COVID-19 Data Mining Lab

Jacob Freiheit

November 2023

1 Abstract

The goal of this lab is to track the relationship between relevance of tweets to certain COVID-19 related topics and the location from which these tweets were made. The first step is to clean and Lemmatize the set of tweets so that they can be more easily parsed and classified by the Cosine Scoring processes. The Lemmatization function used in the lab reduces the words of the tweet to their root form, known as lemma. The cleaning process used eliminates words that can't be used in the text analysis processes used in the lab.

2 Introduction

After cleaning and lemmatizing the tweets in the original dataset, it was necessary to merge superfluous columns. 'state' and 'location' provided the same information, so they were merged. After this, it became a matter of determining the numerical relevance of these tweets to COVID-19 related words. These words have 4 classifications: disinfectant, isolation, medicine, and vaccine. To do this, we were given 4 datasets with lists of words relevant to each of these topics. These lists are then used to count the number of words that are relevant to each classification, which I then used to produce a vector representation of tweet relevance to the COVID-19 topics.

The tweet cleaning took only 34.55 seconds, which was no surprise, given the size of the dataset. The cleaning function implemented ended up with a 75 percent loss of data during the cleaning process. Merging the state and location columns took all of 2 seconds, which was unsurprising, given the simplicity of the process. Lemmatizing all the tweets took longer, at nearly 15 minutes. This was expected as well, given that Lemmatization is a more complicated process and library access is needed for every Lemmatization. The result that wasn't expected was how long it would take to calculate the cosine similarities of every tweet. This process took 9012 seconds, or just over 2 and a half hours. It was expected to take longer, and in retrospect this figure makes sense, given the number and complexity of the cosine similarity calculations, but it wasn't foreseen just how much longer it would take than any of the processes that had been done previously.

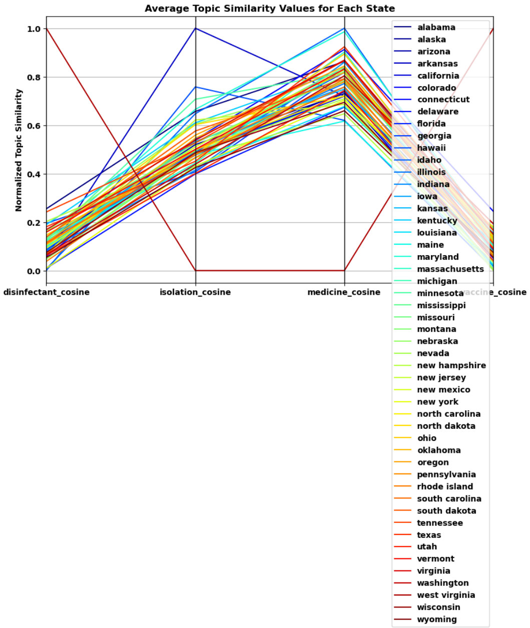
In the interest of not running such

time-intensive code more than once, the cosine similarities were saved into a new csv file that is then accessed by the kmeans and clustering algorithms.

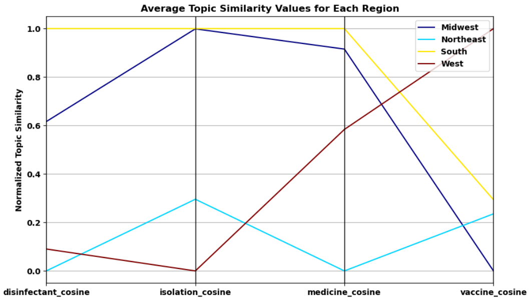
Once these algorithms are run, the Calinski-Harabasz score was taken. These scores represent the separation of clusters from each other (higher scores indicating better defined, more separated clusters). Using the Calinski-Harabasz score, it was found that the highest numbers, or best defined clusters, were found from 4 centers. KMeans came to a highest evaluation of 45858206.175628014 for 4 clusters, and Spectral Clustering found 45858206.17562803. The Spectral Clustering value is negligibly higher in this respect, but a look at other center number evaluations (provided in the python notebook) reveals that KMeans gets better results for more centers. It is for this reason that KMeans in general gets better results.

The idea behind clustering, in this case, is to determine the centers of clusters that best represent the relevance of tweets to the selected COVID-19 terms. The effectiveness of these centers can be determined by how far apart they are from each other. It's in this way that we can determine how different similarly-clustered states are from states in other clusters.

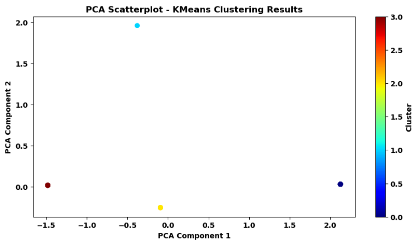
3 Data



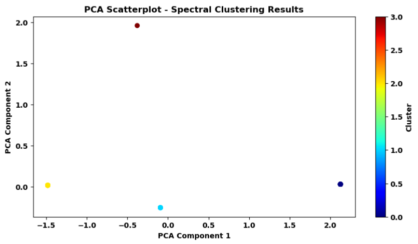
This graph shows the cosine similarity of each state in relation to the COVID-19 terms. This graph was generated from the state topic score data.csv file.



This graph shows the cosine similarity of each region in relation to the COVID-19 terms. This graph was generated from the region topic score data.csv file.



PCA algorithm on KMeans figures



PCA algorithm on Spectral Clustering figures

4 Results

One of the most surprising observations was the difference between the state cosine similarity values and the regional cosine similarity values. The regional cosine similarity values also go against the expected results, those being that the Northeast would rank far higher in all similarity values relative to the other regions.

One of the important things to remember with these graphs is that in the process of normalizing regional and state values, state and regional values are compared to each other, but there must be a state and region whose comparative value must be zero and one for every feature. This means that while these figures might be very close to each other in reality, graphs such as these will make negligible differences seem much larger. This is a major flaw with this form of visualization, and could be improved with a preset smallest and largest value, and then scaled accordingly.

States aren't at all grouped according to their locations. The optimal Calinski-Harabasz values were higher than the sub-optimal ones, often by a factor of over 1,000. The difference in shape between the regional and state graphs shows the lack of correlation between states and their surrounding locations, and and Calinski-Harabasz scores show how far these clusters are from each other.

The PCA graphs for KMeans and Spectral Clustering graphs reveal very little that wasn't already found through the graphs displayed previously. They show the optimal number of clusters, but since essentially the same Calinski-Harabasz score was found for each, the graphs look the same. It's in this way that it is hard to draw a distinction in what is more effective (KMeans or Spectral Clustering) is better based on the information revealed through these graphs.