COVID_19 News Clustering

Team Name: Bogus of the Catastrophic

Nan Gao(ngao), James Li(sli64), Xingjian Gu (xgu2), Hanhui Li(hli34)

Goal

Given the large number of news articles about COVID-19 and the rapid development thereof, it is difficult for audiences to keep up with all the new information. Our project involves using k-means clustering to label new articles based on their topics and contents. The use case for our project would be to help websites recommend to their users news articles that are relevant to what the user has been reading, or help users search for articles that are related to their interests. Hence, the task is as such: given a news article about COVID-19, predict its category (cluster) and return articles with similar contents.

Data

We gathered the news articles from News API. The articles are returned by the API based on the search term ‘coronavirus’, and we use only the news article body for the purpose of this project, meaning that information such as titles, outlets, and time of publication are not included. In total, we used 3566 articles in total, each with a document ID and its contents. Since our task involves unsupervised learning, all data are used in training for the sake of training the model to the best extent. All news articles are in English and are sufficiently clean that some rudimentary preprocessing makes all data usable.

Model+Evaluation Setup

Since the goal is to categorize news articles based on their contents and select articles that are most similar in contents to new testing inputs, k-means clustering is most fitting to the task. In order to properly conduct k-means clustering on our data, we took the following steps. We first used the rudimentary natural language processing techniques – tokenizing, removing stop words, and stemming words – to transform news articles into tokens. We then used tf-idf vectorization to create a matrix to represent how significant each word is to each article. We used principal component analysis to keep the most important features.

With a more workable representation of the initial textual data in the form of a reduced tf-idf matrix, it becomes more viable to apply the k-means model. Since it is hard to predetermine how many topics there are regarding the pandemic, we determined the optimal number of centroids to use in k-means by comparing the distortion between using different numbers of centroids. Finally, we used that number of clusters to categorize all articles, yielding the output model. Given that the situation of the pandemic is ongoing, we trained using all available news articles, and its usage can be tested with articles published thereafter.

Results and Analysis

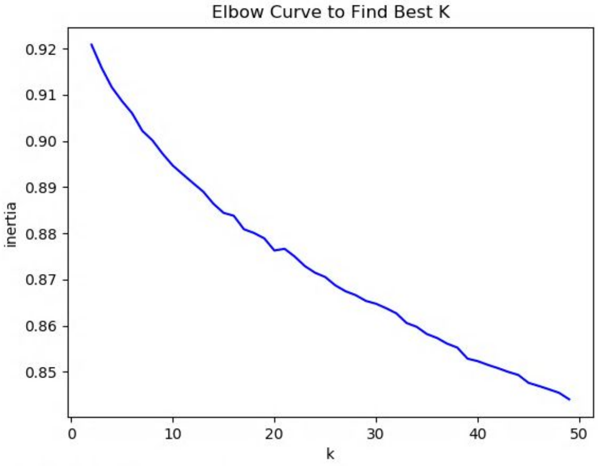
(Good projects should have more than one claim/observation. The first should report the performance of the model against some relevant baseline. The following should offer insight into features, error patterns, overfitting, etc.)

Claim #1: k-means clustering based on tf-idf matrix categorizes similar articles together.

Support for Claim #1: Using Jaccard similarity, we conducted quantitative analysis on the similarity between articles from the same cluster, in comparison to articles from separate clusters. On average, articles from the same cluster have Jaccard similarity of 0.126, as opposed to 0.061 for articles from different clusters, which is relatively distinct.

Claim #2: Using 20 centroids in k-means balances between accuracy and risk of overfitting.

Support for Claim #2: Naturally, more clusters necessarily mean lower loss, and assigning each data point to its own cluster reduce loss to 0. However, having more clusters dilutes the information and we learn less about underlying structures. To find the optimal number of clusters, we plot the inertia of each k, from 2 to 50, and it displays an elbow curve. We choose the k at the inflection point, which is around 20, as the optimal number of k.



Claim #3: Word clouds and scatter plots are suitable methods of visualization for this project

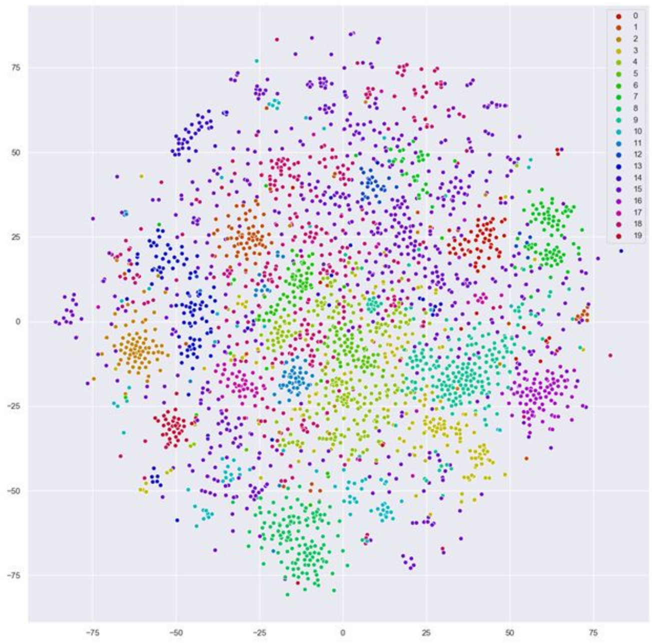
Support for Claim #3: For visualization, we chose word clouds and scatter plots. We created word clouds for the 20 clusters, and compared the differences between word clouds generated from articles of the same cluster and those generated from different clusters (see attached image). T-SNE scatter plot provides a good visual representation of text data, showing data points that are similar in features forming clusters.

Claim #4: Many clusters do capture semantically meaningful topics

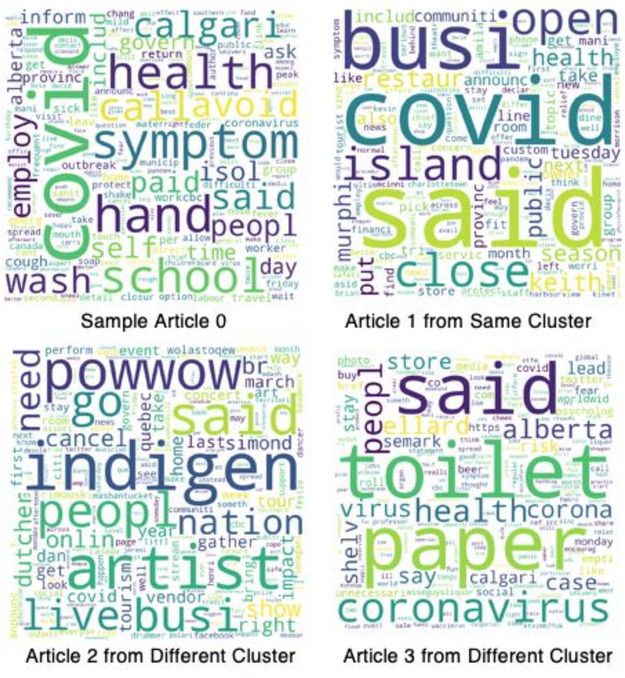
Support for Claim #4: By randomly sampling a few articles from each of the clusters to conduct qualitative evaluation of the efficacy of our model, we found that the clusters do capture meaningful topics (which is rather saliently demonstrated in the word clouds). Some clusters focus on the information about the virus itself (its symptoms and protections against it), while others focus on societal impacts or political responses. This suggests that the unsupervised learning based on relative word frequencies does to some extent captures the topics of the news articles.

Claim #5: Some clusters are less distinct than others

Support for Claim #5: Judging from the t-SNE scatterplot, it appears that there are some data points that are relatively distant from all nearby data points, suggesting that even though it is categorized into a cluster, its topic



might not be as closely related to other articles in the same cluster. By extension, it might suggest that some clusters have a less distinct topic, compared to clusters with data points closely packed together on the scatterplot.



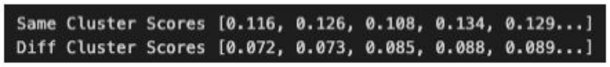
Sample Article 0

Article 1 from Same Cluster

Article 2 from Different Cluster

Article 3 from Different Cluster

Word Cloud Visualization



```
Same Cluster Scores [0.116, 0.126, 0.108, 0.134, 0.129...]
Diff Cluster Scores [0.072, 0.073, 0.085, 0.088, 0.089...]
```

Jaccard Similarity Comparison