BEHIND THE MASK

Tracing the COVID-19 Research Timeline through Topic Modeling and Evolution

You might be thinking:
Is this still relevant?

NEW COVID CASES PER DAY
(PHILSTAR, MAY 2024)

TRACKED COVID VARIANTS (2024)
(BY THE CDC)

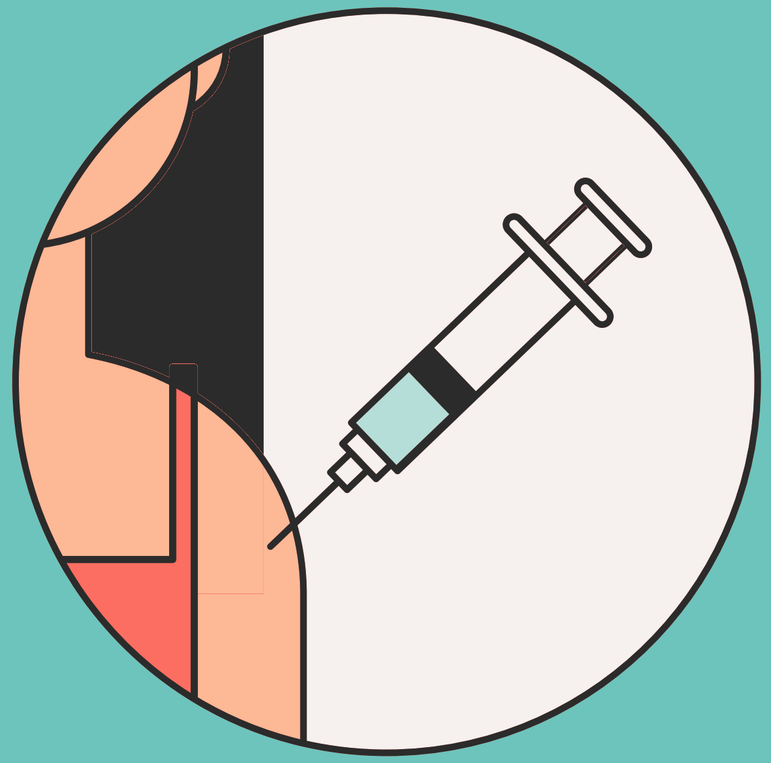
COVID-19 **cases are on the rise again** and are **still being tracked by some organizations**

Our methodology can extend to other fields of study.

(AS OF 2022)

5M

AVERAGE ARTICLES PUBLISHED YEARLY



PROJECT USE CASE

SOLUTION IMPACT

01

PUBLIC HEALTH

Better resource allocation to **policies and studies** for **future pandemics**

02

ACADEMIC RESEARCH

Can be applied to **different categories of research** (ex. physics and engineering)

03

MEDIA INFORMATION

Creation of **informed content** and **summarization** of multiple news

04

POLICY MAKING

Synthesize **news and law articles** to **create necessary policies** for the people.

What seems to be the problem?

We are dealing with

MILLIONS

of articles from 2018 to 2022

Where was the data sourced from?

AWS OPEN DATA
CORD-19 DATASET

Full-text and metadata dataset of **COVID-19** and **coronavirus-related research articles** optimized for machine readability.

FEATURES USED:

YEAR ABSTRACT

WHAT IS THE FILE SIZE?

135.9 GB

QUESTION?

How might we use **topic modeling techniques** to **detect and label nascent research themes** that could become crucial **in managing ongoing and future COVID-19 variants** or other pandemics?

How will we do it?
Project Methodology

STEP 01

TEXT PRE-PROCESS

Tokenize, remove stop words, and **vectorize** words

STEP 02

MODEL TOPICS

Latent Dirichlet Allocation (LDA) to **extract topics**

STEP 03

VISUALIZE RESULTS

Use pyLDAvis to **visualize topics** LDA generated

STEP 01

TOKENIZE TEXT DATA

Breaking down whole text data into **individual words**

[the, antiviral, drug]

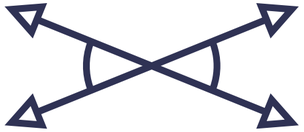
STEP 02

REMOVE STOP WORDS

Remove **extremely rare** and **overly common words**

[antiviral, drug]

STEP 03



VECTORIZE WORDS

Represent words as numbers to match model input

antiviral = [1]

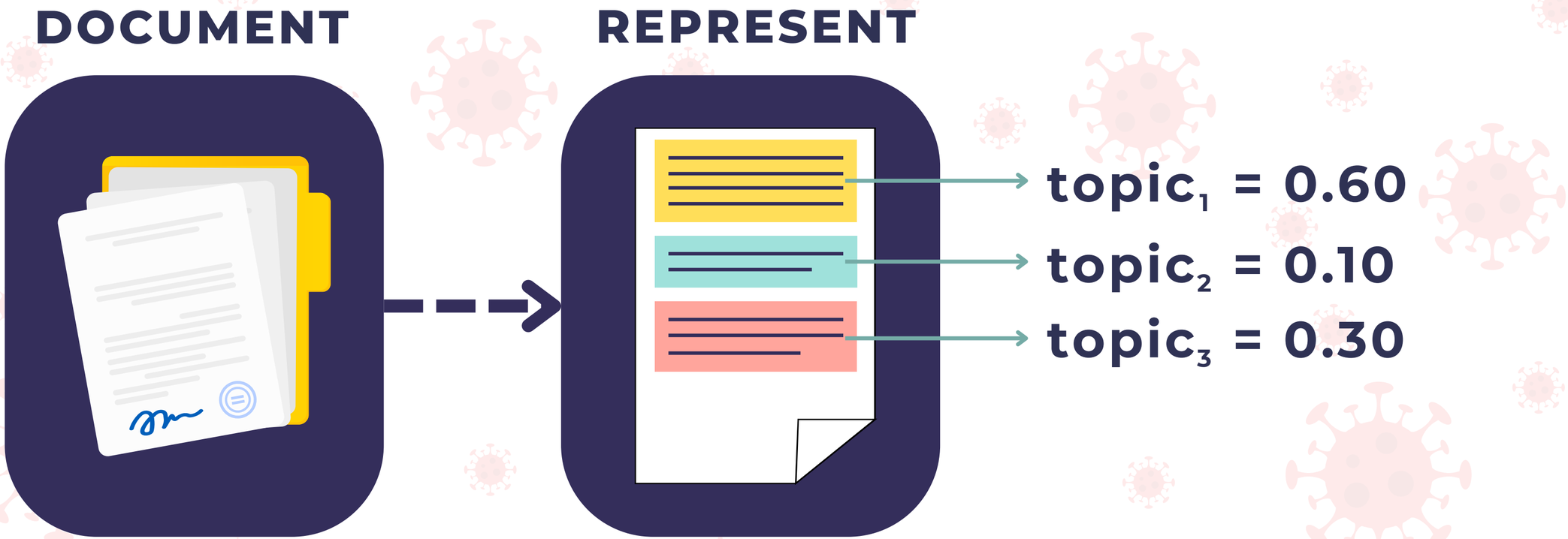
How did the team pre-process the data?

We have research articles and we want to sort them into topics.

LATENT DIRICHLET ALLOCATION

the machine that will extract topics from the millions of articles

Latent Dirichlet Allocation



LDA describes a document as a blend of topics, assigning percentages to reflect each topic's contribution.

TOPIC 1:
MENTAL HEALTH

DOCUMENT 1 CONTAINS:

VACCINE DEVELOPMENT	MENTAL HEALTH	VIRUS TRASMISSION
0.10	0.80	0.10

DOCUMENT 1

DOCUMENT 2 CONTAINS:

VACCINE DEVELOPMENT	MENTAL HEALTH	VIRUS TRASMISSION
0.00	0.50	0.50

DOCUMENT 2

TOPIC 2:
VACCINE DEVELOPMENT

DOCUMENT 3

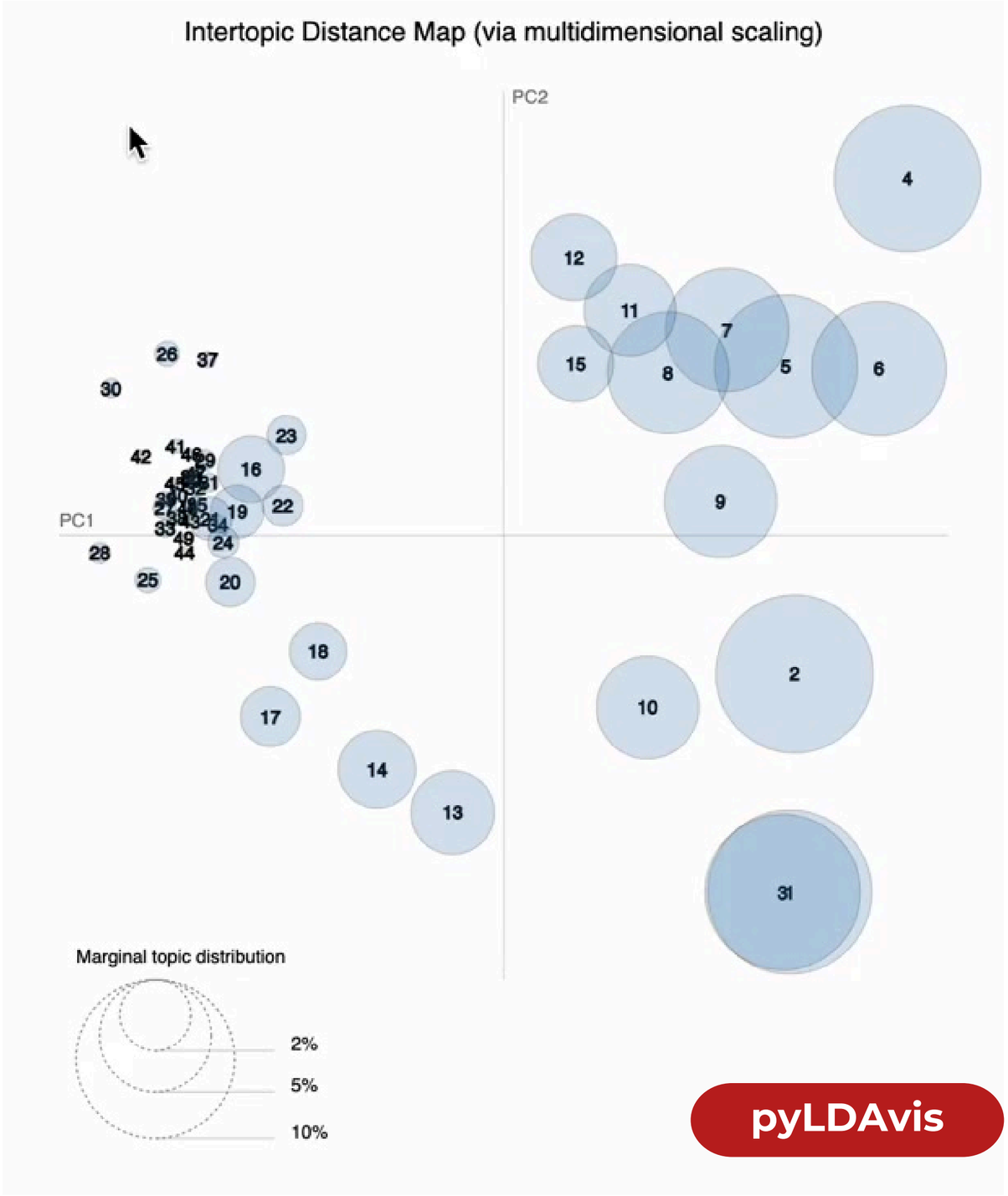
DOCUMENT 4

TOPIC 3:
VIRUS TRANSMISSION

Example: **10%** of Document 1 is about vaccine development, **80%** is about mental health, **10%** is about virus transmission

Latent Dirichlet Allocation

LDA gives us the relative importance or contribution of each word to the topic. The higher the weight, the more important the word is for defining the topic.



Criterion to choose topics to present

TOPIC DISTANCES

Eliminate the risk of getting the **same insight for one year**

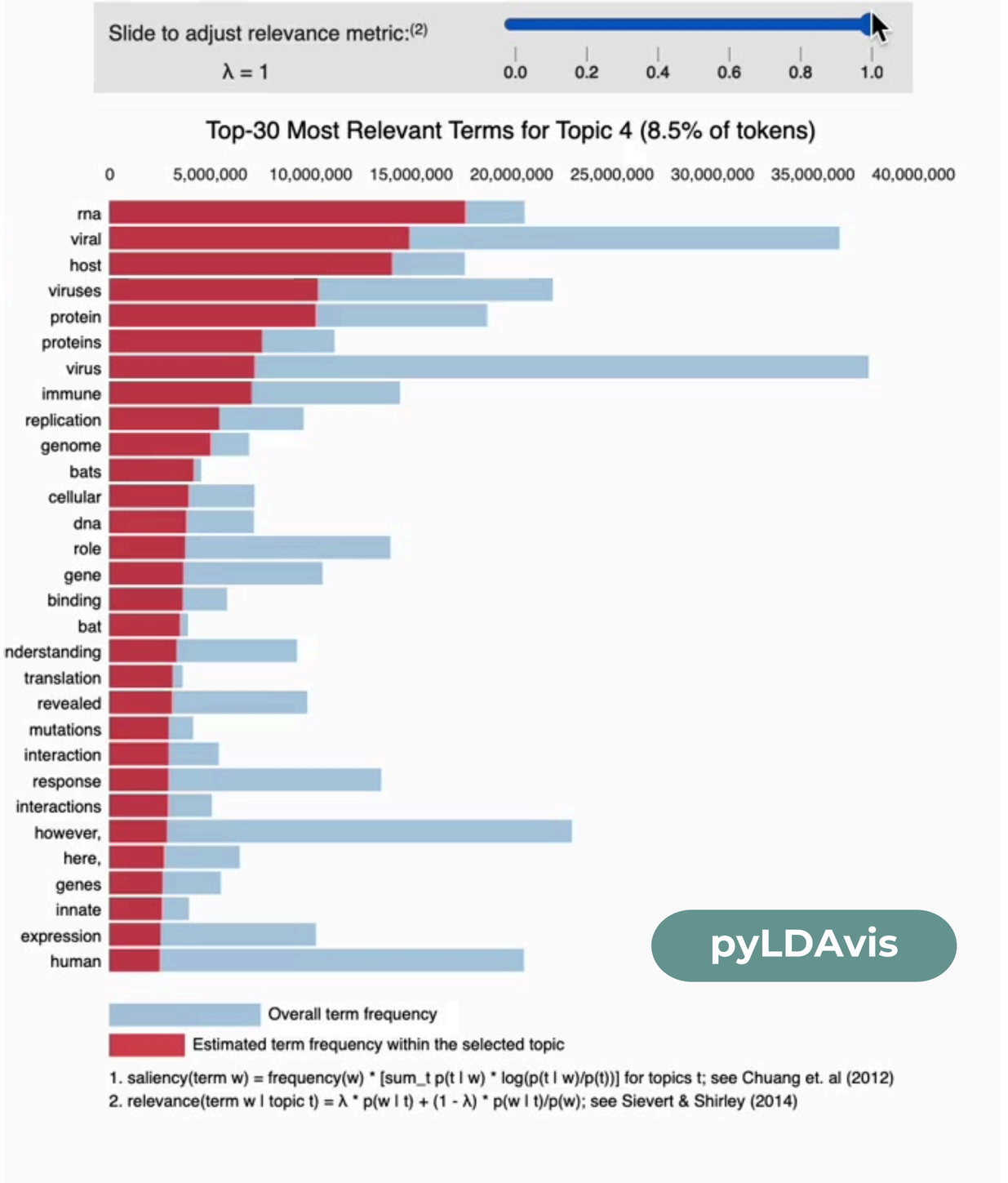
TOPIC SIZES

Maximize marginal topic distribution, larger cluster is better

Words we want to represent topics?

WORD BALANCE

We want words that appear that are **common within a given topic**, while also being relatively **uncommon across all the other topics**.



COVID TIMELINE

2018
PRE-COVID

Research was business-as-usual, like general surgery, immunology, and previous epidemics, among others.

2020
EARLY COVID

Focuses **understanding COVID-19,** its **risks to populations** and impacts on **mental health and educational systems**

2022
LATE COVID

Research was mainly on COVID-19 **recovery**, its **timeline**, and its **global socio-economic impact**.

Timeline illustrates the evolution of research focus from general health topics in 2018, intense COVID-19 studies in 2020, to examining its long-term impacts in 2022.

2018

PRE-COVID

Majority of articles focus on general medical topics such as general surgery and immune system activation

TOPIC 1
GENERAL SURGERY

SURGERY
LAPAROSCOPY
RESECTION
ABLATION
POSTOPERATIVE

A Systematic **Review of the Incidence of and Risk Factors** for **Postoperative** Atrial Fibrillation Following **General Surgery**

TOPIC 4
IMMUNE SYSTEM

CELLS
MACROPHAGES
IMMUNE
ACTIVATIONS
CYTOKINES

H5N1 **influenza virus-specific miRNA-like small RNA** increases cytokine production and mouse **mortality via targeting poly(rC)-binding protein 2**

Minority topics talk about how the Coronavirus has long existed and how it can be a threat in the future

TOPIC 11
MERS-COV OUTBREAK

MERS-COV

INFLUENZA

ZOONOTIC

OUTBREAK

MIDDLE EAST

"It feels like I'm the dirtiest person in the world.": **Exploring the experiences of healthcare providers** who **survived MERS-CoV in Saudi Arabia**

TOPIC 15
CORONA THREAT

THREAT

CORONA VIRUS

SUBJECTS

RECORDED

FUTURE

Detection and Characterization of Distinct **Alphacoronaviruses** in Five Different **Bat Species in Denmark**

2020

EARLY COVID

Most articles focused on understanding COVID-19, its transmission rate and its impacts on different populations and public health systems

TOPIC 1
PUBLIC HEALTH

BARRIERS
COMMUNITY
SERVICES
RESOURCES
ACCESS

Managing enduring public health emergencies such as COVID-19: Lessons from Uganda Red Cross Society's Ebola Virus Disease Response Operation

TOPIC 2
COVID REPORTS

OUTBREAK
CONFIRMED
REPORTED
MONTHS
CASES

Symptoms of COVID-19 Confirmed Cases Presenting to Emergency Department in a Tertiary Care Centre: A Descriptive Cross-sectional Study

TOPIC 3
VULNERABILITIES

AGE
SEX
OLDER
RISK
COMORBIDITIES
ODDS
DIABETES

Are Older Populations at a Disadvantage? County-Level Analysis of Confirmed COVID-19 Cases in Urban and Rural America

Minority topics talk about the impacts of COVID-19 on mental health and educational systems

TOPIC 4
MENTAL HEALTH

PSYCHOLOGICAL
STRESS
ANXIETY
BEHAVIORS
SOCIAL
MENTAL

Impaired mental health status following intensive care unit admission in a patient with COVID-19

TOPIC 5
EDUCATION

STUDENTS
TECHNOLOGY
DIGITAL
TEACHING
ONLINE
REMOTE

Burnout Incidence during Online Lectures in Medical Students in Udayana University during the COVID-19 Pandemic

2022

LATE COVID

Majority of the topics in 2022 were about COVID-19 recovery efforts and the global impacts of the pandemic

TOPIC 1
CLINICAL TRIALS

LEVELS
STATISTICALLY
PATIENTS
DIFFERENCE
SERUMS
CONTROLS

The Fragility of Statistically Significant Results in Randomized Clinical Trials for COVID-19

TOPIC 4
GLOBAL IMPACT

DEVELOPMENT
EFFORTS
SUSTAINABLE
ENERGY
GLOBAL
COUNTRIES

Life in the Time of a Pandemic: Social, Economic, Health and Environmental Impacts of COVID-19—Systems Approach Study

TOPIC 6
COVID TIMELINE

DEATH
YEARS
TRENDS
RATES
INCIDENCES
WAVES
MORTALITY

Impact of COVID-19 on routine immunization in Oyo State, Nigeria: trend analysis of immunization data in the pre- and post-index case period; 2019-2020

Benefits of Topic Modeling and Evolution

THREAT IDENTIFICATION

Checking some of the emerging research topics in the past year

TREND ANALYSIS

Identify popular topics that researchers can focus and collaborate on

FINAL REMARKS!


THANK YOU!