Analyzing COVID-19 Search Trends and Hospitalizations

Connia Ren
Isabela Dragomir
Paloma McAuley-Fernandes
October 21, 2020

1 Abstract

2 Introduction

In late December 2019, doctors in Wuhan,China discovered a novel, pneumonic viral disease that after several months had spread to over 8 million cases worldwide with the largest case numbers in the United States of America. COVID-19 has led to changes in public attitudes and opinion on health, along with a direct increase in related hospitalizations. One way to measure and predict these changes is through popular search engine query data, which has been well attempted for previous viral outbreaks like SARS. (Ginsberg et al., 2009) Although this type of analysis is essential for early detection in pandemic outbreaks, its validity has been critiqued and questioned for its direct relationship to COVID-19 trends. (Sousa-Pinto et al, 2020) To explore this technique further, we will be visualising COVID-19 search trends from the past seven months in various U.S. states and using this data to build supervised and unsupervised machine learning models to predict new regional hospitalizations.

The datasets retrieved for this project comes from Google Trends and includes the US search trends symptom dataset and the global aggregated open COVID-19 dataset. The prior tracks the weekly search trends of various health symptoms from January-September 2020 in sixteen different U.S. states. The data was scaled according to the highest searched symptom in that region from a range of 0-100. The latter is an open-sourced, amalgamated set of various daily COVID-19 related quantities from numerous countries and U.S. states. In this project we will be focussed on the daily new hospitalization counts from March-September 2020 in the U.S. states of interest.

3 Dataset

The first dataset used in this project were the COVID-19 Symptoms search trend dataset, documenting hundreds of broad health symptoms that were generally based on a symptom's prevalence in Google's search engine by region and weekly time intervals from January 6,2020 to September 27,2020. Each symptom was quantified from 0-100, which can represent the normalized popularity of that symptom scaled to the most popular symptom in that region across the entire time range. The second dataset used was the open COVID-19 dataset from Google Research which aggregated many public-sourced COVID-19-related data including cases, deaths and hospitalizations over a daily time series for several countries and regions.

For this project the focus was on combining the hospitalization data from the open dataset with the search trends dataset. Since the open dataset was substantially larger and complex, further exploration of this dataset was done first before any processing. This dataset was filtered by locations with particular emphasis on the United States and regions within it. It was then noted that data was sourced both from aggregate country-wide and state specific data. The data was filtered further by the 16 states of interest from the search dataset and the columns by which related to hospitalization data were given emphasis. The substantial information came from the new and cumulative hospitalizations which were both included in the final project dataset.

In the effort to preprocess the data, insignificant regions were removed which had insufficient hospitalization numbers relative to publicised case numbers such as District of Columbia. Whilst states with few hospitalizations but that were reflective of their reported COVID-19 death toll were kept such as Alaska. This led to a total of

13 states included in our final analysis. The hospitalization data was also aggregated by week to be comparable to the weekly symptom search dataset by summing all values for a given week range found in the search dataset. This left a total of 30 weeks of data kept between March-September 2020. The weekly search trends data was noted to have many symptoms which lacked any sufficient information and made it noisy. It was cleaned by removing insignificantly reported symptoms which had less than 50% of data across all time spans and regions. This left us with 24 significant symptoms with two in particular that had substantial data across all regions. No regions were removed based on search data. The data had to be re-scaled as it was noted the dataset was normalized and scaled specifically to each region making it incomparable across regions. To do this, we took the median value of each symptom for each region and subtracted the median from each datapoint hence standardizing the median for each region to 0. The median was used unconventionally over the mean as it was uncertain whether certain symptoms were less or more popular than others and to prevent any outliers from skewing the data. The cleaned search data and processed hospitalization data were then merged together into one dataset.