**CSEE5590/490: Big Data Programming**

**Final Project Report**

**Due Date: May 7(Fri) 2021, 11:59PM**

**PROJECT TITLE : COVID 19 Sentiment Analysis**

***GROUP 5***

***VYOMA DESAI - 16314631***

***AFFAN CHAROLIA - 16305677***

***ALI ALYAMI- 16293096***

**INTRODUCTION**

COVID-19 has affected every single person around the globe and has had a huge impact on businesses everywhere. People are sounding their emotions on various social networking platforms and one of those platforms is twitter. The hashtag “Covid19” is trending and people are voicing their sentiments on it. Twitter makes public Tweets and replies available to developers. These endpoints can easily be used by people to identify, understand, and counter misinformation around public health initiatives.

**BACKGROUND**

Covid-19 has affected many people in different ways, and to ease up on its effect on them, great minds across the globe have tried to engineer solutions to this problem through vaccines. Although resolutions being available, there are doubts about its effectiveness. To gain insights on this issue, analysis is being made to better understand where the issues lie and to spread knowledge. Twitter has been a common platform where people have sounded their opinions and shared their experience on taking vaccines. Also, dataset has been released by committees giving out vaccine and storing records on its effects on patients.

A. Cotfas et al in their paper entitled ‘The Longest Month: Analyzing Covid-19 Vaccination Opinions Dynamics From Tweets in the Month Following the First Vaccine Announcement’. [1] have used tweets of the first month after release of vaccine to the public and analyzed them through various classification algorithm. They have made comparison between classical machine learning and deep learning algorithm to find which of it would be better suited for this problem and have chosen the best classifier from them based on four performance metrics. 2,349,659 tweets have been used by them for their analysis. The tweets have been classified into favor, against and neutral with an accuracy of 78.94%. The tweets are analyzed with the local news around that location to monitor the evolution of perspective towards Covid-19 Vaccine. In their analysis they found the opinion of people being more towards the neutral side both on daily basis as well as on an overall basis.

The paper ‘Quantifying Covid-19 Content in the Online Health Opinion War Using Machine Learning’ [2] written by R. F. Sear et al talks about the opinions being shared about the vaccine online on various social media platform dividing people into two groups namely pro-vax and anti-vax. The authors have used dataset collected from various social media platforms and used it to divide people into two groups based on their opinions of this subject. They have used machine learning algorithm to cluster people into two groups. They have concluded that the anti-vax group is more vocal on these social media platforms and is made up of a much diverse group as compared to pro-vax whose number are less and are less engaging leading the authors to believe that this may lead the new people entering the discussion getting inclined towards anti-vax group leading to fake news forwarding becoming prevalent. Mechanistic Model has been used to further support the efficiency of the systems used by them.

N. Paul and S. S. Gokhale have talked in their paper entitled ‘Analysis and Classification of Vaccine Dialogue in the Coronavirus Era’ [3] about opinions of people on getting vaccinated. The authors like other have grouped people into ant-vax and pro-vax community. They have used the twitter dataset on Covid-19 vaccine for the analysis. They have used various data mining tool to get hold of this data and have cleaned and analyzed on it. They have done opinion mining to classify people into two communities of pro-vax and anti-vax. They have also used mathematical models to get some numbers to better understand their data. They made a model with an accuracy of 80% and concluded that the anti-vax opinions are more as compared to other group and are leading the authors to believe that this will lead political parties and pharmaceutical companies to cut corners which will lead to adverse effects.

**MOTIVATION**

COVID-19 has affected every single person around the globe and has had a huge impact on businesses everywhere. People are sounding their emotions on various social networking platforms and one of those platforms is twitter. The hashtag “Covid19” is trending, and people are voicing their sentiments on it. Twitter makes public Tweets and replies available to developers. Hence, it allows developers to post Tweets via API. Developers can access Tweets by searching for specific keywords or requesting a sample of Tweets from specific accounts. These endpoints can easily be used by people to identify, understand, and counter misinformation around public health initiatives.

**SIGNIFICANCE**

The pandemic of Covid-19 has taken everyone by surprise. Many people have lost their jobs and their businesses. It's a topic which has grappled everyone and is a point of discussion and debate. There have been cases where people have had breakdowns because of this\environment that has been created because of it. So, it is significant to know how people are feeling about this and what are their sentiments towards it.

**OBJECTIVES**

1. To understand the sentiments of people.

2. To find out which people are affected the most.

3. Which locations are affected the most.

**FEATURES**

We collected tweets on the topic “Covid-19” and will be pre-processing the data to remove unnecessary data. Further we will be using Hadoop and Spark to analyze the data. We will be using various Hadoop tools like MapReduce, sqoop and hive and some of the spark tools to help assist us in this project.

**DATASET**

We are using COVID-19 tweets dataset from all over the world that we got from Kaggle website. Our aim is to perform sentiment analysis on the dataset so that we can conclude about the impact of COVID-19 around the world. Whatever they express on social media like twitter.

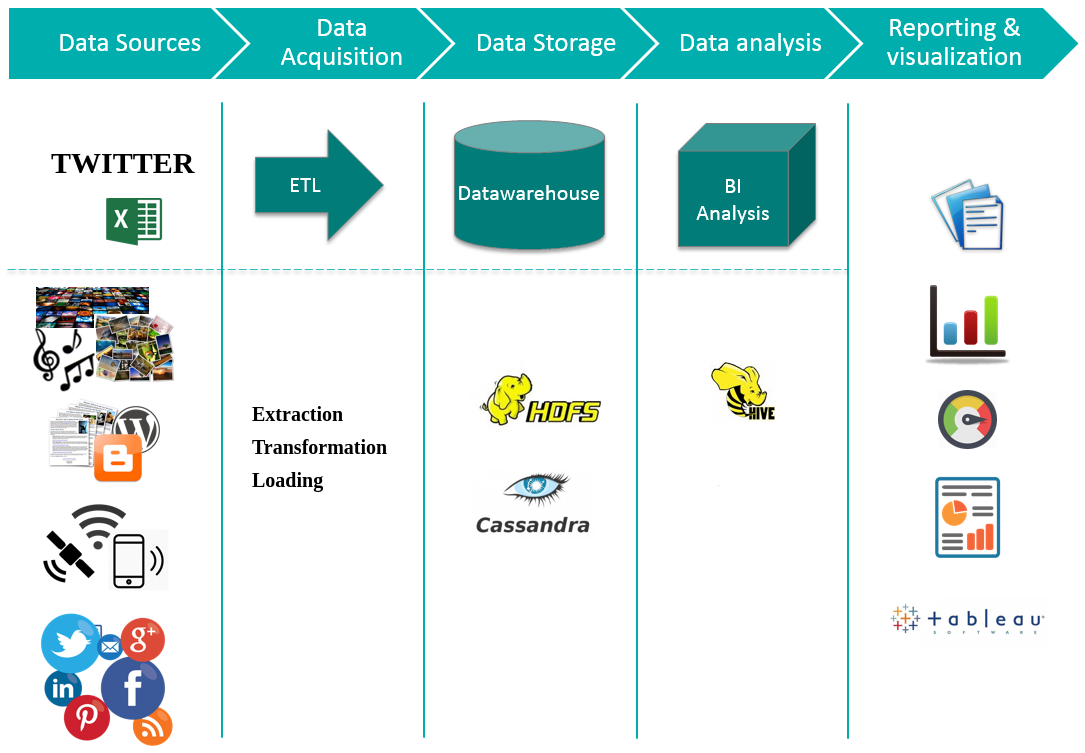
Dataset URL:<https://www.kaggle.com/gpreda/covid19-tweets>

Dataset Features:

Provided dataset contains the following features or tables. User\_name, user\_location, user\_description, user\_created, user\_followers, user\_friends, user\_favourites, user\_verified, date, text, hashtag, source, is\_retweet.

If we consider these fields then we can see that we need very few of them, such as we will be using the “text” field to get the text of tweet, whatever someone has written about the COVID-19. Other tables such as user location can be targeted to find out about the people of a certain area or geo position. We can also use hastags to find out how much the people are expressing the terms.

**PROJECT WORKFLOW**



**Data Analysis: Fields Description:**

|  |  |
| --- | --- |
| Field Title | Description |
| User\_name | Describes the user name of person on twitter |
| User\_location | Describes the location of that person |
| User\_description | Describes the description of that user from twitter account |
| User\_created | Describes when the account of user was created |
| User\_followers | Describes how many followers that user have |
| User\_friends | Describes how many friends the user have |
| User\_favourites | Describes the favorites of the user |
| User\_verified | Describes if the user is verified or not |
| Date | Date when the tweet was made |
| Text | Describes the text that was written in the tweet |
| Hashtag | Describes the hashtags that were used in the tweet |
| Source | Describes the source from where tweet was made such as Phone/Desktop |
| Is\_retweet | Describes if it’s a retweet or not |

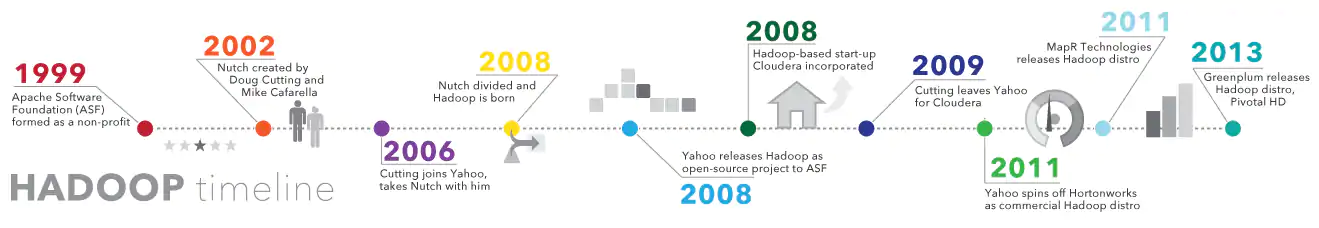
the fields we are looking for are text, date, and user\_location.

**IMPLEMENTATION**

**Hadoop**

Apache Hadoop is an open-source suite of software utilities that make it easy to use a multicomputer network to solve data and computing problems. It provides a software framework for the process of distributing and storing large data using the Mapreduce programming model.

As the World Wide Web grew in the late 1900s and late 2000s, search engines and indexes were developed to help find relevant information between text-pages. In the early years, people returned search results. But as the Internet grew from tens to millions of pages, automation was needed. Web crawlers were developed, most of which were research projects, and new search engines (Yahoo, AltaVista, etc.) began to emerge.



One such project was an open source search engine called Dutch Cutter and Mike Caffrela's Brain Builder. They wanted to quickly return web search results by spreading data and calculations across different computers so that they could perform more than one task at a time. Now, another search engine project called Google is being developed. It was based on the same concept - splitting, automatically storing and processing data so that relevant search results could be returned quickly.

In 2006, Cutting joined Yahoo! and, along with the Nich project, embraced the idea of ​​Google's initial work to automate distributed data storage and processing. The Natch Project split: part of the web crawler remained nuts, and part of the distributed computing and processing became (named after the cutting son's toy elephant). In 2008, Yahoo released Hadoop as an open-source project. Today, the Hadoop Framework and Technology Ecosystem is managed and maintained by the nonprofit Apache Software Foundation (ASF), the global community of software developers and partners.

**Why is Hadoop important?**

Ability to quickly store and process large amounts of any data: This is an important consideration as the amount of data is increasing in various ways, especially from social media and the Internet of Things (IoT).

**Computing power**: The HUDP grid computing model is faster in processing large data. The more compute nodes you use, the more computational power you have.

**Fault tolerance**: is protected from data and application processing hardware failures. If the node fails, the jobs are automatically redirected to other nodes to ensure that grid computing does not fail. Multiple copies of all data are automatically saved.

**Flexibility**: Unlike traditional relative databases, you do not need to pre-process the data before storing it. You can store as much as you want and decide how to use it later. This includes unstructured data such as text, images and videos.

**Low cost:** open-source platform is free and uses common hardware to store large amounts of data.

**Scalability**: You can easily extend your system to handle more data by adding nodes. Requires small administration.

**MapReduce:**

MapReduce is a processing method and programming model for Java-based distributed computing. MapReduce is based on java distributed computing programming model and processing model. The MapReduce algorithm has two main functions, namely mapping and reduction. The map takes one dataset and converts it into another dataset, where the individual elements are divided into topples (key / value pairs).

Steps to complete a MapReduce Task:

**Step 1**: A block is processed by a mapper at the same time. In mapper, the developer can define his proposed logic as needed. As such, the map runs across all cluster nodes and processes data blocks in parallel.

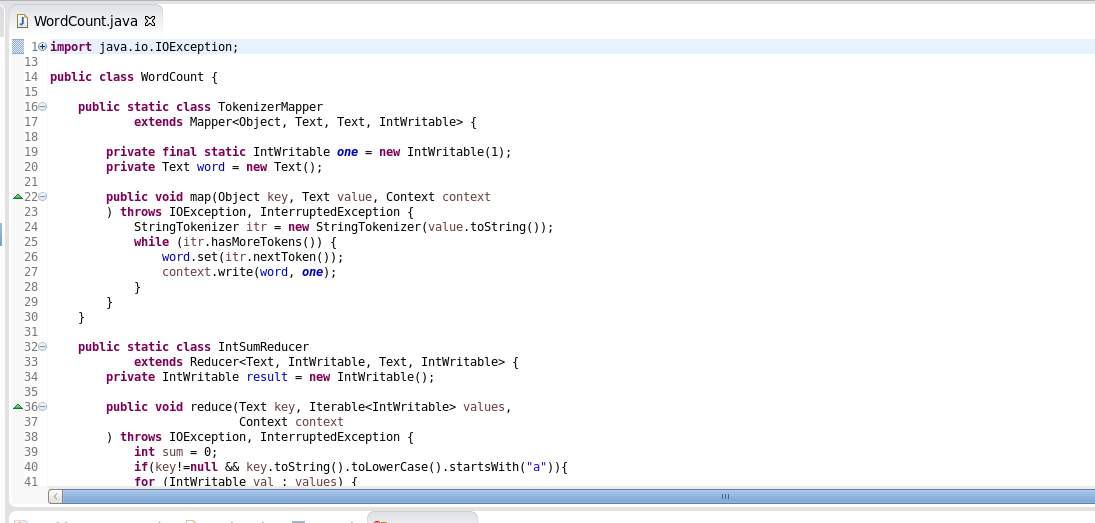
**Step 2**: Mapper output, also called staging output, is written to the local disk. Mapper output is not stored in HDFS because it is temporary data and writing to HDFS will create unnecessary copies.

**Step 3**: Output is converted to a reducer node (which is a common slave node, phase will continue here, hence it is called reducer node). Shuffling / copying is the physical movement of data across a network.

**Step 4:** After all the matches have been completed and their output is adjusted on the reducer nodes, this intermediate output is integrated and configured. Which is then provided as input to reduce the phase.

**Step 5:** Reduce is the second processing step where the user can articulate their business logic as needed. An input to a reducer is provided from all the mappers. The reducer output is the final output listed on the HDFS.

**Java Code (MapReduce WordCount):**

****

**hadoop fs -mkdir /user/cloudera/tweets**

It will create a directory named tweets under user/cloudera in hadoop file system

**hadoop fs -mkdir /user/cloudera/tweets/input**

We will create a sub directory named as input to save our input file here

**hadoop fs -put text.txt /user/cloudera/tweets/input**

It will save the text file in input directory

**hadoop jar WordFreqCount.jar WordCount /user/cloudera/tweets/input /user/cloudera/tweets/output**

It will run the jar file and count word in our input file and store them to ouput which we can either check through

HUE or we can run the command below

**hadoop fs -cat /user/cloudera/tweets/ouput/part-r-00000**

****

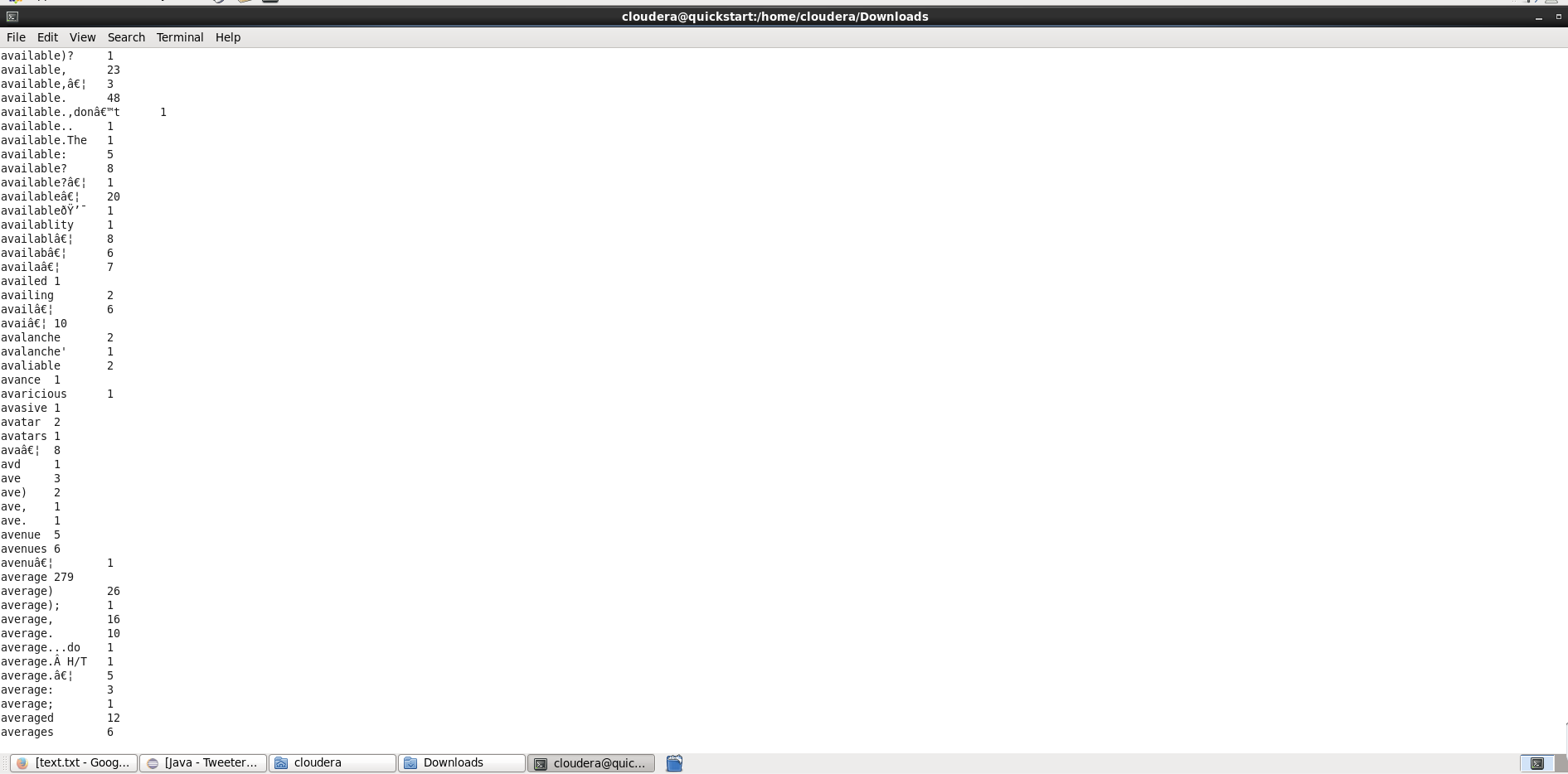
WordCount Output here is showing how many times a certain word was found in the text of tweets we provided from the COVID-19 dataset. From this WordCount we can also create a WordCloud.

**hadoop jar WordFreqaCount.jar WordCount /user/cloudera/tweets/input /user/cloudera/tweets/outputa**

It will run the jar file and count word starting with a letter 'a' in our input file and store them to ouput which

we can either check through HUE or we can run the command below

**hadoop fs -cat /user/cloudera/tweets/ouputa/part-r-00000**



WordCount Output here is showing how many times a certain word that starts with a letter a was found in the text of tweets we provided from the COVID-19 dataset. From this WordCount we can also create a WordCloud.

**CASSANDRA:**

* Covid-19 Twitter Data Implementation

1. We have used Covid-19 Twitter Dataset for our sentimental analysis.
2. Pre-processing of our dataset eliminated special characters.
3. Making use of Cassandra query we have implemented queries to represent data in the form of visualization.

* Dataset:

These tweets are collected using Twitter API and a Python script.

A query for this high-frequency hashtag (#covid19) is run on a daily basis for a certain time period, to collect a larger number of tweets samples.

Our data set mainly contains covid-19 tweets with hashtags #covid-19

* Kaggle Link:

Downloaded KAGGLE Link :

https://www.kaggle.com/gpreda/covid19-tweets

* Data Pre-Processing

1. This Twitter data contained many special characters which failed to upload the large data on cassandra database.
2. Removing special characters from the dataset using python
3. Add csv\_path and define a function clean\_text(string) which will return only those elements which we want.

def clean\_text(string):

return re.sub(r"[^a-zA-Z0-9:/.,@#&]+", ' ', string)

1. For each columns apply lambda function to eliminate all special characters.

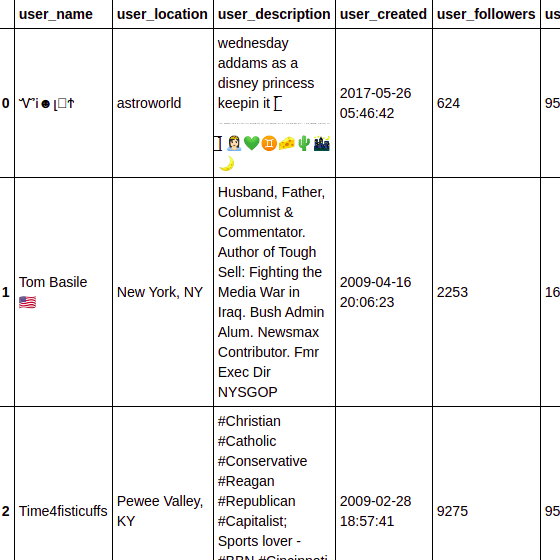
for column in columns:

df[column] = df[column].apply(lambda x: clean\_text(str(x)))

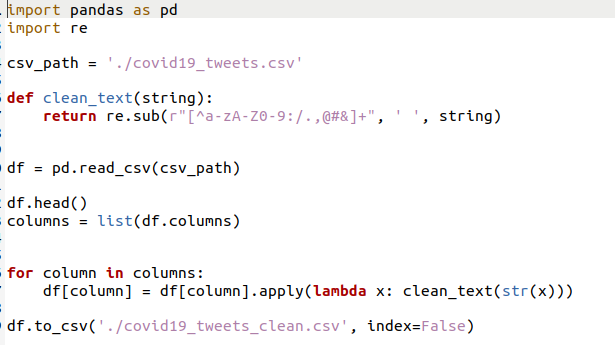
1. Store the new COVID-19 dataset inside covid19\_tweets\_clean.csv

df.to\_csv('./covid19\_tweets\_clean.csv', index=False)

1. We will now use new dataset to run queries for analysing the data.



*Fig : Twitter dataset with special characters.*



*Fig : Preprocessing Twitter Data Set using Python*

* Query Implementation using CQLSH

1. Create a keyspace named covid19\_tweets and create a table named as covid19tweetsdata for a collection column and other columns.

CREATE KEYSPACE covid19tweets

WITH replication = {'class':'SimpleStrategy', 'replication\_factor' : 3};

Describe Keyspace



Using keyspace

cqlsh> USE covid19tweets;

cqlsh:covid19tweets>

1. Create a table and describe table covid19tweetsdata.

Fields Description:

|  |  |
| --- | --- |
| **Field Title** | **Description** |
| User\_name | Describes the user name of person on twitter |
| User\_location | Describes the location of that person |
| User\_description | Describes the description of that user from twitter account |
| User\_created | Describes when the account of user was created |
| User\_followers | Describes how many followers that user have |
| User\_friends | Describes how many friends the user have |
| User\_favourites | Describes the favorites of the user |
| User\_verified | Describes if the user is verified or not |
| Date | Date when the tweet was made |
| Text | Describes the text that was written in the tweet |
| Hashtag | Describes the hashtags that were used in the tweet |
| Source | Describes the source from where tweet was made such as Phone/Desktop |
| Is\_retweet | Describes if it’s a retweet or not |



## Create Primary Key

The primary key is a column that is used to uniquely identify a row. Therefore,defining a primary key is mandatory while creating a table. A primary key is made of one or more columns of a table. You can define a primary key of a table as shown below.

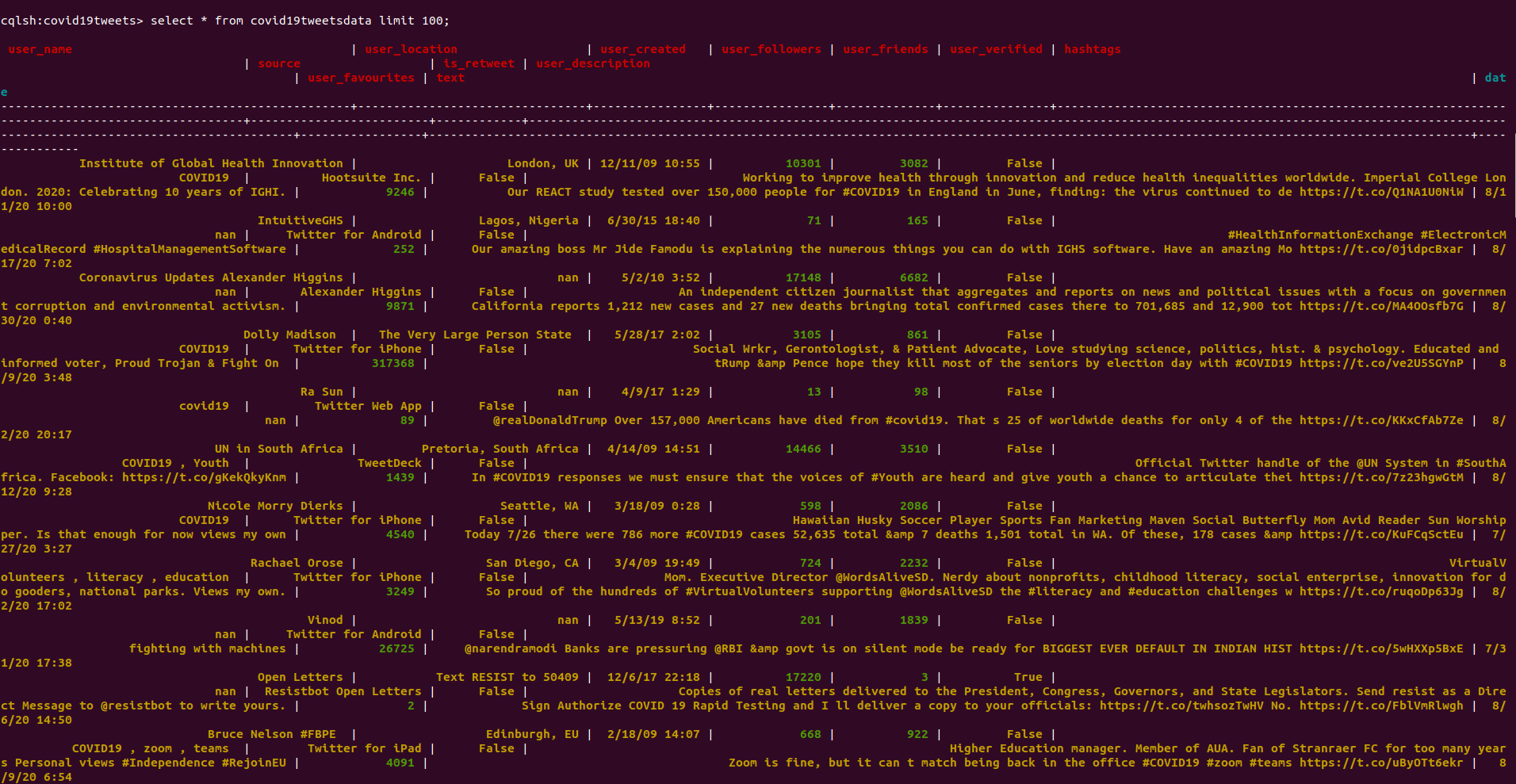
CREATE TABLE tablename(column1 name datatype PRIMARYKEY, column2 name data type, column3 name data type, PRIMARY KEY (column1))

1. Import the CSV values using COPY Command from where the data is stored in home directory covid19\_tweets\_clean,csv using delimiter = “,” and header =true;

COPY covid19tweetsdata (user\_name, user\_location,user\_description, user\_created, user\_followers ,user\_friends , user\_favourites , user\_verified , date , text , hashtags , source ,is\_retweet ) FROM '/home/vyoma/Desktop/BigDataSyllabus/Project Proposal/finalDataset/big\_data\_cassandra/big\_data\_cassandra/covid19\_tweets\_clean.csv' WITH DELIMITER=',' AND HEADER=TRUE;

1. Use a SELECT statement to retrieve a results set from a table as standard output or in JSON format.

Select \* from covid19tweetsdata limit 100;



1. Define a new index on a single column of a table. If data already exists for the column, Cassandra indexes the data during the execution of this statement. After the index is created, Cassandra indexes new data for the column automatically when new data is inserted.

**CREATE** **INDEX** ***IF******NOT******EXISTS*** *index\_name*

**ON** *keyspace\_name.*table\_name (***KEYS*** ( column\_name ) )

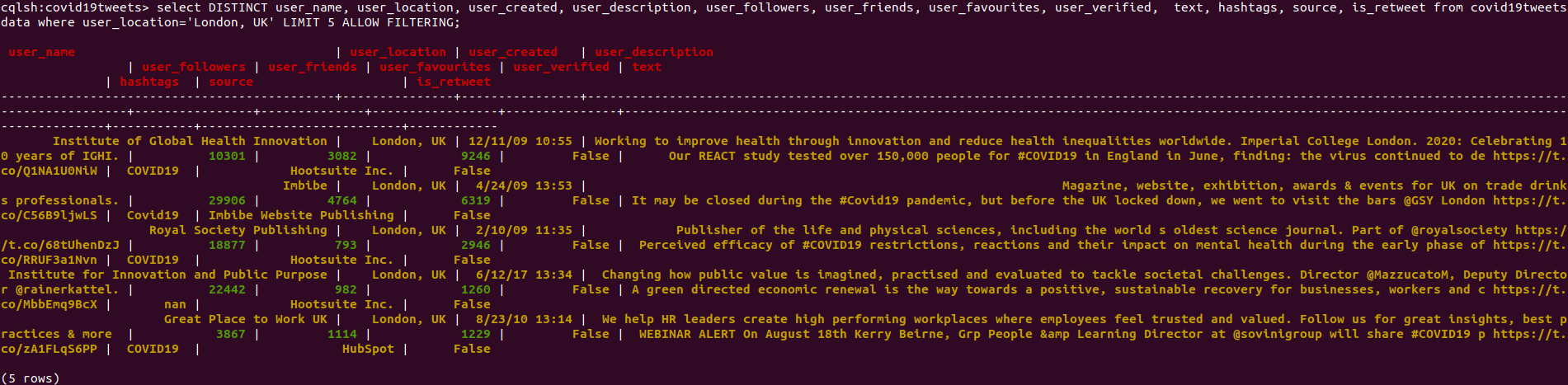
CREATE CUSTOM INDEX fn\_prefix ON covid19tweetsdata (date) using cassandra SASI INDEX



QUERIES ON COVID 19 TWITTER ANALYSIS

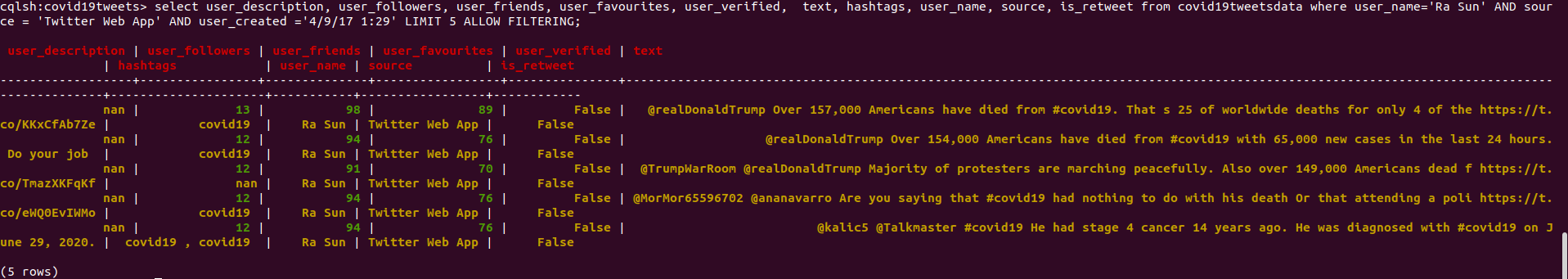
* Select DISTINCT fields form data where place belongs to london

select DISTINCT user\_name, user\_location, user\_created, user\_description, user\_followers, user\_friends, user\_favourites, user\_verified, text, hashtags, source, is\_retweet from covid19tweetsdata where user\_location='London, UK' ALLOW FILTERING;



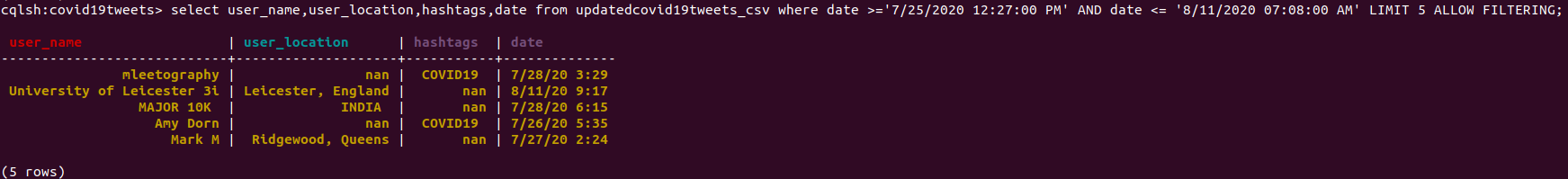
* With twitter web app , display tweets from Ra Sun dated on 4th September?

select user\_description, user\_followers, user\_friends, user\_favourites, user\_verified, text, hashtags, user\_name, source, is\_retweet from covid19tweetsdata where user\_name='Ra Sun' AND source = 'Twitter Web App' AND user\_created ='4/9/17 1:29' ALLOW FILTERING;



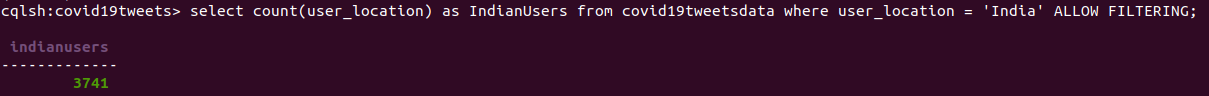
* Display tweets data from 25th July 2020 to 11th Nov 2020

select user\_name,user\_location,hashtags,date from updatedcovid19tweets\_csv where date >='7/25/2020 12:27:00 PM' AND date <= '8/11/2020 07:08:00 AM' ALLOW FILTERING;



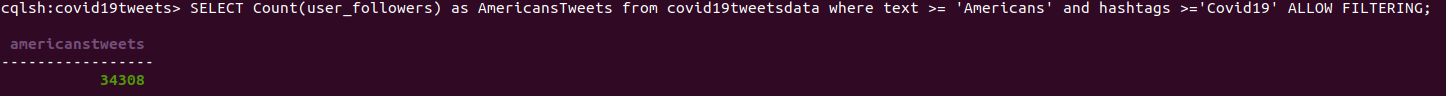
* Retrieve total number of Indians

select count(user\_location) as totalLoc from covid19tweetsdata where user\_location = 'India' ALLOW FILTERING;



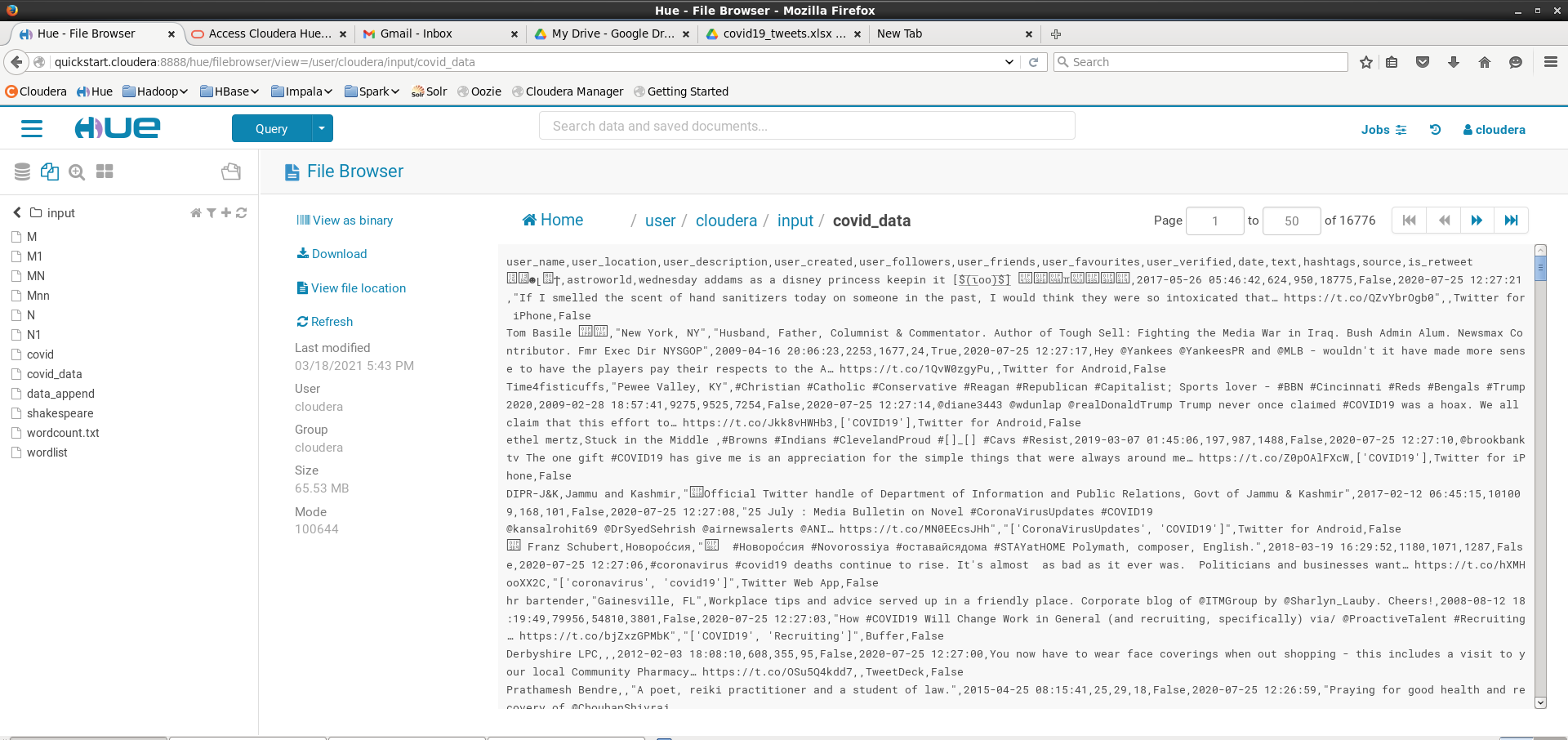
* Americans who twitted about coronavirus

CREATE CUSTOM INDEX comments ON covid19tweets.covid19tweetsdata (text) USING 'org.apache.cassandra.index.sasi.SASIIndex' WITH OPTIONS = {'mode': 'CONTAINS', 'analyzer\_class': 'org.apache.cassandra.index.sasi.analyzer.NonTokenizingAnalyzer', 'case\_sensitive': 'false'};



**Hue:**

Apache hue is used for visualization and querying databases. It is an open source which helps look HDFS file system and manage databases like hive, etc. Where you can visualize your queries. We have added our database to the HDFS and visualize it using hue.



**Hive:**

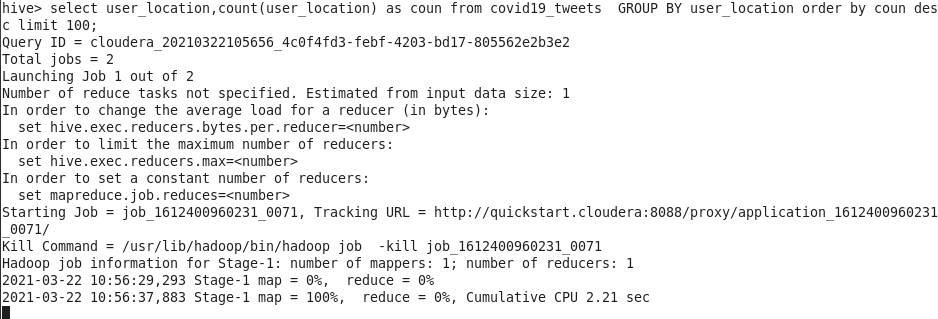
Hive is a data warehouse built on top of Hadoop for structured data. It has SQL-like query and hence is very popular and simple to operate. We have used it to analyze our dataset to find trends in it. We run these queries in hive and visualize it using tableau.

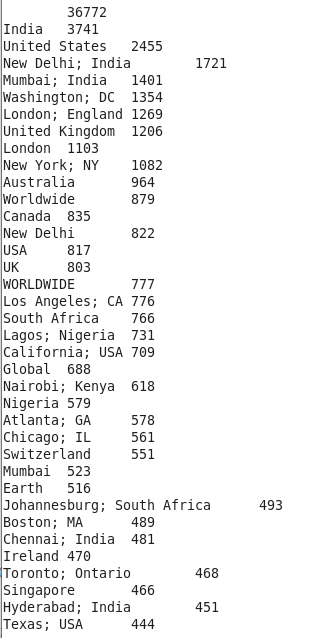
Pseudo-Code:

1. Starting of Hive in the terminal and creating a table named 'covid19\_tweets' along with all its attributes and their data types. The command specifies that each variable is seperated by ',' as it is a csv file and thats how the software should differentiate between different variables and their data.
2. The csv file for hotelbooking is loaded into the table by specifying the file path on the device.
3. Running the queries and getting the results.

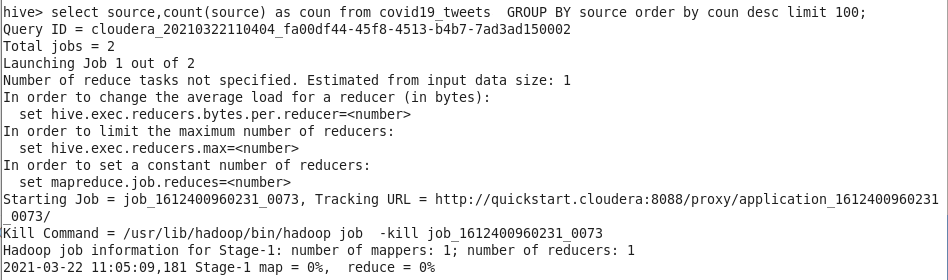
The queries operated on this dataset are:

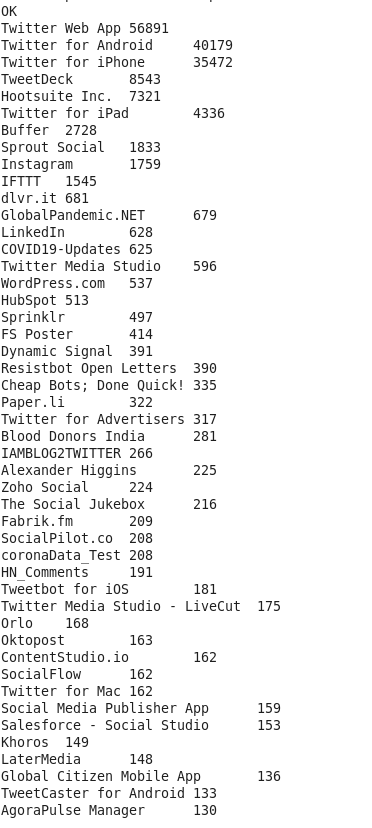
To see from where tweets are coming from majorly.



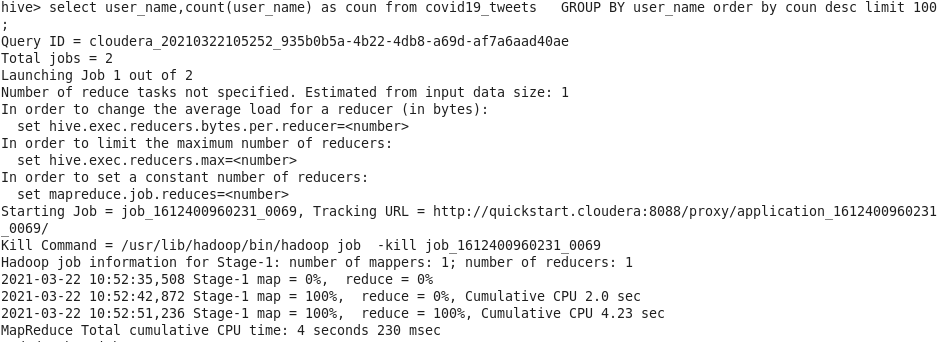


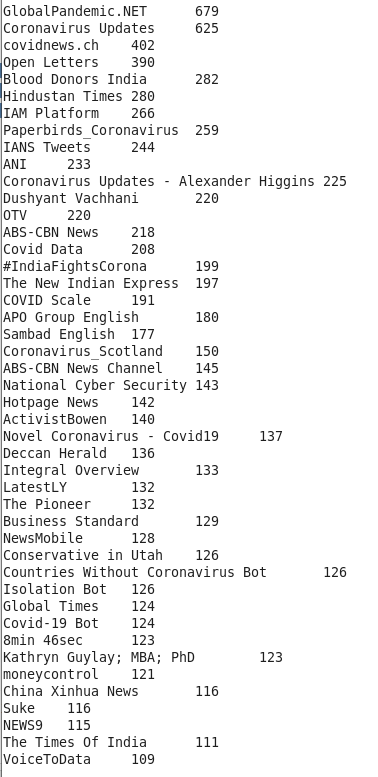
To see from which source are tweets majorly being generated from.



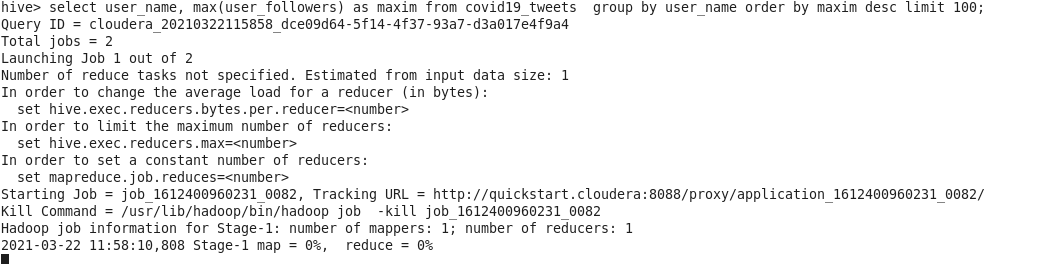


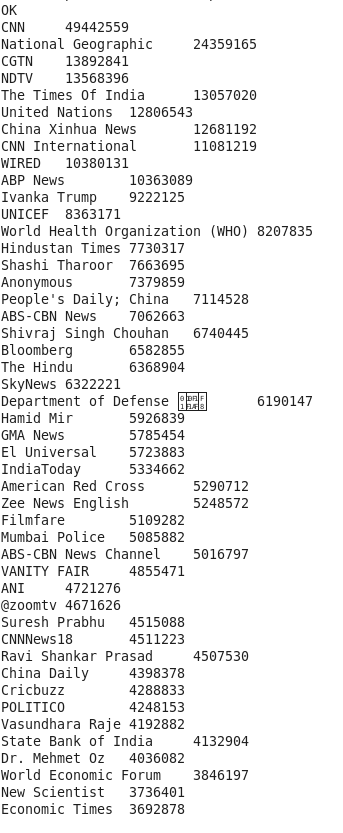
To see from which users have the greatest number of tweets.





To see which users, have the greatest number of followers.





**Spark**

**RDD:**

The code was executed on Google Colab using Apache Spark version 3.0.2. The dataset file and all the intermediate files created were stored in Google Drive. Word Count was done using the help of Apache Spark, and python library ‘pyspark’. Other important python libraries used were ‘nltk’ for removing punctuations, ’re’ for regular expression, and ‘matplotlib’, ‘wordcloud’ for plotting the Word-cloud.

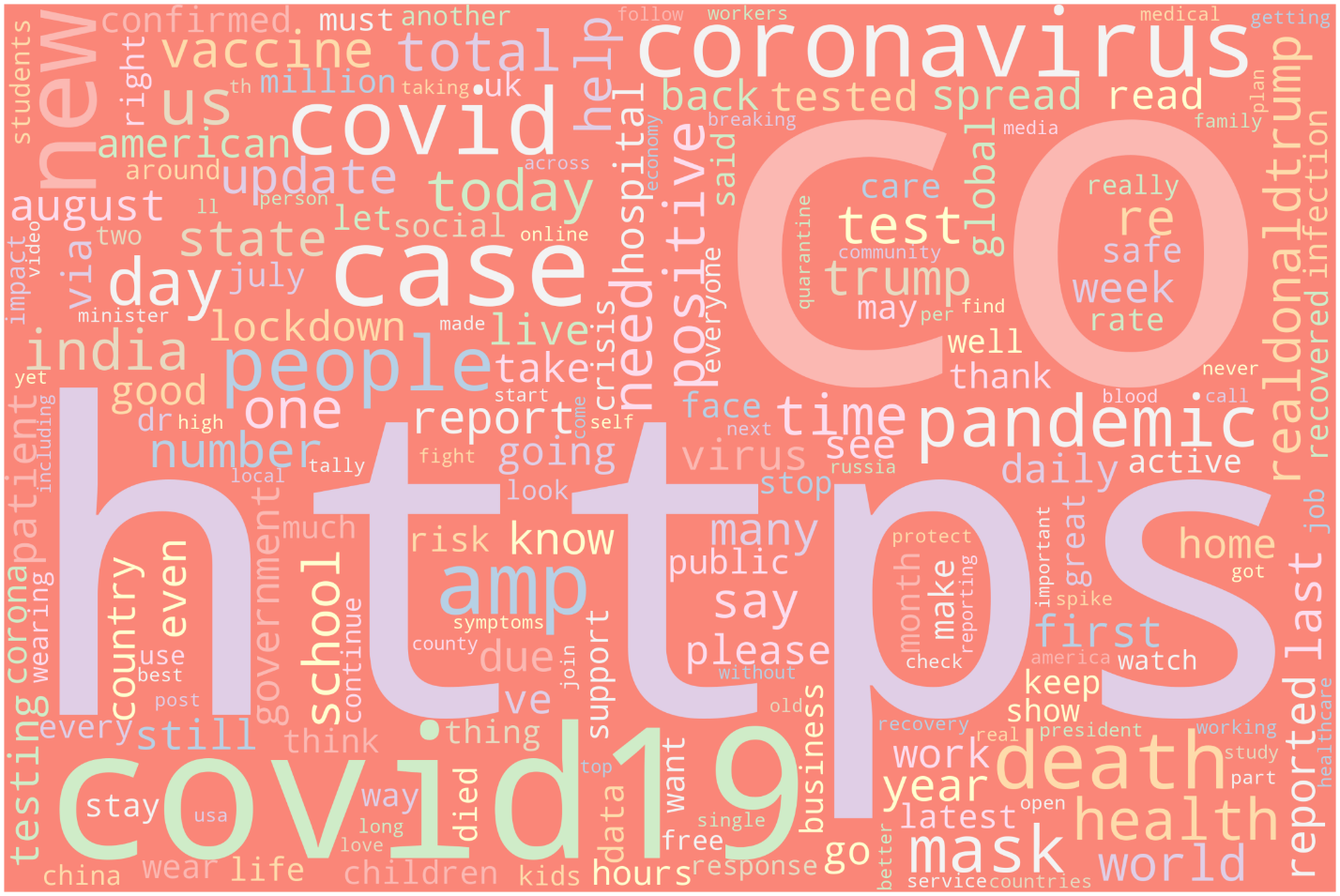
Pseudo-Code:

* The dataset is imported from the google drive and all the punctuations are removed and the output is stored in an intermediate file.
* Spark is loaded in system and the MapReduce function is performed on the intermediate file.
* Before performing Word Count on the data all the stop words are removed using

*‘splitRDD\_no\_stop = words.filter(lambda x: x.lower() not in stop\_words)’* command.

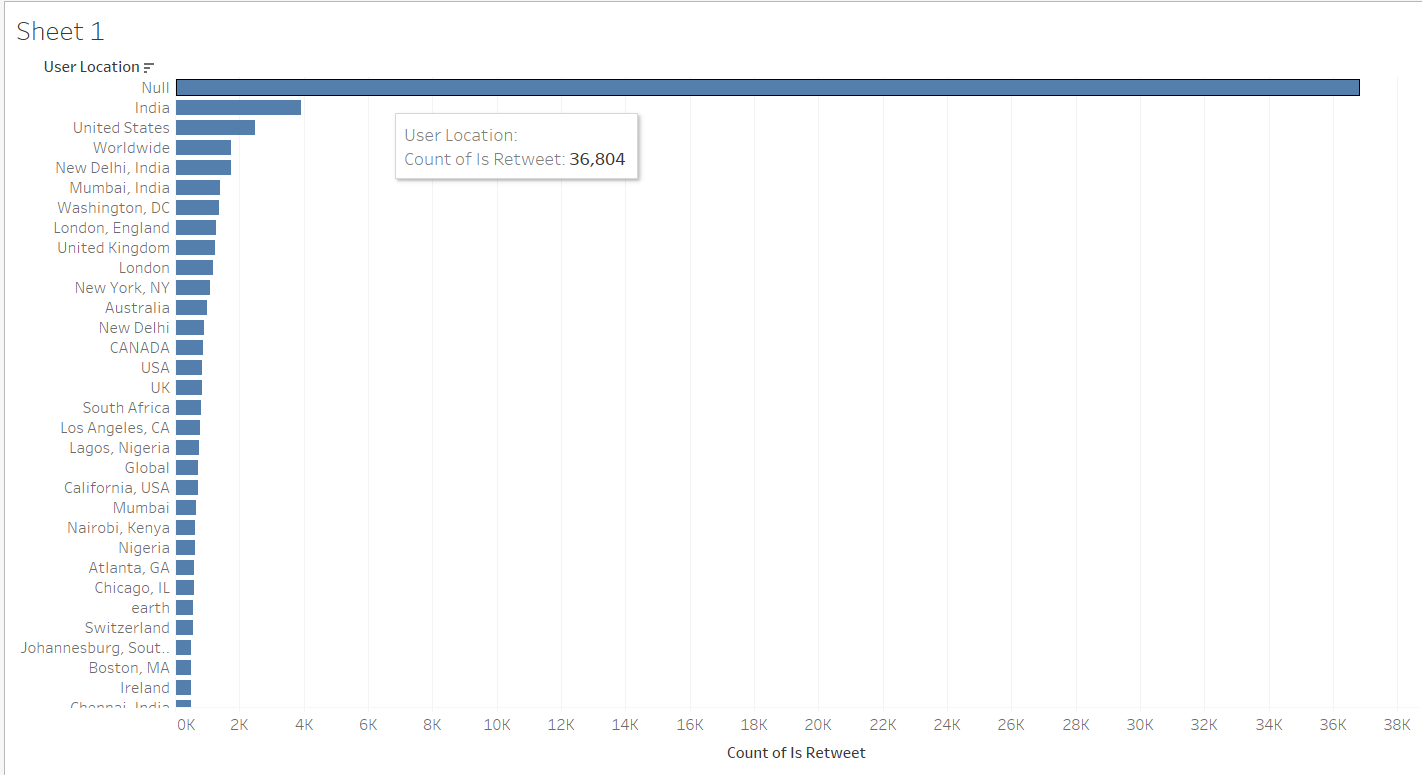
* The Word Count is then performed on the data which returns words as key and their count as value.
* The words are then sorted on basis of their count in the descending order.
* The word and count of the top 200 words are stored in the ‘out\_task1’ file.
* The word cloud is made for the top 200 words using the ‘out\_task1’ file.

**Preliminary Results**



Word Count for all tweets

We have generated Word Cloud from our text dataset from which we can view that letter https occurs most of the time and then CO after that covid19 and coronavirus, list goes on.



To see from which locations tweets are coming from (Viz for hive queries).

There are maximum number of tweets from India followed by United States. The reason behind this maybe due to the population and popularity of the social media application twitter among that location.

Text

Description automatically generated

Word Cloud for tweets originating from India

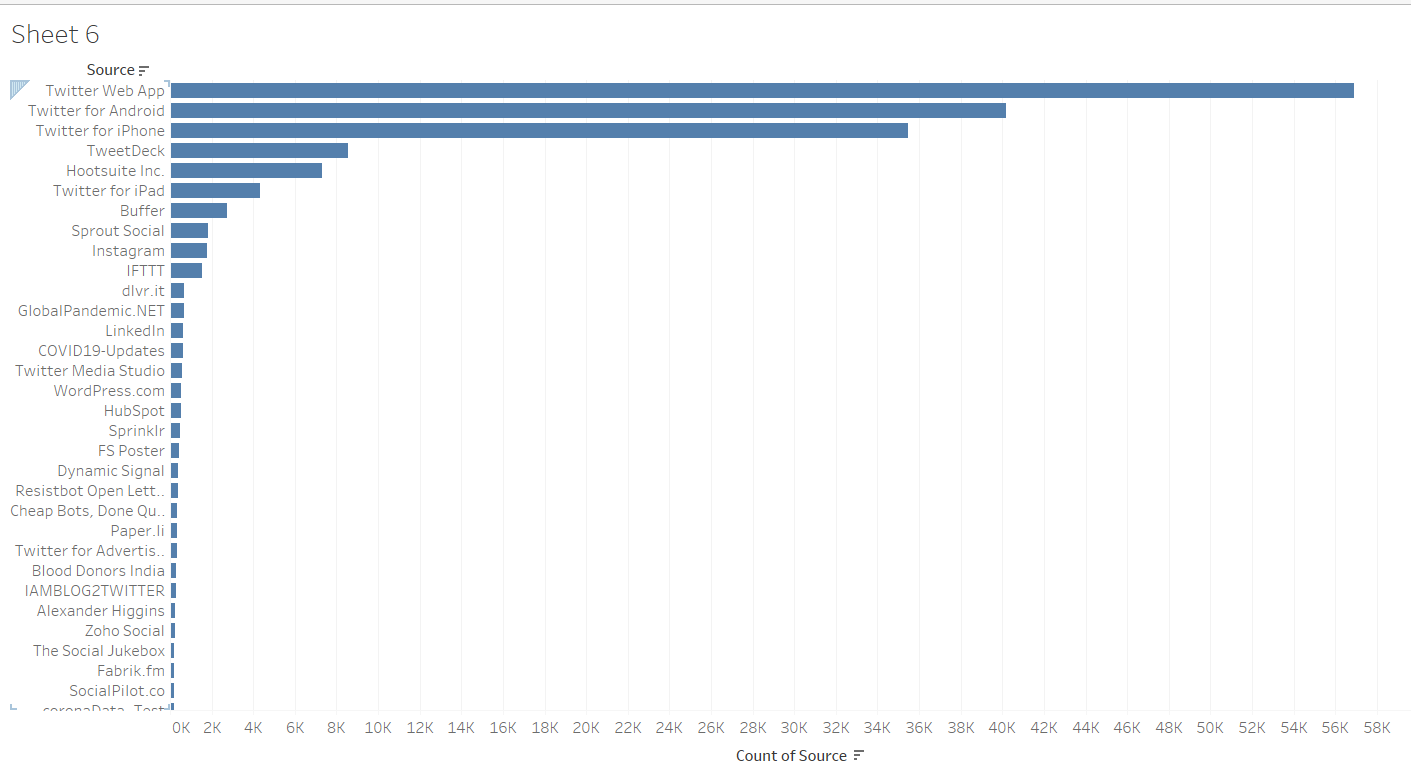
The tweets originating from India have words with states name arising pre-dominantly reporting on the cases in each state. Also, words like hospital, cases, death are pre-dominantly occurring which highlight the conditions in the country. These are terms which indirectly are highlighting how people are feeling in the country. These negative terms have been surrounding people’s life. Also, there are terms like government which were used in the tweets to talk about what the rules and regulations are and how they have been successful or unsuccessful in curbing the problem.

Text

Description automatically generated

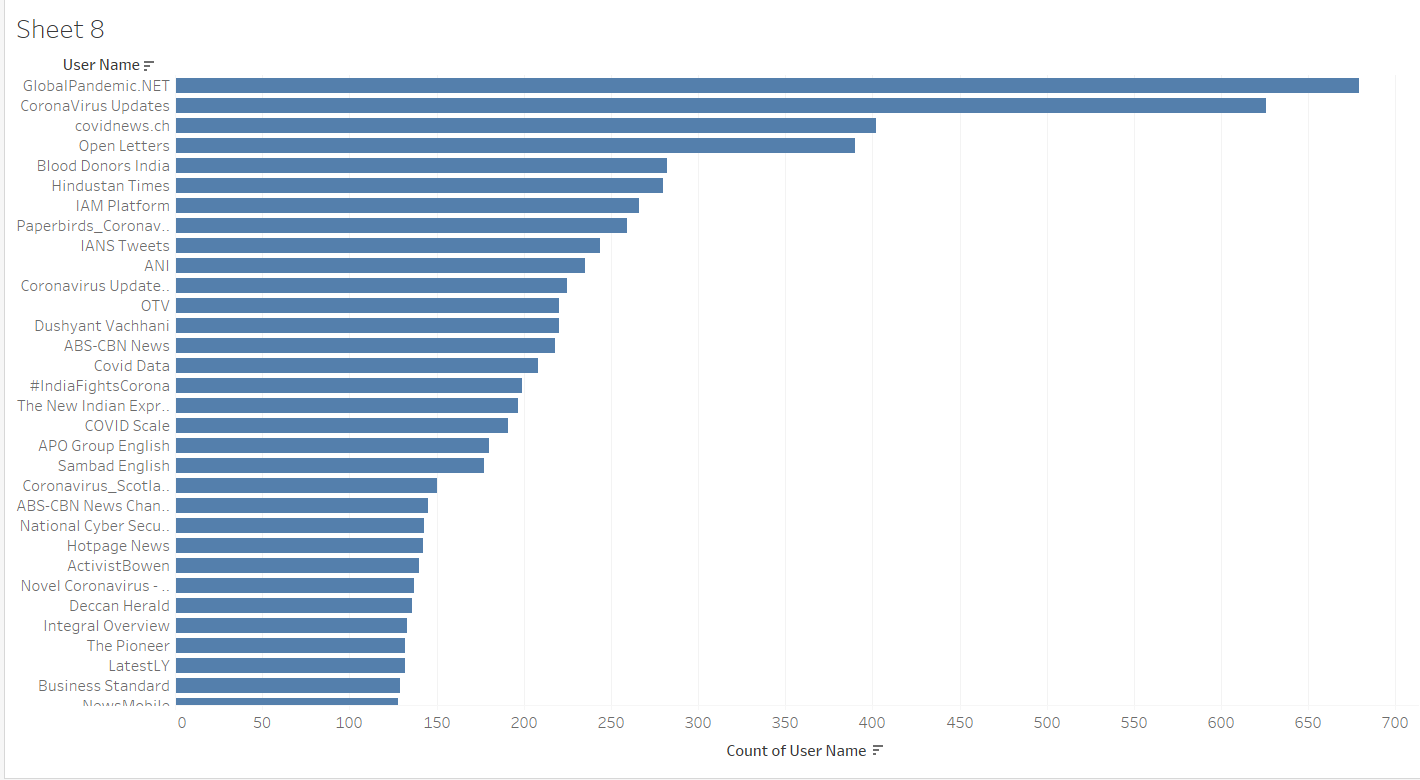
Word Cloud for tweets originating from USA

The tweets originating from USA have terms like Biden and Trump showing people tweeting about the election. Also, the term like business is there which shows how the market is going on. There are also, covid related terms like mask, pandemic, cases to show how people are being affected by it and how they are handling it. Also, the term vaccine is present which shows people first expecting for vaccines to be made and later tweeting about how they felt about talking it.



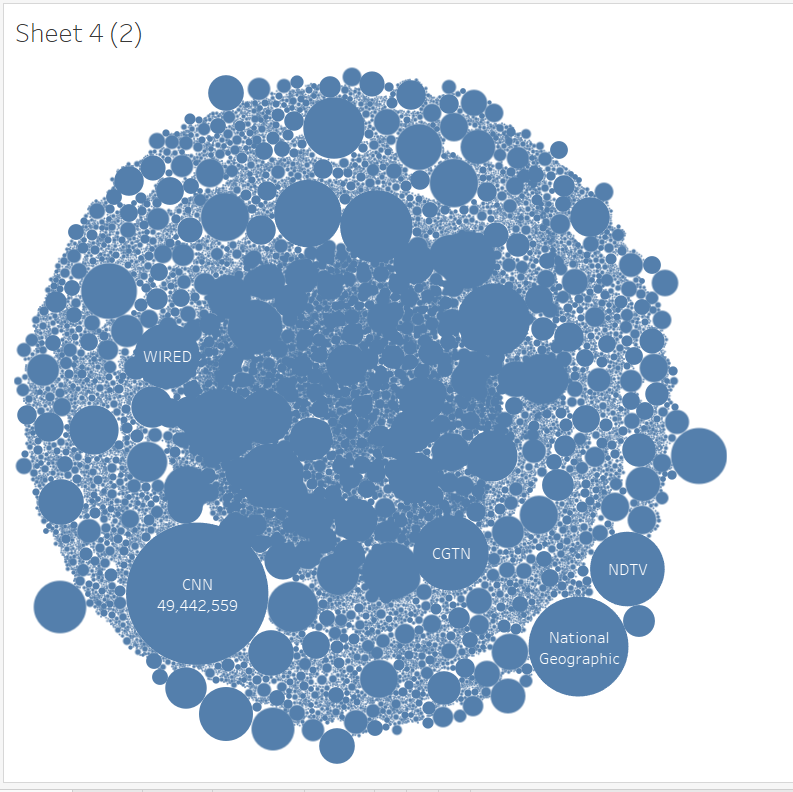
To see from which source are tweets majorly being generated from (Viz for hive queries)

Maximum number of tweets have been from web application followed by android and twitter users. This maybe because a lot of people use their laptop or computers to tweet. Also, there are more android users than iPhone and hence we see this pattern.



To see from which users have the greatest number of tweets (Viz for hive queries)

GlobalPandemic.NET has greatest number of tweets on this pandemic followed by CoronaVirus Updates and covidnews.ch. These users continuously give updates on what is going on around the globe with respect to the current pandemic. How people are reacting and how the pandemic is being administered.



To see which users, have the greatest number of followers (Viz for hive queries)

These are the channels or people with maximum number of followers who have been reporting or tweeting on the covid pandemic. The following shows that these are prominent users who people have faith in on reporting the scenarios correctly.

Text

Description automatically generated

Word Cloud showing tweets from users having more than 7M followers

The word cloud shows terms like country name where the tweeter handle is tweeting on the situations in various country, pre-dominant one being China and India. They are also talking about the cases that are being reported, how many of them are recovering and how many are not. The term vaccine also comes up, tracking of the progress of it across the globe and talking about its effectiveness in its purpose.

Text

Description automatically generated

Word Cloud of tweets that are most popular

The word cloud shows tweets with a lot of negative terms like death, crisis, lockdown, and many other terms showing what the people are pre-dominantly feeling. Also, there is name Donald Trump who was the last US president and used to tweet a lot on this pandemic. He has been popular in people’s tweet because of the election and also because of his ways on handling the spread of the virus in US. Other terms are related to covid like mask, patients which again are negative. There is also vaccine occurring again showing people to be interested in talking about it and closely monitoring its manufacturing and its effectiveness.

**CHAPTER 1(LIFE)**

**WHO:** Covid 19 pandemic has affected everyone around the world in one or the other way.

1. Cities, with their dense international setups were particularly affected which includes markets, business, travel, tourism. These were the common entry points for the virus.
2. Rural areas and regions with high numbers of elderly people were affected as well.
3. The pandemic has deeply impacted jobs and businesses. In the developing world,micro-level, small-level and medium-level enterprises are under intense pressure.

**WHAT:**

1. The pandemic has taken a lot of lives.
2. COVID-19 has rapidly affected businesses, world trade, our day to day life and movements.
3. Presently the impacts of COVID-19 in daily life are extensive and have put tremendous pressure on the healthcare sector and affected the social lives of the people.

**WHEN:** The early traces of the virus were found in November 2019, but it spread rapidly across the world and was declared as a pandemic in March 2020 by the World Health Organization and is still going on.

**WHERE:** 200+ Countries and Territories around the world have reported a total of 113M+cases of the coronavirus COVID-19 that originated from Wuhan, China, and a death toll of2.5M+ (as of this writing)

**WHY:** There is no definite answer to this question. The primary cause of this virus has been linked to the animal “bat” by the mainstream media.

**CHAPTER 2(DATA)**

**WHO:** The dataset covers all the twitter users who have used hashtag Covid19 or used any covid or related material to it in their text. It covers people from every corner across the globe.

**WHAT:** The dataset records all the basic information of the user’s social media handle, which in this case is twitter and along with that it covers the twitter text that the user is sending which is one of the major analysis points in our dataset.

**WHEN:** The collection of datasets began from November 2019 when Covid 19 started spreading from china to everywhere across the globe. The dataset is collected from twitter.

**WHERE**: It was a local topic trending in China, but as it started spreading across the world, people started tweeting about it and that is how the dataset was collected.

**WHY:** The dataset was collected to get an understanding of how people are reacting to this pandemic and how they are being affected by it.

**CHAPTER 3(THE SCIENTIST AND AI)**

**WHO:** The main character in this dataset are the people across the globe who are using tweets to express their feeling? The data scientist that is us, are trying to understand these feelings and report on it.

**WHAT:** Big data tools like Hadoop and Spark were used in the analysis of the dataset. Various modules in Hadoop like map-reduce, hue, hive and Cassandra were used to analyze the data. Also, various spark modules were employed to analyze the data. Tableau was used as one of the visualizing tools.

**WHEN:** The project has been in work for past three months where different tools were used to get perspective of the situation. For better analysis different tools were used so that all the angles were effectively covered, and better analysis can be presented.

**WHERE**: The project was part of the coursework Programming in Hadoop and Spark.

**CHAPTER 4(USERS)**

**WHO:** The main target audience are both the people who are interested in these kind of analysis and programmers or data scientist who are interested in looking into how tools were used to get various answers.

**WHAT:** The tools and visualization gave us the understanding on how people are feeling about this pandemic and situation going around.

**WHEN:** The project is available for anyone to be viewed on GitHub.

**WHY:** The tools can help other data scientist to understand how certain tools can be used to get some inferences.The visualization in turn can help people see what are people tweeting about and how they are feeling.

**CONCLUSION**

We have used various Hadoop, Cassandra, and Spark modules to try and understand the twitter dataset and try to analyze it. We have seen which of the words are frequently used in the twitter texts. Also, we tried to query the tweets based on various factors like location, etc. Finally, we used hive queries and tableau to visualize our dataset and understand more about it. People are using a lot of negative terms in their tweets and they cannot be blamed for that as it is a difficult situation for everyone. They are talking about loss of lives and their livelihood.

**FUTURE WORK**

Further advance to this project can be use of certain deep learning tools to analyze the text further, to better understand what the tweets texts in the twitter dataset is talking about. These can help better classify the problem in more than two terms of positive and negative. It can be used to analyze whether people are angry or sad or if they are happy about something or feeling joy.

**REFERENCES**

[1] L. -A. Cotfas, C. Delcea, I. Roxin, C. Ioanăş, D. S. Gherai and F. Tajariol, "The Longest Month: Analyzing COVID-19 Vaccination Opinions Dynamics From Tweets in the Month Following the First Vaccine Announcement," in *IEEE Access*, vol. 9, pp. 33203-33223, 2021, doi: 10.1109/ACCESS.2021.3059821.

[2] R. F. Sear et al., "Quantifying COVID-19 Content in the Online Health Opinion War Using Machine Learning," in IEEE Access, vol. 8, pp. 91886-91893, 2020, doi: 10.1109/ACCESS.2020.2993967.

[3] N. Paul and S. S. Gokhale, "Analysis and Classification of Vaccine Dialogue in the Coronavirus Era," *2020 IEEE International Conference on Big Data (Big Data)*, Atlanta, GA, USA, 2020, pp. 3220-3227, doi: 10.1109/BigData50022.2020.9377888.

**CONTRIBUTION OF WORK**

**Vyoma Desai ​**

Significance, feature’s part, chapter life. Data research and implementing queries on CASSANDRA, d3.Js.

**Affan Asad Charolia**

​Motivation, background, and objective’s part. Implementing queries using Hue and Hive and representing graphs in Tableau. Doing word count is spark and visualizing it using word cloud. Chapter 3 and 4.

**Ali Alyami**

Project dataset analysis, Conclusion. Implementing dataset wordcount using Hadoop Map Reduce