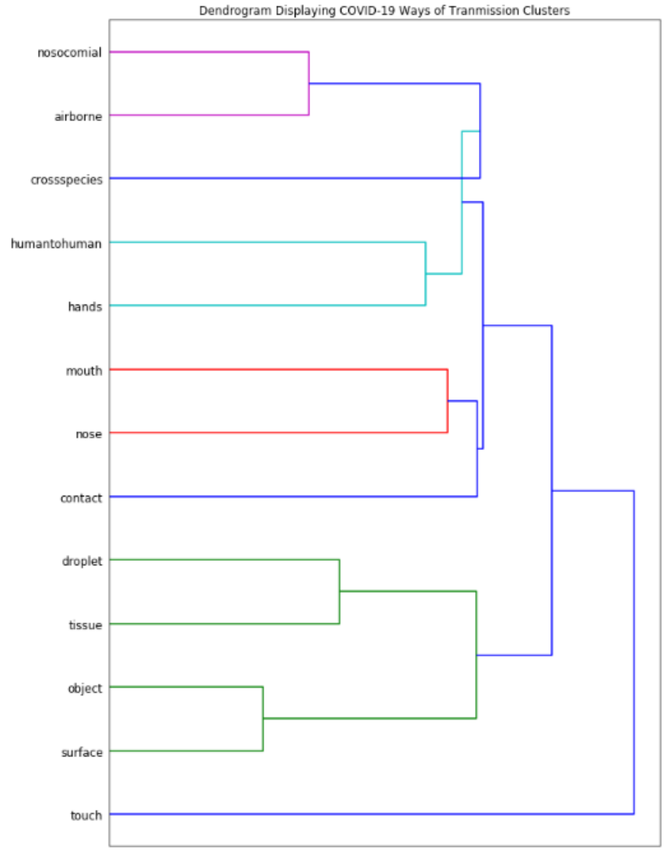
```
Cluster: 3
----------------------------
prevent spread
spread sars
prevent transmission
healthcare workers
transmission potential
saudi arabia
morbidity mortality
humantohuman transmission
risk factors
hand hygiene
----------------------------
```

```
Cluster: 2
----------------------------
hong kong
airborne transmission
spread sars
transmission sars
transmission routes
healthcare workers
close contact
present study
transmission viruses
sars epidemic
----------------------------
```

```
Cluster: 4
----------------------------
wuhan china
spread rapidly
confirmed cases
rapidly spread
hubei province
south korea
humantohuman transmission
reproductive number
province china
highly pathogenic
----------------------------
```

- K-means provides the insight of how texts and phrases related to transmission are clustered and grouped. This helps to understand what types of transmission (e.g., airborne, human to human) would cause people infected by Covid-19.

Dendrogram of Hierarchical Clustering



- The best number of clusters is relatively hard to tell from the dendrogram of the clustering results, probably due to the different word utilization habits of different authors and thus the limitation of the definition of the distance matrix. However, useful insights can be derived to help government, health system and other entities to fight against CONVID-19.

Insights on Policy & Guidance

Useful insights can be derived from the clustering of the ways of transmission of CONVID-19. In order not to provide too much information as well as not to lose too much information, 4 clusters of ways of transmission are discussed for recommendations on Policy and Guidance.

Transmission Media

Keywords: Droplet, Tissue, Object, Surface, Touch

- **Governments** need to organize regular routines to clean and disinfect public places, such as railway station, building lobbies and etc.
- **Industries** need to ensure the supply of hygiene products such as alcohol rubs, masks.
- **Companies** that run places such as shopping malls, grocery stores need to offer proper hygiene at the entrance and ensure regular routines to clean and disinfect as well.

Person to Person Contact

Keywords: Mouth, Nose, Contact, Humantohuman, Hands

- **Governments** need to ask people to perform social distancing and enforce people who had contacts with anyone infected to perform self-quanrantine.
- **Doctors** and **Nurses** need to wear masks, protective suits and gloves when treating infected patientes.
- **Scientists** can study how to reduce the risk of public health by temporarily closing down some businesses and reducing crowds in public places.

Cross species Transmission

Keywords: Crossspecies

- **Governments** need to enforce people not to eat wild animals because they might carry with them diseases.
- **Healthcare professionals** need to take control of possible risks posed by stray cats and dogs if they caught the virus from humans and could continue to pass it off to other humans.

nosocomial transmissions

Keywords: Nosocomial, Airborne

- **Governments** need to make sure that the hospitals and healthcare professionals have enough PPE.
- **Industries** should prioritize the needs from the healthcare system. And doctors and nurses need to performance good hygiene practices.