Eve Kazarian
Math 189Z: Covid-19 Data Analytics and Machine Learning
April 2020

Project Proposal

Team: Alex Bishka, Eve Kazarian
Research question: What is known about a vaccine for COVID-19?

Methods: We plan on looking through the literature on COVID-19, filtering for documents pertaining to vaccination. Once the documents that are irrelevant to our research are filtered out, we plan on performing PCA and LDA on the data.

After using PCA to separate the dataset of research projects into topics, we will choose the number of dimensions of our PCA to be three and get a visualization of the topic distribution in three dimensions. We can then analyze the articles in the topics to vaguely discern each topic's category. If we find a cluster about vaccines, we will look specifically into that cluster to get the latest research in that area.

We are building off an already-existing literature clustering project on Kaggle but modifying it to plot in 3D, changing our choice of stop words and number of clusters, and where exactly in the clusters we do our analysis.

Sources:
Dataset

Article clustering kernel on Kaggle: https://www.kaggle.com/maksimeren/covid-19-literature-clustering

> *Summary:* The author of this project used PCA and LDA to visualize the topic clustering of research articles about COVID-19. After visualizing the distinct clusters, they looked into each of them to assign topics. Topics ranged from surveillance to therapeutics.

3D PCA on Kaggle:
https://www.kaggle.com/luisblanche/cord-19-use-doc2vec-to-match-articles-to-tasks