**Covid 19 – A complete case-study (The past, present and future)**

**Problem Statement:**

**The outbreak of Covid-19 is developing into a major international crisis, and it's starting to impact important aspects of daily life with emergencies and lockdowns across countries.**

**The objective of the project is to build a strong model that predicts the progression of the virus (confirmed cases, recovery cases and deaths) in 2nd half of April 2020 to help with mitigation efforts.**

**Past - Stats of happenings/ events till now**

**Present - Tapping sentiments and social media interactions (twitter)**

**Future - To gauge the impact in the immediate 1 month on 3 metrics: Cases - confirmed, recovered and death**

Background

1. [Primary Stats](https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6)
   1. Total cases thus far - Deaths and recovered
   2. Analysis on active cases
      1. By country
      2. By severity of the cases - mild, serious etc.
   3. Case fatality rate - # of deaths over cases confirmed by country
   4. New cases map by country (world map)
2. Understanding social media interactions from tweets
3. Data extractions with a couple of hashtags - covid19, coronavirus, quarantine life, coronavirusimpact, coronavirusoutbreak, #lockdown, #staysafe, #staysafestayhome, #stayathomechallenge, #remoteworking, #thistooshallpass
4. Twitter has an official API called OAuth, a token-based authentication system that indexes tweets that match a given search string and writes the output to a file.***Limitation -* *only past week tweets***
5. EDA on tweets
   1. Explore themes from other hashtags within the tweets like sympathy, requests for help, instructions, information, cautionary messages, panic, witty etc.
   2. Word frequencies and associations
6. Clustering and topic modelling
   1. To make meaningful interpretations from huge volume of tweets - Cluster similar tweets together through hierarchical / K-means clustering
   2. To deduce themes of text via topic modeling with LDA
7. Sentiment timeline - human behavior (when the west didn't contract yet. <Bonus>
8. Predictions
   1. To gauge COVID-19 cases on three metrics- confirmed cases, recovered cases and death events for the next month using historical data as on a given date.
   2. Modeling:
      1. Technique: Linear Regression, SVM
      2. Metric: R-squared and RMSE

*Deaths are the most useful data points for these analyses. For example, if modelers assume a case-fatality ratio of 1 percent, and that it usually takes 15 days for an infected person to die, then they know a death reported today in a specific region means that 100 people were likely infected there 15 days ago. Add in the time it takes cases to double—Edmunds says it seems to take five days—then modelers can estimate that over those 15 days the number of cases swelled to 800. So, for every death in a region, that means about 800 others are already infected, most of whom will not have been identified. This pattern was verified in Italy, Edmunds says, which as of today has reported* [*12,462 cases and 827 deaths*](https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases)*. When officials tested people living near where someone had died from the disease, in many cases they found hundreds of others were already carrying the virus.*

References: Worldometer

Primary data source - John Hopkins University CoronaVirus data stream

Coding references:

<https://www.analyticsvidhya.com/blog/2016/07/capstone-project/> <chennai floods>

<https://github.com/Jefferson-Henrique/GetOldTweets-python> <chennai floods>

<https://github.com/datasets/covid-19>

<https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data>

<https://www.tableau.com/covid-19-coronavirus-data-resources>

Data github : <https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data>

Data : <https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset>

Data for modeling: <https://www.kaggle.com/c/covid19-global-forecasting-week-4/data>

Demographic related: https://ourworldindata.org/coronavirus

Data visualizations:

<https://towardsdatascience.com/how-i-built-a-dashboard-with-dash-and-plotly-after-being-stuck-in-europes-worst-coronavirus-dc41aaeeca4b>

Demographic data

<https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30026-1/fulltext#seccestitle140>

<https://docs.google.com/spreadsheets/d/1Gb5cyg0fjUtsqh3hl_L-C5A23zIOXmWH5veBklfSHzg/edit#gid=447265963>

Tweets cleaning:

<https://towardsdatascience.com/selenium-tweepy-to-scrap-tweets-from-tweeter-and-analysing-sentiments-1804db3478ac>