FACULTY OF
COMPUTER SCIENCE

Data Science with R - Project Proposal
"ANALYSIS AND PREDICTIONS OF THE IMPACT OF COVID-19"

Anne Rother, Hetti Perera, Mohammed Farhaan Shaikh, Poornima Venkatesha, Shivani Hegde

May 20th, 2020

FACULTY OF
COMPUTER SCIENCE

DATA SCIENCE WITH R - PROJECT PROPOSAL

TEAM MEMBERS

- Anne Rother
- Hetti Maneendra Perera
- Mohammed Farhaan Shaikh
- Poornima Venkatesha
- Shivani Hegde

ANALYSIS AND PREDICTIONS OF THE IMPACT OF COVID-19

BACKGROUND AND MOTIVATION

The novel coronavirus outbreak which originated in Wuhan,China has grasped the entire world. In fact, it has reached every corner of the globe and has brought the world to a standstill. As it has been confirmed in large number of countries the World Health Organization (WHO) has declared COVID-19 as pandemic.

Right now, being the only topic that has captured everyone's heart, mind and soul our team would like to understand the impact of it on different domains. Undoubtedly, it is much more than a health crisis and is creating devastating social, economic and political crisis. This inspires us to predict the number of cases in the future according to the current circumstance which will help the country to take appropriate decisions in order to control the spread of the virus. Also, understanding the cause and drift of the virus on different people is also one of the major concern that everybody is curious to know about. Lastly, we would like to understand the trend of different tweets due to the spread of coronavirus in Twitter to see the opinions and experience of people around the world.

PROJECT OBJECTIVES

Objective 1: Our first objective is to analyse and visualise the patient dataset in order to get better insights of the disease COVID-19 caused by corona virus. It is important to know the symptoms as they are relatively non-specific so that immediate medical attention is given when occurred. We will depict the highest ranked symptoms in the city of origin Wuhan, China. Transmission of the disease is primarily due to close contact with symptomatic people, therefore we will show different countries of the patients who visited Wuhan. People of all ages are infected by the virus. However, the ones with the history of chronic diseases are more vulnerable to becoming severely ill. We will visualize the gender distribution between them. The links between the incidence of chronic disease and chances of death will also be explored. The different age groups of the patients will be compared to understand the severity of health issues and fatal results among them. The road to recovery is not smooth. So, we will analyse the average number of days and age groups taken to recover so that priorities can be set on basis of this in case of lack of medical facilities.

Objective 2: Our second objective is to understand the impact of COVID-19 in Twitter. This is done by retrieving the popular trends from a specific location after the coronavirus outbreak across the globe. It will help us analyse the various responses taken by the administration as well as people's response as the virus unfolded. We will retrieve the tweets using #COVID-19, #Corona etc., in order to get the related posts pertaining to it and analyse the engagement of people in this subject and maximise reach based on similar interests. Retrieving the retweeted tweets will help us understand the effects of COVID-19 among people. We perform sentimental analysis on the tweets in order to understand the sentiments of people is crucial during this time of pandemic.

Objective 3: Our third objective is to predict the number of cases, total deaths and total recovered in a country for the next few days (e.g a week) depending on the current trend. This will give us an idea of the extent to which a country will be affected if no changes are made to improve the current situation. For example, government of a country can decide if there should be a lockdown or not based on the current forecast.

DATASETS

1. Patient Medical Data for Novel Coronavirus COVID-19
(Medical records of patients infected with novel coronavirus COVID-19.This data was imported and made computable on May 27, 2020.) Dataset can be downloaded here: [Patient Medical Data for Novel Coronavirus COVID-19] *https://datarepository.wolframcloud.com/resources/Patient-Medical-Data-for-Novel-Coronavirus-COVID-19*

This dataset mainly includes patient characteritics information and location information. Furthermore, dataset has 23 features which are mentioned below. "Age","Sex","City","AdministrativeDivision","Country","GeoPosition","DateOfOns

2. Twitter Data
We will work on the most recent dataset aggregated from Twitter using tritteR and rtweet libraries within a particular timeframe.

Here twitteR which provides an interface and access to Twitter web API, rtweet which acts as the client for Twitter's REST and stream APIs will be used to retrieve data .

3. Novel Coronavirus 2019 Time Series Data on Cases
(Sourced from the Novel Coronavirus (COVID-19) Cases Data) Dataset can be downloaded here: [Novel Coronavirus (COVID-19) Cases Data]

https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases

There are 3 datasets which consists of time series data that tracks the number of people affected by COVID-19 worldwide and it has following three main information: • Number of confirmed cases of Coronavirus infections • Number of deaths from Coronavirus infection • Number of recovered cases from Coronavirus infection

DESIGN OVERVIEW

Objective 1 Here we will visualize data using the ggplot2 library for understanding the data and for developing an intuition of a COVID-19 patient. For continuous variable such as age, we use histogram. The categorical variables such as symptoms is visualized using pie charts and bar plots. Dot plots are used to see groups of people with similar travel history locations. We will use scatter plots to show the number of days from the onset symptoms till discharge. Also, to see the number of days the patient survived after the symptoms occurred. Since the box plot gave the median, we used it to depict the age group of patients who are most likely to recover. We will show the timeline from the date of onset symptoms and date of discharge as well as date of onset symptoms and date of death.

Objective 2 Perform sentimental analysis on the retrieved tweets in order to analyze how citizens have been impacted all over the world and plot the trends from the specific locations.

Objective 3 We are analysing and predicting time series data using the Autoregressive Integrated Moving Average (ARIMA) model. We visualize the data in the form of a line graph. To check for the accuracy we compare the prediction with a small percentage of days and use mean squared error (MSE) to compare the predictions.

TIME PLAN

TASKS	RESPONSIBILITIES
Brainstorming	All team members
Objective Formulation	All team members
Implementation of Objective 1	Shivani, Poornima
Implementation of Objective 2	Poornima, Shivani
Implementation of Objective 3	Anne, Maneendra, Farhaan
Create Project Website	Farhaan, Maneendra
Create Project Screencast	Shivani, Anne, Poornima
Final Project Presentation	All team members