Table of Contents

TOPIC MODELLING

In the part of topic modelling, we will perform appropriate text pre-processing on the content collected from news sites, containing the term “Monash University”, or tagged with the label “Monash University” and then use LAD and NMF method to find the topic in the content.

Latent Derelicht Analysis (LDA) Method:

In this part, we will use LDA method that a probabilistic model to build the topic model. Besides we will perform two runs of LDA model with different topic number.

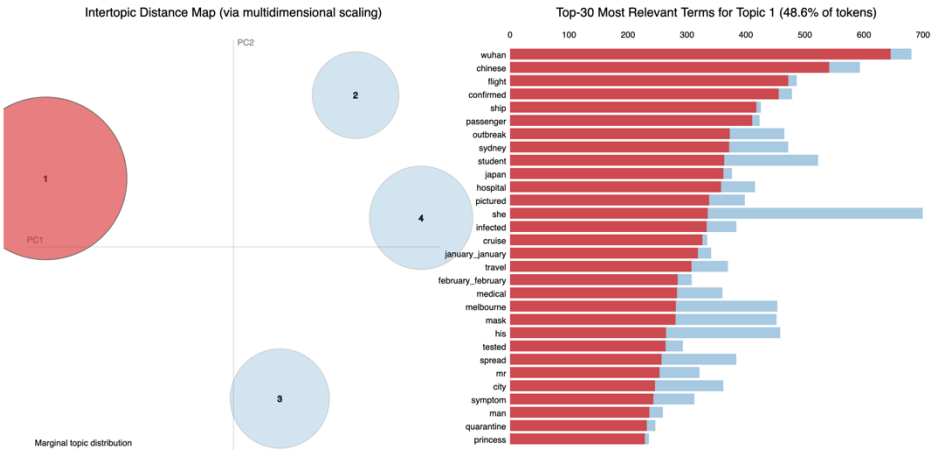
Steps of LDA Method:

1. Preparing Data:
    a. Tokenization: Use spacy library to split the text into sentence and the sentences into words. Besides, lowercase the words and remove punctuation.
    b. Lemmatize the words
    c. Find the bigrams that only appear more than 20 times
    d. Remove rare and common tokens
2. Build the Model and train the model

Results:

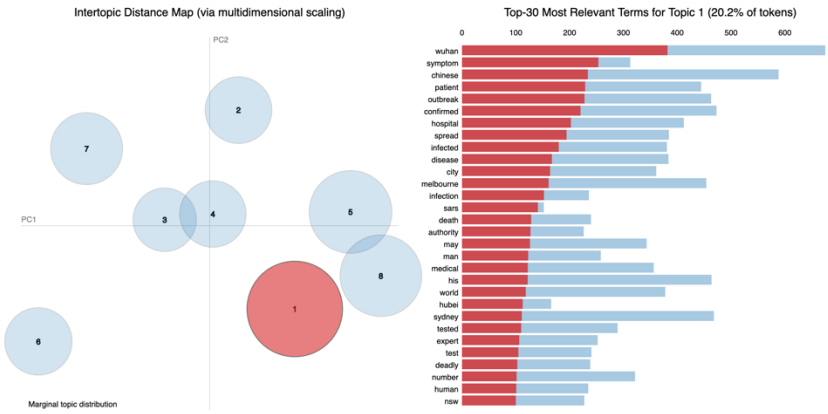
1st run (Topic number = 4):

	Topic 1	Topic 2	Topic 3	Topic 4
0	wuhan	she	fire	you
1	chinese	her	cent	area
2	flight	woman	per	your
3	confirmed	patient	per_cent	work
4	ship	study	climate	say
5	passenger	mask	smoke	student
6	outbreak	face	air	should
7	sydney	his	pandemic	cell
8	student	say	change	school
9	japan	my	bushfires	our
10	hospital	face_mask	animal	such
11	pictured	group	expert	these
12	she	level	world	research
13	infected	hand	say	what
14	cruise	mass	may	home



2nd run (Topic number = 8):

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
0	wuhan	she	fire	school	student	area	say	ship
1	symptom	her	smoke	you	chinese	study	you	cruise
2	chinese	woman	air	should	wuhan	patient	his	passenger
3	patient	mask	bushfires	home	flight	you	change	princess
4	outbreak	face	cent	pandemic	ban	analysis	what	cruise_ship
5	confirmed	face_mask	per	covid	travel	data	climate	diamond_princess
6	hospital	store	per_cent	state	island	research	like	diamond
7	spread	business	million	professor	february_february	using	how	japan
8	infected	hand	bushfire	food	january_january	between	them	flight
9	disease	just	climate	need	pictured	used	specie	board
10	city	child	condition	social	aedt	your	work	tested
11	melbourne	she_wa	quality	work	mr	cell	climate_change	she
12	infection	customer	canberra	spread	christmas	each	my	on_board
13	sars	pictured	city	care	christmas_island	based	your	infected
14	death	his	melbourne	service	confirmed	level	just	sydney



According the above graphs:

1. In the first run which topic number is four, the topic #1shows words may associated with COVID-19, as evident with words such as “Wuhan”, “outbreak”, “infect”, and “confirmed”. However, the topic words in #2, #3, and #4 are difficult to see the pattern of topic. Therefore, we modify the topic number from 4 to 8.
2. In the second run which topic number is eight, we could notice that the topic #1 show words more clearly about COVID-19, along with the words such as “wuhan”, “symptom”, “patient”, “outbreak”, “hospital”, “infected”, and “disease”.
3. In the second run, the topic #3 focuses on bushfires related terms, such as “fire”, “smoke”, “air”, “bushfires”, and “climate”.
4. In the second run, the topic #8 shows words associated diamond princess cruise, as evident with words such as “ship”, “cruise”, “passenger”, “diamond princess”, “japan”, and “board”.

Non-negative Matrix Factorization (NMF) Method:

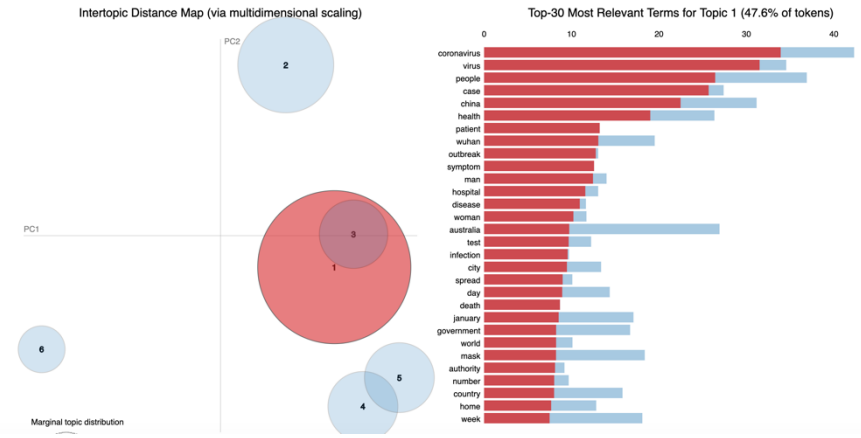
In this part, we will use NMF method that a linear-algeabreic model to build the topic model.

Steps of NMF Method:

1. Preparing Data:
    a. Tokenization: Split the text into sentence and the sentences into words. Besides, lowercase the words and remove punctuation.
    b. Lemmatize the words
    c. Remove stopwords
    d. Keep only noun words
2. Build the Model and train the model

Results:

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
0	coronavirus	climate	student	ship	flight	chemist
1	virus	bushfire	ban	cruise	island	mask
2	people	smoke	university	princess	christma	store
3	case	change	travel	diamond	evacuee	hand
4	china	air	semester	passenger	wuhan	sanitiser
5	health	quality	school	board	zealand	warehouse
6	patient	season	education	japan	passenger	face
7	wuhan	weather	china	yokohama	facility	chatswood
8	outbreak	year	australia	quarantine	qanta	stock
9	symptom	australia	country	vessel	plane	customer
10	man	event	week	health	airline	shelf
11	hospital	temperature	class	coronavirus	detention	supply
12	disease	tennis	sydney	crew	board	demand
13	woman	player	cent	port	air	product
14	australia	pollution	package	food	picture	spokesperson



According the above graphs:
Topic #1 (COVID-19): “coronavirus”, “virus”, “wuhan”, “hospital”, “patient” …
Topic #2 (Bushfire in AU): “climate”, “bushfire”, “smoke”, “air”, “pollution” …
Topic #3 (Student affected by travel ban): “student”, “university”, “semester”, “school”, “education”
Topic #4 (diamond princess cruise): “ship”, “cruise”, “princess”, “diamond” …
Topic #6 (personal protective equipment): “chemist”, “mask”, “sanitizer”, “product” …

Summary:

There are five sorts of news mentioning Monash University. The first is news about COVID-19, second is news about bushfire, third is news about students, fourth is news about diamond princess cruise, and final is news about personal protective equipment. Here are the reasons that the news mentioning the university.

1. When the news about COVID-19 and COVID-19 pandemic on diamond princess cruise mention Monash University is because Monash have many professional professors in epidemiology and preventive medicine. Take < *Coronavirus outbreak: Could the deadly virus reach the UK? Expert issues stark warning*> news for example. This news mentions the Professor Allen Cheng who is from Monash University department of epidemiology and preventive medicine.
2. When the news about Bushfire mention Monash may because Monash also have many professional professors. For example, in the news <*Fires and floods: Australia already seesaws between climate extremes - and there is more to come | Neville Nicholls*>, the author Neville Nicholls is an emeritus professor at Monash University. He spent 35 years with the Bureau of Meteorology, with his research focusing on how and why the climate is changing.
3. When the news about Students affect by travel ban mention Monash is because Monash has a lot of international student. Take the news < *China to relax its internet restrictions for 100,000 students hit by Australia's coronavirus travel ban* > for example. The news mentions Monash already pushed back the start date of its semester by a week.

The advantages of the topic modeling of LDA when we want to find what sorts of news mentioning Monash University is that it is very easy for researcher to know what sorts of news mentioning Monash by analyzing the top topic words. However, here are some disadvantages. The first problem is that it is very difficult to specify the appropriate number of topics in the corpus. The different number of topics could vary the result a lot. The second problem is static. This means there is no evolution of topics over time.