

Multi-Doc Summarization

COVID-19

Sharon Wang
Week 1

COVID-19 News Articles Open Research Dataset

<https://www.kaggle.com/ryanxjhan/cbc-news-coronavirus-articles-march-26/data?select=news.csv>

6,788 news articles about the coronavirus-related infectious diseases

Progress

1. Performed **AutoPhrase** on the corpus
 - a. World Health Organization → World_Health_Organization
2. Filtered out symbols
3. Ran **CatE** on the corpus

CatE Results

Symptom	Transmission	Death	Elderly	Incubation	Asymptomatic
cough	community	fatalities	pre-existing	period	testing
fever	spread	recorded	vulnerable	days	screening
runny	exposure	count	seniors	weeks	showing
breath	evidence	1000-case	immunocompromised	14	positive
shortness	human-to-human	1523	compromised	days	symptomatic
visit	person-to-person	3869	older	time	serological
unexplained	clusters	had	conditions	within	nursing-home
811s	spreading	toll	chronic	cycles	test
should	exposures	died	frail		tests
symptoms		deaths	underlying		
if		fatality			
you					

N95	Ventilator	Vaccine	Mask
masks	machine	treatments	wearing
respirators	ECMO	therapeutics	scarves
surgical	ICU	drugs	bandana
PPE	oxygen	trials	walks
gloves	lungs	antivirals	helmet
gowns	tube	vaccines	smile
nitrile	airway	treatment	gown
Minnesota-based	breathe	antiviral	scissors
	ventilators	anti-viral	wear
	machines		wears

SARS	SARS-CoV-2	MERS
2003	engineered	avian
2002-2003	cultured	EMC
epidemic	novel	camels
influenza	causes	acute
repeat	viruses	coronaviruses
outbreak	genome	bats
H1N1	sequence	syndrome
Ebola	electron	Middle
2002-03	2019-nCoV	syndrome
SARS		respiratory

China	CCP	Wuhan	Origin	Travel	Economy
mainland	ridiculed	Sichuan	virologists	destination	stimulate
countries	puzzled	Hubei	scientists	trips	industries
Korea	infuriated	Tianhe	bioweapons	returning	restart
Taiwan	glaring	epicentre	researchers	non-essential	rebound
Japan	generals	trapped	scientific	advisory	protectionism
Asia	downplaying	repatriating	theory	travelling	third-largest
France	furious	countryside	biologists	travellers	sectors
Europe	conflated	Li	findings	trip	economic
States	deferential	Mei	scientifically	exercise	economies
China	Coons	Zhuo	origins		

Domestic Violence	Protest	Racism	Unemployment
abusers	demonstrators	xenophobia	jobless
perpetrators	protestors	stigmatization	benefits
rise	pro-democracy	discrimination	EI
rates	capitol	misinformation	jobs
empowered	Lansing	bullying	income
gender-based	statehouse	hysteria	savings
empowering	disperse	profiling	1250
violence	mourners	hurtful	qualify
abuse	protested	racial	unemployed
rising	protesting	racist	

CatE Comments

- I noticed that some phrases generated by topics are often synonyms or different versions of each other.
- For example, under the topic **Asymptomatic**, CatE returns **test, tests, testing**.
- I will merge synonyms or different versions of a phrase into a single phrase before a summary is made using the phrases.

Future Steps

- Next step is to work on extracting sentences from the corpus that correspond to the phrases
- Each user-given topic will correspond to 1 section of the summary
- For each phrase, generate ~5 sentences. Can experiment with how many sentences will yield quality results. Maybe 5 can be the default, but the user can specify how many sentences they want. This will result in a summary organized with multiple sections, each about a different topic.
- After this is established, maybe use **JoSH** to create a hierarchical representation of the topics. This will result in a summary organized with multiple sections and subsections.