Erik Lovece

Rabya Ghafoor

**Project Memo**

Executive Summary

This project collects spoken news snippets on COVID-19 and investigated the patterns of media-directed health communications and policy measures that are being discussed in response to the ongoing COVID-19 crisis. It consists of two major parts, both devoted to its own workflow: examining term weights as terms ebb in and out of importance as the pandemic progresses and overall topic relevance comparisons. These parallel workflows are credited to Shravan Kuchkula in his article on topic modeling using NMF and LDA (link in the analysis notebook). In addition to implementing his method, we also extend his observation technique to cover a set of datasets (his example covers only 1 feature matrix) to create time-series observations of term weights and topic relevance.

We find that terms and the topics containing them do not always remain in the same position in the hierarchy of relevance. Term weight varies at least slightly per month, and the topics themselves do not remain at the same levels of importance relative to other topics throughout the observed models. What does remain consistent is the dominance of the top topic of any given month throughout the corpus, whatever its subject matter. Overall, the hierarchy of topics themselves varies through time, but the dominance of the top topic in the corpus does not.

Introduction

As of this writing, the COVID-19 pandemic has infected 2,360,185 people in 148 countries, resulting in 161,985 deaths and now has the potential to reach a relatively high proportion of the global population. The crisis has already had an unprecedented effect on the economic and labor markets, impacting the supply and demand of goods and services. The disruptions in production, that initially started in Asia, has now spread to supply chains across the world. All the businesses regardless of size are facing great challenges with respect to decreasing revenue and job losses in specific sectors. Following travel bans, border closures, and quarantine measures, many worked are unable to move near their places of work and carry out their jobs. This challenge has known effects on incomes especially for hourly wage employees. Given the effects on income consumers are unable or reluctant to purchase goods, causing a delay in investment, production of goods and the hiring of workers.

In order to limit the direct health effects of COVID-19 on workers and families and the indirect effect of the economic fallout, immediate and coordinated policy responses are needed at both at the national and global level. Disease prevention and protecting workers and families from the infection should be a topic priority for all nations across the world. Demand-side measures to protect families who are facing income losses is needed to stimulate the economy. At the community and global level, transparency within the government and social dialogue between governments, workers, and employer organization is key for developing a sustainable solution. Previous crisis such as the Great Recession have shown that we can break through this vicious cycle through coordinated and policy measures.

This topic is of interest because it will not only enable countries to cope with the current crises but also inform our actions in response policy measures, could not only illuminate the severity of the current situation but also bridge the knowledge gap of how countries should operate economically and politically during a global pandemic. This investigation hopes to shed light on several key issues in such as the type of institutional and policy reforms required, multilateral leadership, and social dialogue between national leaders and government agencies. Without understanding the key topics found in textual data and how the topic weights change overtime, it would be exceedingly difficult to assess the changing nature and scale of current crisis in the world due to COVID-19.

Methodology

We first adopted a dataset of coronavirus verbal news mentions from the Global Database of Events, Language, and Tone (GDELT) to extract related news mentions of coronavirus from major press media between January 1, 2020 and March 10, 2020. Data preprocessing consisted of converting the dataframe to a tidy format conforming to conventional practices thereof, then cleaning the textual data in the “snippet” column. A special consideration of ours was controlling for bias in news sources to minimize the nuisance variable of reporting bias by hard left or right political viewpoints incorporated into reporting directly. To best accomplish this, we eliminated observations generated from news sources with known hard-left or right biases, namely MSNBC, Fox News, and Fox Business. We then removed columns that were not pertinent to information organization, like URLs and preview thumbnail image filenames.

Another consideration or ours was counterintuitive; although the data is explicitly about the coronavirus, we decided it would be best to eliminate words like “coronavirus”, “virus,” “covid,” and similar terms from the corpus of snippets. The reason for this was that the snippets are already, by virtue of their source, about the coronavirus. Our aim for this project was to read into ancillary effects of the pandemic and topic model based on those. The implied topics of “coronavirus” and similar take a large precedence in topic weights over other data and detracts from a study of the effects we wish to investigate. That the news reports concern the coronavirus in some way is already abundantly clear, so we only observed terms aside from these implied terms.

Our final consideration was that we wished to study what was considered important in the spoken news media over time. We decided that organizing the data by month was the best way to separate the data with tens of thousands of observations per month. To accommodate this, we split the parent table into 3 smaller tables, one for each of the months of January, February, and March.

After data preprocessing and cleaning, we initiated our workflows, which were as follows:

Cleaned data 🡪 TF-IDF 🡪 NMF 🡪 Visualization

And

Cleaned data 🡪 CV 🡪 LDA 🡪 Visualization

Though different versions of both workflows were tested at each stage.

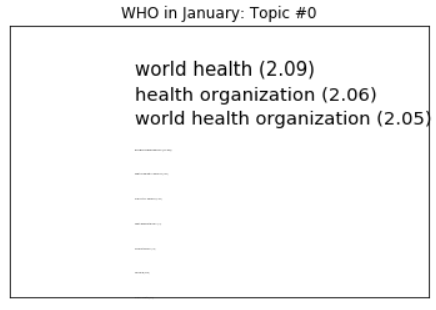
From the cleaned corpus of snippets, we had a total 4 experiments to choose from depending on which sequence of feature extraction and decomposition was chosen. We transformed the textual data corresponding to each month to all 4 possible pipelines. For feature extraction, we tested different ngram ranges for empirical analysis of comprehensiveness and overall ability to communicate an idea of the topics, settling on bigrams and trigrams due to both how uninformative and unhelpful the unigrams were when observed individually and the prevalence in the corpus of multi-word proper nouns such as “Donald Trump” or “World Health Organization,” the latter term being a subject of interest later.

We then fit data from both vectorization methods into decomposition via both NMF and SVD decomposition methods to decide which version of the data we will visualize. After transforming the 6 feature matrices by the decomposition methods (3 from CV, 3 from TF-IDF), we printed the topics with the appropriate weights adapted from their associated topic distribution matrices, and saw how the terms were scored in each topic. From NMF, we noticed that top terms in topics tended to weigh around 4.5 as opposed to around 2.5 from SVD. Since LDA performs calculations using term frequencies we opted to use our CV data for LDA, despite also testing from TF-IDF 🡪 NMF/SVD and CV 🡪 SVD. TF-IDF 🡪 NMF also provided an interesting study for the “World Health Organization” trigram, which we observed through Jan-Mar.

To demonstrate the changing relevance of a term over time, we took this trigram and visualized its output in a matplotlib layout. Using this model, we observed this term decreasing in relevance from Jan to Mar and noted that its weight decreased over time and the topic housing it was superseded by more topics as time passed.

Before we could move forward with LDA analysis, we wished to find the best hyperparameters to set for the model. We tested the CV data from each of the 3 months to find the best fit for LDA, judging by the performance indicators of log likelihood and perplexity. Although the data pertaining to March provided for what was easily the highest-performing model, we could test on all 3. This step would be more helpful had we used a grid search technique to find the exact best parameters to use, which takes a high amount of processing time. In that case, selecting the CV data that performs best would be highly beneficial since it is less likely that an extensive grid search would produce data that is inaccurate or not useful.

Results

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedOur results can be split between the individual term weights from the TF-IDF🡪NMF workflow and the comparative topic relevance from the CV🡪LDA workflow. In the former case, we demonstrated our technique on the “World Health Organization” trigram, which steadily decreased in weight as time passed. Not only that, but the topic the term was housed in also moved down the list of topics, decreasing in importance over time. We suspect this correlates with the changing public perception in the organization over time. At the pandemic’s outset, the WHO was highly sought after as a source of information when the public started to take this issue seriously. As each country proceeded to pass measures suitable to their environments, the need to look to the WHO for direction on how to adapt to the pandemic decreased. The term-weight decrease is not necessarily due to public ignorance, even with the ongoing movement to reopen many non-essential facilities ahead of schedule. We hope that other in-depth inferences could be made using other term weights as well as they too are monitored over time.

Our other main workflow produces a comparison of topic relevance in a chosen month. After fitting our CV data to the LDA model, we produced seaborn count plots of every time a given topic had the highest relevance score according to Kuchkula’s model. As expected, Topic 0 was most often the dominant topic. Whatever the subject of these news snippets, they value Topic 0 at a higher relevance more often than any other individual topic. This model will be much more useful when the corpus is composed of news articles than news snippets, since these will give a broad idea as to the subject matter of each article. Still, it is highly apparent that the spoken news regarding the coronavirus chiefly weighs a single topic so heavily. Further inquiry as to the subject matter of the spoken news each month could be an interesting basis for further study.

