HIERARCHICAL CLUSTERING ON SIMILARITIES OF **COVID** KEYWORDS

BY FENDI HALIM TJOA

HYPOTHESIS

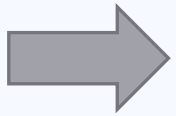
What do abstracts from covid journals say about **PPE**, **medicine**, **disinfect**, **exercise**, and **diet**?

For example, when someone says to use PPE, what kind of PPE should we be expecting to see and what can be inferred from them?

DATA CLEANING ON **ABSTRACT**

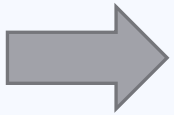
Raw Sentence

OBJECTIVE: This retrospective chart review describes the epidemiology and clinical features of 40 patients with culture-proven Mycoplasma pneumoniae



Regex

Lemmatizing



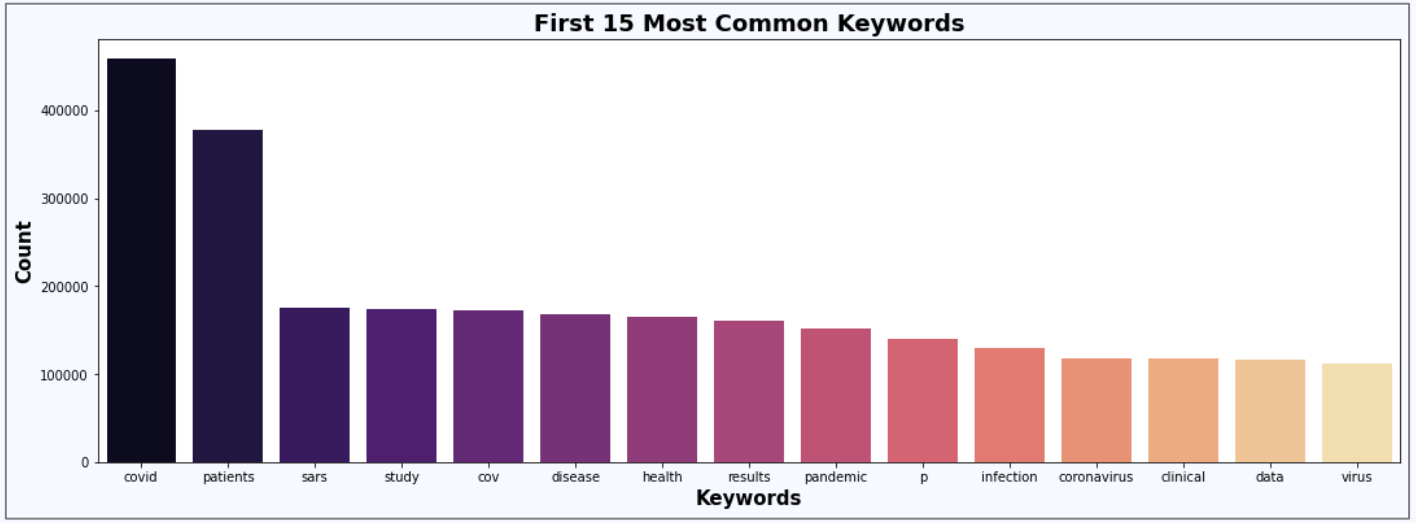
Clean Sentence

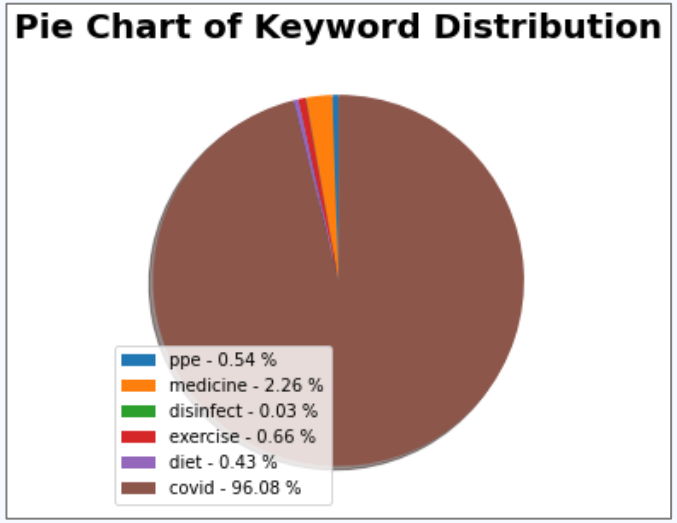
objective this retrospective chart review describes the epidemiology and clinical feature of patient with culture proven mycoplasma pneumoniae

Note: Stopwords were not removed as to keep the dependency information between words for Word2Vec model

DATA VISUALIZATION AND EDA

What are the most common words used? Also, are the keywords selected on the hypothesis show on the most common words?





As expected, **covid** is the highest since the focus of the journal is talking about covid. None of the keywords from the hypothesis are shown in this bar plot.

Keywords from hypothesis are very small compared to **covid**. Based on word count, journals discuss more about symptoms rather than these keywords.

MODEL SELECTION (WORD2VEC)

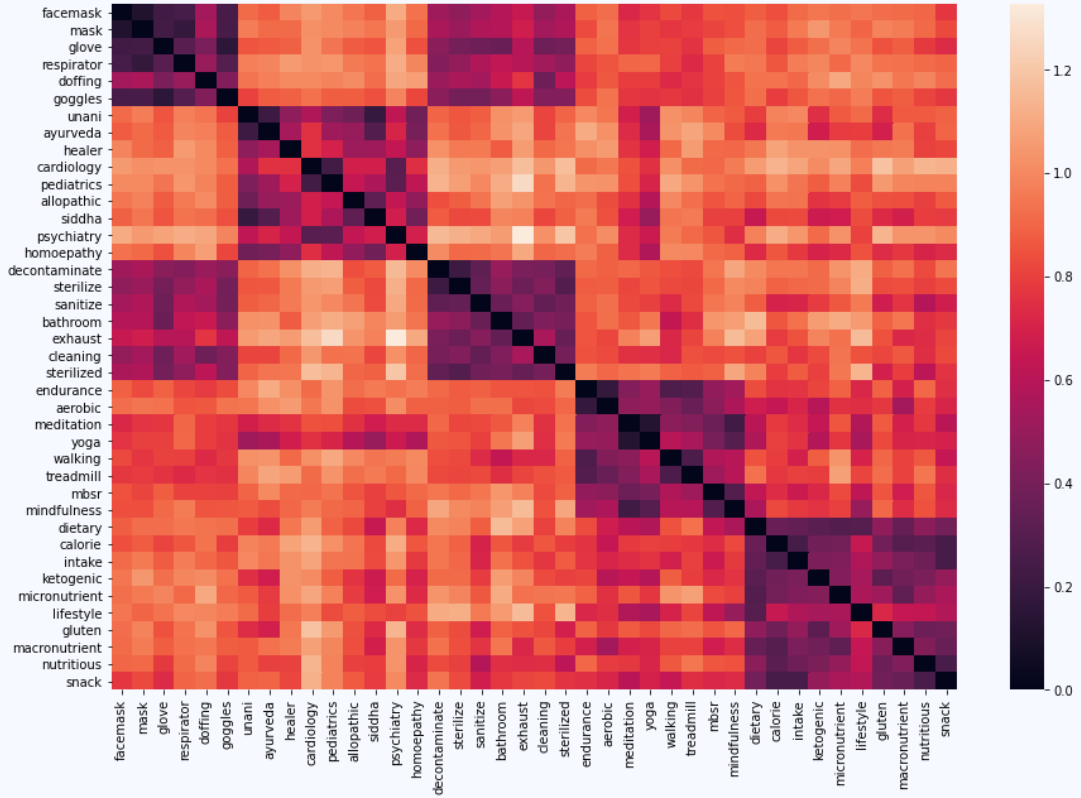
Word2Vec is selected as the model selection for this topic because **Word2Vec** can detect synonymous words. For example, **eat** has a word similarity with **consume**. This model will be utilized with selected keywords from hypothesis.

Word Cloud of Similar Keywords on Each Hypothesis Keyword

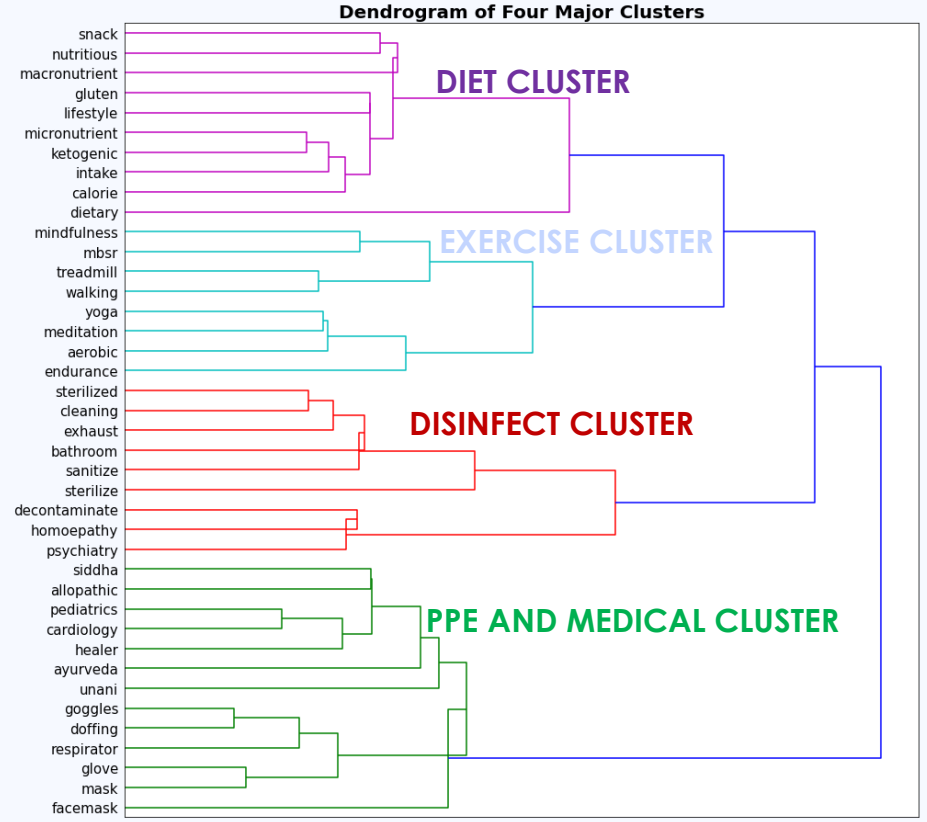
After model is trained, each of the keywords will have a vector. Based on the distance between two word, top 10 of the most similar keywords (per hypothesis keyword) can be compiled and visualized into a word cloud. For example, the keyword **ppe** has similarities with **mask**, **respirator**, and **glove**.

Based on the distance, clusters can be created. If the distance between two keywords are low, it means they have high similarity in terms of meaning

HIERARCHICAL CLUSTERING WITH DENDROGRAM



The heatmap shows cosine distances between words. Dark grid means low similarity and light grid means high similarity. There are *mini* clusters from the heatmap, but the real clusters should be created using dendrogram (hierarchical clustering).



Four major clusters are drawn on the dendrogram. The purple cluster represents **diet**, light blue represents **exercise**, red represents **disinfect**, lastly green represents **ppe** and **medical** combined.

FINDINGS AND INSIGHTS

- **Diet Cluster:**
  - There are many other popular diets such as vegetarian, paleo, raw, but **ketogenic** is the diet that shows up. From this, ketogenic can be a topic that can be further investigate that's related to covid
- **Exercise Cluster:**
  - The exercise mentioned are focused on light activities such as **walking**, **yoga**, **meditation**, and **mbsr**. These give some insights to investigate more towards light activities that easily be done by majority of people.
- **Disinfect Cluster:**
  - This cluster is focused on disinfecting, cleaning, and sterilizing. There is also a word **exhaust** which can be implied there should be ventilation. Researches of these journals can collaborate to further investigate in creating sterilized environments.
- **Medical and PPE Cluster:**
  - **facemask**, **gloves**, **goggles**, and **doffing** are the **PPE** that can be used to prevent getting infected by covid. For medical part of this cluster, there are some alternatives medicine mentioned such as **siddha**, **ayurveda**, **healer**, **unani**. From this, there are researches on alternative medicines to see if they can be scientifically proven to combat covid-19.

Using this dendrogram / hierarchical clustering, anyone can identify which keyword belongs to which cluster. This is useful for researchers as they can collaborate to further research and investigate on similar covid topics (same cluster).