|  |  |
| --- | --- |
| **Insights of Covid-19** | |
| Contact | Name: Runsheng Wang, Jinqi Lu  Email:[runsheng@bu.edu](mailto:runsheng@bu.edu), [jinqilu@bu.edu](mailto:jinqilu@bu.edu)  Cell Phone: N/A |
| Organization | N/A |
| Organization Description | N/A |
| Project Type | Data Science |
| Project Description | We seek to unveil COVID-19 insights through the compilation and analysis of numerous datasets. We’ve divided this project into three parts. The first part is a cluster and classification algorithm that assess the risk level of a patient given its characteristics. We will use a train/test split to validate our results in order to measure our success. We will also create a GUI where risk indices can be calculated from by our algorithm for given inputs. The second part is a sentiment analysis of twitter/reddit posts with a correlation analysis with COVID-19 cases. We will first record the ratio of the sentiments on a daily basis, before attempting to analyze the correlation between cases and sentiments. The third part is a multiple linear regression on covariates that could potentially have an impact on COVID-19 cases. We will evaluate our results using standard methods such as Q-Q plot, residual plot, and construct validation curves to measure our performance. If time permitting, we hope to perform an image recognition analysis using convolutional neural networks. |
| Data Sets | * COVID general case data (e.g. new cases, death, etc) from JHU and CDS, combined with datasets containing our potential covariates of choice (e.g. surface temperature, air humidity, direct sunlight) from Kaggle and NOAA. * Patient specific data from Kaggle and other hospital/research sites that publish open data. These datasets could contain age, race, gender, medical condition, income level, pregnancy, etc * Crawled or existing social media data (twitter & reddit) that contains sentimental data about public’s attitude for COVID-19 (positive vs negative) * COVID-19 data with geographical coordinates to assess the network of spread of the virus within a community. We will also collect geographical information to potentially construct maps to illustrate the spread * Images of people wearing facemasks, for image recognition purposes |
| Suggested Steps | We will first download and scrape the required dataset. We will then perform cleaning and combine everything into one big dataset. We will apply the following six techniques to analyze our input datasets   * clustering * classification * regression * sentiment analysis (news, reddit, twitter, analysis with emojis) * forecast/prediction * Potential image recognition |
| Questions to be answered in Analysis | # Very specific questions that the clients wants answered   * **What factors make a patient more likely to contract COVID-19?**   + Clustering & Classification, k-means++, GMM, kNN, etc   + Nice looking user input table that predicts the risk index for given user input   + Covariates include: Age, Race/ethnicity, Gender, Some medical conditions, Use of certain medications, Poverty and crowding, Certain occupations, Pregnancy * **How do people view COVID-19 on social media platforms?**   + Sentiment analysis of reddit/twitter, crawled data or existing data   + positive to negative ratio, plot with x-axis as time   + Correlation with COVID-19 cases (use data from two weeks in advance to reflect the time lag in attitude, due to 14 day incubation period) * **How does factors contribute to increases in COVID-19 cases?**   + Using COVID-19 cases as the response variable, transform data, apply LASSO feature selection, fit multiple linear regression, evaluate results   + Covariates include: surface temperature, air humidity, direct sunlight, social media sentiments (keep if useful), airport traffic, population density, lockdown status, number of testing (hidden cases)   Potential extensions if time permitting   * **How does covid spread within a community?**   + Using longitudinal and latitudinal data, construct networks with nodes as patients and edges describing relationships between them. Using both real and simulated conditions, assess the path of spread of COVID-19 with in a community, verify with ground truth * **Image recognition**   + Recognizing whether a person is wearing facemasks or not   + Recognizing potential COVID-19 symptoms using CT scan |
| Additional Information | Make inferences/draw conclusions based on results from analysis |