

Master of Technology in Intelligent Systems
Intelligent Reasoning Systems

Team Members

Ng Bo Yan (A0160005B)

Nur Insyirah Binte Mahzan (A0115982Y)

Low Pei Jing (A0131313B)

Apollo News

[HOME](#)[COUNTRY](#)[DISEASE](#)

Rising COVID-19 cases in Singapore driven by XBB subvariants; MOH says infection waves expected 'from time to time'

CLINICS SEEING MORE CASES Clinics are also seeing a spike in patient load but doctors said the jump in cases is expected, now that COVID-19 restrictions have been eased. Unihealth said last week that each of its clinics has been seeing about 15 to 20 COVID-19 patients per day, up by nearly twofold over the past month. "Most of the COVID-19 patients that we see in clinics continue to present with mild symptoms," said Dr Xie Huizhuang, the firm's medical director. Over at Phoenix Medical Group, its clinics have also reported twice as many patients in the past two weeks. Dr Chua Hshan Cher, the group's medical director, said most patients showed "routine respiratory symptoms".

[Original Article](#)

None

First H3N8 bird flu death recorded in China

GENEVA: A woman has died from H3N8 bird flu in China, the World Health Organization reported on Tuesday (Apr 11) - the first known human fatality from the avian influenza strain. H3N8 is known to have been circulating since 2002 after first emerging in North American waterfowl. It had not been detected in humans before two prior non-fatal cases emerged - both also in China - in April and May last year. The woman who died was a 56-year-old from Guangdong province in southeast China. She fell ill on Feb 22, was hospitalised for severe pneumonia on Mar 3 and died on Mar 16, the WHO said.

[Original Article](#)

None

Wildfire in South Korea forces 500 residents to evacuate as rain helps fight flames

SEOUL - More than 500 people evacuated their homes in South Korea's eastern coastal city of Gangneung as strong winds and dry weather fanned a wildfire on Tuesday, officials said, but fears of a further spread eased as rain helped firefighters battle the blaze. The fire injured three people, including two firefighters, and destroyed dozens of buildings, the national forestry agency said. Firefighting crews had struggled to put out the fast-moving blaze due to strong winds, but rain then tamped it down by the afternoon. The fire appears to have started after strong winds blew a tree onto live overhead power cables, igniting flames, Gangwon province's Governor Kim Jin-tae said. South Korean President Yoon Suk-yeol ordered officials to mobilise all available resources to put out the fire as soon as possible and quickly evacuate nearby residents to minimise casualties, his office said.

[Original Article](#)

2023-04-11 12:47:45+08:00

Project Report

Alerts on Potential Pandemic Outbreaks in Singapore

Summary

Singapore is a small and open country with a high population density, at 7,688 persons per sq km in 2022, according to the Department of Statistics Singapore (Department of Statistics, 2023). In addition, before the implementation of border restrictions due to COVID-19, Singapore had cleared 217 million travellers through its checkpoints in 2019 (Immigration and Checkpoints Authority, 2020). This makes Singapore highly susceptible to disease transmissions whenever a global pandemic outbreak occurs.

Studies have indicated that social distancing measures and border restrictions are effective intervention methods to prevent the spread of infectious disease, such as COVID-19 (Haug et al., 2020). Considering the adverse effects of infectious diseases to a nation's economy (Gallup, 2020) and quality of life (Mouratidis, 2021), it is critical for Singapore to ensure that there are sufficient infection control and prevention measures to protect the safety and well-being of the nation.

While there are platforms, such as HealthMap (Clark et al. 2008), that monitor the outbreak of diseases via press releases globally, they are usually not catered specifically to Singapore's context. Thus, relying on such platforms usually requires a lot of manpower to look through all news articles to sieve out information that are relevant to Singapore.

As such, with reference to the adverse effects of the recent COVID-19 pandemic and the previous SARS outbreak on Singapore, there is a necessity to develop an intelligent system to monitor high risk disease outbreaks in countries that are likely to impact Singapore significantly. In addition, similar functions of this intelligent system may also be extended for use by other countries, or extended to other topics that concern Singapore's national security.

Summary	3
Section 1. Background and Objectives	5
1.1 Problem Description	5
1.2 Project objective	5
Section 2. Knowledge Model	6
2.1 Data Sources	6
2.2 Data Identification	7
2.3 Feature Extraction	8
2.4 Ranking/Score	16
2.5 Result Review	21
Section 3. System Design	22
3.1 Article Acquisition Layer	22
3.2 Feature Extraction Layer	22
3.3 Post Processing Layer	23
3.4 Scoring Layer	24
3.5 Display Layer	24
Section 4. Assumptions and Limitations	26
4.1 Assumptions	26
4.2 Limitations and Improvements	26
Section 5. References	28
Appendix A: Interview Transcript	30
Appendix B: Project Proposal	31
Appendix C: Mapped System Functionalities	34
Appendix D: Installation and User Guide	35
1 Download Source Code	35
2 Install Dependencies	35
3 Running Backend (Article & Feature Extraction)	35
4 Starting the Web Application	36
5 Repopulating Knowledge Base and Retraining Model	36
Appendix E: Individual Project Report	37
Ng Bo Yan	37
Nur Insyirah Binte Mahzan	38
Low Pei Jing	39

Section 1. Background and Objectives

1.1 Problem Description

As of today, there are no openly available public health monitoring systems that cater to Singapore specifically. Thus, the current practice of sieving news articles that are disease-related and relevant to Singapore relies on manually reading all news articles in various mainstream news platforms on a daily basis. However, this way of monitoring and picking relevant news articles is both time-consuming and labour-intensive.

Thus, our team proposes to build a semi-automated system that narrows down disease related news articles that are relevant to Singapore. The system should develop the function to monitor and alert users of potential pandemic outbreaks affecting Singapore.

1.2 Project objective

Our objective is to build a system that is able to recommend the most relevant news articles to the users. This will be done based on the relevance score of the article.

To identify key factors to consider in determining relevance of an article, our team held an interview with a Subject Matter Expert (SME), who has the experience of performing tasks such as monitoring and tagging new articles of relevance to Singapore. The SME had provided insights on some of their considerations when identifying important news articles in Singapore's context.

Section 2. Knowledge Model

2.1 Data Sources

What are our data sources?

S/N	Source of Information	Insights from information source	Knowledge acquisition technique
1	News Articles	Trending and up-to-date information about events around the world	<p>Web scraping of news articles via GoogleNews and Newspaper3k python packages.</p> <p>A total of 752 unique news articles from 1 Jan 2023 to 23 Mar 2023 were scrapped. 380 and 372 news articles were manually labelled as disease related and non-disease related news articles accordingly.</p>
2	Subject Matter Expert (SME)	Factors contributing to high relevance to Singapore	Elicitation of tacit knowledge through the conduct of interview
3	Disease Knowledge Base	Variation of disease names, and whether disease is infectious/contagious	Wikidata query service to extract all known available diseases from the free and open crowd-sourced knowledge base.

4	Visitor Arrivals Data	Number of visitor arrivals to ASEAN countries from other countries in 2019 - the latest visitor arrival data that is not affected by travel restrictions and border closures due to COVID-19	Downloaded publicly available data in excel format from ASEANstatsDataPortal
---	-----------------------	--	--

2.2 Data Identification

How do we identify relevant data/knowledge?

Through our interview with the SME, we identified key data and insights that we would need to extract from the unstructured data in news articles. We concluded that the problem can be split into three main parts:

2.2.1 Selection of disease outbreak related articles

Mainstream news sites publish news of different categories everyday. It is critical to identify disease related news articles to scope the number of new articles read daily.

Type of Disease

As compared to diseases, such as chronic diseases or genetic diseases, infectious diseases are considered more critical to Singapore's border health security.

2.2.2 Determine the significance of the disease outbreak based on the country of outbreak

Country of disease outbreak can be evaluated in two different ways:

Geographical location of the country

Countries that are geographically closer to Singapore are inferred to be more relevant to Singapore.

Number of inbound travellers to Singapore from the country

Countries with a greater number of travellers arriving into Singapore are also considered to be more important to Singapore.

2.2.3 Determine the significance of the disease outbreak based other factors

Besides location of the disease outbreaks, other factors are also taken into consideration.

These factors include:

Actions taken by local government

Actions taken indicate that the disease outbreak situation in the country is severe.

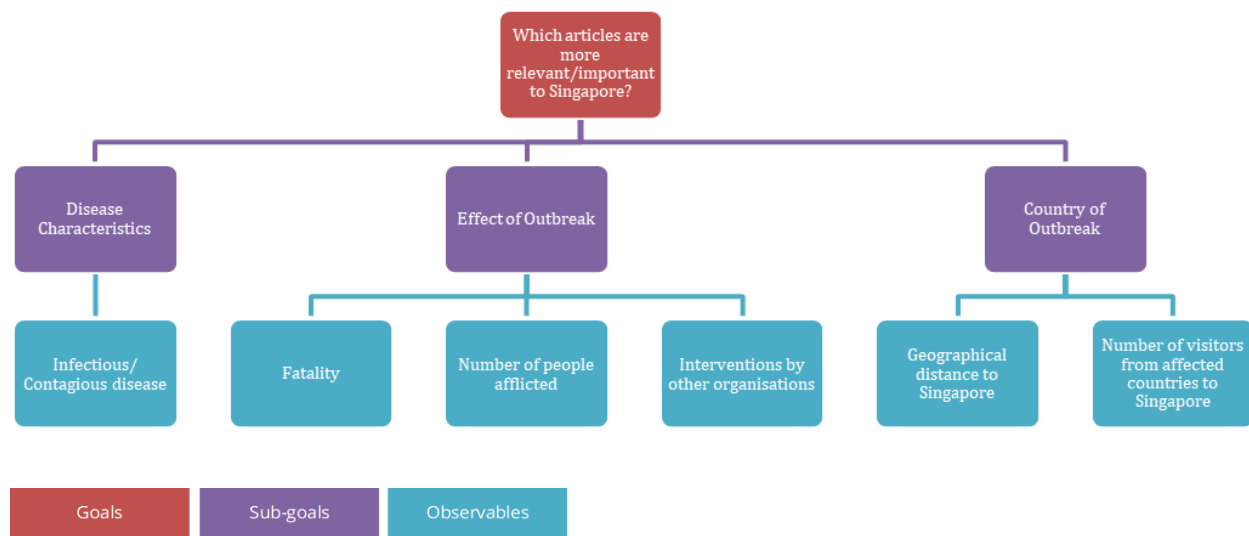
Number of people affected

The higher the number of people affected by the disease, potentially the more contagious the disease.

Number of fatalities

The number of fatalities indicates how deadly the disease is.

The key insights are summarised in the diagram below.



2.3 Feature Extraction

How do we extract the relevant data?

According to a study about the functions of headlines (Scacco & Muddiman, 2015), news headlines serve as a summary of the news articles. Therefore, just the title of the article is sufficient to identify whether the article is about a particular disease outbreak, and where the event occurred. Our SME also confirmed that headlines are generally useful to provide first cut information on the relevance of the articles. As such, only news headlines were used for feature extraction most of the time.

2.3.1 Disease Characteristics

For the scope of this project, we are more concerned about diseases that are infectious and contagious. As some articles do not explicitly indicate whether the disease is infectious/contagious, our team decided to extract the disease names mentioned in the article headlines to infer whether the disease in the article is infectious or not by comparing the extracted disease names to a pre-existing knowledge base of diseases.

Disease Name

To extract the disease name, Named Entity Recognition (NER) technique was used. In particular, we focused on statistics-based NER instead of rules-based NER as we don't have domain or language resources in that area to come up with a set of lexical/linguistic rules to identify disease names.

We manually labelled the 752 article headlines and explored several models to extract the disease name from the unstructured text. This includes:

- custom-built NER model by training the base spacy model using the labelled dataset
- leveraging on a pre-trained scispacy model "en_ner_bc5cdr_md" which was trained on the BC5CDR corpus that consists of 1500 PubMed articles with 4409 annotated chemicals, 5818 diseases, and 3116 chemical-disease interactions (Li et al., 2016).

We split the 752 article headlines into train and test sets. The models were trained/fine-tuned using the train set. We then evaluated the results of the trained model on the test set and chose the best performing model, which was the fine-tuned scispacy model.

Infectious/contagious type

We created a knowledge base of whether a disease is infectious/contagious or not by querying all known diseases from the free and open crowd-sourced WikiData knowledge base (Wikidata, 2023). We generated an indicator of whether the diseases are infectious (indicated as 1) or not (indicated as 0). We stored the list of diseases, their aliases and the infectious indicator in our sqlite database.

When disease names are extracted from the news headlines, we will then query the database for similar diseases. If the disease extracted from the article cannot be matched to any disease in the pre-existing knowledge base, we will assign an infectious score of 0.5 to account for potential new diseases. If there are more than one disease returned, we will take the highest value of the infectious indicator.

2.3.2 Effect of Outbreak

Number of Fatality & Number of Cases

Information extraction technique is utilised to extract data on the number of fatalities and cases from the unstructured text. The methodology focused on the linguistic patterns of how the information is presented in a sentence. The first step involves detecting sentences that mention cases and fatalities using specific keywords.

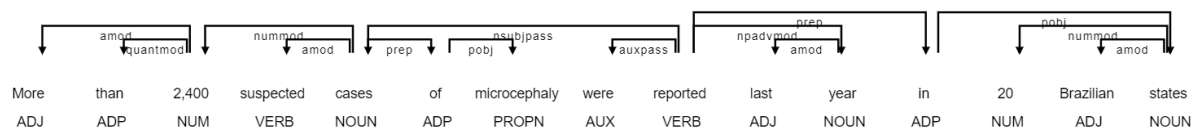
The following table shows the relevant keywords used:

	List of keyword
Fatality	die, death, kill
Cases	case, patient, sick

After detecting the relevant keywords, the next step is to identify the numbers associated with them. Generally, we found two patterns in which these numbers can appear in a sentence.

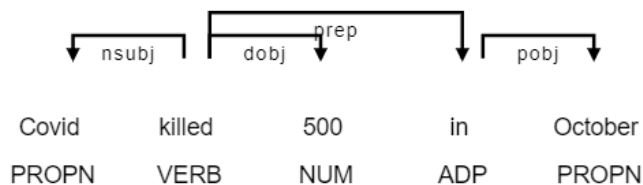
1. Subject of the keyword is a number

Example 1: Simple sentence



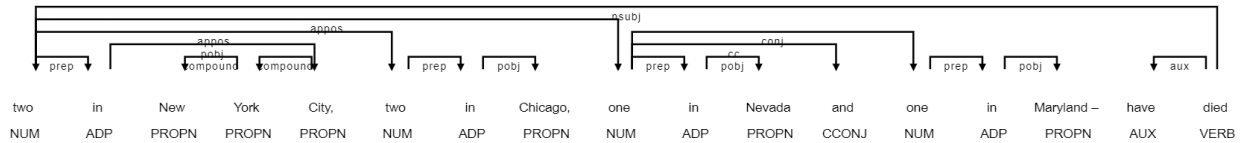
In this example, the keyword "cases" has children "2,400" and "suspected". Thus, we can implement a rule to extract if the child's type is "ORDINAL" or "CARDINAL" or POS tag of "NUM".

Example 2: Simple sentence



In this example, even though the number is after the keyword "kill", it can still be detected that "500" is a child of the keyword.

Example 3: Complex Sentence



In a complex sentence, appositional and conjunction also need to be considered and added together to get the correct number. Hence, we enhanced the rule to consider the children of the number token as well. In this example, the child of the keyword “died” which represents a number “two”, and the children of “two”, is another “two” and “one”. Then the child of “one”, which is another “one” will be added as well.

The following is the pseudo-code of the final rule:

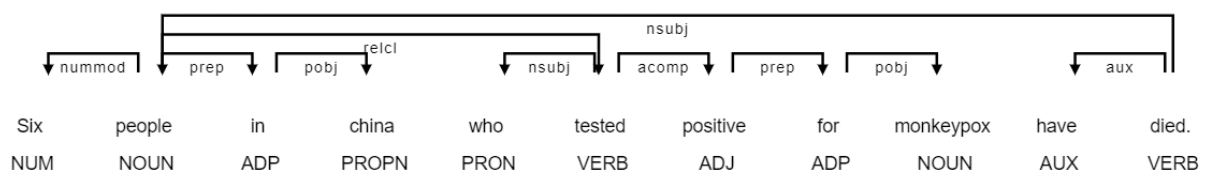
```

1  numbersList -> list
2  keywordList -> list
3  foreach token in sentence
4      if token in listOfKeyword
5          append(keywordList, token)
6
7  foreach keyword in keywordList
8      foreach child in childrenOf(keyword)
9          if isNumber(child)
10             append(numbersList, child)
11
12  foreach number in numbersList
13      for child2 in childrenOf(number)
14          if isNumber(child2)
15             append(numbersList, child2)
16
17  return sum(convertNumber(numbersList))

```

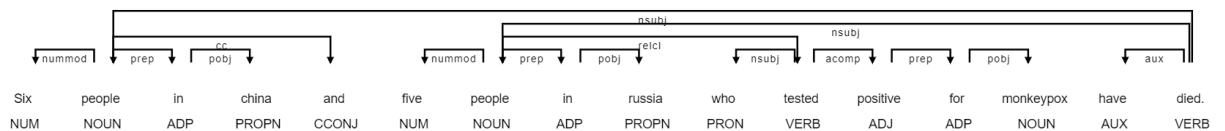
2. Number is attached to the subject of keyword

Example 1: Simple sentence



In a simple sentence, the number related to the keyword can be obtained by looking at the number related to the subject. In this example, the subject of the keyword “died”, has a numeric modifier “Six”.

Example 2: Complex sentence



Similar to Rule 1, the appositional and conjunction also need to be considered to get the correct number. The pseudo-code is as follows:

```

1 numbersList -> list
2 keywordList -> list
3 subjectList -> list
4 foreach token in sentence
5     if token in listOfKeyword
6         append(keywordList, token)
7
8 foreach keyword in keywordList
9     foreach child in childrenOf(keyword)
10        if POS(child) = 'NOUN' and DEP(child) = 'nsubj'
11            append(subjectList, child)
12
13 foreach subject in subjectList
14     for child2 in childrenOf(subject)
15         if isNumber(child2)
16             append(numbersList, child2)
17
18 return sum(convertNumber(numbersList))

```

Organisation NER

To detect if any organisation is mentioned in the headlines, we used existing entity detection capabilities from the Spacy package. The following are some examples of the organisations detected.

Example 1:

The New York City Department of Health and Mental Hygiene **ORG** said it was “deeply saddened by the two **CARDINAL** reported deaths, and our hearts go out to the individuals’ loved ones and community.

Example 2:

Flu activity decreased slightly but remained at high levels across the United States **GPE** during the week ending February 18 **DATE**, according to a report based on preliminary data issued Friday **DATE** by the Centers for Disease Control and Prevention **ORG**.

2.3.3 Country of Outbreak

Location NER

Similar to organisation detection, location can be detected using the entity detection function from Spacy as well.

Example 1:

Health officials in **Puerto Rico** **GPE** reported the island's **first** **ORDINAL** case of Zika, a mosquito-borne virus recently linked to the rise of a serious neurological disorder among newborns in **Brazil** **GPE**.

Example 2:

Flu activity decreased slightly but remained at high levels across **the United States** **GPE** during **the week ending February 18** **DATE**, according to a report based on preliminary data issued **Friday** **DATE** by **the Centers for Disease Control and Prevention** **ORG**.

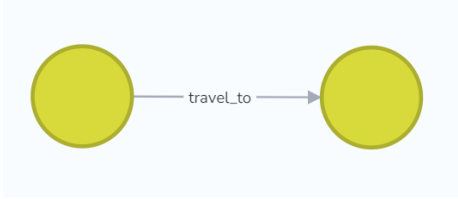
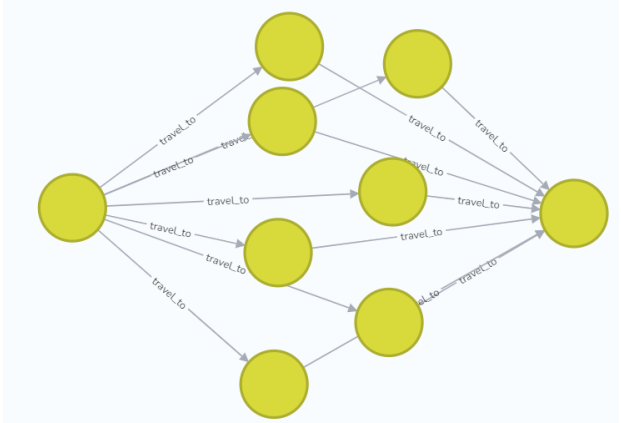
Calculation of Traveller Score

Traveller volumes are determined by traveller arrivals from the detected country to Singapore as the destination. The formula for traveller score considers two factors: one-hop from the detected country to Singapore, and two-hops from the detected country to any ASEAN countries and the ASEAN countries to Singapore as the final destination.

First, traveller volumes collected from 2019 were used as that was the latest available year that was not affected by travel restrictions and border closures (ASEANStatsDataPortal, 2023). The table below shows an extract of the traveller volume data from ASEANStatsDataPortal after pre-processing.

Destination Country Code	Origin Country Code	Visitors
BN	AU	10188
BN	BD	3281
BN	KH	463
BN	CA	2322
BN	CN	74511
BN	DK	312
BN	FR	1381

Next, the tabular information was transformed and stored in the form of a graph database in Neo4j AuraDB. The table below illustrates the one hop and two hop relationship of traveller volume between countries to Singapore based on the traveller volume.

<p>Example 1: Illustration of one hop from a country to Singapore (left: China, right: Singapore)</p> 	<p>Example 2: Illustration of two hops from a country to other ASEAN countries to Singapore (left: China, middle: ASEAN countries, right: Singapore)</p> 
<p>Cypher query: "MATCH (c1 {code :'" f"{country}" '"})-[r:travel_to]-> (c2 {code : 'SG'})" "RETURN c1, c2, r"</p>	<p>Cypher query: "MATCH (c1 {code :'" f"{country}" '"})-[r1:travel_to]->(c2)-[r2:travel_to]-> (c3 {code : 'SG'})" "RETURN c1, c2, r1, r2, c3"</p>

Lastly, the calculation of the traveller score is a three-step process:

Step 1: Normalisation of traveller scores for each hop

Given that there are multiple two-hop relationships, the average of the two-hop traveller score is used for normalisation. For both one-hop and two-hops, traveller scores are normalised using the following equation:

$$s_x^c = \frac{\sum_1^x t_x - t_{min}}{x(t_{max} - t_{min})}$$

s_x^c : Score of x hop from country to country
 t_x : actual traveller volume at x hop
 t_{min} : minimum traveller volume in knowledge graph
 t_{max} : maximum traveller volume in knowledge graph

Step 2: Aggregation of traveller score for one-hop and two-hops together

When aggregating the traveller scores of one-hop and two-hops, a weight of 0.3 was added to the two hops score to reduce the impact of two-hops on the overall traveller volume score. The overall formula used is as follows:

$$S^c = s_1^c + \alpha \frac{\sum_1^n s_2^c}{n}$$

α : weightage of second hop
 S^c : visitor score
 s_x^c : Score of x hop from country c to Singapore
 n : possible number of second hop country

Step 3: Consolidate traveller scores if more than two countries are detected

The average traveller score from the countries detected in the same new article will be used.

Calculation of Distance score

Coordinates (i.e. latitude and longitude) of each country were used to calculate the distances from detected countries in the article to Singapore. Unlike traveller score, only one-hop was used for distance score calculation. Since distance from detected country to Singapore is inversely proportional to relevance to Singapore, the normalisation formula below was used.

$$D_n^c = \frac{d_{max} - \sum_1^n \frac{d_n}{n}}{d_{max}}$$

D_n^c : Distance score of n countries to Singapore
 d_{max} : maximum distance between two places on Earth
 $\sum_1^n \frac{d_n}{n}$: average distance between n countries to Singapore

2.4 Ranking/Score

How do we rank/score the articles?

As the SME could not explicitly define the rules, we used a machine learning algorithm to learn how the SME determined if the news article's topic is disease-related and is relevant to them. Then, we used the trained model to rank the article based on the highest probability.

For the model to learn this decision-making process, the SME had helped to label a set of news articles with 2 indicators: whether an article is disease-related, and subsequently, whether the article is relevant to them.

2.4.1 Disease Topic Classification

Aligned to the feature extraction process, only news headlines were used in the topic classification process. News headlines labelled by the SME were split into a train set and a test set (test size=0.20).

Data Preprocessing using contextualised embeddings

Leveraging on the 'bert-base-uncased' BERT model, contextualised embeddings were performed on news headlines to convert the headlines from text to word token vectors.

Training using Support Vector Machine (SVM)

The word token vectors of the training datasets were trained using SVM modelling. According to scikit-learn documentation (scikit-learn, 2023), SVM modelling is a supervised learning algorithm that supports binary classification (disease vs non-disease news articles) and allows vectors as inputs. The trained SVM model outputs a probability from 0 to 1 for any subsequent word token vectors derived from contextualised embeddings of news articles.

2.4.2 Relevance Classification

After extracting the features in Section 2.3, a method is required to combine the features together to rank the articles. With the data labelled by the SME, we adopted the supervised machine learning approach to train a model that could recommend an article's relevance.

Further pre-processing of the information/features extracted from the title of the articles was done before the model was trained. This includes aggregation of multiple values extracted from the same feature and replacement of null values with zero.

The following are the final features:

Features	Description
has_disease	If one or more disease is mentioned on the headline
has_organisation	If one or more organisation is mentioned on the headline
fatality_count	Number of fatalities detected from the headline
case_count	Number of cases detected from the headline
infectious_score	If the any of the disease(s) mentioned is/are infectious
traveller_score	The score that represent the number of travellers to Singapore
distance_score	The score that represent the distance of the location to Singapore

Random forest algorithm, which is an ensemble algorithm using a basic decision tree, is selected. Random forest is often referred to as a black box model as compared to the basic decision tree algorithm, but some understanding of the model can still be extracted from the model using the feature importance.

We first used the labelled data and split it into a train set and a test set. The random forest model is trained using the train set and eventually tested and evaluated on the test set.

Test Set:

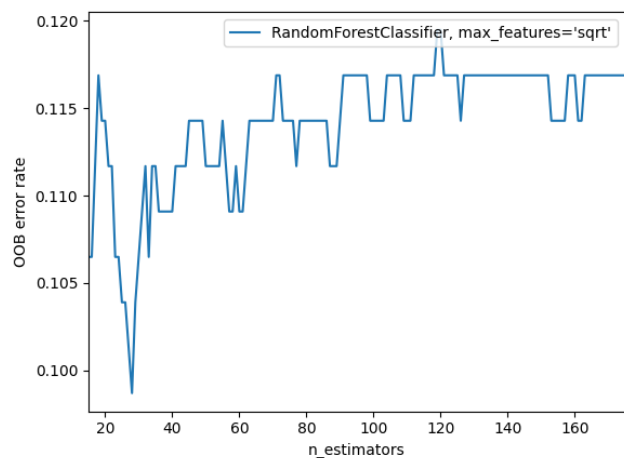
	precision	recall	f1-score	support
False	0.95	0.98	0.96	252
True	0.91	0.75	0.82	56
accuracy			0.94	308
macro avg	0.93	0.87	0.89	308
weighted avg	0.94	0.94	0.94	308

Train Set:

```
[[58  4]
 [ 6  9]]
```

	precision	recall	f1-score	support
False	0.91	0.94	0.92	62
True	0.69	0.60	0.64	15
accuracy			0.87	77
macro avg	0.80	0.77	0.78	77
weighted avg	0.86	0.87	0.87	77

Hyper parameter tuning is done to improve the accuracy of the model. The best performance is when the number of estimators is 30.



After tuning the model, this is the final accuracy from the model:

	precision	recall	f1-score	support
False	0.95	0.98	0.96	252
True	0.91	0.75	0.82	56
accuracy			0.94	308
macro avg	0.93	0.87	0.89	308
weighted avg	0.94	0.94	0.94	308

The random forest model was interpreted using feature importance. The model showed that the SME had put more concern into whether the location mentioned in the article is closer to Singapore in terms of geographical location and connectivity, followed by fatality count and if any infectious disease is mentioned.

Feature	importance
distance_score	0.417962
traveller_score	0.255721
fatality_count	0.099397
infectious_score	0.096254
case_count	0.081181
has_organisation	0.032042
has_disease	0.017442

2.4.3 Final Weighted Score

After obtaining the two classifiers, we combine both 'disease article classification' and 'relevancy classification' to form the final score for ranking/recommendation.

Converting classification to numerical value

Instead of using 1 and 0 to rank the articles, we use the probability from the model as a measure to rank the article. This will ensure the most probable disease related and relevant article will show up first to the user. We leverage the `predict_proba()` function from `sklearn` to provide this probability for both 'disease topic classification' and 'relevancy classification'.

Combining both disease topic probability and relevance probability

After discussing with the subject matter expert, we agree that disease topic probability should have higher weight to relevancy probability, in a 2 to 1 ratio. The formula for the final score is as follows:

$$\text{Score} = 1000 \times (\text{DiseaseProb} + \alpha \text{RelevanceProb})$$

$$\text{where } \alpha = 0.5$$

2.5 Result Review

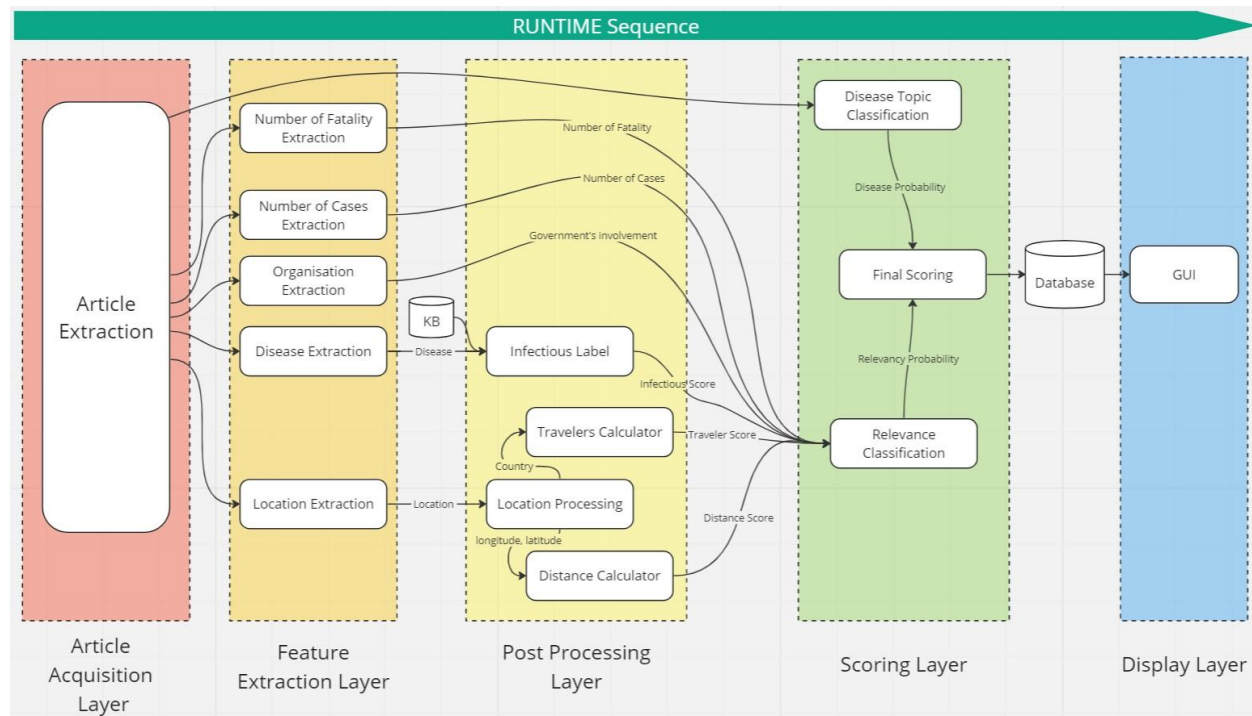
To review our result, we obtained a new dataset to verify if our models and rules were performing as intended. The dataset obtained are new articles from 11 April 2023, where there is an article regarding the death of a H3N8 patient in China.

	index	title	prob
0	18	First H3N8 bird flu death recorded in China	1082.310558
1	261	Rising COVID-19 cases in Singapore driven by X...	1054.481192
2	146	Wildfire in South Korea forces 500 residents t...	703.924324
3	9	China's sandstorm problem spreads to South Kor...	581.412595
4	50	South Korea wildfire forces 500 residents to e...	563.158092
5	102	4 South-east Asian women held captive by sex t...	538.630680
6	264	'Revenge of the geeks': Drones battle on Ukrai...	537.345187
7	250	At Mexico's Chichen Itza, archaeologists disco...	526.282034
8	63	IMF says more flexible BOJ yield control can p...	517.650783
9	87	US stocks mixed as markets try to shrug off do...	490.193306
10	233	Today in Pictures, April 12, 2023	484.823797
11	155	Alibaba to roll out generative AI across apps,...	475.638560
12	178	North Korea parade 'probably oversells' ICBM t...	469.849369
13	121	Swiss researchers use typing, mouse clicks to ...	464.291967
14	125	Eleven International Wins Prestigious Gold Ste...	452.524300
15	25	Explainer: Why a clean energy transition is so...	452.136350

As shown in the result above, the mentioned article is ranked highest by our scoring system, followed by COVID-19 news in Singapore, and lastly articles that are not related to Singapore. The subject matter expert reviewed the result and was satisfied with the system's output.

Section 3. System Design

The figure below shows the flowchart of the various components within Apollo System:



Apollo System consists of 5 different layers where the first 4 layers are designed to be executed daily in the background to populate the database.

3.1 Article Acquisition Layer

This layer is responsible for finding articles from news sites pre-identified by the user. The summary, headline, and publication date from the articles are extracted.

3.2 Feature Extraction Layer

This layer is responsible for extracting relevant information from articles' headlines.

3.2.1 Disease Extraction

Extract any diseases mentioned in the article's headline as defined in Section 2.3.1.

3.2.2 Organisation Extraction

Extract any organisations mentioned in the article using Spacy built-in named entity detection as defined in Section 2.3.1.

3.2.3 Location Extraction

Extract any location mentioned in the article using Spacy built-in named entity detection as defined in Section 2.3.3.

3.2.4 Number of Fatality Extraction

Extract the number of fatalities via rules-based information extraction as defined in Section 2.3.2.

3.2.5 Number of Cases Extraction

Extract the number of cases rules-based information extraction as defined in Section 2.3.2.

3.3 Post Processing Layer

This layer is responsible for converting and processing features extracted to information/inputs required for scoring.

3.3.1 Location Processing

This module is responsible for retrieving the longitude, latitude and country of the location extracted from the article using Google Geolocation API.

3.3.2 Infectious Label

This module infers the disease extracted from the article with the disease knowledge base to find out if the disease in the article is a transmittable disease.

3.3.3 Distance Calculation

This module calculates the distance from the location mentioned in the article to Singapore, using the longitude and latitude extracted.

3.3.4 Travellers Calculation

This module calculates the volume of travellers from the country of a location to Singapore and converts it into a score.

3.4 Scoring Layer

This layer is responsible for handling the scores for each sub-module in the system and the overall scoring mechanism for the final score.

3.4.1 Disease Topic Classification

This module uses the disease topic classification model in Section 2.4.1 to find the probability that an article is a disease-related article.

3.4.2 Relevance Classification

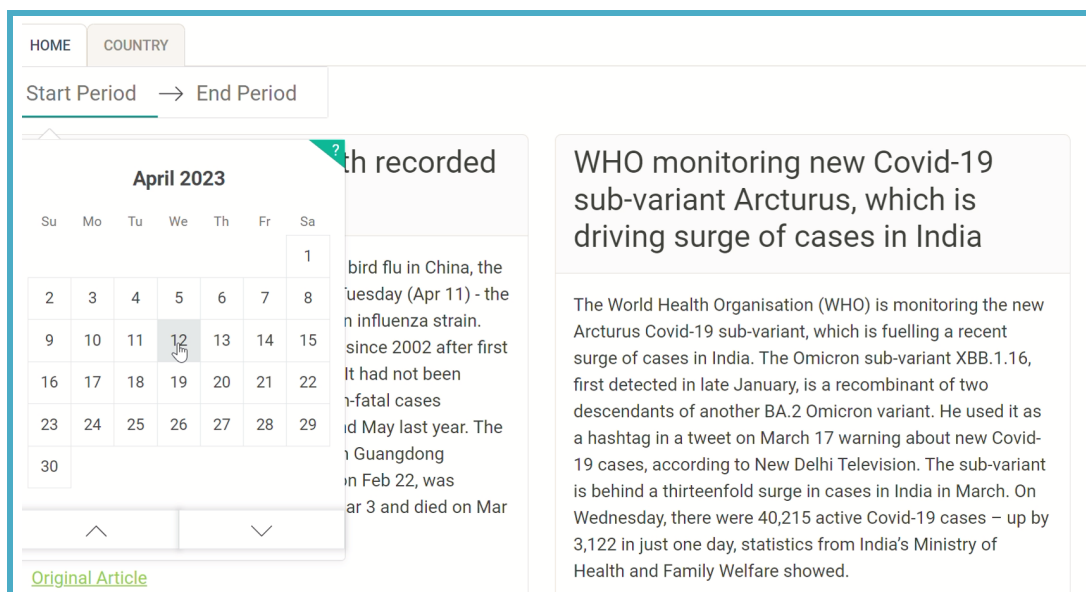
This module uses the relevance classification model in Section 2.4.2 to find the probability that an article is relevant.

3.4.3 Combination of Scoring

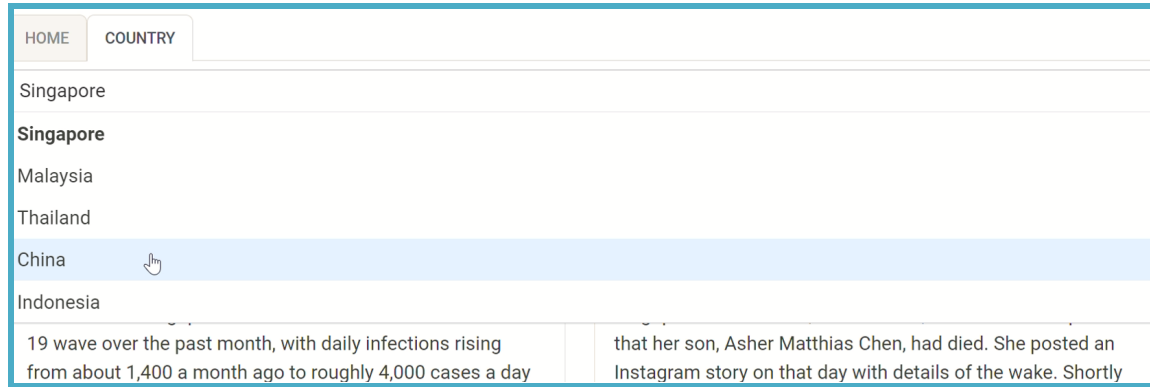
This module uses the formula from Section 2.4.3 to get the final score.

3.5 Display Layer

The user interface layer will rank the articles based on the final score and filter out the top 15 articles to be displayed in the home page. This provides the user with a quick oversight of the most relevant articles. Users may filter by the time period of interest as well. This means that if the user had last checked Apollo System 3 days ago, they can simply filter for the start date to be from 3 days ago so that they can pick up where they left off and be updated with the news during the time period they had missed.



If users are concerned about the events happening in a particular country, they can select the country of interest via the country tab.



Section 4. Assumptions and Limitations

4.1 Assumptions

4.1.1 Little fluctuations to the knowledge databases and graph databases

The creation and maintenance of the knowledge databases and graph databases are based on an underlying assumption that the trends and factors contributing to the relevance to Singapore's context do not change frequently. Thus, the knowledge databases and graph databases are periodically updated.

4.1.2 Period of data the models are trained with is sufficient in the long run

Models were trained on an assumption that the news articles data from 1 January 2023 to 23 Mar 2023 were sufficient to represent the population of news articles that may arise in the future.

4.1.3 Features extracted from news articles for model training

It is assumed that the current set of features extracted, which was recommended by the SME, is adequate to sieve news articles that are disease related and relevant to Singapore's context. In addition, news headlines were taken to be representative of the information contained in the news article.

4.2 Limitations and Improvements

As the timeline to develop the Apollo system is short, there are some limitations to the Apollo system. Therefore, some improvements can be implemented to the Apollo system.

4.2.1 Addition of a feedback mechanism

The Apollo system currently does not allow for users to feedback on the system's output. A feedback mechanism can be developed on Apollo's User Interface for users to rate if the news articles displayed were indeed disease related and whether it is of high relevance to Singapore. This information, collected over time, could then be used to train the Article Topic Classification model and the Relevance Classification model.

4.2.2 Explore different algorithms

Other algorithms can also be explored to compare their performance. For instance, we could have explored using Stanford NLP libraries for Named Entity Recognition and

compare the performance against the current model that is trained on the SciSpacy model. Similarly for the classification models.

4.2.3 Feature Extraction using Deep Learning

We have only interviewed one expert to obtain the knowledge on how they pick a relevant disease-related article, hence the feature extraction we used may be subjective. Due to the lack of labelled data in this project, we were not able to use deep learning for feature extraction. By using BERT or ChatGPT model, it will learn the feature extraction itself based on the labelled data and will be able to improve over time through retraining.

Section 5. References

Population and Population Structure. (n.d.). DOS | SingStat Website - Population and Population Structure - Latest Data. Retrieved May 18, 2023, from <http://www.singstat.gov.sg/find-data/search-by-theme/population/population-and-population-structure/latest-data>

ANNUAL STATISTICS 2019: Volume Of Travellers And Cargoes Cleared At The Checkpoints Increased. (2020). IMMIGRATION & CHECKPOINTS AUTHORITY, SINGAPORE. Retrieved May 18, 2023, from https://www.ica.gov.sg/docs/default-source/ica/stats/annual-stats-report/ica-annual-statistics-2019.pdf?sfvrsn=337746ea_2#:~:text=In%202019%2C%20the%20Immigration%20%26%20Checkpoints,cargoes%20cleared%20compared%20to%202018.

Haug, N., Geyrhofer, L., Londei, A. et al. Ranking the effectiveness of worldwide COVID-19 government interventions. *Nat Hum Behav* 4, 1303–1312 (2020). <https://doi.org/10.1038/s41562-020-01009-0>

Lance, S., Steve, C., Pablo, D., Rebecca, D., Andrew, D., Hania, F., Beatrice, L., Julie, R., Andrew, R., Anne, S. (2020, November 29). Wellcome Global Monitor—How Covid-19 affected people's lives and their views about science. Gallup. Retrieved from <https://cms.wellcome.org/sites/default/files/2021-11/Wellcome-Global-Monitor-Covid.pdf>

Mouratidis, K. How COVID-19 reshaped quality of life in cities: A synthesis and implications for urban planning. *Land use policy*. 2021 Dec;111:105772. doi: 10.1016/j.landusepol.2021.105772. Epub 2021 Sep 20. PMID: 34566233; PMCID: PMC8456312.

Clark C. Freifeld, Kenneth D. Mandl, Ben Y. Reis, John S. Brownstein, HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports, *Journal of the American Medical Informatics Association*, Volume 15, Issue 2, March 2008, Pages 150–157, <https://doi.org/10.1197/jamia.M2544>

Scacco, J., and Muddiman, A. (2015, December). The Current State of News Headlines. Center for Media Engagement. <https://mediaengagement.org/research/the-current-state-of-news-headlines/>

Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C. H., Leaman, R., Davis, A. P., Mattingly, C. J., Wieggers, T. C., & Lu, Z. (2016). BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database : the journal of biological databases and curation, 2016, baw068. <https://doi.org/10.1093/database/baw068>

Wikidata. (n.d.). Wikidata. Retrieved Apr 05, 2023, from https://www.wikidata.org/wiki/Wikidata:Main_Page

Visitor Arrival to ASEAN Member States by Origin Countries (in person)
ASEANStatsDataPortal. (n.d.). Visitor Arrival to ASEAN Member States by Origin Countries (in Person) ASEANStatsDataPortal. Retrieved May 18, 2023, from <https://data.aseanstats.org/visitors>

1.4. Support Vector Machines. (n.d.). Scikit-learn. Retrieved May 16, 2023, from <https://scikit-learn/stable/modules/svm.html>

Appendix A: Interview Transcript

Interview transcript with Subject Matter Expert (SME)

Due to the confidentiality issues, the identity of the SME interviewed has been masked.

Insyirah/Boyan: Thank you for your time to attend the interview. In this interview, we would like to understand more about the requirements for the system, which aims to alert users on potential pandemic outbreaks in Singapore.

SME: Thank you for having me here. To begin, I would like to provide a brief background on how this problem statement arises. Following the COVID-19 pandemic, governments from various countries have identified two effective interventions, border measures and public health measures, to curb disease transmissions in their country. Thus, a monitoring system of the disease outbreak situations worldwide, especially in the order of relevance to Singapore's context, is required.

Insyirah/Boyan: You have mentioned 'the relevance to Singapore's context'. May we understand more about what factors constitute 'relevance' to Singapore?

SME: Definitely. Firstly, from the vast information in news articles, we would like to ensure that the disease mentioned in the articles are infectious, especially when the disease is capable of infecting humans. Next, we will identify the country of the disease outbreak. The importance of the country depends on the geographical distance from Singapore, as well as the inbound traveller volume to Singapore from the country. To add, to understand the significance of disease outbreaks in that country, we will check if there are any fatalities caused by the outbreak, if there are many people who are infected and if the local government has taken actions on the outbreak.

Insyirah/Boyan: Given the widespread interest in this area globally, do you know of any existing systems used by agencies worldwide to monitor the disease outbreak situation?

SME: Indeed, there are existing systems available, such as HealthMap, which was founded as early as 2006, that perform similar functions. However, resources that are currently available online do not cater specifically to Singapore. Thus, we hope to build an enhanced monitoring system that alerts users on disease outbreaks overseas that may potentially result in a pandemic situation in Singapore.

Insyirah/Boyan: We have noted the background and the requirements for the system. Thank you so much for your time.

Appendix B: Project Proposal

GRADUATE CERTIFICATE: Intelligent Reasoning Systems (IRS) PRACTICE MODULE: Project Proposal

Date of proposal: 5 March 2023
Project Title: Alerts on Potential Pandemic Outbreaks in Singapore
Sponsor/Client: <i>(Name, Address, Telephone No. and Contact Name)</i> Institute of Systems Science (ISS) at 25 Heng Mui Keng Terrace, Singapore NATIONAL UNIVERSITY OF SINGAPORE (NUS) Contact: Mr. GU ZHAN / Lecturer & Consultant Telephone No.: 65-6516 8021 Email: zhan.gu@nus.edu.sg
Background/Aims/Objectives: Singapore is a small and open country with a high population density, at 7,688 persons per sq km in 2022, according to the Department of Statistics Singapore. In addition, before the implementation of border restrictions due to Covid-19, Singapore had cleared 217 million travellers through its checkpoints in 2019. This makes Singapore highly susceptible to disease transmissions whenever a global pandemic outbreak occurs. Thus, it is critical to acknowledge the adverse effects of infectious disease and ensure that there are sufficient infection control and prevention measures in Singapore to protect the safety and well-being of the nation. While there are platforms, such as HealthMap, that monitor the outbreak of diseases via press releases globally, they are usually not catered specifically to Singapore's context. Thus, relying on such platforms usually requires a lot of manpower to look through all articles to sieve out information that are relevant to Singapore. As such, the creation of an intelligent system to identify potential high risk diseases would be valuable for the health security of Singapore. Potentially, the intelligent

system may also be extended for use by other countries alike or extended to other topics that concern Singapore's national security.

Requirements Overview:

- Research Ability
- Programming Ability
- System integration ability

Resource Requirements (please list Hardware, Software and any other resources)

Hardware proposed for consideration

- Virtual Server

Software proposed for consideration

- Graph database, eg Neo4j
- Text Processing / NLP technique, eg topic clustering, information extraction
- Reasoning system
- Push notification communication means, eg Telegram, emails

Number of Learner Interns required: (Please specify their tasks if possible)

3

Methods and Standards:

Procedures	Objectives	Key Activities
Requirement Gathering and Analysis	The team should scope the details of the project and ensure the achievement of business objectives.	<ol style="list-style-type: none"> 1. Gather & analyze requirements 2. Define overall system requirements
Literature Review	To research on techniques and methods to fulfill the requirements	<ol style="list-style-type: none"> 1. Research text processing methods 2. Research on methods to define connectedness between locations 3. Research on topics that will constitute disease

		<p>outbreak risk in Singapore</p> <p>4. Research on the factors that contribute to high relevance in Singapore's context</p>
Technical Construction	To develop the code in accordance to the design	<p>1. Develop the code to obtain articles, perform topic classification and extract important information from it</p> <p>2. Construct the knowledge base which specify the relevance between Singapore and other locations</p> <p>3. Develop the scoring algorithm to identify potential pandemic risks to Singapore based on the knowledge base</p> <p>4. Develop the alert / communication means to deliver the information once scoring is above threshold</p>
Acceptance Testing and Delivery	To obtain user acceptance and deploy the solution	<p>1. Planning of the UAT</p> <p>2. Conduct the UAT</p> <p>3. Deploy the solution</p>

Appendix C: Mapped System Functionalities

S/N	Tasks/ System Features	Modular Course	Knowledge/ Techniques/ Skills
1	Data/Information Gathering	Machine Reasoning	Knowledge Elicitation, Knowledge Representation
2	Disease Extraction, Location Extraction, Organisation Extraction	Cognitive System	Named Entity Recognition
3	Case and Fatality Extraction	Cognitive System	Rule-based Information Extraction
4	Disease Article Classification	Cognitive System, Machine Reasoning	Contextualised Embeddings, Machine Learning
5	Traveller Score using data stored in graph database	Machine Reasoning, Reasoning System	Knowledge Representation, Graph Database, Cypher query
6	Article Relevance Classification	Machine Reasoning	Machine Learning
7	Final Scoring	Machine Reasoning	Hybrid Reasoning System - Co-operating Experts
8	Infectious Disease Knowledge Base/Inference	Machine Reasoning	Deductive Reasoning, Relational Database

Appendix D: Installation and User Guide

1 Install Dependencies

1. This is designed using python 3.8, using other version may not work
2. Install packages

```
python -m pip install -r requirements.txt
```

3. Install Spacy package

```
python -m spacy download en_core_web_sm
```

2 Running Backend (Article & Feature Extraction)

Backend Script is called *MainBackEndProcessing.py*. Run this script to start. Offline mode is available as the tool relies on google news to query news articles. Users will get blocked if too many queries are performed within a short period of time.

- Run test mode (test dataset used to develop the tool) (*Note: Test mode will not write data to db*)

```
python MainBackEndProcessing.py -M test
```

- Run production mode (live, using today's article). Use -D to run on specific day of articles (*Note: running in production mode requires internet connection*)

```
python MainBackEndProcessing.py -M prod  
# TO run on specific day  
python MainBackEndProcessing.py -M prod -D 2023-04-11
```

- Run demo mode (using 11 April 2023 dataset)
 - Online Mode (*Note: running in production mode requires internet connection*)

```
python MainBackEndProcessing.py -M prod
```

- Offline Mode

```
python MainBackEndProcessing.py -M prod -O
```

The final output is input into the database under folder "SystemCode\KnowledgeBase\ApolloDM.db" and its snapshot is located at "\output".

In production system, production mode will be run daily using cron job to populate daily data

3 Starting the Web Application

Run *app.py* script

```
python app.py
```

Web is served on <http://127.0.0.1:8050/>

4 Repopulating Knowledge Base and Retraining Model

In case the existing Neo4J AuraDB is not working or an update of the knowledge base is required, there are scripts prepared under the "TrainModelOrBuildKB" folder to repopulate the knowledgebase and machine learning models.

The table below provides a description of the scripts that were prepared:

Script Name	Description
CountryTravelersNetworkBuilding.ipynb	Script to repopulate travellers network graph to neo4j
Disease NER Custom Models.ipynb	Script to retrain disease NER model
DiseaseTopicClassification-SVMContextualizedEmbeddings.ipynb	Script to retrain disease topic classification model
Diseases Knowledge Base Building.ipynb	Script to repopulate disease knowledge base
RelevanceScoreModel.ipynb	Script to retrain relevance score model

Appendix E: Individual Project Report

Ng Bo Yan

Personal Contribution

- Compilation of the whole processing script, designed the software architecture
- Developed the rule based information extraction for fatality count and case count extraction
- Explored extraction of organisation and location extraction using Spacy
- Assisted topic classification on SVM modelling
- Developed final scoring method
- Trained article relevance scoring model using random forest
- Organised team discussion and drive project progress

Useful Learnings

- Natural language processing
 - Text processing
 - Information extraction
 - Rule based
 - Model based (NER Model)
 - Topic classification
 - Neo4j graph db
 - Web scraping
 - Hybrid reasoning

Workplace Application

Working as an engineer in a manufacturing company, I always think that data needs to be labelled in a structured and clean manner. However, there is a lot of text / speech data such as maintenance logs. These are not utilised due to lack of clean labels at the moment. Using NLP techniques like information extraction and topic classification can enable my workplace to use more data for problem classification and root cause finding.

Nur Insyirah Binte Mahzan

Personal Contribution

- Extracted, processed and generated the knowledge databases -- visitor arrival/travellers volume, and disease aliases and infectiousness
- Designed and implemented the web application using Dash Python framework
- Trained custom NER model to detect disease names and explored pre-trained SciSpacy model
- Created the Business Use Case Video

Useful Learnings

- Training a custom NER model
- How to identify relevant data/information from an SME
- Leveraging on crowd-sourced database (Wikidata) for inferencing/deductive reasoning
- Hybrid use of both graph database and relational database within a system

Workplace Application

At my current workplace we mostly use structured data fields to analyse, track and monitor performance of key clinical indicators to ensure quality of patient care. However, a major limitation is that most key information about patients' conditions etc. are usually in unstructured text such as Doctor's Consult Notes, Operation Summary Notes and Discharge Summary Notes.

Given the rich untapped data sources, I can potentially explore building and training NER models to detect medications, diseases, symptoms/complications from unstructured text at my workplace. These information can then be used to supplement current analysis, tracking and monitoring of clinical indicators of concern.

I may also generate a knowledge graph of relationships between diseases, related complications and symptoms of the disease and/or build models to predict potential risks and harm to patients.

Low Pei Jing

Personal Contribution

- Web scraping of news articles from GoogleNews from 1 Jan 2023 to 23 Jan 2023
- Gathered the labelled disease news articles that are relevant in Singapore's context from the pool of web scraped news articles
- Data preprocessing, which includes contextualised embeddings of news headlines, prior to topic classification training using SVM
- Created instance in Neo4J aura for traveller scoring
- Assisted to develop rule based scoring of traveller score and distance score
- Created the System Design Video

Useful Learnings

- Web scraping using a variety of methods, such as GoogleNews and Newspaper3k python packages
- Creating connections from python to other tools, such as Neo4J
- Data preprocessing prior to topic classification modelling
- Machine learning for classification methods using vectors
- Rule based reasoning system

Workplace Application

One of my job responsibilities includes monitoring of unusual events that occur in Singapore or overseas. In fact, the scope of the project has encompassed one of the activities that is required in my workplace by surfacing potential disease outbreaks reported in news articles that may have detrimental impacts to Singapore.

In addition, understanding python's ability to connect to Neo4J has unleashed greater potential and capabilities of graph databases, in consideration that python functions can be used on top of cypher queries for analysis. This is especially useful when conducting person to person investigations, or company linked investigations.

Given that the nature of my job affects the general public to a large extent, there is a strict requirement for any models developed in my work to be explainable. While rule-based reasoning is a familiar topic, there are few chances for me to try machine learning techniques at work. After attempting contextualised embeddings and SVM algorithms during the course of the project, I feel more confident to explore the unstructured data in my workplace and hope that I can try to perform data mining on these unstructured data to gather more insights.