Covid 19 Literature Exploring

Abstract

Since the pandemic started, there have been tons of studies about covid 19. It is difficult for the researcher to keep up with massive information in a short time. The result shows that 26 topics have the highest coherence scores. The keywords for each topic are cancer, financial, economic, vaccine, cell protein, lockdown, symptom, and so on.

Design

This project aims to explore the covid 19 articles and separate the covid papers into different groups based on their abstract. Several representative topics with their keywords are shown in this study. In this way, researchers can get the articles faster and easier based on the keyword they want to search for.

Data

The dataset in the project is from COVID-19 Open Research Dataset Challenge. It has the title, abstract, source, publish time, author, and journal of each paper. There are around 599616 entries/ row in this dataset. The abstract of papers was used for the topic modeling.

Algorithms

Data cleaning- Drop duplicates. Checking the NaN value in the dataset.

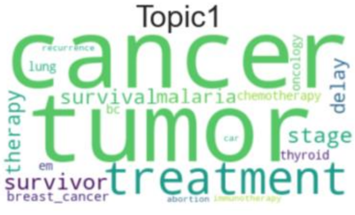
Pre-processing: Langdetect, NLTK- stop word removal, tokenization.

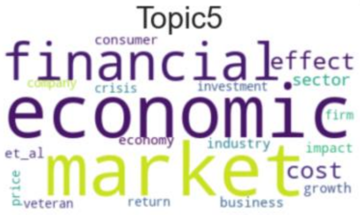
Model- Latent Dirichlet Allocation (LDA)

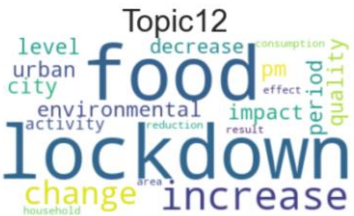
Tools

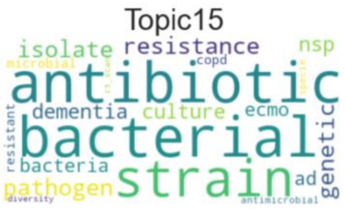
- Pandas and Numpy- data cleaning and modeling
- NLTK, SKLearn, spaCy- removing stopwords and tokenizing
- Gensim: topic modeling
- WorldClouds- visualization

Communication


Topic1
cancer
recurrence
lung
oncology
delay
survival
malaria
chemotherapy
therapy
bc
tumor
stage
car
thyroid
em
survivor
treatment
breast_cancer
abortion
immunotherapy


Topic5
consumer
financial
effect
company
crisis
investment
sector
economic
firm
economy
industry
impact
et_al
market
cost
price
growth
veteran
return
business


Topic12
level
decrease
consumption
food
urban
pm
city
effect
period
quality
environmental
impact
activity
reduction
result
lockdown
area
change
increase
household


Topic15
isolate
resistance
nsp
microbial
copd
antibiotic
ct_scan
specie
dementia
culture
ecmo
resistant
bacterial
genetic
bacteria
pathogen
strain
ad
diversity
antimicrobial